UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES
Mathematical Sciences

Efficient parameterisation of hierarchical Bayesian models for spatially correlated data

by

Mark Bass

Thesis submitted for the degree of Doctor of Philosophy
August 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Doctor of Philosophy

Efficient parameterisation of hierarchical Bayesian models for spatially correlated data

by Mark Bass

Fitting hierarchical Bayesian models to large spatially correlated data sets using MCMC techniques is computationally expensive. Complicated covariance structures of the underlying spatial processes mean that the number of calculations required grows cubically with the number of spatial locations. This necessitates the need for efficient model parameterisations that hasten the convergence and improve the mixing of the associated MCMC algorithms.

We focus on hierarchical centering reparameterisations which act upon the mean structure of the latent spatial processes. For Gaussian data and under the assumption of known variance parameters, we compute the exact convergence rate for the Gibbs samplers emitted by the centred parameterisation (CP) and the non-centred parameterisation (NCP). We analyse the impact of the variance parameters and the correlation structure of the latent variables upon the convergence rate.

The CP and NCP are considered to be opposite extremes of a continuum of partially centred parameterisations (PCPs). By minimising the conditional posterior covariance of the random and global effects for a Gaussian three stage model, we construct a PCP that has zero convergence rate, implying immediate convergence. Where the variance parameters are unknown we provide a dynamically updated PCP and suggest strategies to mitigate its computational expense.

The construction of the PCP requires the computation of the conditional posterior variance of the random effects, which is intractable for non-Gaussian likelihoods. Therefore, an approximation based on the Hessian matrix is used to construct the PCP for spatial Tobit and spatial probit models.

Our work shows that for a Gaussian likelihood and latent spatial processes with exponential correlation functions, that convergence is hastened for the CP when there is stronger spatial correlation, whereas convergence is delayed for the NCP. Simulation studies suggest that these results hold for unknown variance parameters and for non-Gaussian likelihoods. The PCP is shown to outperform the CP and the NCP for Gaussian likelihoods. The pilot adaption schemes that reduce the computational expense of the PCP are shown to inherit the good mixing properties of the PCP.

The work in this thesis extends the current knowledge of hierarchical centering to include the effect that spatial has correlation upon the convergence rate. The development of a dynamically updated PCP provides practitioners with a robust and fully automated algorithm that has better convergence and mixing properties than either the CP or the NCP.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Mark Bass, declare that the thesis entitled "Efficient parameterisation of hierarchical Bayesian models for spatially correlated data" and the work presented in the thesis are both my own, and have 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 Sujit Sahu, whose patient and expert supervision has allowed me to complete this thesis. I would also like to thank Alan Gelfand who has always been very generous with his time and helped enormously in broadening the scope of this work.

Chapter 1

Introduction

1.1 Motivation

Spatially correlated data is prevalent in many of the physical, biological and environmental sciences. It is natural to model these processes in a Bayesian modelling framework, employing Markov chain Monte Carlo (MCMC) techniques for model fitting and prediction. There is a growing interest among researchers in regression models with spatially varying coefficients. Fitting these highly overparameterised and nonstationary models is challenging and computationally expensive. Latent process correlated across space produce dense covariance matrices that require calculations of order $O(n^3)$ to invert, for n spatial locations.

Typically these calculations must be executed many thousands of times when using MCMC sampling techniques to perform Bayesian inference. To mitigate the computational expense practitioners require efficient model fitting strategies that produce Markov chains which converge quickly to the posterior distribution and exhibit low autocorrelation between successive iterates.

It has long been understood that the parameterisation of a hierarchical model affects the performance of the MCMC method used for inference. In particular, high posterior correlations between model parameters can lead to poor mixing and slow convergence. For normal linear hierarchical models (NLHMs) two natural parameterisations emerge; the centred parameterisation (CP) and the non-centred parameterisation (NCP). If we fit a NLHM using the Gibbs sampler the work of Roberts and Sahu (1997) allows us to compute the exact convergence rate, under the assumption of a known posterior precision matrix. In turn, we can show that for independent random effects it is the relative informativity of the data that determines the convergence rate for the CP and the NCP. What is not so well understood is the effect that correlation across the latent variables has upon the convergence rates of the samplers for each parameterisation. Furthermore, how do we decide which parameterisation to use when we do not have access to the exact convergence rate, i.e. when the posterior precision matrix is unknown, as is the case in practice, or when we have a non-Gaussian model for the data.

In this thesis we look to address the following questions:

(i) How does the presence of spatial correlation across the latent variables impact the convergence rate for the CP and the NCP of a hierarchical model?

(ii) Can we develop a robust fitting strategy for hierarchical models that has convergence properties that are independent of the data, and hence can be routinely implemented?

These questions are of particular importance when one considers the proliferation of open source software for fitting spatial models that implement MCMC algorithms.

1.2 Thesis contribution

The contribution of this thesis can be placed into two broad categories in line with the questions posed in Section 1.1.

(i) We extend the current knowledge of the conditions in which hierarchical centering leads to a more efficient Gibbs sampler to include the impact of the correlation structure across the random effects, and therefore add weight to the notion that the CP and the NCP form a complimentary pair.

For NLHMs it is known that the relative informativity of the data dictates which of the CP or the NCP will yield the most efficient Gibbs sampler. When the data precision is relatively high, the CP will perform best and when it is low, the NCP should be implemented. In the case of Gaussian posterior distributions with known precision matrices, we are able to compute the exact convergence rate for the Gibbs sampler. We show that for an exponential correlation function that the convergence rate associated with the CP is hastened with increasing strength of spatial correlation, with the opposing effect seen for the sampler associated with the NCP. We are also able to show that covariance tapering hinders the CP whereas it helps the NCP, and that introducing geometric anisotropy by strengthening the spatial correlation in one direction, helps the CP and hinders the NCP.

When the posterior precision matrices are unknown, or the target distributions are not Gaussian, we run Gibbs samplers on simulated and real data and use well known diagnostic tests to compare the parameterisations. We find that for unknown covariance parameters that the CP is favourable to the NCP not only when the data precision is relatively high, but also when the correlation across the random effects is strong. We see this result for non-Gaussian data as well.

(ii) We develop a dynamically updated, partially centred parameterisation (PCP) that is shown to be robust to the data and outperform the CP and the NCP.

By minimising the posterior covariance of the random and global effects, we are able to parameterise the model in such a way that convergence for the associated Gibbs sampler is immediate. The construction is conditioned on the covariance matrices in the model. When the covariance matrices are known only up to a set of covariance parameters, we show that the parameterisation can be dynamically updated within

the Gibbs sampler. We develop a pilot adapted PCP that reduces the computational expense and still retains the good mixing and convergence properties of the PCP.

1.3 A review of computational strategies for hierarchical Bayesian models

Hierarchical Bayesian models provide a coherent framework to model stochastic processes. They allow for the incorporation of prior knowledge and properly account for uncertainties at different levels of the model. Consequently they have 'taken over the landscape in contemporary stochastic modelling' (Gelfand, 2012).

A commonly used schematic representation of the hierarchical structure follows that laid out by Berliner (1996). Broadly we have the following distributional specifications:

Stage 1. $[data|process, parameters]$

Stage 2. $[process|parameters]$

Stage 3. $[parameters]$.

In general, each stage may have further sub-levels, enriching the model where appropriate.

Bayesian inference allows us to make probability statements about any quantity of interest, but typically we must employ MCMC techniques (Robert and Casella, 2004). High posterior correlation and weak identifiability of the model parameters can lead to slow convergence and poor mixing of the Markov chains. In this section we review some of the techniques that have been developed to address the problems that arise when using MCMC to fit hierarchical models. We begin in Section 1.3.1 with an overview of the fitting strategies that have been used for hierarchical models in a range of contexts. In Section 1.3.2 we consider the response of the statistics community to the particular challenges that are faced by fitting spatial models within a Bayesian framework.

1.3.1 General fitting strategies for hierarchical models

The parameterisation of a hierarchical model plays an important role in the efficiency of any MCMC algorithm employed for inference. Often some form of reparameterisation can help yield better behaved chains, where by *well behaved* we mean chains that explore the parameter space well and converge quickly. Considered part of the art of MCMC, it is often left to the practitioner to discover for themselves the most effective model parameterisation.

Gelfand et al. (1995) show that for NLHMs and linear mixed models (Laird and Ware, 1982) simple hierarchical centering can significantly improve the efficiency of the sampling algorithm. The accompanying paper, Gelfand et al. (1996), extends the result to generalized linear mixed models (GLMMs) (Dey et al., 2000; Breslow and Clayton, 1993), and demonstrates the effectiveness of centering for Poisson and binary data models.

To illustrate centering consider the following simple model taken from Gelfand et al.

(1996, Section 2). Let

$$Y_i = \theta + U_i + \epsilon_i, \tag{1.1}$$

with $U_i \sim N(0, \sigma_u^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independently distributed for all $i = 1, \ldots, n$. The form of the model given by (1.1) is what Gelfand et al. (1995) call *uncentred* but following more recent literature we will refer to it as the non-centred parameterisation (NCP). Consider the variable $\tilde{U}_i = U_i + \theta$. Replacing U_i with $\tilde{U}_i$ in (1.1) gives

$$Y_i = \tilde{U}_i + \epsilon_i, \tag{1.2}$$

and $\tilde{U}_i \sim N(\theta, \sigma_u^2)$, hence $\tilde{U}_i$ is centred on θ and thus (1.2) is referred to as the centred parameterisation (CP).

Assuming a constant prior distribution for θ, and that σ_u^2 and σ_ϵ^2 are known, Gelfand et al. (1996) compute the posterior correlation of the centred random effects $\tilde{U}_i$ and the global effect θ as

$$Corr(\tilde{U}_i, \theta|\boldsymbol{y}) = \left(1 + \frac{n\sigma_u^2}{\sigma_\epsilon^2}\right)^{-1/2} \quad \text{and} \quad Corr(\tilde{U}_i, \tilde{U}_j|\boldsymbol{y}) = \left(1 + \frac{n\sigma_u^2}{\sigma_\epsilon^2}\right)^{-1}, \tag{1.3}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is the vector of observed data. From equation (1.3) we can see that for fixed n the posterior correlations between the model parameters are reduced as the ratio $\sigma_u^2/\sigma_\epsilon^2$ is increased. The equivalent expressions for the NCP are given by

$$Corr(U_i, \theta|\boldsymbol{y}) = -\left(1 + \frac{n\sigma_\epsilon^2}{\sigma_u^2}\right)^{-1/2} \quad \text{and} \quad Corr(U_i, U_j|\boldsymbol{y}) = \left(1 + \frac{n\sigma_\epsilon^2}{\sigma_u^2}\right)^{-1}, \tag{1.4}$$

and here the posterior correlations are reduced by decreasing $\sigma_u^2/\sigma_\epsilon^2$. Gelfand et al. (1995) argue that random effects are included in a model to 'soak up' the variability in the population effects model and so σ_u^2 will typically be greater than σ_ϵ^2, and therefore the CP is preferable to the NCP.

Equations (1.3) and (1.4) highlight two important features of the performance of the CP and the NCP. Firstly, that the ratio of the variance parameters is an important quantity in determining which parameterisation should be employed for model fitting, and secondly, that a change in variance ratio has opposing effects on each of the parameterisations.

Papaspiliopoulos et al. (2003, 2007) consider the NCP and the CP for a broad class of hierarchical models. They define the NCP to be a parameterisation such that the random and global effects are independent *a priori*. This is trivially satisfied by the NCP of model (1.1) but as they admit the construction of an NCP satisfying this condition may be hard to achieve for more complicated models. Roberts et al. (2004) use the NCP of the non-Gaussian Ornstein-Uhlenbeck process (Barndorff-Nielsen and Shephard, 2001) to construct an efficient algorithm for model fitting. Papaspiliopoulos et al. (2003) find that the NCP outperforms the CP for a Cauchy data model with Gaussian latent variables. Papaspiliopoulos and Roberts (2008) further investigate how the model parameterisation and the tail behaviour of the distributions of the data and the latent process all interact to determine the stability of the Gibbs sampler. They look at combinations of Cauchy,

double exponential, Gaussian and exponential power distributions for the CP and the NCP. The heuristic remark that follows from this comparison is that the convergence of the CP is quickest when the data model has lighter tails than that of the latent variables, with the opposite scenario favouring the NCP.

In many cases where one parameterisation does well the other does poorly. To try and take advantage of this dichotomy Yu and Meng (2011) develop a strategy to combine the NCP and the CP. Their *interweaving algorithm* is particularly useful when the practitioner has little knowledge of the convergence properties of either parameterisation. Suppose that the NCP and the CP have associated convergence rates of λ_{nc} and λ_c respectively, where the convergence rate is defined in Section 2.3. Yu and Meng (2011) show that the convergence rate associated with the interweaving algorithm, denoted λ_I, is related to λ_{nc} and λ_c via the following inequality:

$$\lambda_I \leq \mathscr{R}_{nc,c}\sqrt{\lambda_{nc}\lambda_c},$$

where $\mathscr{R}_{nc,c}$ is the maximal correlation between the latent variables under the two parameterisations. This implies that $\lambda_I \leq \max(\lambda_{nc}, \lambda_c)$ and hence the interweaving algorithm is more efficient than the worst of the NCP and the CP.

As an example of the interweaving algorithm, consider the NCP and the CP of the model above, equations (1.1) and (1.2) respectively. A standard Gibbs sampler (see Section 2.2.4) for the NCP alternates between drawing $\boldsymbol{U}^{(t)} \sim \pi(\boldsymbol{U}|\theta^{(t)}, \boldsymbol{y})$ and then drawing $\theta^{(t)} \sim \pi(\theta|\boldsymbol{U}^{(t)}, \boldsymbol{y})$ where $\boldsymbol{U} = (U_1, \ldots, U_n)'$. It is easily shown that

$$\boldsymbol{U}|\theta, \boldsymbol{y} \sim N\left(\frac{\sigma_u^2(\boldsymbol{y} - \theta\mathbf{1})}{\sigma_\epsilon^2 + \sigma_u^2}, \frac{\sigma_\epsilon^2\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2}\boldsymbol{I}\right), \quad \text{and} \quad \theta|\boldsymbol{U}, \boldsymbol{y} \sim N\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - U_i), \frac{\sigma_\epsilon^2}{n}\right),$$

where $\mathbf{1}$ is a $n \times 1$ vector of ones and $\boldsymbol{I}$ is the identity matrix of order n. The equivalent algorithm for the CP draws $\tilde{\boldsymbol{U}}^{(t)} \sim \pi(\tilde{\boldsymbol{U}}|\theta^{(t)}, \boldsymbol{y})$ and then $\theta^{(t)} \sim \pi(\theta|\tilde{\boldsymbol{U}}^{(t)}, \boldsymbol{y})$, where

$$\tilde{\boldsymbol{U}}|\theta, \boldsymbol{y} \sim N\left(\frac{\sigma_u^2\boldsymbol{y} + \sigma_\epsilon^2\theta\mathbf{1}}{\sigma_\epsilon^2 + \sigma_u^2}, \frac{\sigma_\epsilon^2\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2}\boldsymbol{I}\right) \quad \text{and} \quad \theta|\tilde{\boldsymbol{U}}, \boldsymbol{y} \sim N\left(\frac{1}{n}\sum_{i=1}^{n}\tilde{U}_i, \frac{\sigma_u^2}{n}\right).$$

Both parameterisations have the same target distribution, $\pi(\theta|\boldsymbol{y})$, but typically different convergence rates. The interweaving algorithm combines the two sampling algorithms as follows:

Algorithm 1 The interweaving algorithm

Given $\theta^{(t)}$,
Step 1. Draw $\boldsymbol{U}^{(t)} \sim \pi(\boldsymbol{U}|\theta^{(t)}, \boldsymbol{y})$.
Step 2. Draw $\theta^{(t+0.5)} \sim \pi(\theta|\boldsymbol{U}^{(t)}, \boldsymbol{y})$.
Step $\tilde{2}$. Draw $\tilde{\boldsymbol{U}}^{(t+1)} \sim \pi(\tilde{\boldsymbol{U}}|\theta^{(t+0.5)}, \boldsymbol{U}^{(t)}, \boldsymbol{y})$.
Step 3. Draw $\theta^{(t+1)} \sim \pi(\theta|\tilde{\boldsymbol{U}}^{(t+1)}, \boldsymbol{y})$.

The intermediate draw $\theta^{(t+0.5)}$ can be discarded. Often, $\boldsymbol{U}$ and $\tilde{\boldsymbol{U}}$ are related via some

deterministic function and Step $\tilde{2}$ requires a trivial transformation. In this example, $\tilde{\boldsymbol{U}}^{(t+1)} = \boldsymbol{U}^{(t)} + \theta^{(t+0.5)}\mathbf{1}$. It is by re-sampling θ, conditional on the constant vector $\tilde{\boldsymbol{U}} = \boldsymbol{U} + \theta\mathbf{1}$, that the Markovian dependence between successive iterates for θ is reduced.

Papaspiliopoulos (2003) views the CP and the NCP as extremes of a continuum of partially centred parameterisations (PCPs). They develop a PCP for the NLHM and show that for known covariance matrices the resulting Gibbs sampler converges immediately. We let $U_i^w = U_i + w\theta$ for $w \in [0, 1]$. Substituting U_i^w into model (1.1) we have

$$\begin{aligned} Y_i &\sim N((1-w)\theta + U_i^w, \sigma_\epsilon^2) \\ U_i^w &\sim N(w\theta, \sigma_u^2). \end{aligned}$$

Clearly the NCP is recovered when $w = 0$ and the CP recovered when $w = 1$.

It can be shown that

$$Corr(U_i^w, \theta|\boldsymbol{y}) = \frac{w\sigma_\epsilon^2 - (1-w)\sigma_u^2}{([w\sigma_\epsilon^2 - (1-w)\sigma_u^2]^2 + n\sigma_\epsilon^2\sigma_u^2)^{1/2}}. \tag{1.5}$$

We can see from (1.5) that the $Corr(U_i^w, \theta|\boldsymbol{y}) = 0$ when $w\sigma_\epsilon^2 - (1-w)\sigma_u^2 = 0$. This implies that we should set

$$w = \frac{\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2},$$

to minimise the posterior correlation between U_i^w and θ. We discuss PCPs in the context of spatial models in Chapter 5.

1.3.2 Fitting strategies for spatial models

Spatially referenced data sets arise in many diverse areas such as environmetrics/ecology, hydrology, meteorology and many others. The availability of such data has driven a considerable effort to develop statistical models for the observed processes (Cressie and Wikle, 2011; Gelfand et al., 2010; Schabenberger and Gotway, 2004; Banerjee et al., 2003). Bayesian hierarchical modelling provides a natural framework to properly assess the uncertainty in the parameter estimates and spatial predictions. It is common to use a Gaussian processes at the second stage of the hierarchy to model the latent spatial structure in the observable data. As noted by Gelfand et al. (2003) 'the literature here is enormous' and they recommend Cressie (1993) as a place to start.

Conditional independencies determined by the hierarchical structure of the model facilitate the construction of Gibbs sampling type algorithms for model fitting (Gelfand and Smith, 1990). A requirement of these algorithms is the repeated inversion of dense $n \times n$ covariance matrices, an operation of order $O(n^3)$ in computational complexity, for n spatial locations (Cressie and Johannesson, 2008). This, coupled with high posterior correlation between model parameters and weakly identified covariance parameters, means that the problems of MCMC are sharpened for spatial models.

To illustrate some of the fitting strategies developed for spatial models consider the following standard hierarchical model for univariate spatial data, (Cressie, 1993). For

spatial locations $\boldsymbol{s}$ within the spatial domain $\mathscr{D} \subseteq \mathscr{R}^2$, responses $Y(\boldsymbol{s})$ are modelled as

$$Y(\boldsymbol{s}) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\theta} + \beta(\boldsymbol{s}) + \epsilon(\boldsymbol{s}), \tag{1.6}$$

where $\boldsymbol{x}(\boldsymbol{s})$ is a $p \times 1$ vector of spatially referenced covariates and $\boldsymbol{\theta}$ is a vector of p regression coefficients. The residual has two components. The first, $\beta(\boldsymbol{s})$, is a realisation of a zero mean Gaussian process capturing the spatial structure unexplained by the covariates. We have that $E[\beta(\boldsymbol{s})] = 0$, $Var(\beta(\boldsymbol{s})) = \sigma^2_\beta$ and $Cov(\beta(\boldsymbol{s}), \beta(\boldsymbol{s}^*)) = \sigma^2_\beta \rho(\boldsymbol{s}, \boldsymbol{s}^*; \boldsymbol{\phi})$, where $\rho(\cdot, \cdot; \boldsymbol{\phi})$ is a valid two-dimensional correlation function known up to correlation parameters $\boldsymbol{\phi}$. The second, $\epsilon(\boldsymbol{s})$, is a non-spatial, pure error term referred to as the *nugget* in the geostatistics literature, which is assumed to be independent and normally distributed with mean zero and variance σ^2_ϵ for all $\boldsymbol{s}$.

For locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ we have

$$\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))' \quad \text{and} \quad \boldsymbol{\beta} = (\beta(\boldsymbol{s}_1), \ldots, \beta(\boldsymbol{s}_n))'.$$

Putting a flat prior on $\boldsymbol{\theta}$ we can write the NCP of the model (1.6) as

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\theta}, \boldsymbol{\beta} &\sim N(\boldsymbol{X}\boldsymbol{\theta} + \boldsymbol{\beta}, \sigma^2_\epsilon \boldsymbol{I}) \\ \boldsymbol{\beta} &\sim N(\boldsymbol{0}, \sigma^2_\beta \boldsymbol{R}), \end{aligned}$$

where $\boldsymbol{X}$ is an $n \times p$ matrix with ith row equal to $\boldsymbol{x}'(\boldsymbol{s}_i)$, for $i = 1, \ldots, n$, $\boldsymbol{0}$ is an $n \times 1$ vector of zeros and correlation matrix $\boldsymbol{R}$ has ijth entry equal to $\rho(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi})$.

Now define $\tilde{\beta}(\boldsymbol{s}) = \beta(\boldsymbol{s}) + \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\theta}$ or equivalently $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{\theta}$, then the CP for model (1.6) is given by

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}} &\sim N(\tilde{\boldsymbol{\beta}}, \sigma^2_\epsilon \boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}\boldsymbol{\theta}, \sigma^2_\beta \boldsymbol{R}). \end{aligned}$$

It is most common to see the standard spatial model given in its non-centred form (Gelfand et al., 2010; Banerjee et al., 2003), as it is in (1.6). This enables the reader to distinguish between the spatial and non-spatial parts of the residual, but gives no hint as to the best parameterisation to use for model fitting, and so presumably the practitioner will turn to the NCP by default. However, the R package spTimer (Bakar and Sahu, 2015), developed to fit latent Gaussian process regression models to large space-time data sets, use the CP of their model specification. Berrocal et al. (2010) regress ozone concentration data upon the output of a numerical model, allowing for a spatially varying intercept and slope, and they too fit the CP of their model. In either case, the reason for the choice of parameterisation is not stated. Is it ease of programmability or improved convergence or some other reason? The effect of hierarchical centering in the presence of spatial correlated random effects is investigated in Chapters 3 and 4.

The interweaving algorithm described in Section 1.3.1 is designed to work on any two model parameterisations, and so we can use it to interweave the CP and the NCP of the spatial model given in (1.6). Neal and Roberts (2005) combine the CP and the NCP of

stochastic epidemic models such that at each MCMC iterate some of the latent variables are centred and some are non-centred.

An alternative approach is that of *marginalisation*. Due to the normality assumption of both spatial and non-spatial error processes, we can integrate $\boldsymbol{\beta}$ (or $\tilde{\boldsymbol{\beta}}$) out of the likelihood. This reduces the dimension of the joint posterior distribution. It is argued by Banerjee et al. (2003) that marginalisation makes the covariance structure more computationally stable. For either the CP or the NCP the marginalised likelihood is given by

$$\boldsymbol{Y} \sim N(\boldsymbol{X\theta}, \sigma_\epsilon^2 \boldsymbol{I} + \sigma_\beta^2 \boldsymbol{R}).$$

Marginalised likelihoods are used by Gelfand et al. (2003) for fitting spatially varying coefficient regression models and by Banerjee et al. (2008) to implement Gaussian predictive process models.

The R package spBayes (Finley et al., 2007) uses a marginalised likelihood for fitting univariate and multivariate spatial models. Under the reformulation of their core functions they marginalise over all of the random and global effects in the model. Samples are obtained from the marginal posterior distributions of the variance and correlation parameters which are then used to recover samples from the posterior distributions of the global and spatial effects as the user desires, see Finley et al. (2015, Section 2.2) for details.

The strategies discussed so far are concerned with the mean structure. Diggle and Ribeiro Jr (2002) reparameterise the variance parameters to model the relative nugget variance, defined as $v^2 = \sigma_\epsilon^2/\sigma_\beta^2$ and hence a marginalised likelihood of

$$\boldsymbol{Y} \sim N(\boldsymbol{X\theta}, \sigma_\beta^2[v^2 \boldsymbol{I} + \boldsymbol{R}]),$$

is employed. Yan et al. (2007) prefer to consider the relative contribution of the σ_ϵ^2 to the total variation in $\boldsymbol{Y}$. They let $\sigma^2 = \sigma_\epsilon^2 + \sigma_\beta^2$ and $\zeta = \sigma_\epsilon^2/\sigma^2$. Therefore, the marginalised likelihood is written as

$$\boldsymbol{Y} \sim N(\boldsymbol{X\theta}, \sigma^2[\zeta \boldsymbol{I} + (1-\zeta)\boldsymbol{R}]). \tag{1.7}$$

Consequently ζ is bounded to $(0, 1)$ and this facilitates the use of slice sampling (Neal, 2003). The parameterisation given in (1.7) is implemented in the R package Smith et al. (2008) which allows for the joint modelling of data from different spatial scales.

For spatially referenced data for which the normality assumption is inappropriate we have spatial GLMMs (Diggle et al., 1998). In such cases marginalisation is no longer available as the necessary integrals are intractable. Papaspiliopoulos et al. (2003) develop a PCP for application to the Poisson log-normal model (Christensen and Waagepetersen, 2002). Their construction is analogous to the PCP for NLHMs given by Papaspiliopoulos (2003). The approach of Christensen et al. (2006) is to orthogonalise the random effects to remove correlation *a posteriori*. Papaspiliopoulos et al. (2003) and Christensen et al. (2006) rely on an approximation of the conditional posterior covariance matrix of the spatially correlated random effects and both use the quadratic expansion of the log-full conditional distribution for this purpose. We look at the parameterisation of non-Gaussian data models in Chapter 6

Hierarchical centering, interweaving and marginalisation are performed to create faster converging and better mixing Markov chains. When n is very large, say 10^5 or greater, the matrix operations required for model fitting are prohibitive regardless of the parameterisation. To cope with very large data sets some sort of dimension reduction must be performed. For a review of the techniques employed see Sun et al. (2012). A popular method is to use Gaussian predictive processes (GPP) Banerjee et al. (2008). The idea is to use realisations of the Gaussian process at $m << n$ knot locations to approximate $\boldsymbol{\beta}$. For a set of locations $\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_m^*$, where $\boldsymbol{\alpha} = (\alpha(\boldsymbol{s}_1^*), \ldots, \alpha(\boldsymbol{s}_m^*))'$ are realisations of the same process giving rise to $\boldsymbol{\beta}$, we have that $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are jointly distributed as

$$\begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix} \sim N\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \sigma_\beta^2 \begin{bmatrix} \boldsymbol{R}_\alpha & \boldsymbol{C}_{\alpha\beta} \\ \boldsymbol{C}_{\beta\alpha} & \boldsymbol{R} \end{bmatrix} \right),$$

where $\boldsymbol{R}_\alpha$ is an $m \times m$ correlation matrix with ijth entry equal to $\rho(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*; \boldsymbol{\phi})$, for $i, j = 1, \ldots, m$, and $\boldsymbol{C}_{\beta\alpha}$ is an $n \times m$ cross correlation matrix with ijth entry equal to $\rho(\boldsymbol{s}_i, \boldsymbol{s}_j^*; \boldsymbol{\phi})$, for $i = 1, \ldots, n$, $j = 1, \ldots, m$. We estimate $\boldsymbol{\beta}$ by $\boldsymbol{\beta}^*$, where

$$\boldsymbol{\beta}^* = E[\boldsymbol{\beta}|\boldsymbol{\alpha}] = \boldsymbol{C}_{\beta\alpha}\boldsymbol{R}_\alpha^{-1}\boldsymbol{\alpha}.$$

Model (1.6) is then replaced by the GPP model

$$Y(\boldsymbol{s}) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\theta} + \beta^*(\boldsymbol{s}) + \epsilon(\boldsymbol{s}),$$

and consequently we have to invert matrices of order m and not n. Moreover, if the knot locations form a regular lattice, $\boldsymbol{R}_\alpha$ is block circulant and can be inverted efficiently using the discrete Fourier transform (Gray, 2005). Guhaniyogi et al. (2011) model the knots to allow their locations to be stochastically adapted. GPP models are used by Sahu and Bakar (2012) and can be implemented in R packages spBayes and spTimer. GPPs have also been used for nonparameteric regression models (Banerjee et al., 2012).

Another approach is *covariance tapering* (Furrer et al., 2006). Sparsity in the covariance matrix is induced by forcing to zero those entries corresponding to pairs of locations that are separated by a distance greater than some threshold range. The new tapered correlation matrix $\boldsymbol{R}_{Tap}$ is formed by taking the element wise (or Schur) product of the original correlation matrix $\boldsymbol{R}$ and a positive definite tapering matrix $\boldsymbol{T}$, where the ijth entry of $\boldsymbol{T}$ is zero if $\|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ exceeds some desired range. The tapered matrix $\boldsymbol{R}_{Tap}$ is then a banded matrix and can be inverted efficiently using sparse matrix libraries i.e. the R package Koenker and Ng (2003). We look at covariance tapering in greater detail in Chapter 3 and its impact on the efficiency of Gibbs samplers for the CP and the NCP.

The GPP captures large-scale spatial variation whereas the covariance tapering approach captures variation over shorter ranges. Sang and Huang (2012) combine the two methods by writing

$$\beta(\boldsymbol{s}) = \beta^*(\boldsymbol{s}) + \beta_s(\boldsymbol{s}),$$

where $\beta_s(\boldsymbol{s}) = \beta(\boldsymbol{s}) - \beta^*(\boldsymbol{s})$ is the residual of the GPP approximation. They use a tapered

covariance matrix for the residual process $\beta_s(\boldsymbol{s})$ that attempts to capture the small-scale variation that is neglected by the GPP approximation.

Somewhat related to the notion of tapering is the methodology of integrated nested Lapacian approximation (INLA) (Rue et al., 2009). Through an armoury of efficient coding, covariance approximations and deterministic mode finding algorithms, they allow for the fitting of a broad class of latent Gaussian process models to high dimensional data. By using stochastic partial differential equations they represent the Gaussian process as a Gaussian Markov random field, thus creating very sparse correlation matrices (Lindgren et al., 2011). Although INLA does not provide full Bayesian inference it does offer a very powerful tool for fitting spatial models.

1.4 Thesis organisation

The work in this thesis is organised as follows:

- In Chapter 2 we give background information about Bayesian computation and provide details of the three stage normal linear hierarchical model which motivates the work of Chapters 3-5.

- Chapter 3 introduces a general spatial model that has spatially varying coefficients which are realisations of Gaussian processes. We analyse the exact convergence rates of the Gibbs samplers emitted by the CP and the NCP for different covariance structures of the latent variables. We compare the effect of the scale of a covariate upon the convergence rates for the different parameterisations. We also look at the impact of covariance tapering and blocking strategies for updating the model parameters in the Gibbs sampler.

- In Chapter 4 we look at the practical implementation of the CP and the NCP and give details of the full conditional distributions needed to construct the respective Gibbs samplers. Further details are given for the procedure required to sample from the posterior predictive distribution, a procedure necessary in order to construct predictive maps. We use simulated and real data examples to compare the performance of the Gibbs sampler for both parameterisations, where performance is judged by well known diagnostic tests.

- In Chapter 5 we construct a PCP which eliminates the conditional posterior correlation between random and global effects and consequently gives rise to a Gibbs sampler with immediate convergence. We investigate how the weights of partial centering are affected by the covariance parameters and how they vary over space. We show how the PCP can be dynamically updated within the Gibbs sampler and demonstrate the efficacy of pilot adaption schemes that are used to mitigate the computational expense of the PCP. The PCP is compared to the CP and the NCP for both simulated and real data examples and is shown to be robust to changes in the covariance parameters of the data generating mechanism.

- In chapter 6 we look at two frequently used models for non-Gaussian spatial data; the Tobit model, which is applied to censored data, and the probit model, which is applied to binary data. We compare the performance of the CP and the NCP for these models using simulated and real data examples. We fit the Tobit model to New York precipitation data and produce a map of the probability of positive precipitation across New York. The probit model is used to construct a map of the probability that ozone concentrations across California exceed the limits set forth by the U.S. Environmental Protection Agency. We go on to construct a PCP for non-Gaussian data by using the negated Hessian matrix evaluated at the MLE to estimate the conditional posterior covariance of the random effects. The properties of the PCP are investigated and compared to the CP and the NCP for both Tobit and probit models.

- Chapter 7 contains a discussion of the results of the preceding chapters and sets an agenda for future work which includes extensions to spatio-temporal data and multivariate responses.

Chapter 2

Bayesian computation

2.1 Introduction

Bayes theorem is the basic tool for Bayesian statistics. It allows us to update our prior beliefs about a set of unknown parameters $\boldsymbol{\xi} \in \boldsymbol{\Xi} \subset \mathscr{R}^d$ in the light of observed data $\boldsymbol{y} = (y_1, \ldots, y_n)'$. Given a *likelihood* $f(\boldsymbol{y}|\boldsymbol{\xi})$ and a *prior distribution* $\pi(\boldsymbol{\xi})$ for the parameters, Bayes theorem tells us that the *posterior distribution* for $\boldsymbol{\xi}$ is

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) = \frac{f(\boldsymbol{y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi})}{f(\boldsymbol{y})} = \frac{f(\boldsymbol{y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi})}{\int_{\boldsymbol{\Xi}} f(\boldsymbol{y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi})d\boldsymbol{\xi}}, \tag{2.1}$$

where f($\boldsymbol{y}$) is the marginal distribution of the data and the normalising constant for the posterior distribution of $\boldsymbol{\xi}$. Hence we can write

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) \propto f(\boldsymbol{y}|\boldsymbol{\xi})\pi(\boldsymbol{\xi}),$$

which says that the posterior distribution for $\boldsymbol{\xi}$ is proportional to the product of the likelihood and the prior distribution.

Assuming that we can write down a likelihood for the model, we must then choose a prior distribution for $\boldsymbol{\xi}$. Elicitation of prior distributions is an important part of any Bayesian analysis. The choice should reflect all of our prior beliefs about the model parameters. It should respect their support and assign greater mass to intervals of that support where we believe it to be appropriate. In practice, we often have very little prior information on which to make these decisions. This encourages the use of *non-informative* priors, sometimes referred to as *vague* priors.

An example of a vague prior is the uniform prior distribution. We consider all values of $\boldsymbol{\xi}$ equally likely *a priori* such that $\pi(\boldsymbol{\xi}) \propto constant$. One problem with this prior is that under a different parameterisation the prior distribution may no longer be uniform. To combat this one can employ a *Jeffreys prior* which is proportional to the square root of the *Fisher's information matrix* and is invariant to parameterisation. A second problem with a uniform prior is that it is *improper* as it does not integrate to one. Although improper prior distributions can induce proper posterior distributions (see Gelman et al., 2004, Chapter 2) it is not always ensured. For this reason many practitioners use proper

priors with large variability.

For a proper prior distribution, it is computationally convenient to use a *conjugate* prior. Suppose $\pi(\boldsymbol{\xi})$ comes from a family $\mathscr{F}$ of distributions and that we have a random sample from density $f(y|\boldsymbol{\xi})$. If $\pi(\boldsymbol{\xi}|\boldsymbol{y}) \in \mathscr{F}$ for all $f(\cdot|\boldsymbol{\xi})$ we say that $\pi(\cdot)$ is a conjugate prior with respect to f. If f is from the exponential family then we can always find a conjugate prior, but in general one may not exist. A conjugate prior should only be used if it can be reasonably justified in terms of our prior beliefs.

The choice of prior distributions is subjective and as such we must check how sensitive our inference is to this choice. Robustness to prior misspecification can be checked via posterior predictive performance, see Section 2.5.

With a likelihood and prior we can compute the posterior distribution. Access to the posterior distribution allows us to make probability statements about $\boldsymbol{\xi}$. We can compute moments, quantiles and test hypotheses. We often want to calculate posterior expectations of some function $g(\cdot)$. This requires evaluation of the expression

$$E[g(\boldsymbol{\xi})|\boldsymbol{y}] = \frac{\int_{\boldsymbol{\Xi}} g(\boldsymbol{\xi}) f(\boldsymbol{y}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi}}{\int_{\boldsymbol{\Xi}} f(\boldsymbol{y}|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}) d\boldsymbol{\xi}}. \tag{2.2}$$

In most applications closed form solutions for the integrals in (2.2) cannot be found. Numerical integration is an option but becomes unreliable for high-dimensional problems. For this reason we use sampling techniques like Markov chain Monte Carlo (MCMC) to estimate these expectations.

If $\boldsymbol{\xi}^{(t)}$, $t = 1, \ldots, N$, are independent samples from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ then by the law of large numbers

$$\bar{g}_N = \frac{1}{N} \sum_{t=1}^{N} g(\boldsymbol{\xi}^{(t)}) \to E[g(\boldsymbol{\xi})|\boldsymbol{y}] \quad \text{as} \quad N \to \infty. \tag{2.3}$$

Obtaining independent samples from the posterior distribution may not be possible. However, if $\{\boldsymbol{\xi}^{(t)}\}_{t=1}^{N}$ form an ergodic Markov chain with stationary distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$, then under suitable regularity conditions (2.3) still holds (Smith and Roberts, 1993). Therefore, our problem is now to construct a Markov chain with the desired stationary distribution. This can be achieved by using the *Metropolis-Hastings*, *Metropolis* or *Gibbs sampling* algorithms, as described below.

2.2 The Metropolis-Hastings algorithm

A widely used technique for sampling from high-dimensional distributions is the Metropolis-Hastings (M-H) algorithm (Hastings, 1970). It allows us to induce a Markov chain with a stationary distribution equal to the posterior distribution in (2.1). Given the current state of the chain, a candidate value for the next state is drawn from an easy to sample proposal distribution $q(\cdot|\cdot)$. If accepted, the chain jumps to the candidate value. If rejected, the chain remains in its current state. The goal is to explore the parameter space, retaining values in the correct proportions with respect to the target density $\pi(\boldsymbol{\xi}|\boldsymbol{y})$.

The M-H algorithm proceeds as follows:

1. Choose an initial value $\boldsymbol{\xi}^{(0)}$.

2. Given the current value $\boldsymbol{\xi}^{(t)}$, sample a candidate value $\boldsymbol{\xi}^*$ from a proposal distribution $q(\cdot|\boldsymbol{\xi}^{(t)})$.

3. Calculate the acceptance probability

$$\alpha(\boldsymbol{\xi}^{(t)}, \boldsymbol{\xi}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\xi}^*|\boldsymbol{y})q(\boldsymbol{\xi}^{(t)}|\boldsymbol{\xi}^*)}{\pi(\boldsymbol{\xi}^{(t)}|\boldsymbol{y})q(\boldsymbol{\xi}^*|\boldsymbol{\xi}^{(t)})}\right\},$$

4. Draw a value u from a uniform U(0,1) distribution.

5. Let

$$\boldsymbol{\xi}^{(t+1)} = \begin{cases} \boldsymbol{\xi}^* & \text{if } u \leq \alpha(\boldsymbol{\xi}^{(t)}, \boldsymbol{\xi}^*), \\ \boldsymbol{\xi}^{(t)} & \text{if } u > \alpha(\boldsymbol{\xi}^{(t)}, \boldsymbol{\xi}^*). \end{cases}$$

As the acceptance probability is in the form of a ratio, we do not need to compute the normalising constant, $f(\boldsymbol{y})$. Note also that since there is positive probability of rejecting a move the chain will be aperiodic and so the stationary distribution will also be the limiting distribution. Furthermore, it can be shown that as long as $q(\cdot|\cdot)$ and $\pi(\cdot)$ have the same support the resulting chain will have a stationary distribution equal to $\pi(\cdot)$, (see Gilks et al., 1996, Chapter 1). For more details regarding the implementation of the M-H algorithm see Chib and Greenberg (1995).

2.2.1 The Metropolis algorithm

The M-H algorithm is an extension of the Metropolis algorithm, developed for applications in statistical physics (Metropolis et al., 1953). If we consider a symmetric proposal distribution where $q(\cdot|\boldsymbol{\xi}) = q(\boldsymbol{\xi}|\cdot)$, e.g. a multivariate normal distribution, then we recover the Metropolis algorithm. The acceptance probability becomes

$$\alpha(\boldsymbol{\xi}^{(t)}, \boldsymbol{\xi}^*) = \min\left\{1, \frac{\pi(\boldsymbol{\xi}^*|\boldsymbol{y})}{\pi(\boldsymbol{\xi}^{(t)}|\boldsymbol{y})}\right\},$$

and so if a more likely value is proposed it will always be accepted.

2.2.2 Component-wise updating algorithms

It is more computationally efficient to sample from lower dimensional random variates. Therefore it is common to partition $\boldsymbol{\xi}$ into s components of dimension r_i, $i = 1, \ldots, s$, such that $\sum_{i=1}^{s} r_i = d$, and update each of these in turn. Components are usually univariate quantities but highly correlated random variables may be grouped together, a practice known as *blocking*.

The posterior distribution of a component $\boldsymbol{\xi}_i$ given all others, $\boldsymbol{\xi}_{-i}$, where

$$\boldsymbol{\xi}_{-i} = \left(\boldsymbol{\xi}_1', \ldots, \boldsymbol{\xi}_{i-1}', \boldsymbol{\xi}_{i+1}', \ldots, \boldsymbol{\xi}_s'\right)',$$

is known as the *full conditional distribution*, denoted by $\pi(\boldsymbol{\xi}_i|\boldsymbol{\xi}_{-i}, \boldsymbol{y})$. The set of full conditional distributions uniquely determines the joint distribution (Besag, 1974). It follows that if

$$\boldsymbol{\xi}_i^{(t)} \sim \pi(\boldsymbol{\xi}_i|\boldsymbol{\xi}_{-i}, \boldsymbol{y}), \quad i = 1, \ldots, s,$$

then $\left(\boldsymbol{\xi}_1^{(t)}, \ldots, \boldsymbol{\xi}_s^{(t)}\right)'$ is a sample from the joint distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. Moreover, the sequences $\left\{\boldsymbol{\xi}_i^{(t)}\right\}_{t=1}^{N}$ are samples from marginal distributions $\pi(\boldsymbol{\xi}_i|\boldsymbol{y})$.

We can identify the full conditional distributions by noting that

$$\pi(\boldsymbol{\xi}_i|\boldsymbol{\xi}_{-i}, \boldsymbol{y}) = \frac{\pi(\boldsymbol{\xi}|\boldsymbol{y})}{\int_{\boldsymbol{\Xi}} \pi(\boldsymbol{\xi}|\boldsymbol{y}) d\boldsymbol{\xi}_i},$$

and so $\pi(\boldsymbol{\xi}_i|\boldsymbol{\xi}_{-i}, \boldsymbol{y}) \propto \pi(\boldsymbol{\xi})$. In addition, a graphical representation of the model in the form of a *directed acyclic graph* allows us to read off any conditional independencies (Spiegelhalter et al., 1993).

We perform s cycles of the M-H algorithm to obtain a sample from the joint posterior distribution. Let $q_i(\cdot|\cdot)$ be the proposal distribution for the ith component. A candidate value $\boldsymbol{\xi}_i^*$ is then accepted with probability

$$\alpha_i(\boldsymbol{\xi}_i^{(t)}, \boldsymbol{\xi}_i^*) = \min\left\{1, \frac{\pi(\boldsymbol{\xi}_i^*|\boldsymbol{\xi}_{-i}^{(t)}, \boldsymbol{y}) q_i(\boldsymbol{\xi}_i^{(t)}|\boldsymbol{\xi}_i^*, \boldsymbol{\xi}_{-i}^{(t)})}{\pi(\boldsymbol{\xi}_i^{(t)}|\boldsymbol{\xi}_{-i}^{(t)}, \boldsymbol{y}) q_i(\boldsymbol{\xi}_i^*|\boldsymbol{\xi}_i^{(t)}, \boldsymbol{\xi}_{-i}^{(t)})}\right\}.$$

where $\boldsymbol{\xi}_{-i}^{(t)} = \left(\boldsymbol{\xi}_1^{(t+1)}, \ldots, \boldsymbol{\xi}_{i-1}^{(t+1)}, \boldsymbol{\xi}_{i+1}^{(t)}, \ldots, \boldsymbol{\xi}_s^{(t)}\right)'$. Notice that we always condition on the latest values for the other components.

2.2.3 Acceptance rates

The acceptance rate impacts the rate of convergence of the Markov chain to its stationary distribution. By tuning the scale of the proposal distribution we control the proportion of candidate values that are accepted. If the scaling parameter is too small then we are more likely to propose small jumps. These will be accepted with high probability. However, many iterations will be needed to explore the entire parameter space. On the other hand, if the scale is too large then many of the proposed jumps will be to areas of low posterior density and will be rejected. This again will lead to slow exploration of the space.

For Metropolis algorithms, if $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ can be factorised into IID components we have the asymptotic result that as $d \to \infty$ the optimal acceptance rate is 0.234 (Roberts et al., 1997; Gelman et al., 1996). Neal and Roberts (2006) show that if we partition $\boldsymbol{\xi}$ into components of equal dimension i.e. $r = r_1 = r_2 = \ldots = r_s$, then the optimal acceptance rate is independent of r. For $r = 1$ Gelman et al. (1996) showed that the optimal acceptance rate is 0.44, and so for univariate components-wise updating algorithms we look to tune the proposals distribution to achieve this rate.

2.2.4 The Gibbs sampler

If the full conditional distributions can be sampled from directly then the algorithm is referred to as a *Gibbs sampler* (Gelfand and Smith, 1990; Geman and Geman, 1984). As

pointed out by Gelman (1992) the Gibbs sampler can be considered a special case of the M-H algorithm where the proposal density is equal to that of the target density, hence candidate values are accepted with probability one. This makes Gibbs samplers easy to program and quick to run. Given starting values $\boldsymbol{\xi}^{(0)}$, the Gibbs sampler cycles through the following steps for $t = 0, \ldots, N-1$:

$$\begin{aligned}
\boldsymbol{\xi}_1^{(t+1)} &\sim \pi(\boldsymbol{\xi}_1 | \boldsymbol{\xi}_2^{(t)}, \boldsymbol{\xi}_3^{(t)}, \ldots, \boldsymbol{\xi}_s^{(t)}, \boldsymbol{y}) \\
\boldsymbol{\xi}_2^{(t+1)} &\sim \pi(\boldsymbol{\xi}_2 | \boldsymbol{\xi}_1^{(t+1)}, \boldsymbol{\xi}_3^{(t)}, \ldots, \boldsymbol{\xi}_s^{(t)}, \boldsymbol{y}) \\
&\vdots \\
\boldsymbol{\xi}_s^{(t+1)} &\sim \pi(\boldsymbol{\xi}_s | \boldsymbol{\xi}_1^{(t+1)}, \boldsymbol{\xi}_2^{(t+1)}, \ldots, \boldsymbol{\xi}_{s-1}^{(t+1)}, \boldsymbol{y}),
\end{aligned}$$

and hence $\boldsymbol{\xi}^{(t+1)} = \left(\boldsymbol{\xi}_1^{(t+1)}, \boldsymbol{\xi}_2^{(t+1)}, \ldots, \boldsymbol{\xi}_s^{(t+1)}\right)'$.

It may be the case that some components have full conditional distributions that cannot be sampled from directly. Such components are updated using M-H steps. This leads to a hybrid sampling algorithm known as *Metropolis-Hastings within Gibbs*.

2.3 Convergence rates

Any practitioner of MCMC methods faces the problem of knowing when the chain has converged to the stationary distribution. For a square integrable function g, let $E_\pi[g(\boldsymbol{\xi})]$ be the expectation of $g(\boldsymbol{\xi})$ under the target distribution for $\boldsymbol{\xi}$. We define the convergence rate λ to be the minimum number such that for all square integrable functions g, and for all $r > \lambda$,

$$\lim_{t \to \infty} \left(E_\pi\left[g(\boldsymbol{\xi}^{(t)}) | \boldsymbol{\xi}^{(0)}\right] - E_\pi\left[g(\boldsymbol{\xi})\right] \right)^2 r^{-t} = 0. \tag{2.4}$$

The convergence rate λ is bounded by the interval $[0, 1]$, with $\lambda = 0$ indicating immediate convergence and $\lambda = 1$ indicating subgeometric convergence (Meyn and Tweedie, 1993). This form of convergence is considered by Amit (1991) and Roberts and Sahu (1997) among others.

It is standard practice to discard an initial portion of the Markov chain and make inferences based on simulations after that. A difficulty here is on deciding the length of this initial portion, the so called *burn-in* period. The burn-in required will be problem specific as the convergence rate is a measure of the discrepancy between the transition kernel $P(\boldsymbol{\xi}^{(t)} | \boldsymbol{\xi}^{(0)})$ and the stationary distribution $\pi(\boldsymbol{\xi})$. Attempts to predetermine the burn-in have had limited success. Upper bounds for the distance of the chain to stationarity, after a certain number of iterations, are difficult to compute and are often too large of be of any practical use, see Roberts and Rosenthal (1998) and references therein. Many tests rely on an analysis of the output of the chain to diagnose convergence.

2.3.1 Diagnostic tests

A straightforward diagnostic test is to plot the values of the chain on a graph and see how many iterations it takes for the chain to 'settle down'. So called trace plots are a useful

visual tool but can be misleading. A chain may appear to have settled and be mixing well but may be merely be stuck in a local mode. For an example of this behavior see Ripley and Kirkland (1990). It is therefore recommended to overlay the trace plots of many chains with widely spread starting points. The object here is to identify the number of iterations it takes for the chains to overlap and 'forget' where they had begun. A quantitative application of this principle underpins many of the post sampling statistical tests of convergence (Gelman and Rubin, 1992; Johnson, 1996; Liu et al., 1992; Roberts, 1996). The difficulty with these measures is choosing start values that are overdispersed with respect to the unknown posterior distribution.

A popular measure from Gelman and Rubin (1992) compares the variance across the chains to that of the variance within the chains. It is computed as follows. Consider the scalar parameter of interest, ξ, assumed to be *a posteriori* normally distributed with mean μ and variance σ^2. Run $l \geq 2$ independent chains of length $2N$, discarding the first N iterations. Let ξ^{jt} be the tth iteration for the jth chain, $t = 1 \ldots, N$, $j = 1, \ldots, l$. The between chain variance K/N is calculated as

$$K/N = \frac{1}{l-1}\sum_{j=1}^{l}(\xi^{j\cdot} - \bar{\xi})^2 \quad \text{where} \quad \xi^{j\cdot} = \frac{1}{N}\sum_{t=1}^{N}\xi^{jt}, \quad \bar{\xi} = \frac{1}{l}\sum_{j=1}^{l}\xi^{j\cdot},$$

and the within chain variance M is given by

$$M = \frac{1}{l}\sum_{j=1}^{l}s_j^2, \quad \text{where} \quad s_j^2 = \frac{1}{N-1}\sum_{t=1}^{N}(\xi^{jt} - \xi^{j\cdot})^2.$$

A weighted average of K and M gives an estimator for σ^2

$$\sigma_+^2 = \frac{K}{N} + \frac{N-1}{N}M.$$

If samples are from the stationary distribution then σ_+^2 is unbiased. As starting points are overdispersed, if the chain has not yet converged then σ_+^2 will overestimate σ^2. On the other hand, if the whole parameter space is yet to be explored then M will underestimate σ^2. Adjusting for the sampling variability in the estimator for μ gives a pooled variance estimator $\widehat{V} = \sigma_+^2 + K/lN$. The statistic is then given by

$$R = \sqrt{\frac{(e+3)\widehat{V}}{(e+1)M}},$$

where $e = 2\widehat{V}^2/Var(\widehat{V})$ is the estimated degrees of freedom for a t-distribution with mean $\hat{\mu}$ and variance $\widehat{V}$.

The value R is referred to as the potential scale reduction factor (PSRF), the factor by which the width of the credible intervals for μ could be reduced if N were increased. Values much greater than 1 indicate a failure to converge, with values less than 1.1 or 1.2 considered low enough to be satisfied that convergence has been achieved (Gilks et al., 1996, Chapter 8).

The PSRF is extended to a multivariate convergence diagnostic by Brooks and Gelman (1998). The multivariate potential scale reduction factor (MPSRF) compares estimates of covariance matrices. If $\boldsymbol{\xi}^{jt} = \left(\xi_1^{jt}, \xi_2^{jt}, \ldots, \xi_s^{jt}\right)'$ is a vector containing the tth iteration of the jth chain for scalar components ξ_i, $i = 1 \ldots, s$, then

$$\boldsymbol{M} = \frac{1}{l(N-1)} \sum_{j=1}^{l} \sum_{t=1}^{N} (\boldsymbol{\xi}^{jt} - \boldsymbol{\xi}^{j\cdot})(\boldsymbol{\xi}^{jt} - \boldsymbol{\xi}^{j\cdot})',$$

and

$$\boldsymbol{K}/N = \frac{1}{l-1} \sum_{j=1}^{l} (\boldsymbol{\xi}^{j\cdot} - \bar{\boldsymbol{\xi}})(\boldsymbol{\xi}^{j\cdot} - \bar{\boldsymbol{\xi}})',$$

where $\boldsymbol{\xi}^{j\cdot} = \left(\xi_1^{j\cdot}, \xi_2^{j\cdot}, \ldots, \xi_s^{j\cdot}\right)'$ and $\bar{\boldsymbol{\xi}} = \left(\bar{\xi}_1, \bar{\xi}_2, \ldots, \bar{\xi}_s\right)'$. The MPSRF is given by

$$R_s = \frac{N-1}{N} + \left(\frac{l+1}{l}\right)\gamma,$$

where γ is the largest eigenvalue of $\boldsymbol{M}^{-1}\boldsymbol{K}/n$, with values of R_s substantially above one indicating a failure to converge.

It must be noted that diagnostic tests are used to indicate a failure to converge as we can never know if convergence has truly been achieved. In their comparative review of convergence diagnostics, Cowles and Carlin (1996) find that all of the methods they consider can fail to detect a lack of convergence. Consequently they advise employing a battery of tests. It is also warned by Cowles et al. (1999) that the act of employing diagnostic tests to assess the number of iterations to be discarded, can itself induce biases into the estimator (2.3). It is therefore recommended that several pilot chains are used to determine the length of the burn-in period and then inference is based on one long separate chain.

2.3.2 Autocorrelation

Markov chains that exhibit high autocorrelation will mix more slowly and hence take longer to converge. The autocorrelation plot is sometimes used to determine the *thinning interval*. The idea is to try and achieve close to independent samples values by retaining every mth value where the autocorrelation at lag m falls below some tolerance level. Although this may seem reasonable, MacEachern and Berliner (1994) show that by throwing away information the variance of the mean of the samples can only be increased. It is far better to use the concept of an *effective sample size* or ESS (Robert and Casella, 2004, Chapter 12). The ESS is computed by dividing the number of post burn-in samples N by an estimate of the autocorrelation time κ, where

$$\kappa = 1 + 2\sum_{k=1}^{\infty} \rho(k),$$

and $\rho(k)$ is the autocorrelation of at lag k. We can estimate κ by using the sample autocorrelations of the chain and truncate the infinite sum when the autocorrelation falls below

some threshold. This may lead to a biased estimate for κ (Banerjee et al., 2003, Chapter 4). Another method of estimating κ is to estimate the spectral density at frequency zero. If we consider scalar quantities such that $\bar{g}_N = N^{-1}\sum_{t=1}^{N} g(\xi^{(t)})$, we have from Ripley (1987, Chapter 6) that

$$NVar(\bar{g}_N) \to \upsilon^2\kappa = 2\pi f(0),$$

where υ^2 is the variance and $f(0)$ the spectral density of the chain $\{g(\xi^{(t)})\}_{t=1}^{N}$. So asymptotically $N/\kappa = N\upsilon^2/2\pi f(0)$. It is this method that is used in the R package CODA (Plummer et al., 2006) to estimate the ESS.

2.3.3 Convergence rates for the Gibbs sampler

For Gibbs samplers with Gaussian target distributions with known precision matrices we have analytical results for the exact convergence rate (Roberts and Sahu, 1997, Theorem 1). Convergence here is defined in terms of how rapidly the expectations of square integrable functions approach their stationary values, see (2.4). If $\boldsymbol{\xi}|\boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ be the posterior precision matrix. To compute the convergence rate first partition $\boldsymbol{Q}$ according to the s blocks used for updating, i.e.,

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} & \cdots & \boldsymbol{Q}_{1s} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} & \cdots & \boldsymbol{Q}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{Q}_{s1} & \boldsymbol{Q}_{s2} & \cdots & \boldsymbol{Q}_{ss} \end{pmatrix}. \tag{2.5}$$

Let $\boldsymbol{A} = \boldsymbol{I} - \text{diag}(\boldsymbol{Q}_{11}^{-1}, \ldots, \boldsymbol{Q}_{ss}^{-1})\boldsymbol{Q}$ and $\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{L}_A)^{-1}\boldsymbol{U}_A$, where $\boldsymbol{L}_A$ is the block lower triangular matrix of $\boldsymbol{A}$, and $\boldsymbol{U}_A = \boldsymbol{A} - \boldsymbol{L}_A$. Roberts and Sahu (1997) show that the Markov chain induced by the Gibbs sampler with components block updated according to matrix (2.5), has a Gaussian transition density with mean $E[\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)}] = \boldsymbol{F}\boldsymbol{\xi}^{(t)} + \boldsymbol{f}$, where $\boldsymbol{f} = (\boldsymbol{I} - \boldsymbol{F})\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} - \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}'$. Their observation leads to the following theorem:

Theorem 2.3.1 *(Roberts and Sahu, 1997) A Markov chain with transition density*

$$N(\boldsymbol{F}\boldsymbol{\xi}^{(t)} + \boldsymbol{f}, \boldsymbol{\Sigma} - \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}'),$$

has a convergence rate equal to the maximum modulus eigenvalue of $\boldsymbol{F}$.

Corollary 2.3.2 *If we update* $\boldsymbol{\xi}$ *in two blocks so that* $s = 2$ *then*

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix}, \quad \textit{and} \quad \boldsymbol{F} = \begin{pmatrix} \boldsymbol{0} & -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} \\ \boldsymbol{0} & \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} \end{pmatrix},$$

and the convergence rate is the maximum modulus eigenvalue of $\boldsymbol{F}_{22} = \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}$.

We will make repeated use of Theorem 2.3.1 in Chapter 3. As an example of its application we return to the elementary model (1.1). Recall that the CP of the model

has $Y_i = \tilde{U}_i + \epsilon_i$, $i = 1, \ldots, n$, with $\tilde{U}_i \sim N(\theta, \sigma_u^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ all independent. A flat prior is given to θ. For the CP it can be shown that the joint posterior distribution $\pi(\tilde{\boldsymbol{U}}, \theta|y)$ is multivariate normal with precision matrix $\boldsymbol{Q}^c$, where

$$\boldsymbol{Q}^c = \begin{pmatrix} (1/\sigma_\epsilon^2 + 1/\sigma_u^2)\boldsymbol{I} & -1/\sigma_u^2\mathbf{1} \\ -1/\sigma_u^2\mathbf{1}' & n/\sigma_u^2 \end{pmatrix}.$$

Applying Theorem 2.3.1 we find that the convergence rate for the CP is

$$\lambda_c = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_u^2}. \tag{2.6}$$

By definition a smaller value of λ_c indicates faster convergence. Therefore, convergence is hastened for the CP when the random effects variance dominates the data variance.

The NCP is found by letting $U_i = \tilde{U}_i - \theta$. The joint posterior distribution $\pi(\boldsymbol{U}, \theta|y)$ is multivariate normal with precision matrix $\boldsymbol{Q}^{nc}$, given by

$$\boldsymbol{Q}^{nc} = \begin{pmatrix} (1/\sigma_\epsilon^2 + 1/\sigma_u^2)\boldsymbol{I} & 1/\sigma_\epsilon^2\mathbf{1} \\ 1/\sigma_\epsilon^2\mathbf{1}' & n/\sigma_\epsilon^2 \end{pmatrix}.$$

The convergence rate for the NCP is

$$\lambda_{nc} = \frac{\sigma_u^2}{\sigma_\epsilon^2 + \sigma_u^2}, \tag{2.7}$$

and we see that in constrast to the CP, convergence for the NCP is hastened as the variance for the random effects shrinks compared to that of the data.

It is also worth noting that $\lambda_{nc} = 1 - \lambda_c$. Although this relationship does not hold if θ is given a proper prior distribution, it sharply demonstrates that the two parameterisations are complementary, where one does well the other does poorly.

Roberts and Sahu (2001) consider the problem of predetermining which parameterisation to use in the absence of a known precision matrix. They suggest approximating (2.5) by evaluating the negative Hessian matrix of the posterior distribution at the posterior mode, where the mode is found using the EM algorithm (Dempster et al., 1977). We use a similar approach in Section 6.4 to estimate the conditional posterior covariance of the random effects for non-Gaussian data.

2.4 The three stage linear model

In this section we consider the following hierarchically centred three stage model:

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3). \end{aligned} \tag{2.8}$$

where $\boldsymbol{Y}$ is a $G \times 1$ vector of responses, $\boldsymbol{X}_1$ is $G \times P$ design matrix and $\tilde{\boldsymbol{\beta}}$ is $P \times 1$ vector of centred random effects. The random effects are modelled jointly as multivariate normal with mean $\boldsymbol{X}_2\boldsymbol{\theta}$ where $\boldsymbol{X}_2$ is a $P \times H$ design matrix and $\boldsymbol{\theta}$ is $H \times 1$ vector of global effects.

Model (2.8) is a general set up and many models can be written in this form. For example, suppose that $\boldsymbol{Y} = (\boldsymbol{Y}_1', \ldots, \boldsymbol{Y}_n')'$ and $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$, for $i = 1, \ldots, n$, then $G = \sum_{i=1}^{n} n_i$. Further, let $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1', \ldots, \tilde{\boldsymbol{\beta}}_n')'$ where $\tilde{\boldsymbol{\beta}}_i$ is a $p \times 1$ vector, hence $P = np$. If we assume that $Cov(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = Cov(\tilde{\boldsymbol{\beta}}_i, \tilde{\boldsymbol{\beta}}_j) = \boldsymbol{0}$, for $i \neq j$ and $\boldsymbol{C}_3^{-1} = \boldsymbol{0}$ then we have the model considered by Papaspiliopoulos (2003, Section 2.4). If in addition $Var(\boldsymbol{Y}_i) = \sigma_i^2 \boldsymbol{I}_{n_i}$ then we have the model considered by Gelfand et al. (1995, Section 2). Furthermore, the spatial models we consider in Chapters 3, 4 and 5 can be written in this form, hence we give some of its properties in the remainder of this section.

The NCP for model (2.8) is found by letting

$$\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2\boldsymbol{\theta}. \tag{2.9}$$

Substituting (2.9) into model (2.8) gives the NCP for the three stage model as

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\theta} &\sim N(\boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_1) \\ \boldsymbol{\beta} &\sim N(\boldsymbol{0}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3). \end{aligned}$$

2.4.1 Conditional posterior distributions for the CP and the NCP of the three stage model

The work of Lindley and Smith (1972) for the Bayesian analysis of the normal linear hierarchical model is relevant here and we make repeated use of the following lemma:

Lemma 2.4.1 *(Lindley and Smith, 1972) If* $\boldsymbol{Y} \sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1)$ *and* $\tilde{\boldsymbol{\beta}} \sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2)$ *for* $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, $\boldsymbol{C}_1$ *and* $\boldsymbol{C}_2$ *all known, the conditional posterior distribution of* $\tilde{\boldsymbol{\beta}}$ *is*

$$\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{y}) \sim N(\boldsymbol{B}\boldsymbol{b}, \boldsymbol{B}), \tag{2.10}$$

where

$$\boldsymbol{B}^{-1} = \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1} \quad \textit{and} \quad \boldsymbol{b} = \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\theta}.$$

We can immediately apply Lemma 2.4.1 to find the conditional posterior distribution of $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\beta}}, \boldsymbol{y}) \sim N(\tilde{\boldsymbol{m}}^*, \tilde{\boldsymbol{C}}_3^*), \tag{2.11}$$

where

$$\tilde{\boldsymbol{C}}_3^* = \left(\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1} \quad \text{and} \quad \tilde{\boldsymbol{m}}^* = \tilde{\boldsymbol{C}}_3^* \left(\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\tilde{\boldsymbol{\beta}} + \boldsymbol{C}_3^{-1}\boldsymbol{m}\right).$$

Similarly, we find the conditional posterior distributions for the NCP. For $\boldsymbol{\beta}$ we have

$$\pi(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y}) \sim N(\boldsymbol{B}\boldsymbol{d}, \boldsymbol{B}), \tag{2.12}$$

where $\boldsymbol{d} = \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta})$, and the conditional posterior distribution for $\boldsymbol{\theta}$ for the NCP is given by

$$\pi(\boldsymbol{\theta}|\boldsymbol{\beta}, \boldsymbol{y}) \sim N(\boldsymbol{m}^*, \boldsymbol{C}_3^*), \tag{2.13}$$

where

$$\boldsymbol{C}_3^* = \left((\boldsymbol{X}_1\boldsymbol{X}_2)'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1}$$

and

$$\boldsymbol{m}^* = \boldsymbol{C}_3^*\left((\boldsymbol{X}_1\boldsymbol{X}_2)'\boldsymbol{C}_1^{-1}(\boldsymbol{y} - \boldsymbol{X}_1\boldsymbol{\beta}) + \boldsymbol{C}_3^{-1}\boldsymbol{m}\right).$$

2.4.2 Posterior covariance matrices for the CP and the NCP of the three stage model

In this section we derive the posterior variances and covariance for random and global effects for the different model parameterisations. Computations are similar in spirit to those conducted by Gelfand et al. (1995) and we make repeated use of the results for the two stage model given by Lemma 2.4.1. We remind ourselves that covariance matrices $\boldsymbol{C}_1$, $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ are assumed to be known. We begin by considering the CP and use the relationship given in equation (2.9) to find the equivalent results for the NCP.

First we find the marginal posterior distribution of $\boldsymbol{\theta}$. Marginalising over the $\tilde{\boldsymbol{\beta}}$'s we get a likelihood of

$$\boldsymbol{Y}|\boldsymbol{\theta} \sim N\left(\boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{\Sigma}_{Y|\theta}\right),$$

where $\boldsymbol{\Sigma}_{Y|\theta} = \boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1'$. Then we find the marginal posterior distribution of $\boldsymbol{\theta}$ to be

$$\boldsymbol{\theta}|\boldsymbol{y} \sim N\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\Sigma}_{\theta|y}\right),$$

where

$$\begin{aligned}
\boldsymbol{\Sigma}_{\theta|y} &= \left((\boldsymbol{X}_1\boldsymbol{X}_2)'\boldsymbol{\Sigma}_{Y|\theta}^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1} \\
\hat{\boldsymbol{\theta}} &= \boldsymbol{\Sigma}_{\theta|y}\left((\boldsymbol{X}_1\boldsymbol{X}_2)'\boldsymbol{\Sigma}_{Y|\theta}^{-1}\boldsymbol{y} + \boldsymbol{C}_3^{-1}\boldsymbol{m}\right).
\end{aligned}$$

To compute the marginal posterior distribution for $\tilde{\boldsymbol{\beta}}$ we marginalise the conditional posterior distribution given in (2.10) over $\boldsymbol{\theta}$. We have that $\tilde{\boldsymbol{\beta}}|\boldsymbol{y}$ is normally distributed with expectation

$$\begin{aligned}
E[\tilde{\boldsymbol{\beta}}|\boldsymbol{y}] &= E\left[E[\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{y}]\right] \\
&= \boldsymbol{B}E[\boldsymbol{b}] \\
&= \boldsymbol{B}(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{X}_2E[\boldsymbol{\theta}|\boldsymbol{y}]) \\
&= \boldsymbol{B}\hat{\boldsymbol{b}},
\end{aligned}$$

where $\hat{\boldsymbol{b}} = \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{X}_2\hat{\boldsymbol{\theta}}$, and variance

$$\begin{aligned} Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{y}) &= E\big[Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta},\boldsymbol{y})\big] + Var\big(E[\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta},\boldsymbol{y}]\big) \\ &= E[\boldsymbol{B}] + Var(\boldsymbol{B}\boldsymbol{b}) \\ &= \boldsymbol{B} + Var(\boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\theta}) \\ &= \boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B}. \end{aligned}$$

We now compute the posterior covariance of $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}$. We have

$$\begin{aligned} Cov(\tilde{\boldsymbol{\beta}},\boldsymbol{\theta}|\boldsymbol{y}) &= E\big[\{\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}]\}\{\boldsymbol{\theta} - E[\boldsymbol{\theta}]\}'|\boldsymbol{y}\big] \\ &= E\big[\{\tilde{\boldsymbol{\beta}} - E[\tilde{\boldsymbol{\beta}}]\}\boldsymbol{\theta}'|\boldsymbol{y}\big] \\ &= E\big[\tilde{\boldsymbol{\beta}}\boldsymbol{\theta}'|\boldsymbol{y}\big] - E\big[E[\tilde{\boldsymbol{\beta}}]\boldsymbol{\theta}'|\boldsymbol{y}\big] \\ &= E\big[E[\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta},\boldsymbol{y}]\boldsymbol{\theta}'|\boldsymbol{y}\big] - \boldsymbol{B}\hat{\boldsymbol{b}}E[\boldsymbol{\theta}'|\boldsymbol{y}] \\ &= \boldsymbol{B}E\big[(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\theta})\boldsymbol{\theta}'|\boldsymbol{y}\big] - \boldsymbol{B}\big(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y} + \boldsymbol{C}_2^{-1}\boldsymbol{X}_2\hat{\boldsymbol{\theta}}\big)\hat{\boldsymbol{\theta}}' \\ &= \boldsymbol{B}\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y}\hat{\boldsymbol{\theta}}' + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2E\big[\boldsymbol{\theta}\boldsymbol{\theta}'|\boldsymbol{y}\big] - \boldsymbol{B}\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{y}\hat{\boldsymbol{\theta}}' + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}' \\ &= \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2E\big[\boldsymbol{\theta}\boldsymbol{\theta}'|\boldsymbol{y}\big] - \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\hat{\boldsymbol{\theta}}\hat{\boldsymbol{\theta}}' \\ &= \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}. \end{aligned}$$

We now turn our attention to the NCP. Recall from (2.9) that $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2\boldsymbol{\theta}$. The posterior covariances are

$$\begin{aligned} Var(\boldsymbol{\beta}|\boldsymbol{y}) &= Cov(\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2\boldsymbol{\theta}|\boldsymbol{y}) \\ &= Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{y}) - Cov(\tilde{\boldsymbol{\beta}},\boldsymbol{\theta}|\boldsymbol{y})\boldsymbol{X}_2' - \boldsymbol{X}_2Cov(\boldsymbol{\theta},\tilde{\boldsymbol{\beta}}|\boldsymbol{y}) + \boldsymbol{X}_2Var(\boldsymbol{\theta}|\boldsymbol{y})\boldsymbol{X}_2' \\ &= \boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2' - \boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} \\ &\quad + \boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2', \end{aligned}$$

and

$$\begin{aligned} Cov(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) &= Cov(\tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2,\boldsymbol{\theta}|\boldsymbol{y}) \\ &= Cov(\tilde{\boldsymbol{\beta}},\boldsymbol{\theta}|\boldsymbol{y}) - \boldsymbol{X}_2Var(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y} - \boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}. \end{aligned}$$

2.4.3 Convergence rates for the CP and the NCP of the three stage model

Suppose that we run a Gibbs sampler updating the random effects as one block and global effects as another. We continue to assume that all prior covariance matrices are known. For the CP we have $\boldsymbol{\xi} = (\tilde{\boldsymbol{\beta}}', \boldsymbol{\theta}')'$. Given $\boldsymbol{\xi}^{(t)}$ we obtain $\boldsymbol{\xi}^{(t+1)}$ as follows:

1. Draw $\tilde{\boldsymbol{\beta}}^{(t+1)} \sim \pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}^{(t)},\boldsymbol{y})$.

2. Draw $\boldsymbol{\theta}^{(t+1)} \sim \pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\beta}}^{(t+1)},\boldsymbol{y})$,

where $\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta},\boldsymbol{y})$ and $\pi(\boldsymbol{\theta}|\tilde{\boldsymbol{\beta}},\boldsymbol{y})$ are given in (2.10) and (2.11) respectively. To compute the convergence rate of the Gibbs sampler we first compute the posterior precision matrix of $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}$. The matrix is determined by model (2.8) and can be identified by writing

$$\begin{aligned}
\pi(\tilde{\boldsymbol{\beta}},\boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}})\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
&\propto \exp\Big\{-\frac{1}{2}\Big[(\boldsymbol{Y}-\boldsymbol{X}_1\tilde{\boldsymbol{\beta}})'\boldsymbol{C}_1^{-1}(\boldsymbol{Y}-\boldsymbol{X}_1\tilde{\boldsymbol{\beta}})+(\tilde{\boldsymbol{\beta}}-\boldsymbol{X}_2\boldsymbol{\theta})'\boldsymbol{C}_2^{-1}(\tilde{\boldsymbol{\beta}}-\boldsymbol{X}_2\boldsymbol{\theta}) \\
&\quad +(\boldsymbol{\theta}-\boldsymbol{m})'\boldsymbol{C}_3^{-1}(\boldsymbol{\theta}-\boldsymbol{m})\Big]\Big\} \\
&= \exp\Big\{-\frac{1}{2}\Big[\ldots+\tilde{\boldsymbol{\beta}}'(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1})\tilde{\boldsymbol{\beta}}-2\tilde{\boldsymbol{\beta}}'\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\theta} \\
&\quad +\boldsymbol{\theta}'(\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{X}_2+\boldsymbol{C}_3^{-1})\boldsymbol{\theta}+\ldots\Big]\Big\},
\end{aligned}$$

where the last equation only includes the terms containing both $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}$. Therefore the posterior precision matrix for the CP is given by

$$\boldsymbol{Q}^c=\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1} & -\boldsymbol{C}_2^{-1}\boldsymbol{X}_2 \\ -\boldsymbol{X}_2'\boldsymbol{C}_2^{-1} & \boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{X}_2+\boldsymbol{C}_3^{-1}\end{pmatrix}.$$

By Corollary 2.3.2 the convergence rate of the Markov chain induced by the Gibbs sampler under the CP is given by the maximum modulus eigenvalue of

$$\boldsymbol{F}_{22}^c=\left(\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{X}_2+\boldsymbol{C}_3^{-1}\right)^{-1}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1}\right)^{-1}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2.$$

For the NCP we cycle between drawing $\boldsymbol{\beta}^{(t+1)}\sim\pi(\boldsymbol{\beta}|\boldsymbol{\theta}^{(t)},\boldsymbol{y})$ and $\boldsymbol{\theta}^{(t+1)}\sim\pi(\boldsymbol{\theta}|\boldsymbol{\beta}^{(t+1)},\boldsymbol{y})$, where the relevent distributions are given in (2.12) and (2.13) respectively. The posterior precision matrix for the NCP is found by writing

$$\begin{aligned}
\pi(\boldsymbol{\beta},\boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\beta},\boldsymbol{\theta})\pi(\boldsymbol{\beta}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
&\propto \exp\Big\{-\frac{1}{2}\Big[(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}-\boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta})'\boldsymbol{C}_1^{-1}(\boldsymbol{Y}-\boldsymbol{X}_1\boldsymbol{\beta}-\boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta})+ \\
&\quad +\boldsymbol{\beta}'\boldsymbol{C}_2^{-1}\boldsymbol{\beta}+(\boldsymbol{\theta}-\boldsymbol{m})'\boldsymbol{C}_3^{-1}(\boldsymbol{\theta}-\boldsymbol{m})\Big]\Big\} \\
&= \exp\Big\{-\frac{1}{2}\Big[\ldots+\boldsymbol{\beta}'(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1})\boldsymbol{\beta}+2\boldsymbol{\beta}'\boldsymbol{X}_1'\boldsymbol{C}_2^{-1}\boldsymbol{X}_1\boldsymbol{X}_2\boldsymbol{\theta} \\
&\quad +\boldsymbol{\theta}'(\boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2+\boldsymbol{C}_3^{-1})\boldsymbol{\theta}+\ldots\Big]\Big\},
\end{aligned}$$

and hence we have

$$\boldsymbol{Q}^{nc}=\begin{pmatrix}\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1} & \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 \\ \boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 & \boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2+\boldsymbol{C}_3^{-1}\end{pmatrix}.$$

By Corollary 2.3.2, the convergence rate of the Gibbs sampler for the NCP is the maximum modulus eigenvalue of

$$\begin{aligned}\boldsymbol{F}_{22}^{nc} &= \left(\boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2+\boldsymbol{C}_3^{-1}\right)^{-1}\boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1}\right)^{-1}\\ &\quad \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2.\end{aligned}$$

2.5 Criteria for calibrating out of sample predictions

We use three criteria to assess the accuracy of spatial predictions. The first two, the mean absolute prediction error (MAPE) and the root mean squared prediction error (RMSPE), compare point estimates with observed values for the validation data set. A third measure, the continuous ranked probability score (CRPS), compares the CDF of the posterior predictive distribution with the known validation observations. The CRPS is a more appropriate measure in the Bayesian setting as it takes into account the sharpness and not just the location of the posterior predictive density.

Let $y_i, i = 1, \ldots, m$, denote the m known validation observations. These are estimated by $\hat{y}_i$ which may be the mean or mode of the post burn-in samples $y_i^{(t)}, t = 1, \ldots, N$, which are drawn from $\pi(Y_i|\boldsymbol{y})$. We calculate the MAPE as

$$\text{MAPE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - \hat{y}_i|,$$

and the RMSPE as

$$\text{RMSPE} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2}.$$

If F is the CDF of the posterior predictive distribution, Gneiting and Raftery (2007) show that the CRPS can be written 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 CDF F and finite first moment. Note that if F were a point estimate the CRPS would equal MAPE. Given MCMC samples $y^{(t)}, t = 1, \ldots, N$, we can estimate the CRPS as

$$\widehat{\text{crps}}(F, y) = \frac{1}{N}\sum_{t=1}^{N}|y^{(t)} - y| - \frac{1}{N^2}\sum_{t=1}^{N}\sum_{u=1}^{N}|y^{(t)} - y^{(u)}|.$$

For m validation sites we take the overall measure to be

$$\text{CRPS} = \frac{1}{m}\sum_{i=1}^{m}\widehat{\text{crps}}(F_i, y_i).$$

2.6 Summary

In this chapter we have provided a short overview of Bayesian computation, focusing on those aspects of the field which are used in the remaining chapters of this thesis. We have provided details of Metropolis Hastings and Gibbs sampling algorithms and discussed convergence and convergence diagnostics. In particular, we have given the a result that allows us to compute the exact convergence rate for Gibbs samplers with Gaussian target distributions with known precision matrices, a result we make use of in Chapter 3. We have also provided details of the potential scale reduction factor and the effective sample size, which we make use of in Chapters 4, 5 and 6.

We have given details and properties of the three stage normal linear hierarchical model and expressions for the exact convergence rate of the CP and the NCP of models that can be written in this general form. We have also given three criteria for assessing the accuracy of out of sample predictions. These criteria are used in the real data examples of Chapters 4 and 6.

Chapter 3

Exact convergence rates for the CP and the NCP

3.1 Introduction

Spatially varying coefficient (SVC) models (Gelfand et al., 2003) are being widely applied by researchers looking to understand observed processes that exhibit spatial dependency. Berrocal et al. (2010) include a spatially varying intercept and slope to correct for the bias in the numerical model output that is used as the covariate in their downscaler model for tropospheric ozone concentrations. Hamm et al. (2015) take a similar approach to model the concentration of particulate matter across Europe. Wheeler et al. (2014) use SVC models to analyse housing sale prices in Toronto, Canada for January 2001 and Finley et al. (2011) construct an SVC model for continuous forest variables, e.g. biomass.

In Section 1.3.2 we give details of the standard Gaussian process model for point referenced spatial data, which is given by

$$Y(\boldsymbol{s}) = \boldsymbol{x}'(\boldsymbol{s})\boldsymbol{\theta} + \beta(\boldsymbol{s}) + \epsilon(\boldsymbol{s}). \tag{3.1}$$

Model (3.1) takes a normal linear regression model and includes a Gaussian process to capture the spatial association unexplained by the covariates. Re-writing the model as

$$Y(\boldsymbol{s}) = \theta_0 + \beta(\boldsymbol{s}) + \sum_{k=1}^{p-1} x_k(\boldsymbol{s})\theta_k + \epsilon(\boldsymbol{s}), \tag{3.2}$$

we can consider $\beta(\boldsymbol{s})$ as a local adjustment to the global intercept θ_0, where $\beta(\boldsymbol{s})$ is modelled as a realisation of a Gaussian process at location $\boldsymbol{s}$. Alternatively, we can view $\tilde{\beta}(\boldsymbol{s}) = \theta_0 + \beta(\boldsymbol{s})$ as a random intercept process. Gelfand et al. (2003) extend model (3.2) to allow all of the coefficients of the explanatory variables to vary locally, envisioning a spatial surface for each coefficient and thus providing a flexible class of non-stationary models.

In Section 2.3.3 it is shown that for the simple independent random effects model given in (1.1), that the ratio of the variance parameters is important for determining the conver-

gence rate. In this chapter we consider SVC models and use Theorem 2.3.1 to investigate the additional role that correlation across the random effects plays in determining the convergence rate of the Gibbs samplers for the CP and the NCP.

The rest of this chapter is organised as follows: In Section 3.2 we give details of the set up for regression models with spatially varying coefficients and compute the posterior precision matrices for the CP and the NCP which are needed to calculate the convergence rate of their respective Gibbs samplers. In Section 3.3 we look at how the variance components and the strength of correlation between random effects impacts upon the convergence rates for equi-correlated and then spatially correlated random effects. Section 3.4 looks at covariance tapering and in Section 3.5 we look at how the convergence rate is affected by the scale of the covariate. In Section 3.6 we consider different correlation functions from a family of isotropic functions widely applied in spatial statistics and in Section 3.7 we construct anisotropic correlation functions and assess their impact on the convergence rates of the CP and the NCP. Section 3.8 considers the effect of blocking on the convergence rate and we close in Section 3.9 with some summary remarks.

3.2 A general spatial model

Following Gelfand et al. (2003, Section 3) we consider the following normal linear model with spatially varying regression coefficients

$$Y(\boldsymbol{s}_i) = \theta_0 + \beta_0(\boldsymbol{s}_i) + \{\theta_1 + \beta_1(\boldsymbol{s}_i)\}x_1(\boldsymbol{s}_i) + \ldots + \{\theta_p + \beta_p(\boldsymbol{s}_i)\}x_{p-1}(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \quad (3.3)$$

for $i = 1, \ldots, n$. We model errors $\epsilon(\boldsymbol{s}_i)$ as independent and normally distributed with mean zero and variance σ^2_ϵ. Spatially indexed observations $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$ are conditionally independent and normally distributed as

$$Y(\boldsymbol{s}_i) \sim N(\boldsymbol{x}'(\boldsymbol{s}_i)\{\boldsymbol{\theta} + \boldsymbol{\beta}(\boldsymbol{s}_i)\}, \sigma^2_\epsilon),$$

where $\boldsymbol{x}(\boldsymbol{s}_i) = (1, x_1(\boldsymbol{s}_i), \ldots, x_{p-1}(\boldsymbol{s}_i))'$ is a vector containing covariate information for site $\boldsymbol{s}_i$ and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ is a vector of global regression coefficients. The kth element of $\boldsymbol{\theta}$ is locally perturbed by a realisation of a zero mean Gaussian process, denoted $\beta_k(\boldsymbol{s}_i)$, which are collected into a vector $\boldsymbol{\beta}(\boldsymbol{s}_i) = (\beta_0(\boldsymbol{s}_i), \ldots, \beta_{p-1}(\boldsymbol{s}_i))'$. The n realisations of the Gaussian process associated with the kth covariate are given by

$$\boldsymbol{\beta}_k = (\beta_k(\boldsymbol{s}_1), \ldots, \beta_k(\boldsymbol{s}_n)) \sim N(0, \boldsymbol{\Sigma}_k),$$

where

$$\boldsymbol{\Sigma}_k = \sigma^2_k \boldsymbol{R}_k, \quad \text{and} \quad (\boldsymbol{R}_k)_{ij} = corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\}.$$

The CP is found by introducing the variables $\tilde{\beta}_k(\boldsymbol{s}_i) = \theta_k + \beta_k(\boldsymbol{s}_i)$, for $k = 0, \ldots, p-1$, and $i = 1, \ldots, n$. Therefore

$$\tilde{\boldsymbol{\beta}}_k = (\tilde{\beta}_k(\boldsymbol{s}_1), \ldots, \tilde{\beta}_k(\boldsymbol{s}_n)) \sim N(\theta_k \mathbf{1}, \boldsymbol{\Sigma}_k).$$

Global effects $\boldsymbol{\theta}$ are assumed to be multivariate normal *a priori* and so we write model (3.3) in its hierarchically centred form as

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned}$$

where $\boldsymbol{C}_1 = \sigma_\epsilon^2\boldsymbol{I}$ and $\boldsymbol{X}_1 = (\boldsymbol{I}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{p-1})$ is the $n \times np$ design matrix for the first stage where $\boldsymbol{D}_k$ is a diagonal matrix with entries $\boldsymbol{x}_k = (x_k(\boldsymbol{s}_1), \ldots, x_k(\boldsymbol{s}_n))'$. We denote by $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0', \ldots, \tilde{\boldsymbol{\beta}}_{p-1}')'$ the $np \times 1$ vector of centred, spatially correlated random effects.

The design matrix for the second stage, $\boldsymbol{X}_2$, is a $np \times p$ block diagonal matrix, the blocks made of vectors of ones of length n,

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}.$$

The p processes are assumed independent *a priori* and so $\boldsymbol{C}_2$ is block diagonal where the kth block is given by $\boldsymbol{\Sigma}_k$, and so

$$\boldsymbol{C}_2 = \begin{bmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{p-1} \end{bmatrix}.$$

The global effects $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{p-1})'$ are assumed to be independent *a priori* with the kth element assigned a Gaussian prior distribution with mean m_k and variance $\sigma_k^2 v_k$, hence we write $\theta_k \sim N(m_k, \sigma_k^2 v_k)$, for $k = 1, \ldots, n$. Therefore $\boldsymbol{m} = (m_0, \ldots, m_{p-1})'$ and

$$\boldsymbol{C}_3 = \begin{bmatrix} \sigma_0^2 v_0 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 v_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{p-1}^2 v_{p-1} \end{bmatrix}.$$

The hierarchical form of model (3.3) is identical to that of the three stage model introduced in Section 2.4, where here we have $G = n$, $P = np$ and $H = p$. Therefore we can use the results of that section to compute the posterior covariance matrices $Var(\boldsymbol{\beta}|\{\boldsymbol{C}_i\}_{i=1,2,3}, \boldsymbol{y})$, $Cov(\boldsymbol{\beta}, \boldsymbol{\theta}|\{\boldsymbol{C}_i\}_{i=1,2,3}, \boldsymbol{y})$, $Var(\tilde{\boldsymbol{\beta}}|\{\boldsymbol{C}_i\}_{i=1,2,3}, \boldsymbol{y})$, $Cov(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}|\{\boldsymbol{C}_i\}_{i=1,2,3}, \boldsymbol{y})$ and $Var(\boldsymbol{\theta}|\{\boldsymbol{C}_i\}_{i=1,2,3}, \boldsymbol{y})$.

Model (3.3) extends many of the models that appear in the literature. If we let $\beta_1(\boldsymbol{s}_i) = \ldots = \beta_{p-1}(\boldsymbol{s}_i) = 0$, then the model collapses to model (3.1), the standard Gaussian process geostatistical model found in, for example, Banerjee et al. (2003, Chapter 5) or Schabenberger and Gotway (2004, Chapter 6). In much of this chapter we use

model (3.1) with $\boldsymbol{x}(\boldsymbol{s}_i) = 1$ to investigate how the strength of correlation between the random effects impacts the convergence rate for the different parameterisations. This same simplified model was used by Banerjee et al. (2008) to investigate the efficacy of Gaussian predictive process models.

If we have just one covariate and it has a spatially varying coefficient, i.e. in model (3.3) we let $p = 2$, then we have

$$Y(\boldsymbol{s}_i) = \theta_0 + \beta_0(\boldsymbol{s}_i) + \{\theta_1 + \beta_1(\boldsymbol{s}_i)\}x_1(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \tag{3.4}$$

which is the model used by Berrocal et al. (2010) to model ground-level ozone concentrations for the eastern states of the U.S. We will revisit model (3.4) in Section 4.5 as we use it to model Californian ozone concentration data. In Section 3.5 we investigate the effect of covariate information upon the convergence rates for the different parameterisations by using model (3.4) with $\theta_0 = \beta_0(\boldsymbol{s}_i) = 0$.

The general form of the posterior precision matrices for the CP and the NCP of the three stage model are given in Section 2.4.3. Using those results we can write down the posterior precision matrix for the CP of model (3.3) as

$$\boldsymbol{Q}^c = \begin{pmatrix} \boldsymbol{Q}^c_{\tilde{\beta}} & \boldsymbol{Q}^c_{\tilde{\beta}\theta} \\ \boldsymbol{Q}^c_{\theta\tilde{\beta}} & \boldsymbol{Q}^c_{\theta} \end{pmatrix}, \tag{3.5}$$

where $\boldsymbol{Q}^c_{\tilde{\beta}}$ is an $np \times np$ block matrix with blocks corresponding to the processes $\tilde{\boldsymbol{\beta}}_k$, for $k = 0, \ldots, p-1$. The klth $n \times n$ block of $\boldsymbol{Q}^c_{\tilde{\beta}}$ is given by

$$(\boldsymbol{Q}^c_{\tilde{\beta}})_{kl} = \begin{cases} \boldsymbol{D}_k \boldsymbol{C}_1^{-1} \boldsymbol{D}_k + \boldsymbol{\Sigma}_k^{-1} & \text{if } l = k, \\ \boldsymbol{D}_k \boldsymbol{C}_1^{-1} \boldsymbol{D}_l & \text{if } l \neq k, \end{cases}$$

for $k, l = 0, \ldots, p-1$, and we define $\boldsymbol{D}_0 = \boldsymbol{I}$. The submatrix $\boldsymbol{Q}^c_{\tilde{\beta}\theta} = (\boldsymbol{Q}^c_{\theta\tilde{\beta}})'$ is block diagonal with the kth block equal to

$$(\boldsymbol{Q}^c_{\tilde{\beta}\theta})_k = -\boldsymbol{\Sigma}_k^{-1}\mathbf{1}, \quad k = 0, \ldots, p-1.$$

The submatrix $\boldsymbol{Q}^c_{\theta}$ is diagonal with the kth diagonal entry equal to

$$(\boldsymbol{Q}^c_{\theta})_k = \mathbf{1}'\boldsymbol{\Sigma}_k^{-1}\mathbf{1} + 1/(\sigma_k^2 v_k), \quad k = 0, \ldots, p-1.$$

The form of $\boldsymbol{Q}^c$ indicates the conditional independence between the $\tilde{\boldsymbol{\beta}}_k$ and θ_l and between θ_k and θ_l for $l \neq k$, given the rest of the parameters in the model. We can write these statements as

$$\tilde{\boldsymbol{\beta}}_k \perp\!\!\!\perp \theta_l | \tilde{\boldsymbol{\beta}}_{-k}, \boldsymbol{\theta}_{-l}, \quad k, l = 0, \ldots, p-1, \quad l \neq k.$$

$$\theta_k \perp\!\!\!\perp \theta_l | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}_{-k,l}, \quad k, l = 0, \ldots, p-1, \quad l \neq k.$$

The posterior precision matrix for the NCP is written as

$$\boldsymbol{Q}^{nc} = \begin{pmatrix} \boldsymbol{Q}^{nc}_{\beta} & \boldsymbol{Q}^{nc}_{\beta\theta} \\ \boldsymbol{Q}^{nc}_{\theta\beta} & \boldsymbol{Q}^{nc}_{\theta} \end{pmatrix}, \tag{3.6}$$

where the $\boldsymbol{Q}^{nc}_{\beta} = \boldsymbol{Q}^{c}_{\tilde{\beta}}$. The submatrix $\boldsymbol{Q}^{nc}_{\beta\theta} = (\boldsymbol{Q}^{nc}_{\theta\beta})'$ is a $np \times np$ block matrix with klth $n \times n$ block equal to

$$(\boldsymbol{Q}^{nc}_{\beta\theta})_{kl} = \boldsymbol{D}_k \boldsymbol{C}_1^{-1} \boldsymbol{x}_l, \quad k, l = 0, \ldots, p-1,$$

where we define $\boldsymbol{x}_0 = \mathbf{1}$. The submatrix $\boldsymbol{Q}^{nc}_{\theta}$ is a $p \times p$ matrix with klth entry equal to

$$(\boldsymbol{Q}^{nc}_{\theta})_{kl} = \begin{cases} \boldsymbol{x}_k \boldsymbol{C}_1^{-1} \boldsymbol{x}_k' + 1/(\sigma_k^2 v_k) & \text{if } l = k, \\ \boldsymbol{x}_k \boldsymbol{C}_1^{-1} \boldsymbol{x}_l' & \text{if } l \neq k, \end{cases}$$

for $k, l = 0, \ldots, p-1$.

It is shown in Section 2.3.3 that given the posterior precision matrix of a normal linear model we can compute the convergence rate for the associated Gibbs sampler. By Corollary 2.3.2 we have that for a 2×2 block precision matrix

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{pmatrix},$$

the convergence rate is the maximum modulus eigenvalue of the matrix

$$\boldsymbol{F}_{22} = \boldsymbol{Q}_{22}^{-1} \boldsymbol{Q}_{21} \boldsymbol{Q}_{11}^{-1} \boldsymbol{Q}_{12}.$$

Therefore, the convergence rate of the CP of model (3.3) with precision matrix (3.5) is given by the maximum modulus eigenvalue of

$$\boldsymbol{F}^{c}_{22} = (\boldsymbol{Q}^{c}_{\theta})^{-1} \boldsymbol{Q}^{c}_{\theta\tilde{\beta}} (\boldsymbol{Q}^{c}_{\tilde{\beta}})^{-1} \boldsymbol{Q}^{c}_{\tilde{\beta}\theta},$$

and the convergence rate of the NCP of model (3.3) with precision matrix (3.6) is given by the maximum modulus eigenvalue of

$$\boldsymbol{F}^{nc}_{22} = (\boldsymbol{Q}^{nc}_{\theta})^{-1} \boldsymbol{Q}^{nc}_{\theta\beta} (\boldsymbol{Q}^{nc}_{\beta})^{-1} \boldsymbol{Q}^{nc}_{\beta\theta}.$$

3.3 Convergence rates of the CP and the NCP in the presence of correlated random effects

To investigate the effect of correlation between the realisations of the latent processes upon the convergence rate of the different parameterisations, we let $p = 1$ and therefore model

(3.3) reduces to

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0 &\sim N(\tilde{\boldsymbol{\beta}}_0, \sigma_\epsilon^2 \boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}_0|\theta_0 &\sim N(\theta_0 \mathbf{1}, \boldsymbol{\Sigma}_0) \\ \theta_0 &\sim N(m_0, \sigma_0^2 v_0). \end{aligned} \tag{3.7}$$

Recalling that $\boldsymbol{\Sigma}_0 = \sigma_0^2 \boldsymbol{R}_0$, the posterior precision matrices for the CP and the NCP are given by

$$\boldsymbol{Q}^c = \begin{pmatrix} 1/\sigma_\epsilon^2 \boldsymbol{I} + 1/\sigma_0^2 \boldsymbol{R}_0^{-1} & -1/\sigma_0^2 \boldsymbol{R}_0^{-1}\mathbf{1} \\ -1/\sigma_0^2 \mathbf{1}'\boldsymbol{R}_0^{-1} & 1/\sigma_0^2 \mathbf{1}'\boldsymbol{R}_0^{-1}\mathbf{1} + 1/(\sigma_0^2 v_0) \end{pmatrix},$$

and

$$\boldsymbol{Q}^{nc} = \begin{pmatrix} 1/\sigma_\epsilon^2 \boldsymbol{I} + 1/\sigma_0^2 \boldsymbol{R}_0^{-1} & 1/\sigma_\epsilon^2 \mathbf{1} \\ 1/\sigma_\epsilon^2 \mathbf{1}' & n/\sigma_\epsilon^2 + 1/(\sigma_0^2 v_0) \end{pmatrix},$$

respectively. As θ_0 is a scaler the $\boldsymbol{F}_{22}$ matrices are also scalers and so the respective convergence rates are

$$\lambda_c = \left(1/\sigma_0^2 \mathbf{1}'\boldsymbol{R}_0^{-1}\mathbf{1} + 1/(\sigma_0^2 v_0)\right)^{-1} 1/\sigma_0^2 \mathbf{1}'\boldsymbol{R}_0^{-1} \left(1/\sigma_\epsilon^2 \boldsymbol{I} + 1/\sigma_0^2 \boldsymbol{R}_0^{-1}\right)^{-1} 1/\sigma_0^2 \boldsymbol{R}_0^{-1}\mathbf{1}, \tag{3.8}$$

and

$$\lambda_{nc} = \left(n/\sigma_\epsilon^2 + 1/(\sigma_0^2 v_0)\right)^{-1} 1/\sigma_\epsilon^2 \mathbf{1}' \left(1/\sigma_\epsilon^2 \boldsymbol{I} + 1/\sigma_0^2 \boldsymbol{R}_0^{-1}\right)^{-1} 1/\sigma_\epsilon^2 \mathbf{1}. \tag{3.9}$$

In the rest of this section we investigate how the structure of the correlation matrix $\boldsymbol{R}_0$ affects the values for λ_c and λ_{nc}.

3.3.1 Convergence rates for equi-correlated random effects

To illustrate how changing the strength of correlation between the random effects influences the convergence rates of the different parameterisations, we begin by assuming a equi-correlation model. We suppose that

$$(\boldsymbol{R}_0)_{ij} = \begin{cases} \rho & \text{if } i \neq j \\ 1 & \text{if } i = j, \end{cases} \tag{3.10}$$

for $0 \leq \rho < 1$.

To assist in the computation of convergence rates λ_c and λ_{nc} we make use of the following two matrix inversion lemmas.

Lemma 3.3.1 *(Woodbury, 1950)Let $\boldsymbol{N}$ be an $n \times n$ matrix, $\boldsymbol{U}$ be an $n \times m$ matrix, $\boldsymbol{M}$ be an $m \times m$ matrix and $\boldsymbol{V}$ be an $m \times n$ matrix, then*

$$(\boldsymbol{N} + \boldsymbol{UMV})^{-1} = \boldsymbol{N}^{-1} - \boldsymbol{N}^{-1}\boldsymbol{U}(\boldsymbol{M}^{-1} + \boldsymbol{VN}^{-1}\boldsymbol{U})^{-1}\boldsymbol{VN}^{-1}.$$

Lemma 3.3.2 *If $\boldsymbol{I}$ is the $n \times n$ identity matrix and $\boldsymbol{J}$ is an $n \times n$ matrix of ones, then*

$$(a\boldsymbol{I} + b\boldsymbol{J})^{-1} = \frac{1}{a}\boldsymbol{I} - \frac{b}{a(a+nb)}\boldsymbol{J},$$

for constants $a > 0$, $b \neq -(a/n)$.

This can be easily checked by direct multiplication and noting that

$$\boldsymbol{J}\boldsymbol{J} = \mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}' = \mathbf{1}n\mathbf{1}' = n\boldsymbol{J}.$$

Also Lemma 3.3.2 follows from Lemma 3.3.1 if we set $\boldsymbol{N} = a\boldsymbol{I}$, $\boldsymbol{U} = \mathbf{1}$, $\boldsymbol{M} = b\boldsymbol{I}$ and $\boldsymbol{V} = \mathbf{1}'$.

To compute the convergence rates given in (3.8) and (3.9) we must invert matrices $\sigma_0^2\boldsymbol{R}_0$ and $(1/\sigma_\epsilon^2\boldsymbol{I} + 1/\sigma_0^2\boldsymbol{R}_0^{-1})$. Using Lemma 3.3.1 we see that

$$(1/\sigma_\epsilon^2\boldsymbol{I} + 1/\sigma_0^2\boldsymbol{R}_0^{-1})^{-1} = \sigma_\epsilon^2\boldsymbol{I} - \sigma_\epsilon^2\boldsymbol{I}(\sigma_\epsilon^2\boldsymbol{I} + \sigma_0^2\boldsymbol{R}_0)^{-1}\sigma_\epsilon^2\boldsymbol{I}.$$

For $\boldsymbol{R}_0$ defined by (3.10) we write

$$\sigma_0^2\boldsymbol{R}_0 = \sigma_0^2(1-\rho)\boldsymbol{I} + \sigma_0^2\rho\boldsymbol{J}, \tag{3.11}$$

and

$$\sigma_\epsilon^2\boldsymbol{I} + \sigma_0^2\boldsymbol{R}_0 = (\sigma_\epsilon^2 + \sigma_0^2(1-\rho))\boldsymbol{I} + \sigma_0^2\rho\boldsymbol{J}. \tag{3.12}$$

Applying Lemma 3.3.2 to invert matrix (3.11) we have the following restrictions on ρ:

$$\sigma_0^2(1-\rho) \neq 0 \implies \rho \neq 1, \tag{3.13}$$

and

$$\sigma_0^2\rho \neq -\frac{\sigma_0^2(1-\rho)}{n} \implies \rho \neq -\frac{1}{n-1}. \tag{3.14}$$

Applying Lemma 3.3.2 to invert matrix (3.12) we have the following further restrictions on ρ:

$$\sigma_\epsilon^2 + \sigma_0^2(1-\rho) \neq 0 \implies \rho \neq \frac{\sigma_\epsilon^2 + \sigma_0^2}{\sigma_0^2}, \tag{3.15}$$

and

$$\sigma_0^2\rho \neq -\frac{\sigma_0^2(1-\rho) - \sigma_\epsilon^2}{n} \implies \rho \neq -\frac{\sigma_\epsilon^2 + \sigma_0^2}{(n-1)\sigma_0^2}, \tag{3.16}$$

Restriction (3.13) is satisfied by insisting that $\rho < 1$ as we have done in (3.10), and restriction (3.15) is trivially satisfied for $\sigma_\epsilon^2 > 0$. Restrictions (3.14) and (3.16) correspond to negative values of ρ and are non-linear functions of other parameters. For values of ρ around $\rho = -1/(n-1)$ or $\rho = -(\sigma_\epsilon^2 + \sigma_0^2)/(n-1)\sigma_0^2$ the matrix inversions are unstable and so are the subsequent calculations of the convergence rates. Hence we restrict ρ to take only non-negative values, as is usual in spatial data modelling.

After some cancellation we find the convergence rates for the CP and the NCP to be

$$\lambda_c = \frac{nv_0}{\sigma_0^2(1-\rho) + n\sigma_0^2\rho + nv_0}\left(\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho}\right), \tag{3.17}$$

and

$$\lambda_{nc} = \frac{n\sigma_0^2v_0}{\sigma_\epsilon^2 + n\sigma_0^2v_0}\left(\frac{\sigma_0^2(1-\rho) + n\sigma_0^2\rho}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho}\right). \tag{3.18}$$

Note that if we have a flat prior on θ_0, achieved by letting $1/v_0 = 0$, then

$$\lambda_c = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho},$$

and

$$\lambda_{nc} = \frac{\sigma_0^2(1-\rho) + n\sigma_0^2\rho}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho}. \tag{3.19}$$

and then $\lambda_c + \lambda_{nc} = 1$. For $\rho = 0$ we recover the rates for the independent random effects model, given in equations (2.6) and (2.7). Also note from equations (3.17) and (3.18) that if we use of a proper prior we speed up convergence.

We let $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ be the ratio of the variance parameters. Figure 3.1 shows convergence rates for the CP for $0 \leq \rho < 1$ and when $1/v_0 = 0$ for $n = 20, 50, 100, 250$, for five levels of $\delta_0 = 0.01, 0.1, 1, 10, 100$. The equivalent plot for the NCP is given in Figure 3.2. Recall that a lower rate indicates faster convergence. We can see clearly that for fixed n and ρ, increasing δ_0 improves the performance of the CP but worsens the performance of the NCP. For a fixed δ_0 and sample size n, increasing the strength of correlation reduces the convergence rate for the CP but increases it for the NCP. We also see that for a fixed ρ increasing n helps the CP but hinders the NCP.

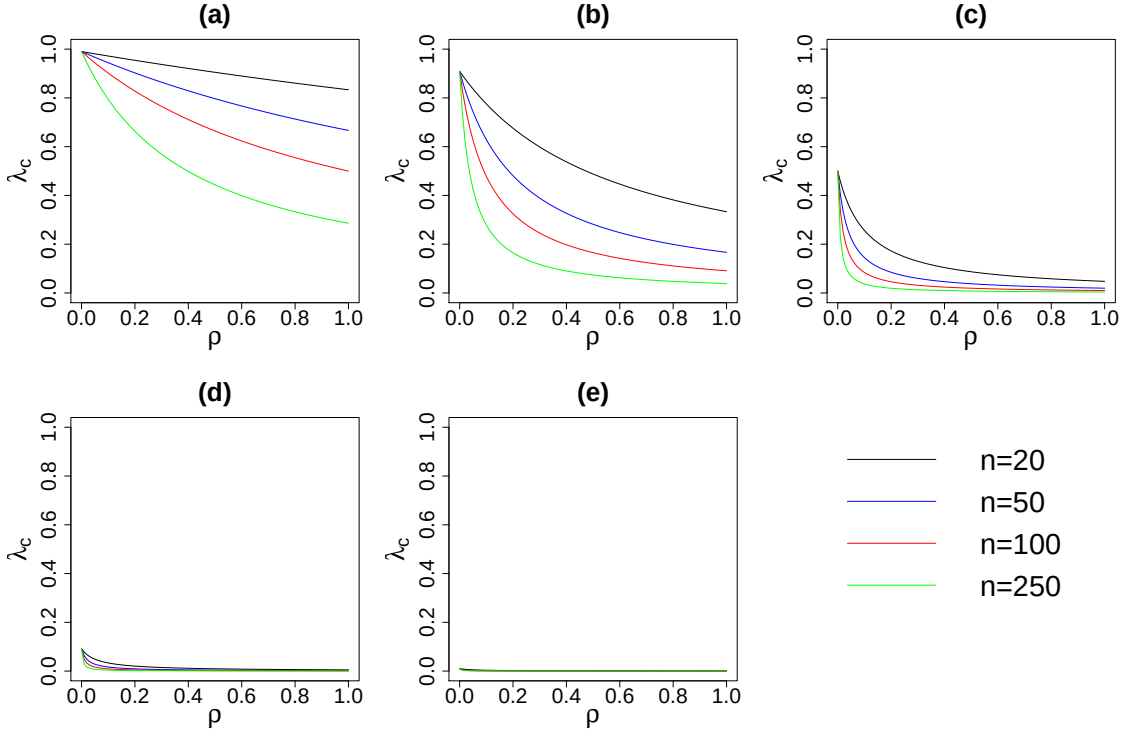

Figure 3.1: Convergence rates for the CP of the equi-correlation model for $n = 20, 50, 100, 250$, for different values of δ_0. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

The effect of a change of correlation on the two different parameterisations can be

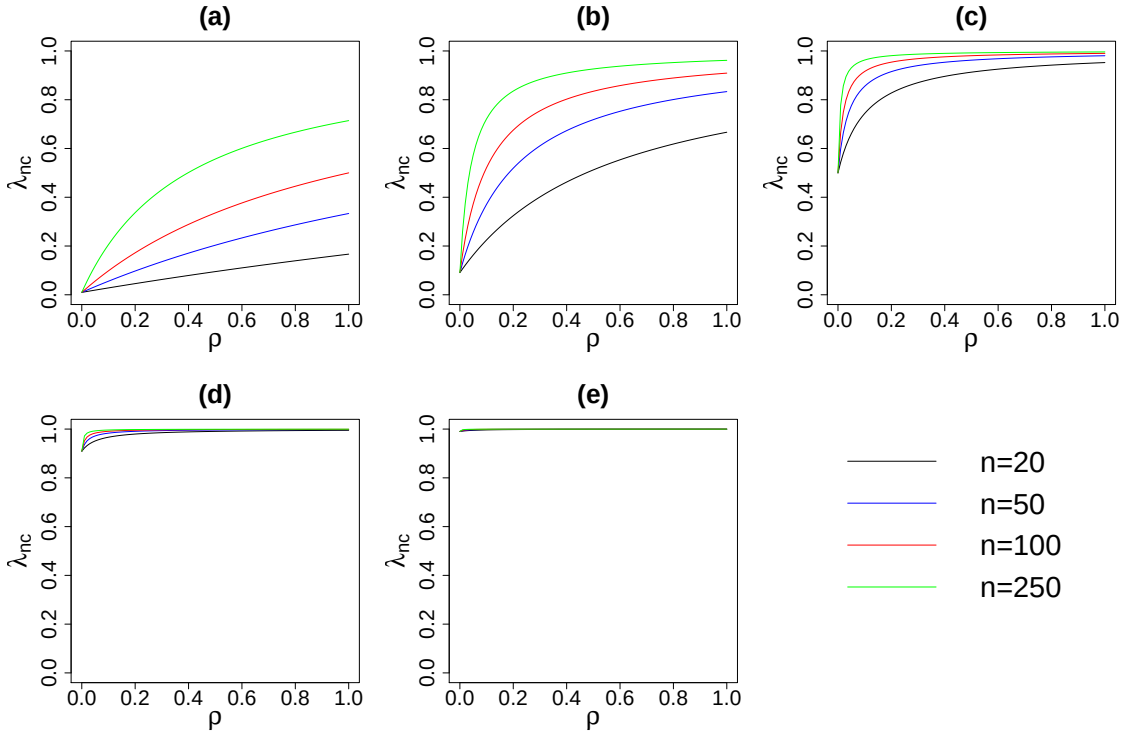

Figure 3.2: Convergence rates for the NCP of the equi-correlation model for $n = 20, 50, 100, 250$, for different values of δ_0. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

examined through the partial derivatives of the convergence rates with respect to ρ.

$$\frac{\partial \lambda_c}{\partial \rho} = -\frac{n(n-1)\sigma_\epsilon^2 v_0 \left[\sigma_\epsilon^2 + 2(\sigma_0^2(1-\rho) + n\sigma_0^2\rho) + n\sigma_0^2 v_0\right]}{\left((1-\rho) + n\rho + nv_0\right)^2 \left(\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho\right)^2} < 0, \tag{3.20}$$

$$\frac{\partial \lambda_{nc}}{\partial \rho} = \frac{n(n-1)\sigma_\epsilon^2 \sigma_0^4 v_0}{(\sigma_\epsilon^2 + n\sigma_0^2 v_0)\left(\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho\right)^2} > 0. \tag{3.21}$$

Equations (3.20) and (3.21) show that λ_c is monotonic decreasing function of ρ, and λ_{nc} is monotonically increasing in ρ.

When $1/v_0 = 0$ we have

$$\frac{\partial \lambda_c}{\partial \rho} = -\frac{(n-1)\sigma_\epsilon^2 \sigma_0^2}{(\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho)^2} < 0,$$

$$\frac{\partial \lambda_{nc}}{\partial \rho} = \frac{(n-1)\sigma_\epsilon^2 \sigma_0^2}{(\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho)^2} > 0,$$

and trivially $\partial\lambda_c/\partial\rho + \partial\lambda_{nc}/\partial\rho = 0$.

3.3.2 Convergence rates for spatially correlated random effects

In spatial modelling the correlation between two realisations of a latent process is usually assumed to be a function of their separation. Here we look at a commonly used correlation

function, namely the exponential function. The entries of the correlation matrix are given by

$$(\boldsymbol{R}_0)_{ij} = \exp\{-\phi d_{ij}\}, \tag{3.22}$$

where ϕ controls the rate of decay of correlation between sites random effects at $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and d_{ij} denotes the distance between them. By employing an exponential correlation function there is always non-zero correlation between any two realisations of $\boldsymbol{\beta}_0$, no matter how great the distance between them. This gives rise to the notion of an *effective range*, defined as the distance such that correlation falls to 0.05. For the exponential correlation function the effective range, d_0, is given by

$$\exp\{-\phi d_0\} = 0.05 \implies d_0 = -\log(0.05)/\phi \approx 3/\phi.$$

The exponential correlation function is a special case of the Matérn class of correlation functions, which we will revisit in Section 3.6.

We cannot compute explicit expressions for the entires of $\boldsymbol{R}_0^{-1}$ and hence we cannot find expressions for the convergence rate in terms of ϕ. Therefore we use a simulation approach to investigate how the convergence rates for the CP and the NCP are affected by changes in σ_0^2, σ_ϵ^2 and ϕ.

We randomly select $n = 40$ points in the unit square, which is taken to be spatial domain, see Figure 3.3. We compute the convergence rates given in equations (3.8) and (3.9)

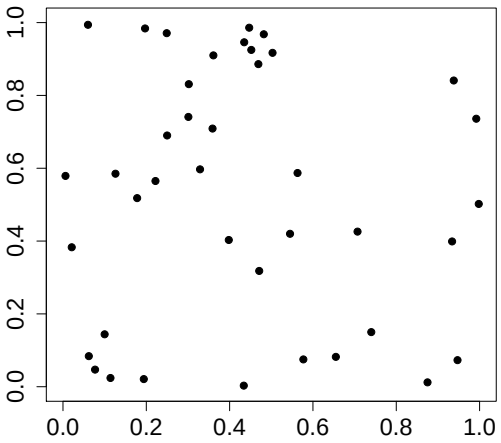

Figure 3.3: Points in the unit square used as sampling locations for simulating data from model (3.3).

for different variance ratios and for effective ranges between zero (no spatial correlation) and $\sqrt{2}$ (the maximum possible separation of two points in the domain). Again we let $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ and set δ_0 equal to 0.01, 0.1, 1, 10 and 100. Convergence rates are plotted against the effective range, d_0, for the CP and the NCP in Figure 3.4, where a lower rate indicates faster convergence. For a fixed d_0 we can see that increasing δ_0 decreases the convergence rate for the CP but increases it for the NCP. We also observe that for a fixed

level of δ_0 increasing d_0, thus increasing the strength of correlation between the random effects, decreases the convergence rate for the CP and increases it for the NCP.

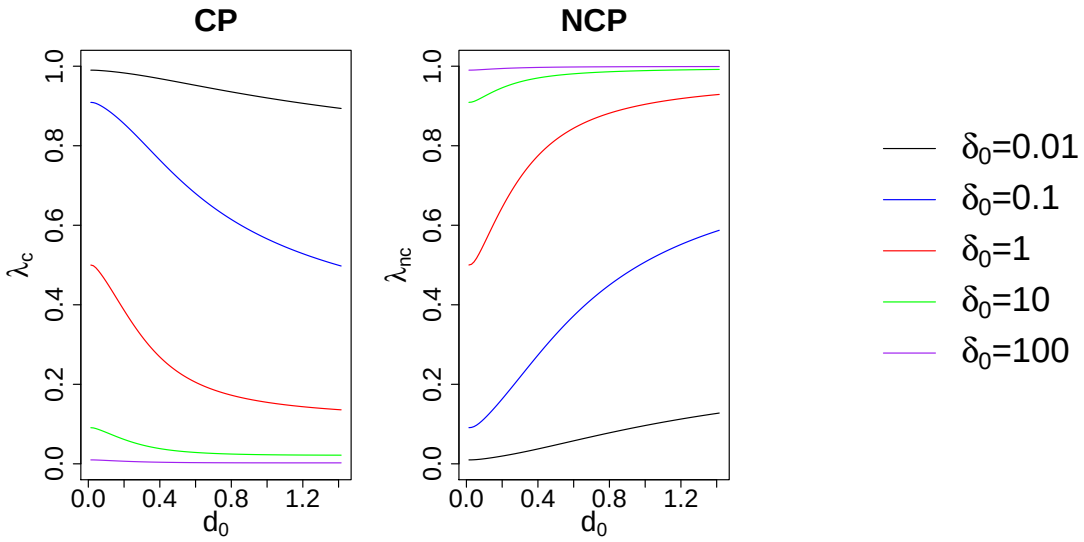

Figure 3.4: Convergence rate against effective range for the CP and the NCP at different levels of δ_0.

The convergence rates computed here are dependent on the set of locations given in Figure 3.3. For a different set of locations the rates are changed but the overall picture is not; increasing δ_0 or d_0 quickens convergence for the CP and slows convergence for the NCP.

3.4 Tapered covariance matrices

When spatial association is modelled as a Gaussian process the resulting covariances matrices are dense and inverting them can be slow or even infeasible for large n. In Section 1.3.2 we discuss covariance tapering (Furrer et al., 2006; Kaufman et al., 2008). The idea is to force to zero the entries in the covariance matrix that correspond to pairs of locations that are separated by a distance greater than a predetermined range. This results in sparse matrices that can be inverted more quickly than the original. In this section we investigate the effect covariance tapering on the convergence rates for the CP and the NCP. We take model (3.7) with an exponential correlation function for $\boldsymbol{R}_0$ and compare the convergence rates given in Section 3.3.2 with those computed when we use a tapered covariance matrix.

The tapered correlation matrix, $\boldsymbol{R}_{Tap}$, is the element wise product of the original correlation matrix $\boldsymbol{R}_0$ and the tapering correlation matrix $\boldsymbol{T}$, where $\boldsymbol{T}$ is a sparse matrix with ijth entry equal to zero if d_{ij} is greater than some threshold distance. Positive definiteness of $\boldsymbol{R}_{Tap}$ is assured if $\boldsymbol{T}$ is positive definite (Horn and Johnson, 2012, Theorem 7.5.3).

Given that our original correlation function is an exponential one, we follow Furrer

et al. (2006) and use a spherical tapering function such that

$$\boldsymbol{T}_{ij} = \begin{cases} 1 - \dfrac{3d_{ij}\chi}{2} + \dfrac{d_{ij}^3\chi^3}{2} & \text{if } d_{ij} < 1/\chi, \chi > 0 \\ 0 & \text{otherwise,} \end{cases}$$

with decay parameter χ, where $1/\chi$ is equal to the effective range, so that here we have $\chi = -\phi/\log(0.05)$. Therefore

$$(\boldsymbol{R}_{Tap})_{ij} = \begin{cases} \exp\{-\phi d_{ij}\}\left(1 - \dfrac{3d_{ij}\chi}{2} + \dfrac{d_{ij}^3\chi^3}{2}\right) & \text{if } d_{ij} < d_0, \phi > 0, \chi > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $d_0 = -\log(0.05)/\phi$ is the effective range.

As in Section 3.3.2 we use the $n = 40$ locations given in Figure 3.3 and let $\delta_0 = 0.01, 0.1, 1, 10$ and 100 and vary d_0 between 0 and $\sqrt{2}$.

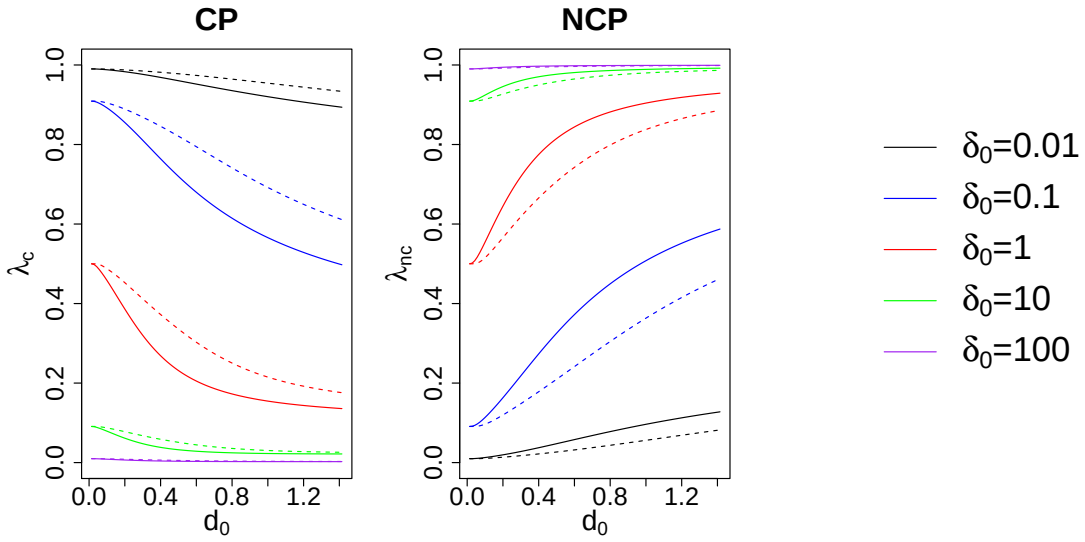

Figure 3.5: Convergence rates with tapered covariance matrices for the CP and the NCP at different levels of δ_0.

The convergence rates for the CP and the NCP are given in Figure 3.5. The dashed line represents the use of the tapered correlation matrix. The solid line for comparison are the rates achieved using the original correlation matrix $\boldsymbol{R}_0$ and are identical to those given in Figure 3.4. Convergence rates are slowed by tapering for the CP and hastened for the NCP. Intuitively we can say that the under the CP stronger correlation is desirable and tapering reduces that, with the opposite being true for the NCP.

We can illustrate this effect by considering a spatial model with just two locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ such that $\boldsymbol{s}_1 \neq \boldsymbol{s}_2$. Let $0 \leq corr(\beta(\boldsymbol{s}_1), \beta(\boldsymbol{s}_2)) = \rho < 1$. Suppose that we use a tapering function that takes values ρ^* if $d_{12} < d_0$ and zero otherwise, where $0 \leq \rho^* < 1$. The tapered correlation is

$$\rho_{Tap} = \begin{cases} \rho\rho^* & \text{if } d_{12} < d_0 \\ 0 & \text{otherwise.} \end{cases}$$

Therefore $\rho_{Tap} \leq \rho$, with equality attained only when $\rho = 0$. We know from equations (3.20) and (3.21) that

$$\frac{\partial \lambda_c}{\partial \rho} < 0 \quad \text{and} \quad \frac{\partial \lambda_{nc}}{\partial \rho} > 0,$$

and so for $n = 2$ tapering can only increase the convergence rate for the CP and only decrease it for the NCP.

3.5 Covariates and convergence rates

In this section we investigate the effect of the covariates upon the convergence rate. We consider the following model

$$Y(\boldsymbol{s}_i) = \{\theta_1 + \beta_1(\boldsymbol{s}_i)\}x_1(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \quad i = 1, \ldots, n, \tag{3.23}$$

which may be found by letting $k = 1, \ldots, p-1$, and $p = 2$ in model (3.3). Recalling that $\tilde{\boldsymbol{\beta}}_1 = (\tilde{\beta}_1(\boldsymbol{s}_1), \ldots, \tilde{\beta}_1(\boldsymbol{s}_n))'$, where $\tilde{\beta}_1(\boldsymbol{s}_i) = \beta_1(\boldsymbol{s}_i) + \theta_1$, and $\boldsymbol{x}_1 = (x_1(\boldsymbol{s}_1), \ldots, x_1(\boldsymbol{s}_n))'$ and $\boldsymbol{D}_1 = \text{diag}(\boldsymbol{x}_1)$, we can write model (3.23) in the following form

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_1 &\sim N(\boldsymbol{D}_1\tilde{\boldsymbol{\beta}}_1, \sigma_\epsilon^2\boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}_1|\theta_1 &\sim N(\theta_1\mathbf{1}, \sigma_1^2\boldsymbol{R}_1) \\ \theta_1 &\sim N(m_1, \sigma_1^2 v_1). \end{aligned} \tag{3.24}$$

Given that we consider only one covariate in the rest of this section we drop the subscript from $\boldsymbol{D}_1$ and $\boldsymbol{x}_1$

Using the results of Section 2.4.3 we can immediately write down the posterior precision matrix for the CP as

$$\boldsymbol{Q}^c = \begin{pmatrix} 1/\sigma_\epsilon^2\boldsymbol{D}\boldsymbol{D} + 1/\sigma_1^2\boldsymbol{R}_1^{-1} & -1/\sigma_1^2\boldsymbol{R}_1^{-1}\mathbf{1} \\ -1/\sigma_1^2\mathbf{1}'\boldsymbol{R}_1^{-1} & 1/\sigma_1^2\mathbf{1}'\boldsymbol{R}_1^{-1}\mathbf{1} + 1/(\sigma_1^2 v_1) \end{pmatrix}.$$

The equivalent matrix for the NCP is given by

$$\boldsymbol{Q}^{nc} = \begin{pmatrix} 1/\sigma_\epsilon^2\boldsymbol{D}\boldsymbol{D} + 1/\sigma_1^2\boldsymbol{R}_1^{-1} & 1/\sigma_\epsilon^2\boldsymbol{D}\boldsymbol{x} \\ 1/\sigma_\epsilon^2\boldsymbol{x}'\boldsymbol{D} & 1/\sigma_\epsilon^2\boldsymbol{x}'\boldsymbol{x} + 1/(\sigma_1^2 v_1) \end{pmatrix}.$$

3.5.1 Convergence rates for independent random effects

Suppose that random effects are independent. This can be considered the limiting case for weakening spatial correlation. The posterior precision matrix for the CP is

$$\boldsymbol{Q}^c = \begin{pmatrix} 1/\sigma_\epsilon^2\boldsymbol{D}\boldsymbol{D} + 1/\sigma_1^2\boldsymbol{I} & -1/\sigma_1^2\mathbf{1} \\ -1/\sigma_1^2\mathbf{1}' & n/\sigma_1^2 + 1/(\sigma_1^2 v_1) \end{pmatrix}.$$

For the sake of notational clarity, under the assumption of spatial independence we write $x(\boldsymbol{s}_i) = x_i$, for $i = 1, \ldots, n$. Therefore, the convergence rate for the CP is given by

$$\lambda_c = \frac{1}{n + 1/v_1} \sum_{i=1}^{n} \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_1^2 x_i^2}.$$

Letting $1/v_1 = 0$, we can write λ_c as

$$\lambda_c = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + (\sigma_1^2/\sigma_\epsilon^2) x_i^2}. \tag{3.25}$$

We introduce the variable $\delta_1 = \sigma_1^2/\sigma_\epsilon^2$. For fixed $\boldsymbol{x}$, we can see that as δ_1 tends to zero the convergence rate for the CP of model (3.23) tends to one. As δ_1 gets larger the convergence rate goes to zero.

To see the effect of the scale of $\boldsymbol{x}$ we introduce variables u_i, where

$$u_i = \frac{x_i - \bar{\boldsymbol{x}}}{sd_x}, \quad i = 1, \ldots, n, \tag{3.26}$$

and $\bar{\boldsymbol{x}}$ and sd_x are the sample mean and sample standard deviation of $\boldsymbol{x}$ respectively. Substituting equation (3.26) into equation (3.25) we have

$$\lambda_c = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + (\sigma_1^2/\sigma_\epsilon^2) \left(u_i sd_x + \bar{\boldsymbol{x}}\right)^2}.$$

We suppose that the x_i's have already been centred on zero and so $\bar{\boldsymbol{x}} = 0$. For fixed variance parameters, the effect of the scale of $\boldsymbol{x}$ is clear; an increase in sd_x results in a decrease in the convergence rate and vice versa.

For independent random effects the posterior precision matrix for the NCP becomes

$$\boldsymbol{Q}^{nc} = \begin{pmatrix} 1/\sigma_\epsilon^2 \boldsymbol{D}\boldsymbol{D} + 1/\sigma_1^2 \boldsymbol{I} & 1/\sigma_\epsilon^2 \boldsymbol{D}\boldsymbol{x} \\ 1/\sigma_\epsilon^2 \boldsymbol{x}'\boldsymbol{D} & 1/\sigma_\epsilon^2 \boldsymbol{x}'\boldsymbol{x} + 1/(\sigma_1^2 v_1) \end{pmatrix},$$

and the convergence rate is given by

$$\lambda_{nc} = \frac{1}{\sum_{i=1}^{n} x_i^2 + \sigma_\epsilon^2/(\sigma_1^2 v_1)} \sum_{i=1}^{n} \frac{\sigma_1^2 x_i^4}{\sigma_\epsilon^2 + \sigma_1^2 x_i^2}.$$

Letting $1/v_1 = 0$, we can write λ_{nc} as

$$\lambda_{nc} = \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} \frac{x_i^4}{(\sigma_\epsilon^2/\sigma_1^2) + x_i^2}. \tag{3.27}$$

For fixed $\boldsymbol{x}$, if $\sigma_\epsilon^2/\sigma_1^2$ goes to zero then λ_{nc} goes to one. Contrastingly, as the data variance dominates that of the random effects the convergence rate falls.

To see the effect of the scale of $\boldsymbol{x}$ upon λ_{nc} we substitute equation (3.26) into equation

(3.27). Then we have

$$\lambda_{nc} = \frac{1}{\sum_{i=1}^{n}(u_i sd_x + \bar{\boldsymbol{x}})^2} \sum_{i=1}^{n} \frac{(u_i sd_x + \bar{\boldsymbol{x}})^4}{(\sigma_\epsilon^2/\sigma_1^2) + (u_i sd_x + \bar{\boldsymbol{x}})^2}.$$

Again, assuming $\bar{\boldsymbol{x}} = 0$, we get

$$\begin{aligned} \lambda_{nc} &= \frac{1}{\sum_{i=1}^{n}(u_i sd_x)^2} \sum_{i=1}^{n} \frac{(u_i sd_x)^4}{(\sigma_\epsilon^2/\sigma_1^2) + (u_i sd_x)^2} \\ &= \frac{1}{\sum_{i=1}^{n} u_i^2} \sum_{i=1}^{n} \frac{u_i^4}{(\sigma_\epsilon^2/\sigma_1^2 sd_x^2) + u_i^2}. \end{aligned}$$

Fixing σ_ϵ^2 and σ_1^2, as sd_x tends to infinity, λ_{nc} tends to 1, as sd_x tends to zero, λ_{nc} tends to 0.

3.5.2 Convergence rates for spatially correlated random effects

In this section we investigate the effect that increasing the strength of correlation between realisations of the slope surface has upon the performance of the CP and the NCP. We let $(\boldsymbol{R}_1)_{ij} = \exp\{-\phi d_{ij}\}$ and so the effective range $d_1 = -\log(0.05)/\phi$. We use the $n = 40$ locations given in Figure 3.3. To generate the values of $\boldsymbol{x}$ we select a point $\boldsymbol{s}_x$, which we may imagine to be the site of a source of pollution. We assume that the value for the observed covariate at site $\boldsymbol{s}$ decays exponentially at a rate ϕ_x with increasing separation from $\boldsymbol{s}_x$, so that

$$x(\boldsymbol{s}_i) = \exp\{-\phi_x \|\boldsymbol{s}_i - \boldsymbol{s}_x\|\}, \quad i = 1, \ldots, n.$$

The spatial decay parameter ϕ_x is chosen such that there is an effective spatial range of $\sqrt{2}/2$, i.e. if $\|\boldsymbol{s} - \boldsymbol{s}_x\| = \sqrt{2}/2$ then $x(\boldsymbol{s}) = 0.05$. The values of $\boldsymbol{x}$ are standardised by subtracting their sample mean and dividing by their sample standard deviation.

We compute the convergence rate for the CP and the NCP for model (3.24) for five values of $\delta_1 = 0.01, 0.1, 1, 10, 100$, and for an effective range d_1 between 0 and $\sqrt{2}$. Results are given in Figure 3.6.

We see that for the CP for a fixed d_1, increasing δ_1 achieves faster convergence. If we fix δ_1 the performance of the CP is improved as the effective range is increased. The opposite is seen for the NCP, whose performance is improved by decreasing δ_1 or shortening the effective range.

3.6 The effect of the correlation function upon the convergence rate

The exponential correlation function is a special case of a more general class of correlations functions; the Matérn class (Handcock and Stein, 1993; Matérn, 1986). The correlation

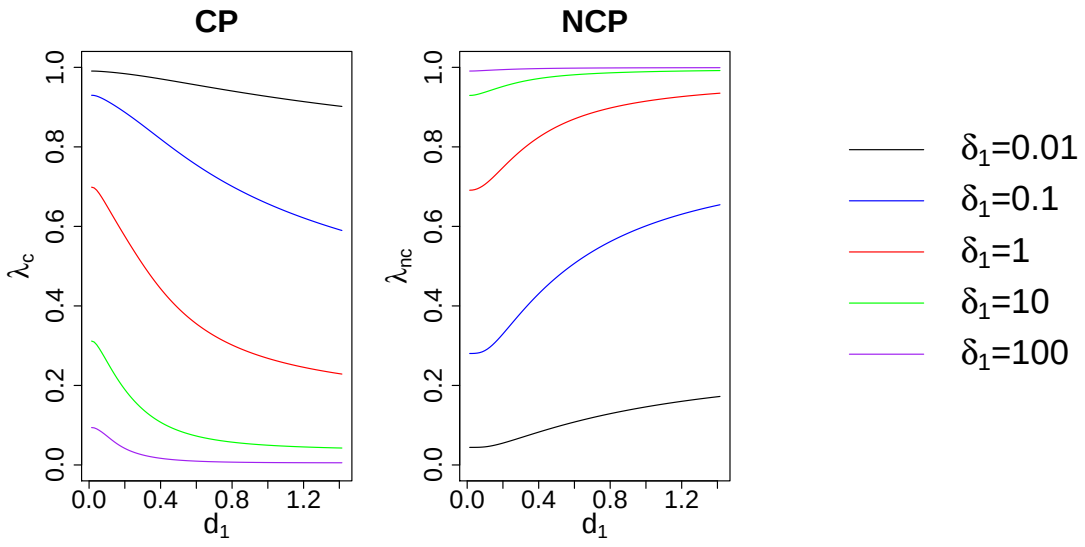

Figure 3.6: A comparison of convergence rates for the CP and the NCP at different levels of δ_1.

between realisations of the Gaussian process at two sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is given by

$$\rho(d_{ij}, \phi, \nu) = \frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\phi d_{ij})^{\nu} K_{\nu}(\sqrt{2\nu}\phi d_{ij}), \quad \phi > 0, \nu > 0, \tag{3.28}$$

where $\Gamma(\cdot)$ is the gamma function and $K_{\nu}(\cdot)$ is the modified Bessel function of the second kind of order ν (Abramowitz and Stegun, 1972, Section 9.6). The parameter ϕ controls rate of decay of the correlation between two points as their separation increases. The smoothness of the realised random field is controlled by ν, as the process realisations are $\lfloor \nu \rfloor$-times mean-square differentiable. Again d_{ij} is the distance between sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$.

A number of parameterisations of the Matérn correlation function exist, for examples see Schabenberger and Gotway (2004, Section 4.7.2). The form given in (3.28) is taken from Rasmussen and Williams (2006, Section 4.2.1), and has the advantage that for $\nu = 0.5$ it is identical to the exponential correlation function given in (3.22). To see this we can use the following results given by Schabenberger and Gotway (2004, Section 4.3.2)

$$\Gamma(0.5) = \sqrt{\pi}, \quad K_{0.5}(t) = \sqrt{\frac{\pi}{2t}} e^{-t},$$

and substituting into (3.28) we get

$$\begin{aligned} \rho(d_{ij}, \phi, \nu) &= \left(\frac{2}{\pi}\right)^{0.5} (\phi d_{ij})^{0.5} \left(\frac{\pi}{2\phi d_{ij}}\right)^{0.5} \exp\{-\phi d_{ij}\} \\ &= \exp\{-\phi d_{ij}\}. \end{aligned} \tag{3.29}$$

Another advantage of this parameterisation is that the decay parameter is not a function of ν as it is in the parameterisation given in Handcock and Wallis (1994) and employed in the R package geoR (Ribeiro Jr and Diggle, 2001).

When ν is a half integer, such that $\nu = b + 0.5$ where $b = 0, 1, 2, \ldots$, the correlation

function takes on a simpler form, so that

$$\rho(d_{ij},\phi,\nu)=\exp\{-\sqrt{2\nu}\phi d_{ij}\}\frac{\Gamma(b+1)}{\Gamma(2b+1)}\sum_{i=0}^{b}\frac{(b+i)!}{i!(b-i)!}(\sqrt{8\nu}\phi d_{ij})^{b-i}. \tag{3.30}$$

In particular when $\nu=1.5$ the correlation function is

$$\rho(d_{ij},\phi,\nu)=(1+\sqrt{3}\phi d_{ij})\exp\{-\sqrt{3}\phi d_{ij}\}, \tag{3.31}$$

and when $\nu=2.5$ it becomes

$$\rho(d_{ij},\phi,\nu)=\left(1+\sqrt{5}\phi d_{ij}+\frac{5\phi^2 d_{ij}^2}{3}\right)\exp\{-\sqrt{5}\phi d_{ij}\}. \tag{3.32}$$

As $\nu\to\infty$ the correlation function goes to

$$\rho(d_{ij},\phi,\nu)=\exp\left\{-\frac{\phi^2 d_{ij}^2}{2}\right\}, \tag{3.33}$$

which is sometimes known as the squared exponential or Gaussian correlation function.

We return to model (3.7), which we recall is

$$\begin{aligned}\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0 &\sim N(\tilde{\boldsymbol{\beta}}_0,\sigma_\epsilon^2\boldsymbol{I})\\ \tilde{\boldsymbol{\beta}}_0|\theta_0 &\sim N(\theta_0\mathbf{1},\sigma_0^2\boldsymbol{R}_0)\\ \theta_0 &\sim N(m_0,v_0).\end{aligned}$$

In what follows we compare the convergence rates for the CP and the NCP rates for the correlation functions given in equations (3.29)–(3.33) for the $n=40$ locations given in Figure 3.3.

In earlier sections we have considered the strength of correlation in terms of the effective range, which for the exponential correlation function is $-\log(0.05)/\phi$. In terms of ϕ the effective range for the Gaussian correlation function is given by $\sqrt{-2\log(0.05)}/\phi$. For other members of the Matérn class there is no closed form expression for the effective range. Therefore, for the cases when ν is equal to 1.5 and 2.5, we take an effective range d_0 and search for the value of ϕ that solves

$$\rho(d_0,\phi,\nu)-0.05=0,$$

where $\rho(d_0,\phi,\nu)$ is given by functions (3.31) and (3.32) respectively.

Convergence rates are computed for each parameterisation for effective ranges between 0 and $\sqrt{2}$ and for five values of $\delta_0=0.01,0.1,1,10,100$. The results for the CP are given in Figure 3.7. We see that for fixed ν and ϕ, increasing the δ_0 reduces the convergence rate. Also we see that for fixed ϕ and δ_0, the convergence rate is slowed when ν is increased, except for the $\delta_0=0.1$ case where the ordering only becomes apparent as the effective range is increased. Unlike in the case for $\nu=0.5$, increasing the effective range does not

reduce the convergence rate for other values of ν.

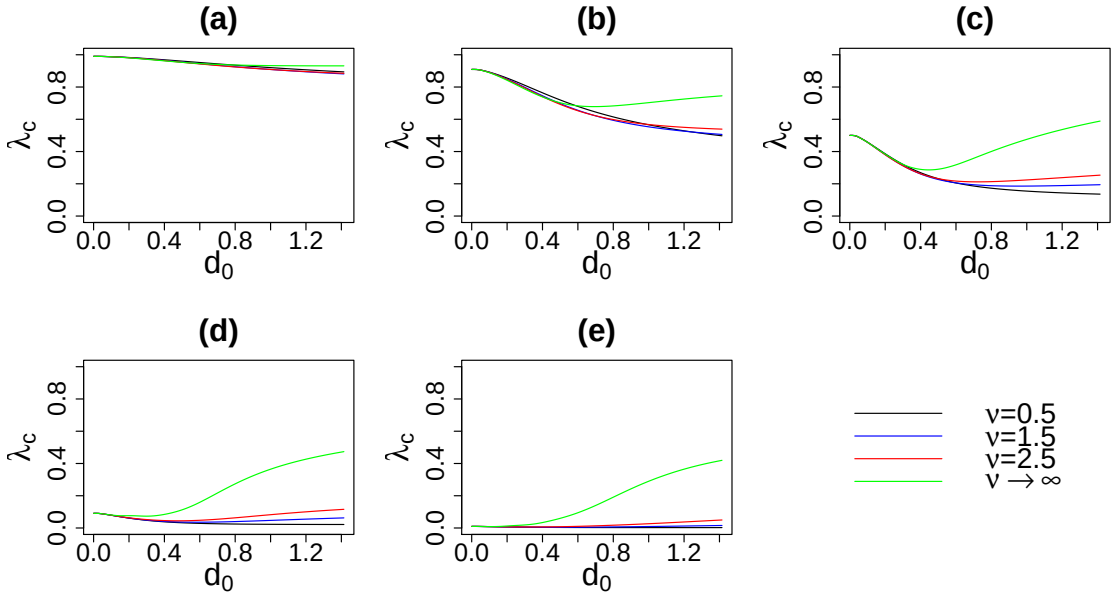

Figure 3.7: Convergence rates for the CP of model (3.7) for different values of ν. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

The equivalent plot for the NCP is given in Figure 3.8. For fixed ν and ϕ, increasing δ_0 increases the convergence rate. For fixed ϕ and δ_0, increasing ν slows convergence as it does for the CP. The convergence rate is monotonically increasing with increasing effective range for all four correlation functions. We also note that convergence rates for the NCP are not as sensitive to changes in ν as they are for the CP.

3.7 Geometric anisotropy

The class of Matérn correlation functions is isotropic. This means that the correlation between the random variables at any two points, $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, depends on the distance between them $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ (and parameters ϕ and ν) and hence the contours of iso-correlation are circular. The assumption that spatial dependence is the same in all directions is not always appropriate and therefore we may seek an *anisotropic* specification for the correlation structure.

Anisotropic correlation functions are widely used and have been employed to model, for example, scallop abundance in the North Atlantic (Ecker and Gelfand, 1999), extreme precipitation in Western Australia (Apputhurai and Stephenson, 2013) and the phenotypic traits of trees in northern Sweden (Banerjee et al., 2010).

Different forms of anisotropy exist, see Zimmerman (1993), but we consider only geometric anisotropy. Geometric anisotropic correlation functions can be constructed from isotropic correlation functions by taking a linear transformation of the lag vector $\boldsymbol{s}_i - \boldsymbol{s}_j$. Let

$$d^*_{ij} = \|\boldsymbol{G}(\boldsymbol{s}_i - \boldsymbol{s}_j)\|, \tag{3.34}$$

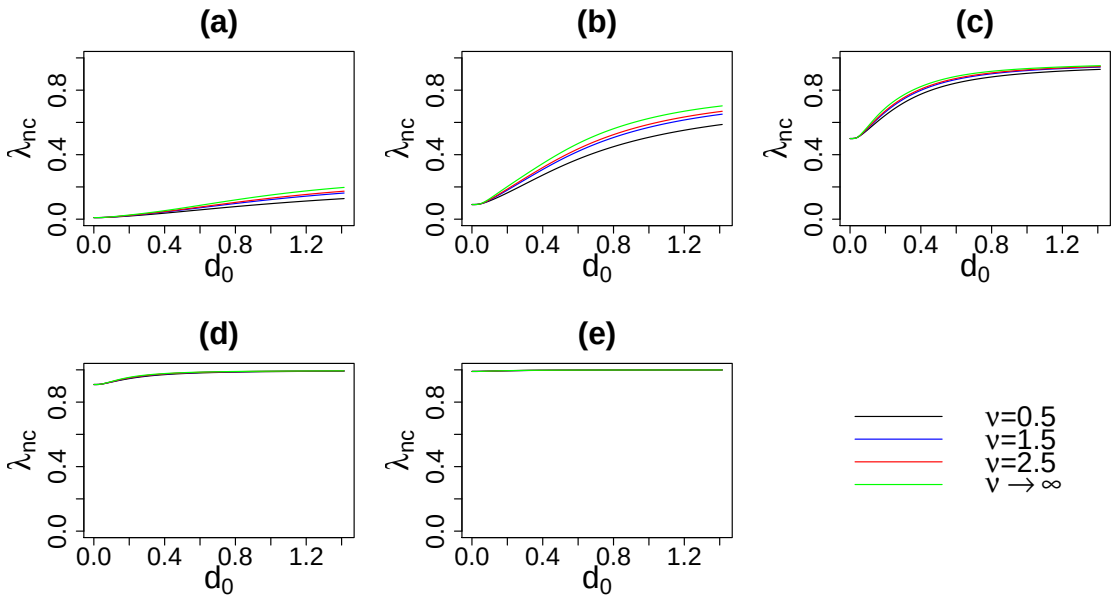

Figure 3.8: Convergence rates for the NCP of model (3.7) for different values of ν.(a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

where $\boldsymbol{G}$ is a 2×2 transformation matrix. In Euclidean space (3.34) is equivalent to

$$d_{ij}^* = \left[(\boldsymbol{s}_i - \boldsymbol{s}_j)' \boldsymbol{H} (\boldsymbol{s}_i - \boldsymbol{s}_j)\right]^{1/2},$$

where $\boldsymbol{H} = \boldsymbol{G}'\boldsymbol{G}$. The matrix $\boldsymbol{H}$ must be positive definite, i.e. $d_{ij}^* > 0$ for $\boldsymbol{s}_i \neq \boldsymbol{s}_j$, which is ensured if $\boldsymbol{G}$ is non-singular, see Harville (1997, Corrollary 14.2.14). By replacing d_{ij} with d_{ij}^* in (3.28) we have a geometric anisotropic Matérn correlation function with elliptical contours of iso-correlation.

Following Schabenberger and Gotway (2004, Chapter 4) we let

$$\boldsymbol{G} = \begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{bmatrix} = \begin{bmatrix} \alpha\cos\psi & \alpha\sin\psi \\ -\sin\psi & \cos\psi \end{bmatrix}, \tag{3.35}$$

hence the axis are rotated anti-clockwise through an angle ψ and then stretched in the direction of the x-axis by a factor $1/\alpha > 0$. The determinant of $\boldsymbol{G}$ is α and so it is non-singular for $\alpha \neq 0$, hence $\boldsymbol{H}$ is positive definite as required.

For $\boldsymbol{G}$ given in (3.35) we have

$$\boldsymbol{H} = \boldsymbol{G}'\boldsymbol{G} = \begin{bmatrix} \alpha^2\cos^2\psi + \sin^2\psi & (\alpha^2 - 1)\cos\psi\sin\psi \\ (\alpha^2 - 1)\cos\psi\sin\psi & \cos^2\psi + \alpha^2\sin^2\psi \end{bmatrix}.$$

If $\alpha = 1$, then $\boldsymbol{H}$ is the identity matrix and isotropy is recovered. If $\psi = 0 \pm 2\pi m$, m=1,2,..., then

$$\boldsymbol{H} = \begin{bmatrix} \alpha^2 & 0 \\ 0 & 1 \end{bmatrix}$$

which is equivalent to just a stretch of the x-axis by $1/\alpha$.

To illustrate the effect of the transformation matrix $\boldsymbol{G}$, we consider $\alpha = 0.5, 1, 2$ and $\psi = 0, \pi/4, \pi/2$ with an anisotropic exponential correlation function such that

$$\rho(d^*_{ij}, \phi) = \exp\{-\phi d^*_{ij}\}. \tag{3.36}$$

We take the point $\boldsymbol{s}^* = (0.5, 0.5)'$ in the unit square and fix decay parameter $\phi = 1$. We then compute the correlation between $\boldsymbol{s}^*$ and all points on a 20×20 grid, according to the correlation function given in (3.36). The values are then smoothed to produce a correlation surface. This is repeated for each of the nine combinations of α and ψ and displayed in Figure 3.9.

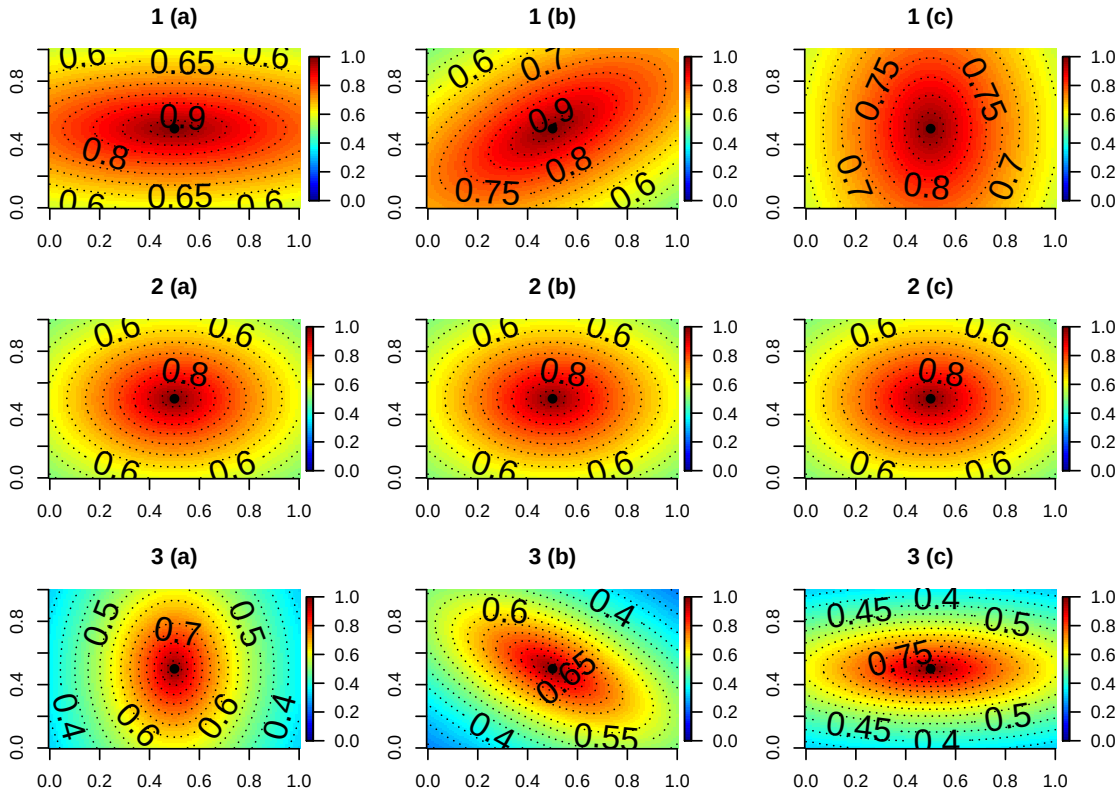

Figure 3.9: Correlation surface for $\beta(\boldsymbol{s}^*)$, $\boldsymbol{s}^* = (0.5, 0.5)'$, for exponential anisotropic correlation functions with transformation matrix $\boldsymbol{G}$ given in (3.35). Panels are given an alpha-numeric label. Numbers refer to three values of $\alpha = 0.5, 1, 2$. Letters (a), (b) and (c) refer to three values of $\psi = 0, \pi/4, \pi/2$.

We can see that setting $\alpha = 0.5$ strengthens correlation in the x-direction. This is because for the purposes of computing correlation, the separation of two points in the x-direction is halved. When $\alpha = 1$, the angle of rotation ψ does not effect the contours as they are circular.

To assess the impact of anisotropy on the convergence rates for the CP and the NCP we return to model (3.7) and the $n = 40$ locations given in Figure 3.3. We consider an anisotropic exponential correlation function for the spatial process and so replace d_{ij} with

d^*_{ij} in (3.22) such that

$$(\boldsymbol{R}_0)_{ij} = Corr(\beta(\boldsymbol{s}_i), \beta(\boldsymbol{s}_j)) = \exp\{-\phi d^*_{ij}\},$$

where d^*_{ij} is given by equation (3.34).

We begin by fixing $\psi = 0$ and letting $\alpha = 0.5, 1, 2$. This corresponds to panels 1 (a), 2 (a), and 3 (a), in Figure 3.9. We use five values for $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$ and vary ϕ such that $3/\phi \in (0, \sqrt{2}]$. Here, the effective range is direction dependent so we no longer refer to $3/\phi$ as the effective range. We compute convergence rates for the CP and the NCP as given by expressions (3.8) and (3.9) with $1/v_0 = 0$. Results for the CP and NCP are plotted in Figures 3.10 and 3.11 respectively. As α is reduced we increase the strength of correlation in the x-direction. This result is faster convergence for the CP and slower convergence for the NCP. This is consistent with the results of Section 3.3.2 which shows that increasing the effective range of an isotropic exponential correlation function, thus strengthening the correlation in all directions, helps the CP and hinders the NCP.

We now look at the effect of rotating the axis. If $\alpha = 1$ then a rotation has no impact on the correlation function as $\boldsymbol{G}$ is the identity. We consider four combinations of $\alpha = 0.5, 2$ and $\psi = \pi/4, \pi/2$. These values correspond to panels 1 (b) and 1 (c) for $\alpha = 0.5$, and 3 (b) and 3 (c) for $\alpha = 2$ in in Figure 3.9. Again, we let $\delta_0 = 0.01, 0.1, 1, 10, 100$ and vary ϕ such that $3/\phi \in (0, \sqrt{2}]$.

The results for the CP and the NCP are given in Figures 3.12 and 3.13 respectively. We can see that changing ψ has very little effect on the convergence rates of either parameterisation. Further investigation is needed to determine whether the same holds for patterned sampling locations.

The results given are for only one set of randomly selected locations and although the actual convergence rates are different for different sets of randomly selected locations, the picture remains the same.

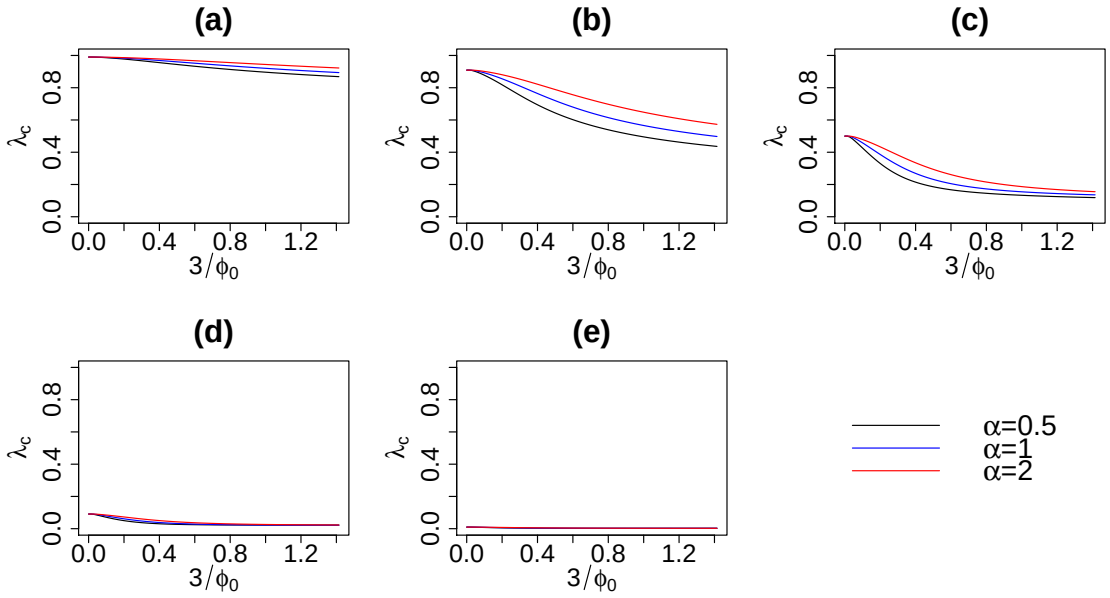

Figure 3.10: Convergence rates for the CP of model (3.7) with an anisotropic exponential correlation function for different values of α. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

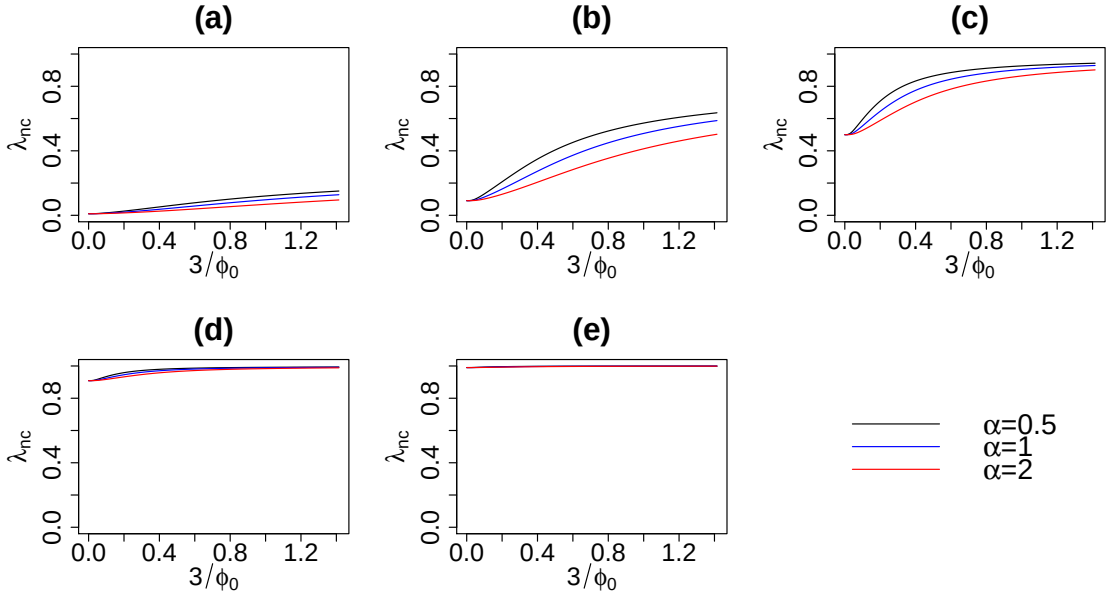

Figure 3.11: Convergence rates for the NCP of model (3.7) with an anisotropic exponential correlation function for different values of α. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

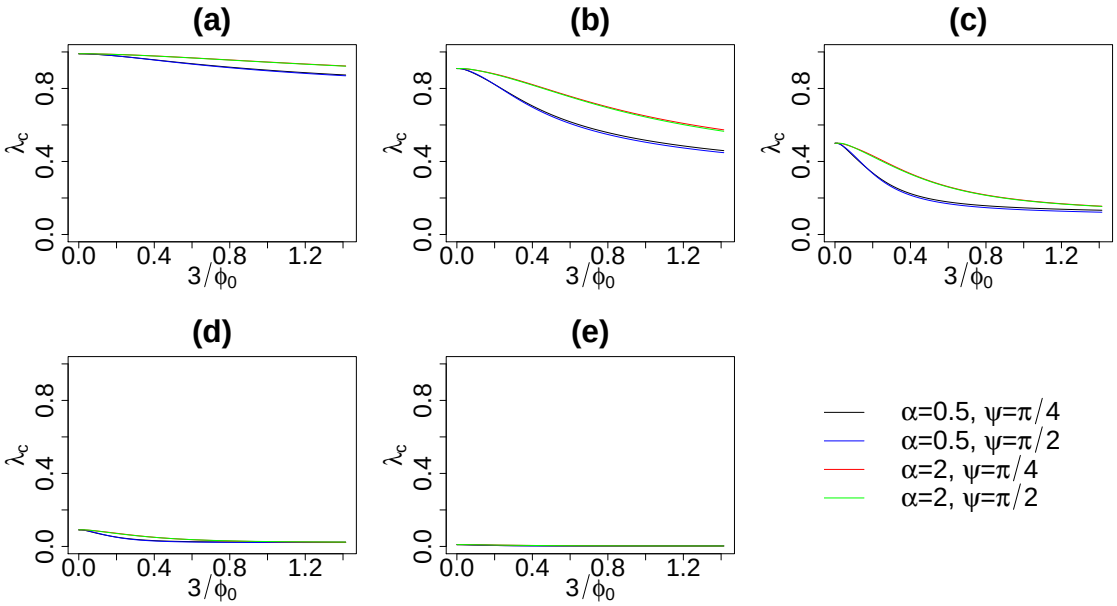

Figure 3.12: Convergence rates for the CP of model (3.7) with an anisotropic exponential correlation function for different values of α and ψ. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

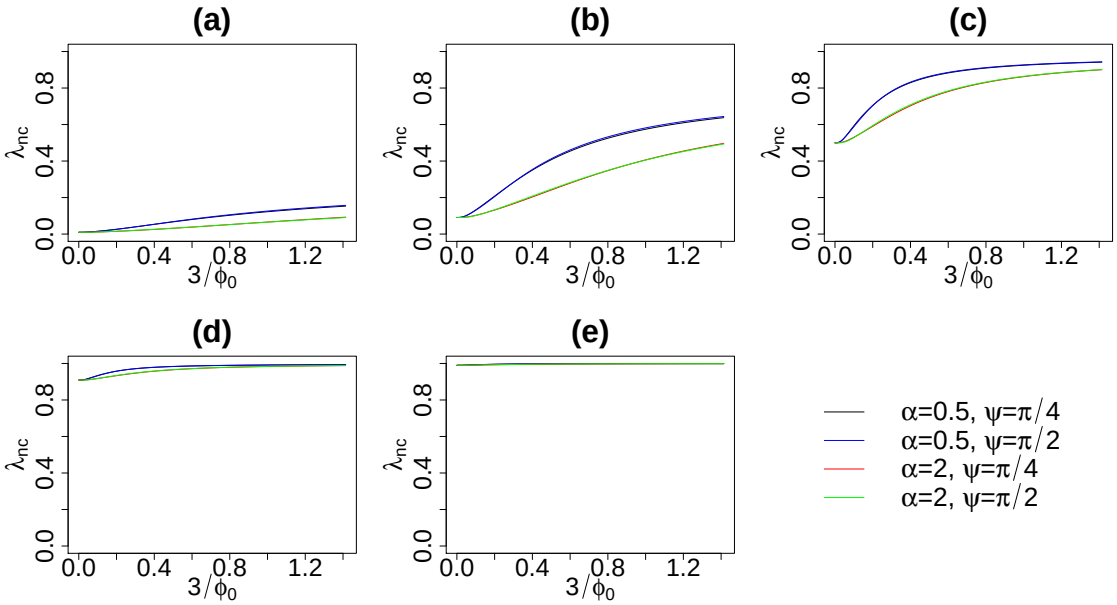

Figure 3.13: Convergence rates for the NCP of model (3.7) with an anisotropic exponential correlation function for different values of α and ψ. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

3.8 Blocking

So far we have considered the convergence rates when the random effects, the realisations of a Gaussian process at the sampling locations, have been updated all at once, or in other words in one *block*. In this section we consider the impact on the partitioning the set of random effects and updating them in two or more blocks.

It is acknowledged that jointly updating highly correlated variates may improve convergence (Gilks et al., 1996, Chapter 1). Therefore one may, as we have done, consider the random effects as one component. However, this comes at a computational cost. For example, in order to update an $n-$dimensional multivariate normal component, a Cholesky decomposition of an $n \times n$ matrix is performed. This operation is of cubic order in computational complexity. Therefore, if we can partition the component and update a series of lower dimensional components, we may achieve a more efficient Gibbs sampler. This section looks at how we might partition the random effects and the impact this has upon the convergence rate. In all that follows we use an exponential correlation function.

3.8.1 Blocking by location

In the spatial setting we assume that the correlation between random effects increases as the distance between the locations at which they are realised is shortened. Therefore we update the random effects according to their location by partitioning the spatial domain.

In this subsection we utilise model (3.7), which we recall is

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0 &\sim N(\tilde{\boldsymbol{\beta}}_0, \sigma_\epsilon^2 \boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}_0|\theta_0 &\sim N(\theta_0 \mathbf{1}, \sigma_0^2 \boldsymbol{R}_0) \\ \theta_0 &\sim N(m_0, v_0), \end{aligned}$$

and hence we have one spatial process and one global effect to update at each iteration of the Gibbs sampler. We consider $n = 200$ sampling locations randomly selected across the unit square.

We update the 200 random effects in one block and then in 2, 4, 8 and 16 blocks according to the partitioning of the unit square given in Figure 3.14, and hence locations that lie within the same section are updated together. We let $1/v_0 = 0$ and compare the convergence rates for the CP and the NCP for different values of $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ and for an effective range d_0 between 0 and $\sqrt{2}$, shown in Figures 3.15 and 3.16 respectively.

For both parameterisations we see that for a fixed δ_0 and at a fixed d_0, increasing the number of blocks slows convergence, as expected. As δ_0 increases, the convergence rate becomes less sensitive to the blocking strategy until at $\delta_0 = 100$ when the difference in convergence rates for different blocking strategies is negligible. In Section 3.3.2 we see that for model (3.7) with an exponential correlation function, increasing the effective range hastens convergence for the CP. Here we see that if more than one block is used, for a particular blocking strategy and δ_0, there is a critical effective range such that further increasing the strength of correlation results in a slower convergence. Beyond this

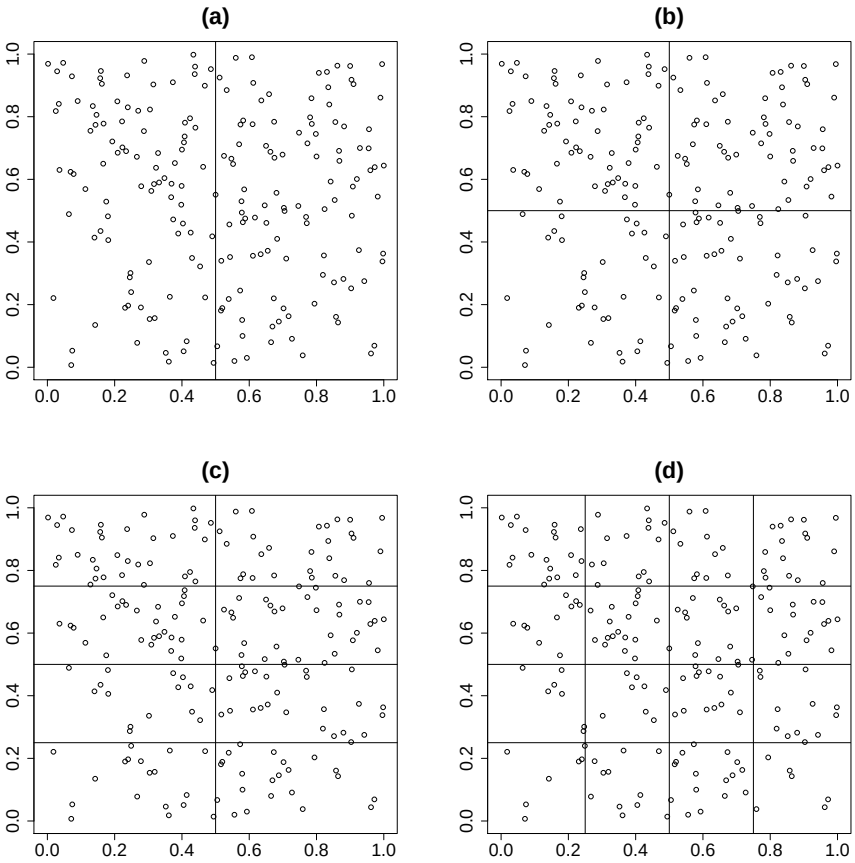

Figure 3.14: Partitioning of sampling locations used for blocking; (a) 2 blocks, (b) 4 blocks, (c) 8 blocks, (d) 16 blocks.

range, the penalty for partitioning $\tilde{\boldsymbol{\beta}}_0$, when there is inter-block correlation, overwhelms the improvement in convergence that is achieved with strengthening correlation. For the NCP we see that increasing the number of blocks exaggerates the effect of increasing the strength of correlation, which we see in Section 3.3.2 also slows convergence.

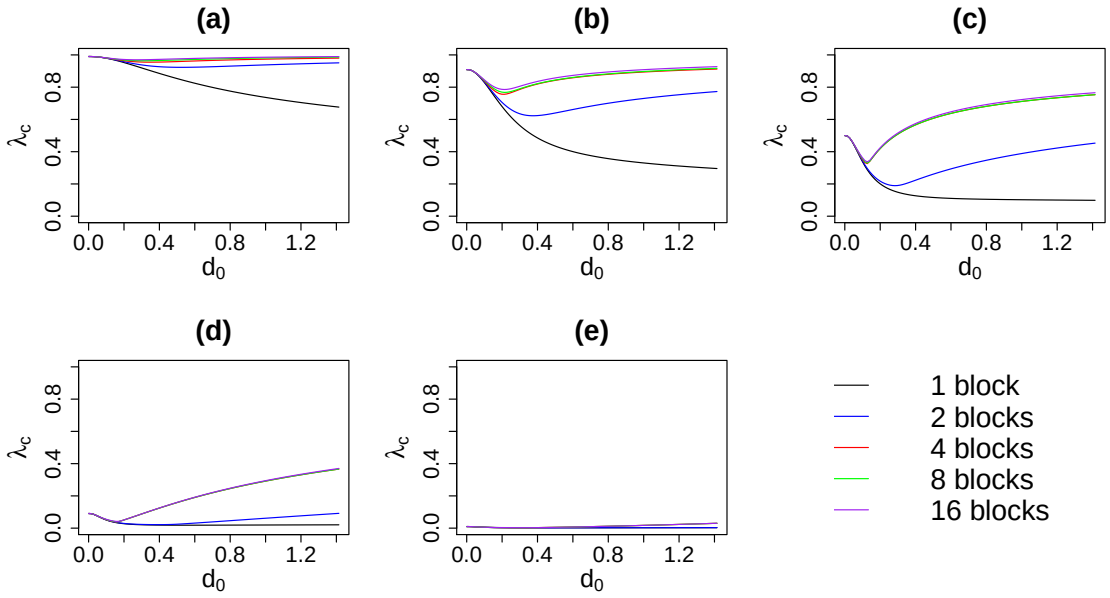

Figure 3.15: Convergence rates for the CP with blocking according to Figure 3.14 for effective ranges between 0 and $\sqrt{2}$. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

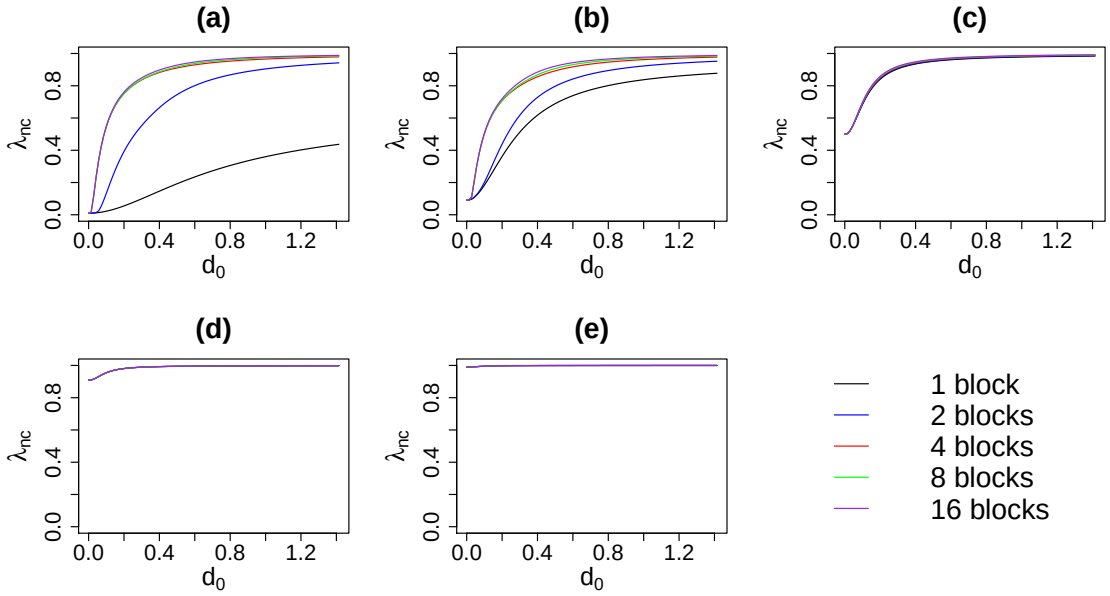

Figure 3.16: Convergence rates for the NCP with blocking according to Figure 3.14 for effective ranges between 0 and $\sqrt{2}$. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

3.8.2 Blocking by cluster

Instead of sampling uniformly over the unit square we now sample according to the pattern given in Figure 3.17, which is to divide the unit square into a 3×3 grid. Locations are chosen uniformly within the nine sub-squares as follows; 100 top left; 25 in top middle, middle left and middle middle; five top right, middle right and bottom third. We use three blocking strategies: one block to update all random effects together, two blocks to update the cluster of 100 locations separately from the rest, and nine blocks to update the random effects according to the blocks used to create the pattern of sampling locations.

We compare convergence rates for different blocking strategies for different ratios of the variance components and for effective ranges between 0 and $\sqrt{2}$. The rates for the CP and the NCP can be found in Figures 3.18 and 3.19 respectively.

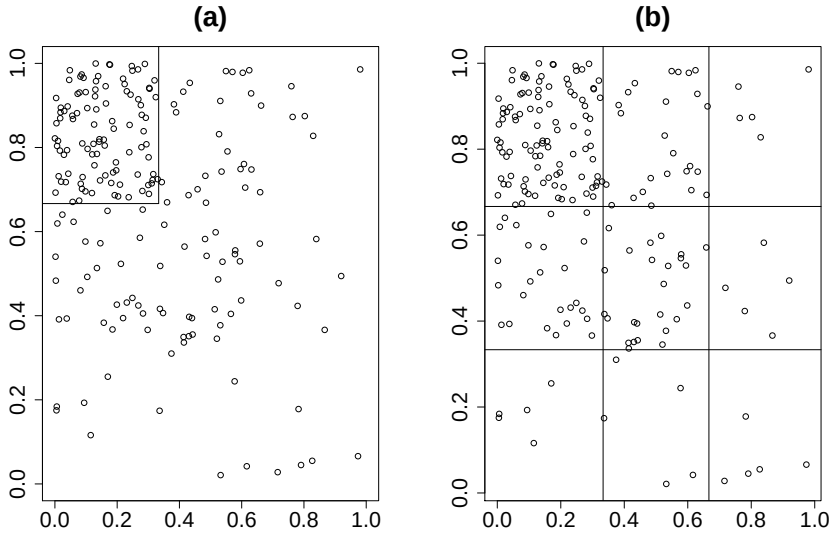

Figure 3.17: Locations chosen as follows: (a) two blocks, one for the dense cluster in the top left, and one for all other locations; (b) nine blocks according to the pattern of locations.

We see a similar result here as in Section 3.8.1. Given that there is spatial correlation between realisations of the Gaussian process at any pair of sampling locations, using more than one block to update the random effects will only slow convergence. Even for short effective ranges, at which the convergence rate for the CP is robust to the blocking strategy employed, the convergence rate for the NCP can be sorely effected if all random effects are not jointly updated. However it should be noted that the convergence rates computed here do not reflect the cost in computation time that is incurred by jointly updating all random effects. It may be more computationally efficient to use lower dimensional components for the Gibbs sampler, an issue we revisit in Chapter 4.

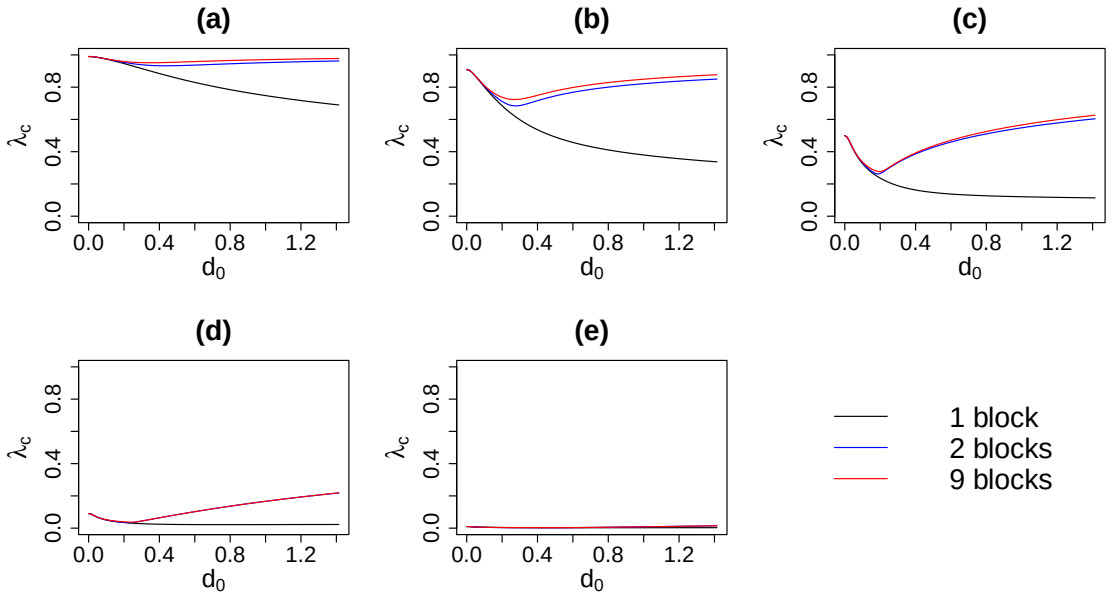

Figure 3.18: Convergence rates for the CP with blocking according to Figure 3.17 for effective ranges between 0 and $\sqrt{2}$. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

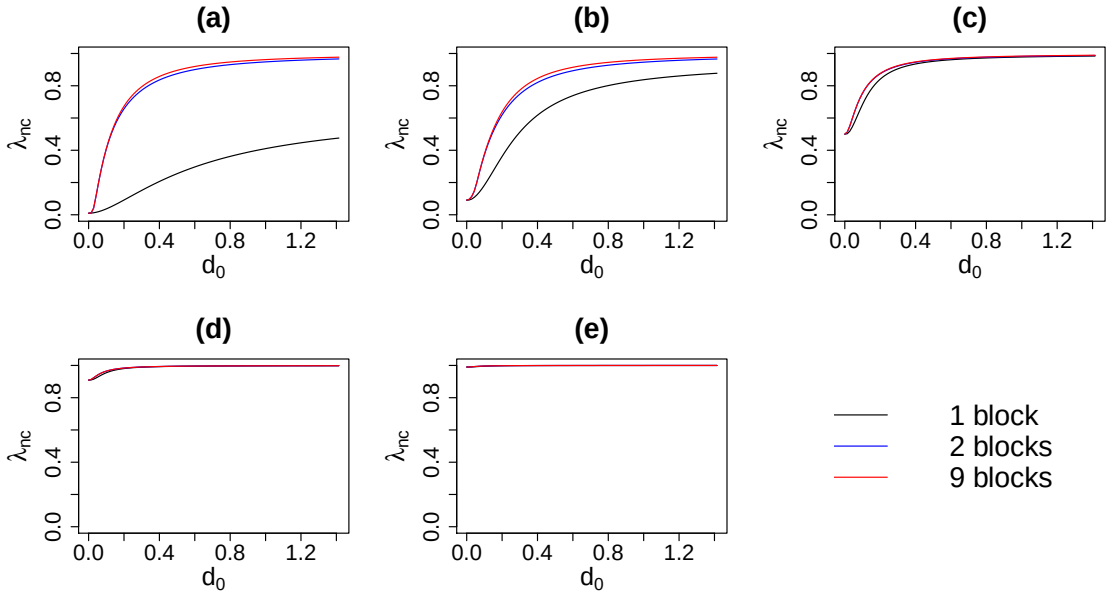

Figure 3.19: Convergence rates for the NCP with blocking according to Figure 3.17 for effective ranges between 0 and $\sqrt{2}$. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

3.8.3 Blocking and tapering

We have seen in Sections 3.8.1 and 3.8.2 that where there is correlation between all pairs of random effects we should use one block to jointly update all of the random effects. Here we use a tapered covariance matrix to induce independence between clusters of sampling locations. We might imagine that the clusters are separated by some geographical feature and for some reason we believe that there is negligible correlation between them.

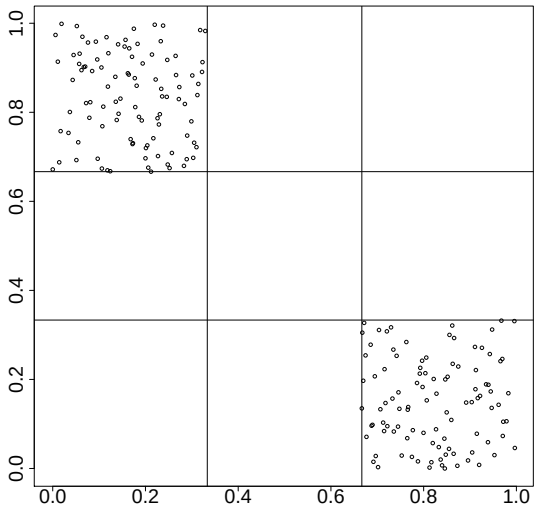

Figure 3.20: Pattern of $n = 200$ sampling locations split into two clusters of $n = 100$.

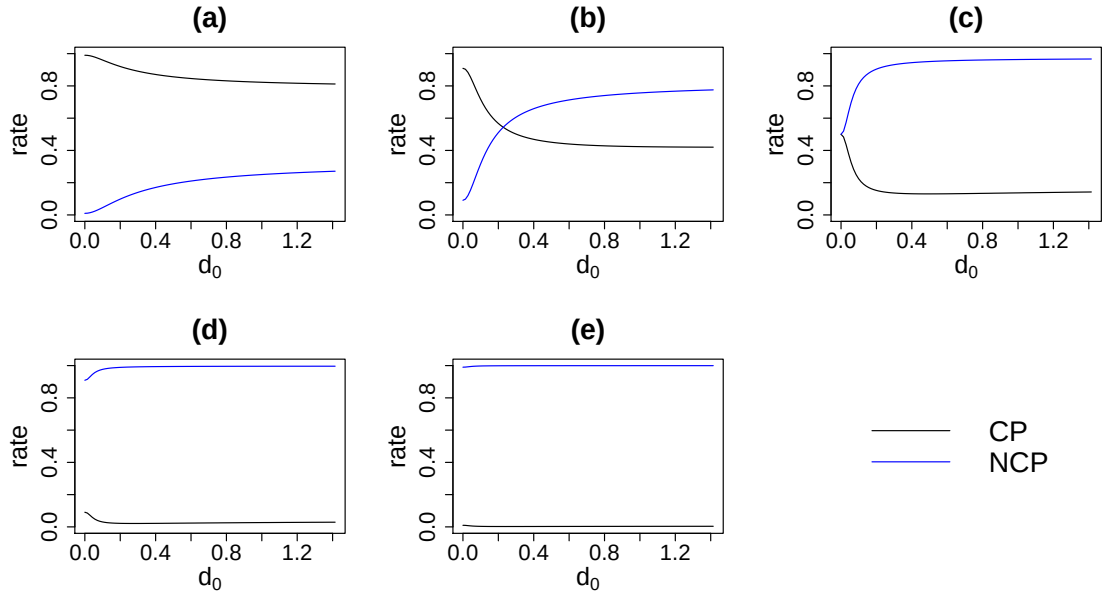

Figure 3.21: Convergence rates for the CP and the NCP for tapered covariance matrices for sampling locations given in Figure 3.20.

We place over the unit square a 3×3 grid and select 100 locations in the top left and bottom right sub-squares. The locations are given in Figure 3.20. We use a spherical

tapering function as described in Section 3.4, with a range of $1/\chi = \sqrt{2}/3$. Therefore, there is only non-zero correlation between pairs of realisations of the Gaussian process if they lie within the same cluster.

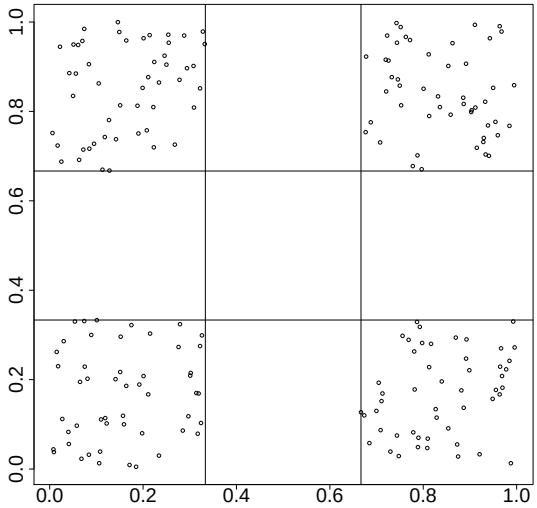

Figure 3.22: Pattern of $n = 200$ sampling locations split into four clusters of $n = 50$.

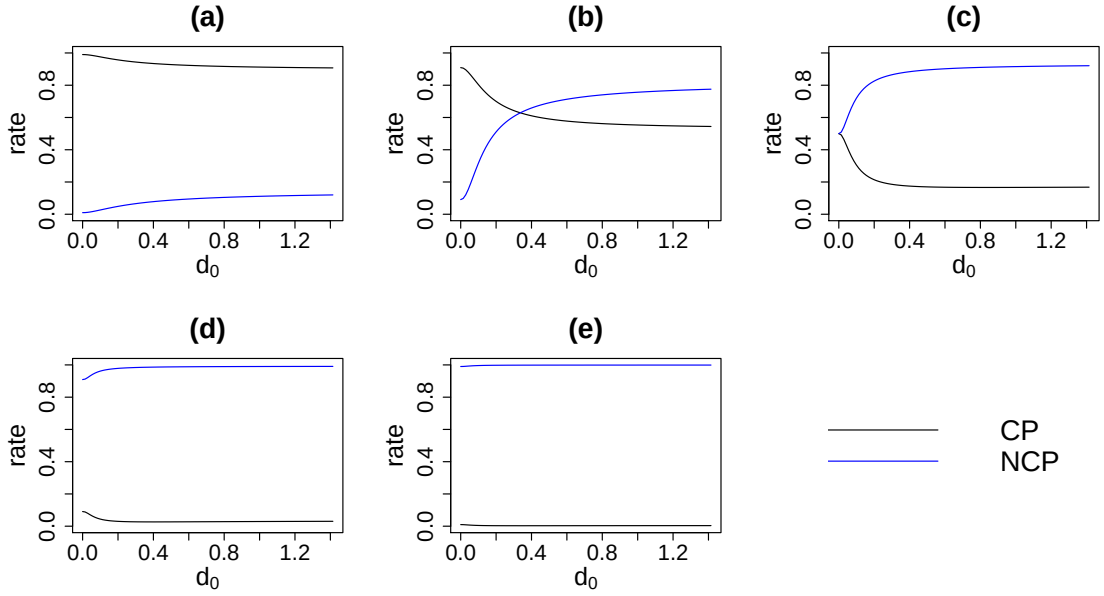

Figure 3.23: Convergence rates for the CP and the NCP for tapered covariance matrices for sampling locations given in Figure 3.22.

The convergence rates for the CP and the NCP are given in Figure 3.21. We find that whether the random effects are jointly updated or updated according to the two clusters, the convergence rate remains the same.

We repeat this approach with four clusters, see Figure 3.22, but now the range of the spherical tapering function is equal to 1/3, thus ensuring independence across the clusters. In this case whether one block or four blocks are used to update the random effects, the

convergence rates for both the CP and the NCP, given in Figure 3.23, remain unaffected.

3.8.4 Blocking by process

In Sections 3.8.1–3.8.3 we concerned ourselves with one spatial process and saw that in the presence of spatial correlation an *update all at once* approach for the random effects is best in terms of the convergence rate. Suppose that we have more than one process at the same set of locations. Consider model (3.3) where $p = 2$, so that we have the following hierarchically centred model

$$\begin{aligned}
\boldsymbol{Y} &\sim N(\tilde{\boldsymbol{\beta}}_0 + \boldsymbol{D}_1\tilde{\boldsymbol{\beta}}_1, \sigma_\epsilon^2\boldsymbol{I}) \\
\begin{pmatrix}\tilde{\boldsymbol{\beta}}_0 \\ \tilde{\boldsymbol{\beta}}_1\end{pmatrix} &\sim N\left(\begin{pmatrix}\mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}\end{pmatrix}\begin{pmatrix}\theta_0 \\ \theta_1\end{pmatrix}, \begin{pmatrix}\boldsymbol{\Sigma}_0 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_1\end{pmatrix}\right) \\
\begin{pmatrix}\theta_0 \\ \theta_1\end{pmatrix} &\sim N\left(\begin{pmatrix}m_0 \\ m_1\end{pmatrix}, \begin{pmatrix}\sigma_0^2 v_0 & 0 \\ 0 & \sigma_1^2 v_1\end{pmatrix}\right).
\end{aligned}$$

Using the results of Section 3.2 we can write down the posterior precision matrix for $(\tilde{\boldsymbol{\beta}}_0', \tilde{\boldsymbol{\beta}}_1', \theta_0, \theta_1)'$ as

$$\boldsymbol{Q}^c = \begin{pmatrix}
1/\sigma_\epsilon^2\boldsymbol{I} + \boldsymbol{\Sigma}_0^{-1} & 1/\sigma_\epsilon^2\boldsymbol{D}_1 & -\boldsymbol{\Sigma}_0^{-1}\mathbf{1} & \mathbf{0} \\
1/\sigma_\epsilon^2\boldsymbol{D}_1' & 1/\sigma_\epsilon^2\boldsymbol{D}_1'\boldsymbol{D}_1 + \boldsymbol{\Sigma}_1^{-1} & \mathbf{0} & -\boldsymbol{\Sigma}_1^{-1}\mathbf{1} \\
-\mathbf{1}'\boldsymbol{\Sigma}_0^{-1} & \mathbf{0} & \mathbf{1}'\boldsymbol{\Sigma}_0^{-1}\mathbf{1} + 1/(\sigma_0^2 v_0) & 0 \\
\mathbf{0} & -\mathbf{1}'\boldsymbol{\Sigma}_1^{-1} & 0 & \mathbf{1}'\boldsymbol{\Sigma}_1^{-1}\mathbf{1} + 1/(\sigma_1^2 v_1)
\end{pmatrix},$$

and where the random effects have zero mean *apriori* we have a posterior precision matrix of

$$\boldsymbol{Q}^{nc} = \begin{pmatrix}
1/\sigma_\epsilon^2\boldsymbol{I} + \boldsymbol{\Sigma}_0^{-1} & 1/\sigma_\epsilon^2\boldsymbol{D}_1 & 1/\sigma_\epsilon^2\mathbf{1} & 1/\sigma_\epsilon^2\boldsymbol{x} \\
1/\sigma_\epsilon^2\boldsymbol{D}_1' & 1/\sigma_\epsilon^2\boldsymbol{D}_1'\boldsymbol{D}_1 + \boldsymbol{\Sigma}_1^{-1} & 1/\sigma_\epsilon^2\boldsymbol{D}_1'\mathbf{1} & 1/\sigma_\epsilon^2\boldsymbol{D}_1'\boldsymbol{x} \\
1/\sigma_\epsilon^2\mathbf{1}' & 1/\sigma_\epsilon^2\mathbf{1}'\boldsymbol{D}_1 & n/\sigma_\epsilon^2 + 1/(\sigma_1^2 v_0) & 1/\sigma_\epsilon^2\mathbf{1}'\boldsymbol{x} \\
1/\sigma_\epsilon^2\boldsymbol{x}' & 1/\sigma_\epsilon^2\boldsymbol{x}'\boldsymbol{D}_1 & 1/\sigma_\epsilon^2\boldsymbol{x}'\mathbf{1} & 1/\sigma_\epsilon^2\boldsymbol{x}'\boldsymbol{x} + 1/(\sigma_1^2 v_1)
\end{pmatrix}.$$

For the CP, θ_0 and θ_1 are conditionally independent. If we transform the values $\boldsymbol{x}$ so that they have zero mean then $\mathbf{1}'\boldsymbol{x} = 0$ and θ_0 and θ_1 are also conditionally independent for the NCP. This in turn means that the convergence rate does not depend on whether the global effects are updated together or separately. To see this suppose that we update all random effects as one block and all global effects as another. Then the posterior precision matrix has the form

$$\boldsymbol{Q} = \begin{pmatrix}\boldsymbol{Q}_\beta & \boldsymbol{Q}_{\beta\theta} \\ \boldsymbol{Q}_{\theta\beta} & \boldsymbol{Q}_\theta\end{pmatrix},$$

and by Theorem 2.3.1 the convergence rate is the maximum modulus eigenvalue of

$$\boldsymbol{F} = \begin{pmatrix}\mathbf{0} & -\boldsymbol{Q}_\beta^{-1}\boldsymbol{Q}_{\beta\theta} \\ \mathbf{0} & \boldsymbol{Q}_\theta^{-1}\boldsymbol{Q}_{\theta\beta}\boldsymbol{Q}_\beta^{-1}\boldsymbol{Q}_{\beta\theta}\end{pmatrix}. \tag{3.37}$$

Now suppose instead that we partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$. We write the precision matrix as

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{\beta} & \boldsymbol{Q}_{\beta\theta_1} & \boldsymbol{Q}_{\beta\theta_2} \\ \boldsymbol{Q}_{\theta_1\beta} & \boldsymbol{Q}_{\theta_1} & \boldsymbol{Q}_{\theta_1\theta_2} \\ \boldsymbol{Q}_{\theta_2\beta} & \boldsymbol{Q}_{\theta_1\theta_2} & \boldsymbol{Q}_{\theta_2} \end{pmatrix}.$$

If $\boldsymbol{Q}_{\theta_1\theta_2} = \mathbf{0}$ it can be shown that the convergence rate is the maximum modulus eigenvalue of

$$\boldsymbol{F} = \begin{pmatrix} \mathbf{0} & -\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_1} & -\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_2} \\ \mathbf{0} & \boldsymbol{Q}_{\theta_1}^{-1}\boldsymbol{Q}_{\theta_1\beta}\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_1} & \boldsymbol{Q}_{\theta_1}^{-1}\boldsymbol{Q}_{\theta_1\beta}\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_2} \\ \mathbf{0} & \boldsymbol{Q}_{\theta_2}^{-1}\boldsymbol{Q}_{\theta_2\beta}\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_2} & \boldsymbol{Q}_{\theta_2}^{-1}\boldsymbol{Q}_{\theta_2\beta}\boldsymbol{Q}_{\beta}^{-1}\boldsymbol{Q}_{\beta\theta_2} \end{pmatrix}, \tag{3.38}$$

and matrices (3.37) and (3.38) are equal. Therefore, only the blocking structure of the random effects will impact upon the convergence rate.

We compare the convergence rates when all of the random effects are updated together and when they are updated according to the process of which they are realisations. We again use the $n = 40$ locations given in Figure 3.3 and generate values for $\boldsymbol{x}$ as described in Section 3.5.

Both processes are assigned exponential correlation functions with decay parameter ϕ and hence we let $d = d_0 = d_1$ be the common effective range. We compute the convergence rates for the CP and the NCP for five levels of variance ratios such that $\delta = \delta_0 = \delta_1 = 0.01, 0.1, 1, 10, 100$ at effective ranges between 0 and $\sqrt{2}$. For each parameterisation convergence rates are computed for the two blocking strategies; jointly updating all random effects in one block, labelled CP_1 and NCP_1, or partitioning the random effects according to their process, labelled CP_2 and NCP_2. Results are given in Figure 3.24.

For the CP, updating the random effects according to process has the effect of slowing convergence. The penalty for not jointly updating all random effects increases for larger values of δ and d. For the NCP convergence can be hastened by blocking by processes, but this improvement is slight and becomes even smaller as δ increases.

3.9 Summary

In this chapter we have compared the CP and the NCP of spatial models with known covariance parameters via the exact convergence rates of Gibbs samplers constructed under the different parameterisations. We find that in addition to the ratio of the variance parameters, the correlation structure between the random effects play a key role in determining the convergence rate. We have shown that for spatially correlated random effects with an exponential correlation function, increasing the relative informativity of the data about the latent surface, as well as increasing the strength of correlation, works to hasten the convergence of the CP but slows the convergence of the NCP.

However, when the covariance matrix is tapered to remove long range correlation, convergence for the CP is hindered but convergence for the NCP is helped. Introducing geometric anisotropy to strengthen the correlation in one direction has, for randomly selected locations, a similar effect to strengthening it in all directions; the CP is helped

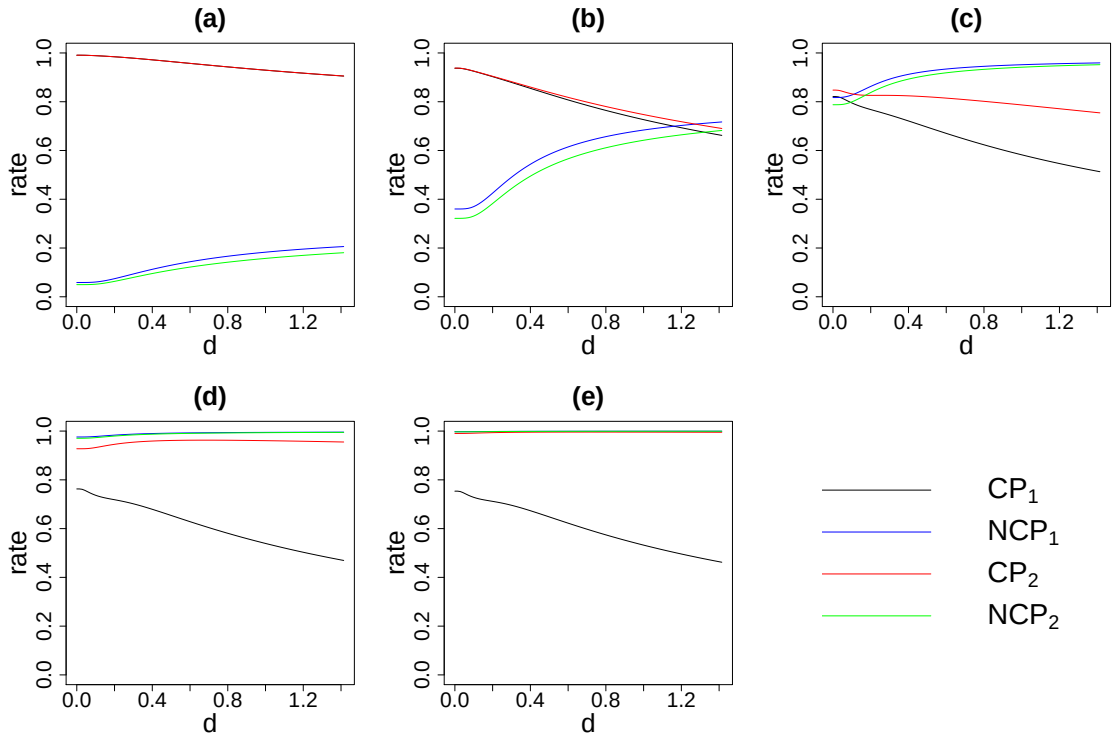

Figure 3.24: Convergence rates for the CP and the NCP for a model with two processes. CP_1 and NCP_1 indicate that the random effects are updated all at once. CP_2 and NCP_2 indicate that the random effects are blocked according to their process.

and the NCP hindered. Both of these results are consistent with the notion that the performance of CP is improved in the presence of greater spatial correlation but the performance of the NCP is worsened.

We have seen that as the smoothness parameter in the Matérn correlation function is increased both the CP and the NCP are slower to converge. Also, if there is any spatial correlation across the random effects then they should be updated together. To use more than one block results in a slower to converge Gibbs sampler for both the CP and the NCP. When we considered a model with two processes we saw that while it makes little difference to the NCP, it can be greatly beneficial to jointly update all of the random effects for the CP.

Chapter 4

Efficiency of the CP and the NCP for spatial models with unknown variance components

4.1 Introduction

In this chapter we focus on the practical implementation of the Gibbs sampler for the CP and the NCP for spatially varying coefficient models. The joint posterior distribution is unaffected by hierarchical centering and so inferential statements are the same under either parameterisation. However, what is affected is the efficiency of the Gibbs sampler used to make those statements.

In Chapter 3 the CP and the NCP are compared in terms of the exact convergence rate of the associated Gibbs sampler. The key assumption needed to compute these rates is that the joint posterior distribution is Gaussian with known precision matrix. Here we allow for the more common scenario that the precision matrix is known only up to a set of covariance parameters. In this case we cannot compute the exact convergence rate. Therefore, we use the MCMC samples to assess the efficiency of the Gibbs samplers induced by the CP and the NCP. Performance is judged by the (multivariate) potential scale reduction factor and the effective sample size, see Section 2.3 for details.

The rest of this chapter is organised as follows: In Section 4.2 we give the full conditional distributions that are needed to run the Gibbs samplers for the CP and the NCP. In Section 4.3 we give details of how to sample from the posterior predictive distributions. Section 4.4 contains a simulation study and Section 4.5 applies the different model parameterisations to ozone concentration data from California. The chapter is concluded with some closing remarks in Section 4.6.

4.2 Model specification and posterior distributions

In this section we give details of the model, specify the prior distributions and derive the joint posterior and full conditional distributions for the CP and the NCP. We have the

following normal linear model with spatially varying regression coefficients

$$Y(\boldsymbol{s}_i) = \theta_0 + \beta_0(\boldsymbol{s}_i) + \{\theta_1 + \beta_1(\boldsymbol{s}_i)\}x_1(\boldsymbol{s}_i) + \ldots + \{\theta_{p-1} + \beta_{p-1}(\boldsymbol{s}_i)\}x_{p-1}(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \quad (4.1)$$

for $i = 1, \ldots, n$. We model errors $\epsilon(\boldsymbol{s}_i)$ as independent and normally distributed with mean zero and variance σ^2_ϵ. Spatially indexed observations $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$ are conditionally independent and normally distributed

$$Y(\boldsymbol{s}_i) \sim N(\boldsymbol{x}'(\boldsymbol{s}_i)\{\boldsymbol{\theta} + \boldsymbol{\beta}(\boldsymbol{s}_i)\}, \sigma^2_\epsilon),$$

where $\boldsymbol{x}(\boldsymbol{s}_i) = (1, x_1(\boldsymbol{s}_i), \ldots, x_{p-1}(\boldsymbol{s}_i))'$ is a vector containing covariate information for site $\boldsymbol{s}_i$ and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ is a vector of global regression coefficients. The kth element of $\boldsymbol{\theta}$ is locally perturbed by a realisation of a zero mean Gaussian process, denoted $\beta_k(\boldsymbol{s}_i)$, which are collected into a vector $\boldsymbol{\beta}(\boldsymbol{s}_i) = (\beta_0(\boldsymbol{s}_i), \ldots, \beta_{p-1}(\boldsymbol{s}_i))'$.

The n realisations of the Gaussian process associated with the kth covariate are given by

$$\boldsymbol{\beta}_k = (\beta_k(\boldsymbol{s}_1), \ldots, \beta_k(\boldsymbol{s}_n))' \sim N(0, \boldsymbol{\Sigma}_k),$$

for $k = 0, \ldots, p-1$. We discuss the structure of the $\boldsymbol{\Sigma}_k$ in Section 4.2.1. The CP is induced by introducing the variables $\tilde{\beta}_k(\boldsymbol{s}_i) = \theta_k + \beta_k(\boldsymbol{s}_i)$, for $k = 0, \ldots, p-1$, and $i = 1, \ldots, n$. Therefore

$$\tilde{\boldsymbol{\beta}}_k = (\tilde{\beta}_k(\boldsymbol{s}_1), \ldots, \tilde{\beta}_k(\boldsymbol{s}_n))' \sim N(\theta_k \mathbf{1}, \boldsymbol{\Sigma}_k).$$

Global effects $\boldsymbol{\theta}$ are assumed to be multivariate normal *a priori* (see Section 4.2.1 for details) and so we write model (4.1) in its hierarchically centred form as

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned} \quad (4.2)$$

where $\boldsymbol{C}_1 = \sigma^2_\epsilon \boldsymbol{I}$ and $\boldsymbol{X}_1 = [\boldsymbol{I}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{p-1}]$ is the $n \times np$ design matrix for the first stage with $\boldsymbol{D}_k = \text{diag}(\boldsymbol{x}_k)$ where $\boldsymbol{x}_k = (x_k(\boldsymbol{s}_1), \ldots, x_k(\boldsymbol{s}_n))'$. We denote by $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}'_0, \ldots, \tilde{\boldsymbol{\beta}}'_{p-1})'$ the $np \times 1$ vector of centred, spatially correlated random effects. The design matrix for the second stage, $\boldsymbol{X}_2$, is a $np \times p$ block diagonal matrix, the blocks made of vectors of ones of length n,

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}.$$

The p processes are assumed independent *a priori* and so $\boldsymbol{C}_2$ is block diagonal where

the kth block is given by $\boldsymbol{\Sigma}_k$, therefore

$$\boldsymbol{C}_2 = \begin{bmatrix} \boldsymbol{\Sigma}_0 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_1 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Sigma}_{p-1} \end{bmatrix}.$$

4.2.1 Prior distributions

We complete the model specification by assigning prior distributions to the model parameters. The global effects $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ are assumed to be independent *a priori* with the kth element assigned a Gaussian prior distribution with mean m_k and variance $\sigma_k^2 v_k$, hence we write $\theta_k \sim N(m_k, \sigma_k^2 v_k)$, for $k = 0, \ldots, p-1$, and so $\boldsymbol{m} = (m_0, \ldots, m_{p-1})'$ and

$$\boldsymbol{C}_3 = \begin{bmatrix} \sigma_0^2 v_0 & 0 & \cdots & 0 \\ 0 & \sigma_1^2 v_1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_{p-1}^2 v_{p-1} \end{bmatrix}.$$

The realisations of the kth non-centred Gaussian process, $\boldsymbol{\beta}_k$, have a prior covariance matrix given by $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{R}_k$. This prior covariance matrix is shared by the kth centred Gaussian process, $\tilde{\boldsymbol{\beta}}_k$. The prior distributions for the variance parameters are given by

$$\sigma_k^2 \sim IG(a_k, b_k), \text{ for } k = 0, \ldots, p-1, \text{ and } \sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon),$$

where we write $X \sim IG(a, b)$ if X has a density proportional to $x^{-(a+1)} e^{-b/x}$. The entries of the $\boldsymbol{R}_k$ are given by

$$(\boldsymbol{R}_k)_{ij} = corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\} = \rho_k(d_{ij}; \phi_k, \nu_k)$$

where $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ denotes the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and ρ_k is a correlation function from the Matérn family, see Section 3.6.

We wish investigate the effect of the variance and decay parameters upon the performance of the CP and the NCP, and so henceforth we fix $\nu_k = 0.5$, for $k = 0, \ldots, p-1$. Therefore we have an exponential correlation function, i.e.

$$\rho_k(d_{ij}; \phi_k) = \exp\{-\phi_k d_{ij}\}.$$

These are used widely in applications (Sahu et al., 2010; Berrocal et al., 2010; Sahu et al., 2007; Huerta et al., 2004). The effective range for the kth process is the distance, d_k, such that $corr(\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)) = 0.05$. For an exponential correlation function we have that

$$d_k = -\log(0.05)/\phi_k \approx 3/\phi_k.$$

Decay parameters are given uniform prior distributions, i.e.

$$\phi_k \sim U(l_k, u_k), \quad k = 0, \ldots, p-1,$$

where the lower bound, l_k, and the upper bound, u_k, are suitably chosen non-negative values. It is with respect to the scale of the domain and the effective range that we choose the upper and lower bounds for the uniform prior distributions that are placed upon the decay parameters. The problem with this approach is that we are excluding the possibility that the parameters lie outside support of the prior distribution. However, we cannot use a non-informative prior distribution as we are unable to consistently estimate both variance and decay parameters under weak prior information (Zhang, 2004). It is common in applications to estimate the decay parameters by performing a grid search over a small number of values (Sahu et al., 2011; Berrocal et al., 2010; Sahu et al., 2007), choosing values that minimise some calibration criterion, like those described in Section 2.5. This approach is employed in Section 4.5.1.

4.2.2 Posterior distributions for the CP

In this section we give the joint posterior and full conditional distributions for the CP of model (4.1). We denote by $\boldsymbol{\sigma^2} = (\sigma_0^2, \ldots, \sigma_{p-1}^2)'$ the vector containing the variance parameters of the random effects and let $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_{p-1})'$ contain the decay parameters. We let $\boldsymbol{\xi} = (\tilde{\boldsymbol{\beta}}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2\prime}}, \sigma_\epsilon^2, \boldsymbol{\phi}')'$ contain all np random effects, p global effects, $p+1$ variance parameters and p decay parameters. The joint posterior distribution of $\boldsymbol{\xi}$ is given by

$$\begin{aligned}
\pi(\boldsymbol{\xi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}, \sigma_\epsilon^2)\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\sigma^2})\pi(\boldsymbol{\sigma^2})\pi(\sigma_\epsilon^2)\pi(\boldsymbol{\phi}) \\
&\propto \prod_{k=0}^{p-1} \left(\sigma_k^2\right)^{-(n/2+1/2+a_k+1)} |\boldsymbol{R}_k|^{-1/2} \left(\sigma_\epsilon^2\right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \boldsymbol{D}_k\tilde{\boldsymbol{\beta}}_k\right)'\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \boldsymbol{D}_k\tilde{\boldsymbol{\beta}}_k\right) + 2b_\epsilon\right]\right\} \\
&\quad \exp\left\{-\frac{1}{2}\sum_{k=0}^{p-1}\left(\tilde{\boldsymbol{\beta}}_k - \theta_k\mathbf{1}\right)'\boldsymbol{\Sigma}_k^{-1}\left(\tilde{\boldsymbol{\beta}}_k - \theta_k\mathbf{1}\right)\right\} \\
&\quad \exp\left\{-\frac{1}{2}\sum_{k=0}^{p-1}\frac{1}{\sigma_k^2}\left(\frac{(\theta_k - m_k)^2}{v_k} + 2b_k\right)\right\}\prod_{k=0}^{p-1}\pi(\phi_k), \qquad (4.3)
\end{aligned}$$

where $\boldsymbol{D}_0$ is defined to be the identity matrix $\boldsymbol{I}$.

We use Gibbs sampling (see Section 2.2.4) to sample from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ for the CP, given in (4.3). We assume that the random effects will be block updated according to their process, i.e. we jointly update the n-dimensional vector $\boldsymbol{\beta}_k$, for $k = 0, \ldots, p-1$. All other parameters in $\boldsymbol{\xi}$ are updated as single univariate components. The full conditional distributions we need for the CP are given below.

- The full conditional distribution for the centred spatially correlated random effects

$\tilde{\boldsymbol{\beta}}_k$, $k = 0, \ldots, p-1$, is given by

$$\tilde{\boldsymbol{\beta}}_k | \tilde{\boldsymbol{\beta}}_{-k}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim N(\boldsymbol{m}^*_k, \boldsymbol{\Sigma}^*_k),$$

where we denote by $\tilde{\boldsymbol{\beta}}_{-k}$ the vector of all random effects $\tilde{\boldsymbol{\beta}}$ without the realisations of the kth process $\tilde{\boldsymbol{\beta}}_k$ and

$$\begin{aligned}
\boldsymbol{\Sigma}^*_k &= \left(\frac{1}{\sigma^2_\epsilon} \boldsymbol{D}'_k \boldsymbol{D}_k + \boldsymbol{\Sigma}^{-1}_k \right)^{-1} \\
\boldsymbol{m}^*_k &= \boldsymbol{\Sigma}^*_k \left[\frac{1}{\sigma^2_\epsilon} \boldsymbol{D}_k \left(\boldsymbol{y} - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{D}_j \tilde{\boldsymbol{\beta}}_j \right) + \boldsymbol{\Sigma}^{-1}_k \theta_k \mathbf{1} \right].
\end{aligned}$$

- The full conditional distribution for the global effects θ_k, $k = 0, \ldots, p-1$, for the CP is given by

$$\theta_k | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}_{-k}, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim N(m^*_k, v^*_k),$$

where

$$\begin{aligned}
v^*_k &= \left(\mathbf{1}' \boldsymbol{\Sigma}^{-1}_k \mathbf{1} + \frac{1}{\sigma^2_k v_k} \right)^{-1}, \\
m^*_k &= v^*_k \left(\mathbf{1}' \boldsymbol{\Sigma}^{-1}_k \tilde{\boldsymbol{\beta}}_k + \frac{m_k}{\sigma^2_k v_k} \right).
\end{aligned}$$

- The full conditional distribution for the random effects variance σ^2_k, $k = 0, \ldots, p-1$, for the CP is given by

$$\sigma^2_k | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}_{-k}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim IG \Bigg\{ \frac{n+1}{2} + a_k, \quad \frac{1}{2} \Big[\left(\tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1} \right)' \boldsymbol{R}^{-1}_k \left(\tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1} \right) + \frac{(\theta_k - m_k)^2}{v_k} + 2b_k \Big] \Bigg\}.$$

- The full conditional distribution for data variance σ^2_ϵ for the CP is given by

$$\sigma^2_\epsilon | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \boldsymbol{y} \sim IG \left\{ \frac{n}{2} + a_\epsilon, \frac{1}{2} \left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \boldsymbol{D}_k \tilde{\boldsymbol{\beta}}_k \right)' \left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \boldsymbol{D}_k \tilde{\boldsymbol{\beta}}_k \right) + 2b_\epsilon \right] \right\}.$$

- The full conditional distribution for the decay parameter ϕ_k, $k = 0, \ldots, p-1$, for the CP is given by

$$\pi(\phi_k | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}_{-k}, \boldsymbol{y}) \propto |R_k|^{-1/2} \exp \left\{ -\frac{1}{2\sigma^2_k} \left(\tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1} \right)' \boldsymbol{\Sigma}^{-1}_k \left(\tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1} \right) \right\},$$

for $l_k < \phi_k < u_k$, zero otherwise.

We note that the mean and variance parameters are conditionally conjugate, and so their full conditional distributions can be sampled from directly. The form of the full conditional distributions for ϕ_k is not one belonging to a known distribution. Therefore we use a Metropolis-Hastings step within the Gibbs sampler, see Section 2.2. Moreover, if we sample ϕ_k on the log-scale we can use a Gaussian proposal distribution centred on the current value $\log \phi_k^{(t)}$. This proposal distribution is symmetric in its arguments and so we use a Metropolis step, see Section 2.2.1.

Given $\phi_k^{(t)}$ we obtain $\phi_k^{(t+1)}$ as follows:

1. Sample a candidate value $\log \phi_k^* \sim N(\log \phi_k^{(t)}, \sigma_\phi^2)$ where σ_ϕ^2 is the tuning parameter.

2. Calculate the acceptance probability

$$\alpha(\log \phi_k^{(t)}, \log \phi_k^*) = \min \left\{ 1, \frac{\pi(\phi_k^* | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2, \boldsymbol{\phi}_{-k}, \boldsymbol{y}) \exp\{\log \phi_k^*\}}{\pi(\phi_k^{(t)} | \tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2, \boldsymbol{\phi}_{-k}, \boldsymbol{y}) \exp\{\log \phi_k^{(t)}\}} \right\},$$

   where the terms $\exp\{\log \phi_k^*\}$ and $\exp\{\log \phi_k^{(t)}\}$ appear in the quotient due to the change of variables.

3. Draw a value u from a uniform U(0,1) distribution

4. Let

$$\log \phi_k^{(t+1)} = \begin{cases} \log \phi_k^* & \text{if } u \leq \alpha(\log \phi_k^{(t)}, \log \phi_k^*) \text{ and } \phi_k^* \in (l_k, u_k) \\ \log \phi_k^{(t)} & \text{otherwise,} \end{cases}$$

   The value $\phi_k^{(t+1)}$ is stored to be used to sample from the other variables.

4.2.3 Posterior distributions for the NCP

We now look at the joint posterior and full conditional distributions of the model parameters for the NCP. For the NCP we have $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2\prime}}, \sigma_\epsilon^2, \boldsymbol{\phi}')'$, and

$$\begin{aligned}
\pi(\boldsymbol{\xi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\epsilon^2)\pi(\boldsymbol{\beta}|\boldsymbol{\sigma^2}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\sigma^2})\pi(\boldsymbol{\sigma^2})\pi(\sigma_\epsilon^2)\pi(\boldsymbol{\phi}) \\
&\propto \prod_{k=0}^{p-1} (\sigma_k^2)^{-(n/2+1/2+a_k+1)} |\boldsymbol{R}_k|^{-1/2} \left(\sigma_\epsilon^2\right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp\left\{ -\frac{1}{2\sigma_\epsilon^2} \left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} (\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k) \right)' \left(\boldsymbol{Y} - \sum_{k=0}^{p-1} (\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k) \right) + 2b_\epsilon \right] \right\} \\
&\quad \exp\left\{ -\frac{1}{2} \sum_{k=0}^{p-1} \boldsymbol{\beta}_k' \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\beta}_k \right\} \exp\left\{ -\frac{1}{2} \sum_{k=0}^{p-1} \frac{1}{\sigma_k^2} \left(\frac{(\theta_k - m_k)^2}{v_k} + 2b_k \right) \right\} \\
&\quad \prod_{k=0}^{p-1} \pi(\phi_k),
\end{aligned} \tag{4.4}$$

where we define $\boldsymbol{x}_0$ to be the vector of ones.

In Section 4.1 we note that the equivalence of the joint posterior distribution under the CP and the NCP means that inference is unaffected by reparameterisation. It is clear

that (4.3) and (4.4) are equivalent when we recall that $\boldsymbol{\beta}_k = \tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1}$ and $\boldsymbol{D}_k \mathbf{1} = \boldsymbol{x}_k$.

The full conditional distributions for the NCP are given below.

- The full conditional distribution for the non-centred spatially correlated random effects $\boldsymbol{\beta}_k$, $k = 0, \ldots, p-1$, is given by

$$\boldsymbol{\beta}_k | \boldsymbol{\beta}_{-k}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim N(\boldsymbol{m}^*_k, \boldsymbol{\Sigma}^*_k),$$

  where

$$\begin{aligned}
\boldsymbol{\Sigma}^*_k &= \left(\frac{1}{\sigma^2_\epsilon} \boldsymbol{D}'_k \boldsymbol{D}_k + \boldsymbol{\Sigma}^{-1}_k \right)^{-1} \\
\boldsymbol{m}^*_k &= \boldsymbol{\Sigma}^*_k \left[\frac{1}{\sigma^2_\epsilon} \boldsymbol{x}'_k \left(\boldsymbol{y} - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{D}_j \boldsymbol{\beta}_j - \sum_{j=0}^{p-1} \boldsymbol{x}_j \theta_j \right) \right].
\end{aligned}$$

- The full conditional distribution for the global effects θ_k, $k = 0, \ldots, p-1$, for the NCP is given by

$$\theta_k | \boldsymbol{\beta}, \boldsymbol{\theta}_{-k}, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim N(m^*_k, v^*_k),$$

  where

$$\begin{aligned}
v^*_k &= \left(\frac{1}{\sigma^2_\epsilon} \boldsymbol{x}'_k \boldsymbol{x}_k + \frac{1}{\sigma^2_k v_k} \right)^{-1}, \\
m^*_k &= v^*_k \left[\frac{1}{\sigma^2_\epsilon} \boldsymbol{x}'_k \left(\boldsymbol{y} - \sum_{j=0}^{p-1} \boldsymbol{D}_j \boldsymbol{\beta}_j - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{x}_j \theta_j \right) + \frac{m_k}{\sigma^2_k v_k} \right].
\end{aligned}$$

- The full conditional distribution for the random effects variance σ^2_k, $k = 0, \ldots, p-1$, for the NCP is given by

$$\sigma^2_k | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}_{-k}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim IG \left\{ \frac{n+1}{2} + a_k, \frac{1}{2} \left(\boldsymbol{\beta}'_k \boldsymbol{R}^{-1}_k \boldsymbol{\beta}_k + \frac{(\theta_k - m_k)^2}{v_k} + 2b_k \right) \right\}.$$

- The full conditional distribution for the data variance σ^2_ϵ for the NCP is given by

$$\begin{aligned}
\sigma^2_\epsilon | \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \boldsymbol{y} \sim IG \Bigg\{ & \frac{n}{2} + a_\epsilon \\
& \frac{1}{2} \left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \left(\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k \right) \right)' \left(\boldsymbol{Y} - \sum_{k=0}^{p-1} \left(\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k \right) \right) + 2b_\epsilon \right] \Bigg\}.
\end{aligned}$$

- The full conditional distribution for the decay parameter ϕ_k, $k = 0, \ldots, p-1$, for

the NCP is given by

$$\pi(\phi_k|\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{\sigma^2},\sigma_\epsilon^2,\boldsymbol{\phi}_{-k},\boldsymbol{y}) \propto |R_k|^{-1/2}\exp\left\{-\frac{1}{2\sigma_k^2}\boldsymbol{\beta}_k'\boldsymbol{\Sigma}_k^{-1}\boldsymbol{\beta}_k\right\},$$

for $l_k < \phi_k < u_k$, zero otherwise.

As with the CP, we use a Metropolis step to update each ϕ_k within the Gibbs sampler.

4.3 Predictive distributions

The advantage of process based models is that we can obtain predictions at highest possible resolution and interpolate the surface to produce predictive maps. In a fully Bayesian framework this is done by evaluating the posterior predictive distribution (PPD), $\pi(Y(\boldsymbol{s}^*)|\boldsymbol{y})$, where $\boldsymbol{s}^*$ is a location at which data has not been observed. The PPD gives us access to full predictive inference, i.e. we may compute point estimates, $\widehat{Y}(\boldsymbol{s}^*)$, or probabilities such as $P(Y(\boldsymbol{s}^*) > c)$, where c may be some threshold value of interest such as those given by air quality directives for concentrations of pollutants.

The PPD also provides a natural way to judge the performance of the model and robustness to prior specifications. In Section 4.5.1 we use the PPD to estimate the decay parameters. By partitioning the data into training and validation sets we can obtain numerical measures for prediction error such as those described in Section 2.5.

4.3.1 Posterior predictive distribution for the CP

The centred form of model (4.1) implies that

$$Y(\boldsymbol{s}^*) = \tilde{\beta}_0(\boldsymbol{s}^*) + \tilde{\beta}_1(\boldsymbol{s}^*)x_1(\boldsymbol{s}^*) + \ldots + \tilde{\beta}_{p-1}(\boldsymbol{s}^*)x_{p-1}(\boldsymbol{s}^*) + \epsilon(\boldsymbol{s}^*),$$

for any new location $\boldsymbol{s}^*$. The set of model parameters $\boldsymbol{\xi} = (\tilde{\boldsymbol{\beta}}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2\prime}}, \sigma_\epsilon^2, \boldsymbol{\phi}')'$ is augmented by the realisations of the p spatial processes at location $\boldsymbol{s}^*$. We denote the augmented set by $\boldsymbol{\xi}^* = (\tilde{\boldsymbol{\beta}}(\boldsymbol{s}^*), \boldsymbol{\xi}')'$ where $\tilde{\boldsymbol{\beta}}(\boldsymbol{s}^*) = (\tilde{\beta}_0(\boldsymbol{s}^*), \ldots, \tilde{\beta}_{p-1}(\boldsymbol{s}^*))'$. To find the posterior predictive distribution we must evaluate the following integral

$$\begin{aligned}
\pi(Y(\boldsymbol{s}^*)|\boldsymbol{y}) &= \int \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*,\boldsymbol{y})\pi(\boldsymbol{\xi}^*|\boldsymbol{y})d\boldsymbol{\xi}^* \\
&= \int \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*,\boldsymbol{y})\pi(\tilde{\boldsymbol{\beta}}(\boldsymbol{s}^*)|\boldsymbol{\xi})\pi(\boldsymbol{\xi}|\boldsymbol{y})d\boldsymbol{\xi}^* \\
&= \int \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*,\boldsymbol{y})\prod_{k=0}^{p-1}\pi(\tilde{\beta}_k(\boldsymbol{s}^*)|\boldsymbol{\xi})\pi(\boldsymbol{\xi}|\boldsymbol{y})d\boldsymbol{\xi}^*, \qquad (4.5)
\end{aligned}$$

where the last step is possible due to the assumption of prior independence across the p processes.

We estimate (4.5) by composition sampling, (Banerjee et al., 2003, Chapter 5). If a posterior sample $\boldsymbol{\xi}^{(t)} \sim \pi(\boldsymbol{\xi}|\boldsymbol{y})$ and $\tilde{\beta}_k^{(t)}(\boldsymbol{s}^*) \sim \pi(\tilde{\beta}_k(\boldsymbol{s}^*)|\boldsymbol{\xi}^{(t)})$, for $k = 0, \ldots, p-1$, then draws $Y^{(t)}(\boldsymbol{s}^*) \sim \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^{*(t)},\boldsymbol{y})$ have the marginal distribution $\pi(Y(\boldsymbol{s}^*)|\boldsymbol{y})$. We

receive samples from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ when we sample from the full conditional distributions given in Section 4.2.2.

The conditional distributions for the other elements of the integrand in (4.5) are found using standard results for multivariate normal distributions. The joint distribution of $\tilde{\beta}_k(\boldsymbol{s}^*)$ and $\tilde{\boldsymbol{\beta}}_k$ is given by

$$\begin{pmatrix} \tilde{\beta}_k(\boldsymbol{s}^*) \\ \tilde{\boldsymbol{\beta}}_k \end{pmatrix} \sim N\left\{ \begin{pmatrix} \theta_k \\ \theta_k \mathbf{1} \end{pmatrix}, \sigma_k^2 \begin{pmatrix} 1 & \boldsymbol{c}_k' \\ \boldsymbol{c}_k & \boldsymbol{R}_k \end{pmatrix} \right\},$$

where $\boldsymbol{c}_k$ is an n-dimensional vector whose elements are given by $\rho_k(\|\boldsymbol{s}_i - \boldsymbol{s}^*\|; \phi_k, \nu_k)$. It follows that

$$\tilde{\beta}_k(\boldsymbol{s}^*)|\boldsymbol{\xi} \sim N\left(\theta_k + \boldsymbol{c}_k' \boldsymbol{R}_k^{-1}(\tilde{\boldsymbol{\beta}}_k - \theta_k \mathbf{1}), \sigma_k^2(1 - \boldsymbol{c}_k' \boldsymbol{R}_k^{-1} \boldsymbol{c}_k)\right).$$

Using a similar approach we write the joint distribution of $Y(\boldsymbol{s}^*)$, and $\boldsymbol{Y}$ as

$$\begin{pmatrix} Y(\boldsymbol{s}^*) \\ \boldsymbol{Y} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{x}'(\boldsymbol{s}^*)\tilde{\boldsymbol{\beta}}(\boldsymbol{s}^*) \\ \boldsymbol{X}_1 \tilde{\boldsymbol{\beta}} \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \right\},$$

where $\boldsymbol{x}'(\boldsymbol{s}^*) = (1, x_1(\boldsymbol{s}^*), \ldots, x_k(\boldsymbol{s}^*))$. Hence, the conditional distribution for the data at location $\boldsymbol{s}^*$ is

$$Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*, \boldsymbol{y} \sim N(\boldsymbol{x}'(\boldsymbol{s}^*)\tilde{\boldsymbol{\beta}}(\boldsymbol{s}^*), \sigma_\epsilon^2).$$

Therefore, for each post burn-in sample from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ we can obtain draws from the PPD for any number of out of sample locations.

4.3.2 Posterior predictive distribution for the NCP

For the NCP we follow the same procedure. Model (4.1) implies that

$$Y(\boldsymbol{s}^*) = \theta_0 + \beta_0(\boldsymbol{s}^*) + \{\theta_1 + \beta_1(\boldsymbol{s}^*)\}x_1(\boldsymbol{s}^*) + \ldots + \{\theta_{p-1} + \beta_{p-1}(\boldsymbol{s}^*)\}x_{p-1}(\boldsymbol{s}^*) + \epsilon(\boldsymbol{s}^*).$$

The set of model parameters for the NCP, $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2\prime}}, \sigma_\epsilon^2, \boldsymbol{\phi}')'$, is augmented by the realisation of the p zero mean spatial processes at location $\boldsymbol{s}^*$ and so $\boldsymbol{\xi}^* = (\boldsymbol{\beta}(\boldsymbol{s}^*), \boldsymbol{\xi}')'$ where $\boldsymbol{\beta}(\boldsymbol{s}^*) = (\beta_0(\boldsymbol{s}^*), \ldots, \beta_{p-1}(\boldsymbol{s}^*))'$. We can write the PPD as

$$\pi(Y(\boldsymbol{s}^*)|\boldsymbol{y}) \quad = \quad \int \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*, \boldsymbol{y}) \prod_{k=0}^{p-1} \pi(\beta_k(\boldsymbol{s}^*)|\boldsymbol{\xi}) \pi(\boldsymbol{\xi}|\boldsymbol{y}) d\boldsymbol{\xi}^*. \tag{4.6}$$

We evaluate the integral (4.6) in the same way as we did integral (4.5), by composition sampling. Given a sample $\boldsymbol{\xi}^{(t)} \sim \pi(\boldsymbol{\xi}|\boldsymbol{y})$, we draw $\beta_k^{(t)}(\boldsymbol{s}^*) \sim \pi(\beta_k(\boldsymbol{s}^*)|\boldsymbol{\xi}^{(t)})$, for $k = 0, \ldots, p-1$. We use these samples to generate $Y^{(t)}(\boldsymbol{s}^*) \sim \pi(Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^{*(t)}, \boldsymbol{y})$ which has the marginal distribution $\pi(Y(\boldsymbol{s}^*)|\boldsymbol{y})$. We receive samples from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ when we sample from the full conditional distributions given in Section 4.2.3.

The conditional distributions for the other elements of the integrand in (4.6) are found using standard results for multivariate normal distributions. The joint distribution of

$\beta_k(\boldsymbol{s}^*)$ and $\boldsymbol{\beta}_k$ is given by

$$\begin{pmatrix} \beta_k(\boldsymbol{s}^*) \\ \boldsymbol{\beta}_k \end{pmatrix} \sim N\left\{ \begin{pmatrix} 0 \\ \mathbf{0} \end{pmatrix}, \sigma_k^2 \begin{pmatrix} \sigma_k^2 & \boldsymbol{c}_k' \\ \boldsymbol{c}_k & \boldsymbol{R}_k \end{pmatrix} \right\},$$

and the conditional distribution of $\beta_k(\boldsymbol{s}^*)$ is given by

$$\beta_k(\boldsymbol{s}^*)|\boldsymbol{\xi} \sim N\left(\boldsymbol{c}_k'\boldsymbol{R}_k^{-1}\boldsymbol{\beta}_k, \sigma_k^2(1 - \boldsymbol{c}_k'\boldsymbol{R}_k^{-1}\boldsymbol{c}_k)\right).$$

The joint distribution of $Y(\boldsymbol{s}^*)$ and $\boldsymbol{Y}$ for the NCP is

$$\begin{pmatrix} Y(\boldsymbol{s}^*) \\ \boldsymbol{Y} \end{pmatrix} \sim N\left\{ \begin{pmatrix} \boldsymbol{x}'(\boldsymbol{s}^*)(\boldsymbol{\beta}(\boldsymbol{s}^*) + \boldsymbol{\theta}) \\ \boldsymbol{X}_1(\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\theta}) \end{pmatrix}, \sigma_\epsilon^2 \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \right\},$$

Hence, the conditional distribution for the data at location $\boldsymbol{s}^*$ is

$$Y(\boldsymbol{s}^*)|\boldsymbol{\xi}^*, \boldsymbol{y} \sim N(\boldsymbol{x}'(\boldsymbol{s}^*)(\boldsymbol{\beta}(\boldsymbol{s}^*) + \boldsymbol{\theta}), \sigma_\epsilon^2).$$

4.4 CP versus NCP: A simulation study

In this section we examine the convergence and mixing properties of the Gibbs samplers induced by the CP and the NCP. We let $p = 1$ in model (4.1) and so we have the following hierarchically centred model

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0 &\sim N(\tilde{\boldsymbol{\beta}}_0, \sigma_\epsilon^2\boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}_0|\theta_0 &\sim N(\theta_0\mathbf{1}, \sigma_0^2\boldsymbol{R}_0) \\ \theta_0 &\sim N(m_0, \sigma_0^2 v_0). \end{aligned} \tag{4.7}$$

We use an exponential correlation function, so that

$$\boldsymbol{R}_0 = \exp\{-\phi_0 d_{ij}\}.$$

We generate data from model (4.7) as described in Section 4.4.1 for different combinations of σ_ϵ^2, σ_0^2 and ϕ_0. Two simulation studies are conducted. The first fixes the variance and decay parameters at their true values and so we only sample from the joint posterior distributions of the global effect, θ_0, and the random effects, $\tilde{\boldsymbol{\beta}}_0$ or $\boldsymbol{\beta}_0$. We use the MCMC output for θ_0 to compute diagnostic statistics for comparing the performance of the CP and the NCP. The first statistic we use is based on the potential scale reduction factor (PSRF), described in Section 2.3.1. We define the $\text{PSRF}_M(1.1)$ to be the number of iterations required for the PSRF to fall below 1.1. To compute the $\text{PSRF}_M(1.1)$ we run five chains of length 25,000 from widely dispersed starting values. In particular, we take values that are outside of the intervals described by pilot chains. Moreover, the same starting values are used for both the CP and the NCP. At every fifth iteration the PSRF is calculated and number of iterations for its value to first drop below 1.1 is the value that we record. The second statistic we use is the effective sample size (ESS) of θ_0, see Section

2.3.2. The ESS is computed using all 125,000 MCMC samples and gives us a measure of the Markovian dependence between successive MCMC iterates, with values of 125,000 indicating independence.

By fixing the variance and decay parameters we have known posterior precision matrices. Therefore we are able to compare the measures of efficiency used here with the exact convergence rates computed in Chapter 3.

The second simulation study drops the assumption of known variance parameters, fixing only the decay parameter. In this case we judge performance by the $\text{MPSRF}_M(1.1)$ which we define to be the number of iteration needed for the multivariate PSRF (see Section 2.3.1) to fall below 1.1. When we sample from the variance parameters we record the ESS of θ_0, σ_0^2 and σ_ϵ^2. In both simulation studies we let hyperparameters $m_0 = 0$ and $v_0 = 10^4$.

4.4.1 Data generation

We simulate data from model (4.7) for $n = 40$ randomly chosen locations across the unit square, see Figure 4.1. These are the same locations used in Chapter 3. We set

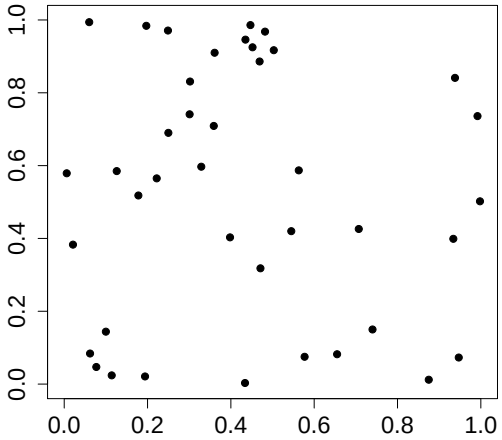

Figure 4.1: Points in the unit square used as sampling locations for simulating data from model (4.1).

$\theta_0 = 0$ and generate data with five variance parameter ratios such that $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$. This is done by letting $\sigma_0^2 = 1$ and varying σ_ϵ^2 accordingly. For each of the five levels of δ_0 we have four values of the decay parameter ϕ_0, chosen such that there is an effective range, denoted d_0, of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$, where $\sqrt{2}$ is the maximum possible separation of two points in the unit square. Hence there are 20 combinations of σ_0^2, σ_ϵ^2 and ϕ_0 in all. Each of these combinations is used to simulate 20 datasets, and so there are 400 data sets in total.

4.4.2 Known variance parameters

Initially we set the variance parameters equal to the values used to generate the data and we sample from the spatially correlated random effects ($\tilde{\boldsymbol{\beta}}_0$ or $\boldsymbol{\beta}_0$) and global effect θ_0. For each of the 400 data sets, five chains of length 25,000 are produced under the CP and the NCP. We are only sampling from one global parameter and so the efficiency of the Gibbs sampler is assessed in terms of the $\text{PSRF}_M(1.1)$ and the ESS of θ_0. There is a negligible difference in the run times for the CP and the NCP and so we do not adjust these measures by computation time.

Figure 4.2 shows boxplots of the $\text{PSRF}_M(1.1)$ (top row) and the ESS of θ_0 (bottom row) for the CP. Each panel contains the results for a fixed value of δ_0, increasing from 0.01 on the left to 100 on the right. Each panel contains four boxplots corresponding to the four effective ranges of 0, $x/3$, $2x/3$, and x, where $x = \sqrt{2}$. As the effective range increases we have stronger spatial correlation between the random effects. Each boxplot is produced from the 20 values obtained for a given combination of δ_0 and ϕ_0.

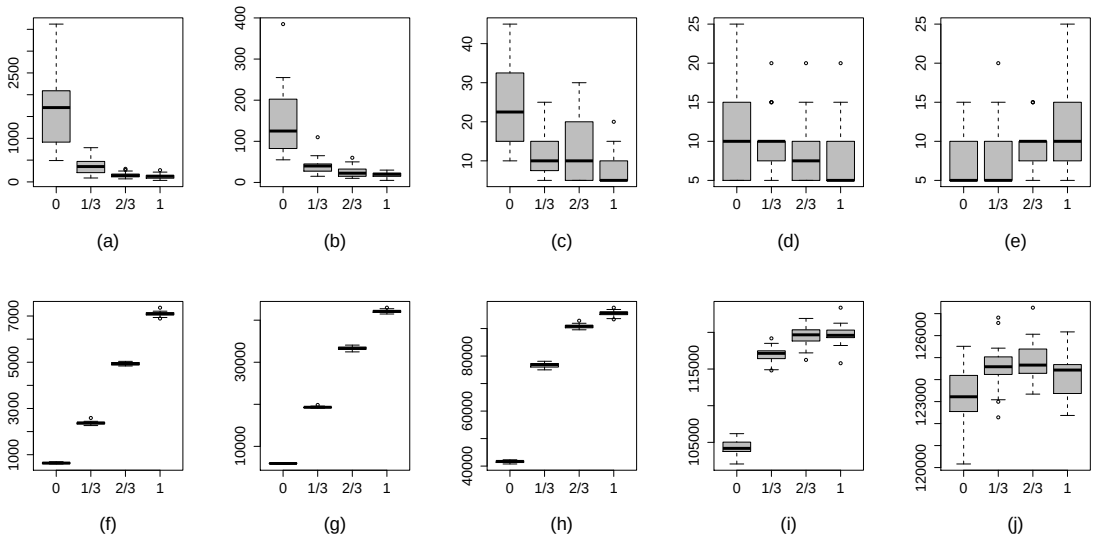

Figure 4.2: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the CP of the Gaussian model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

We see that convergence is hastened and the ESS increased as we increase δ_0. We can also see improvement in the performance of the sampler with increasing spatial correlation. The pattern reversed for the NCP, see Figure 4.3, whose performance degrades with increasing δ_0 or increasing spatial correlation.

We can see that the observed measures of convergence and mixing employed here are in agreement with the convergence rates for the CP and the NCP given in Figure 3.4. This give us confidence to use these measures to judge the performance of the samplers when we do not have access to the exact convergence rate, i.e. when the variance parameters are unknown.

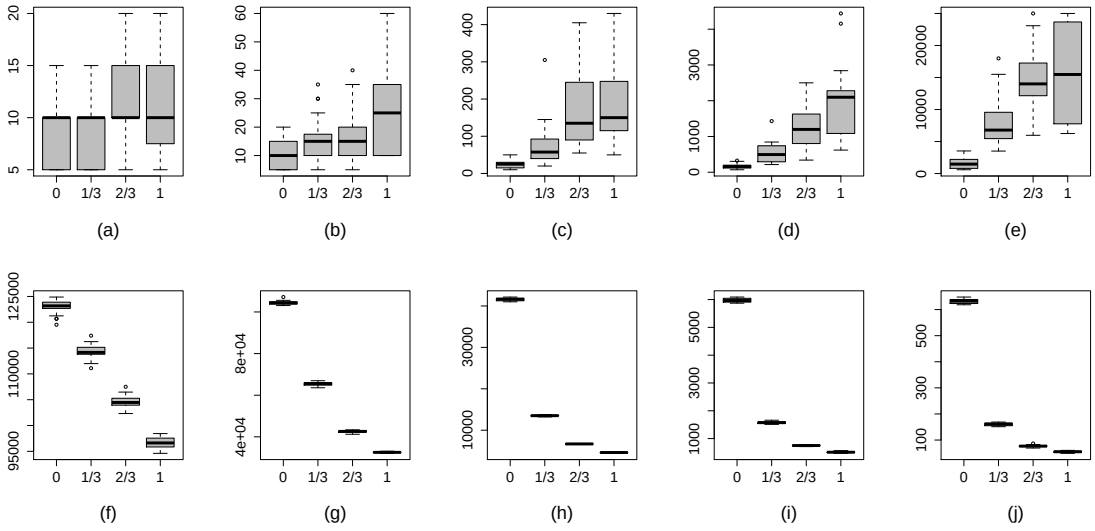

Figure 4.3: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the Gaussian model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

4.4.3 Unknown variance parameters

In this section we drop the assumption that the variance parameters are known and we sample from their full conditional distributions given in Section 4.2. Recall that the variance parameters are given inverse gamma prior distributions with $\pi(\sigma_0^2) = IG(a_0, b_0)$ and $\pi(\sigma_\epsilon^2) = IG(a_\epsilon, b_\epsilon)$. We let $a_0 = a_\epsilon = 2$ and $b_\epsilon = b_\epsilon = 1$, implying a prior mean of one and infinite prior variance for σ_0^2 and σ_ϵ^2. These are common hyperparameters for inverse gamma prior distributions, see Sahu et al. (2010, 2007); Gelfand et al. (2003).

As we are sampling from more than one parameter, we measure efficiency by the $\text{MPSRF}_M(1.1)$ and the ESS of θ_0, σ_0^2 and σ_ϵ^2. The results here are represented in the same way as in Figures 4.2 and 4.3. A row contains the results for either the $\text{MPSRF}_M(1.1)$ or ESS of a parameter. Panels in each row correspond to a fixed value of the true variance ratio δ_0, and a boxplot within a panel is made of 20 results for a fixed δ_0 and effective range d_0.

Figure 4.4 gives the $\text{MPSRF}_M(1.1)$ and ESS of θ_0 for the CP. We can see that the performance of the CP improves with increasing δ_0 and also with increasing strength of correlation between the random effects. The equivalent plot for the NCP is given in Figure 4.6. We can see a reverse of the pattern displayed by the CP. The performance of the NCP is worsened as δ_0 increases and the detrimental effect of increasing the strength of correlation between the random effects is also clearly evident. Therefore, δ_0 and the d_0 have the same influence on the CP and the NCP as we saw for the case when the variance parameters were assumed to be known in Section 4.4.2.

Figure 4.5 gives the ESS of σ_0^2 (top row) and the ESS of σ_ϵ^2 (bottom row) for the CP. We can see a general increasing trend in the ESS of σ_0^2 for increasing δ_0, but a downward trend is seen for σ_ϵ^2. However, for a fixed value of δ_0 we can see an improvement as the

effective range increases, particularly in σ_ϵ^2. This is because for the case when there is zero effective range, marginally the data variance is $(\sigma_0^2 + \sigma_\epsilon^2)\boldsymbol{I}$, and so increasing the effective range moves us away from the unidentifiable case which can result in poor mixing of the chains.

Figure 4.7 shows the ESS of σ_0^2 and σ_ϵ^2 for the NCP. We see that the ESS of σ_0^2 is stable under changes in δ_0 and d_0, with the exception being the case where $\delta_0 = 0.1$ and $d_0 = 0$. In this case the results are again explained by the lack of identifiability of the variance parameters for independent random effects. The ESS of σ_ϵ^2 is reduced by increasing δ_0. For a fixed value of δ_0 we can see an improvement in the ESS as d_0 increases. This was also observed σ_ϵ^2 under the CP and is similarly explained.

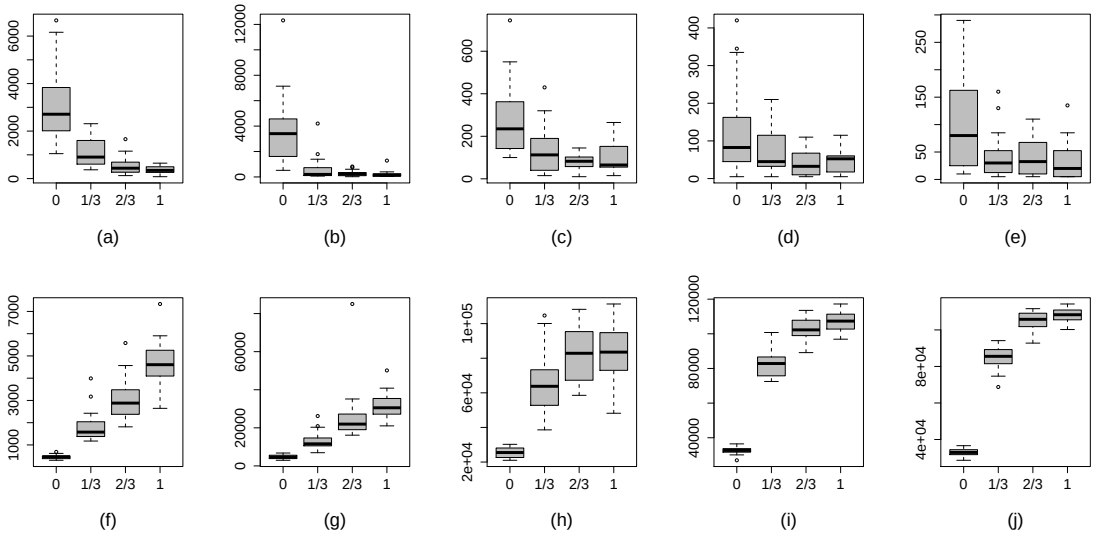

Figure 4.4: MPSRF$_M(1.1)$ and the ESS of θ_0 for the CP of the Gaussian model with unknown variance parameters Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

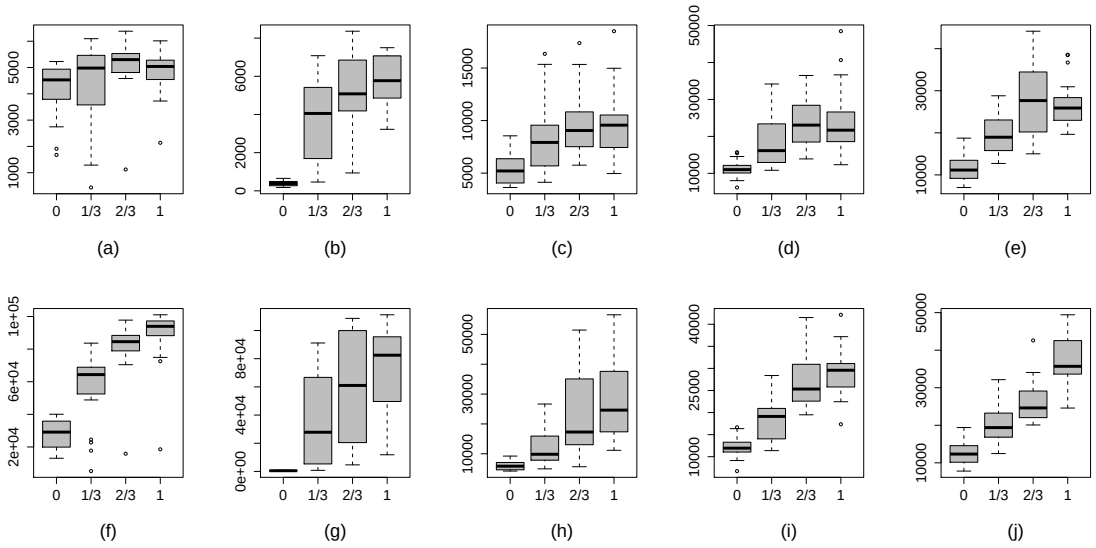

Figure 4.5: ESS of σ_0^2 and σ_ϵ^2 for the CP of the Gaussian model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

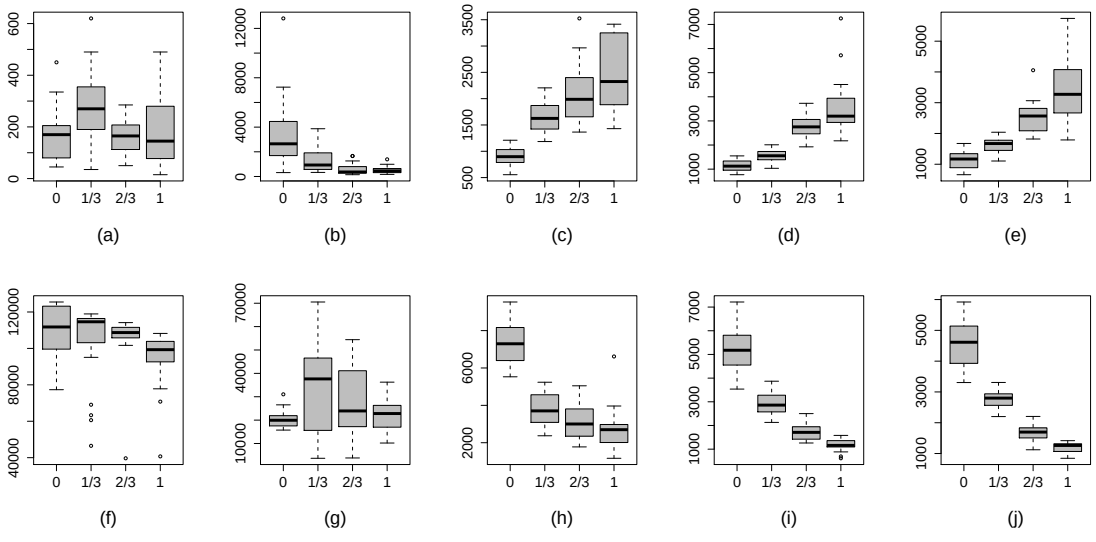

Figure 4.6: $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the Gaussian model with unknown variance parameters. Plots (a)–(e) give the $\text{MPSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

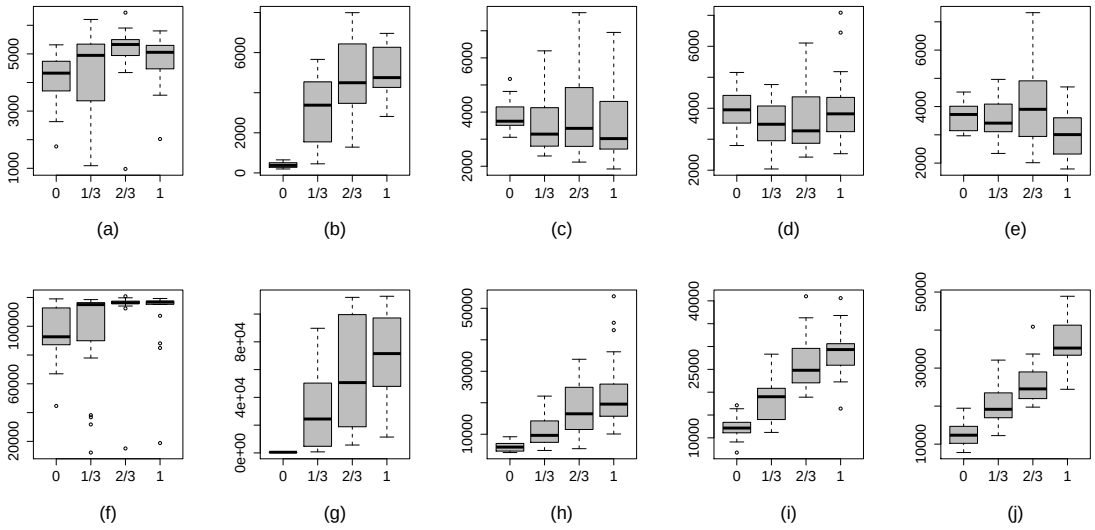

Figure 4.7: ESS of σ_0^2 and σ_ϵ^2 for the NCP of the Gaussian model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

4.5 Californian ozone concentration data

In Section 4.4 we saw how varying the values of the variance and decay parameters used to generate data affects the efficiency of Gibbs samplers constructed under the CP and the NCP. In this section we fit the CP and the NCP of model (4.1) to a real data set. We have ozone concentration data from the State of California. It is a spatial data set with values, in parts per billion (ppb), of the annual fourth highest daily maximum eight-hour average. The eight-hour average for the current hour is the mean concentration of the last four hours, the current hour and the future three hours. The annual fourth highest eight-hour average is the key measure used by the U.S. Environmental Protection Agency for monitoring ozone concentrations. The current standard, set in 2008, is 75 ppb [1] down from the 80 ppb standard set in 1997. Proposals are in place to bring the standard within the range 65-70 ppb.

Data are collected at 176 irregularly spaced locations across California. We fit the model using data from 132 sites, leaving out 44 sites for validation, see Figure 4.8. The mean and standard deviation for the 132 data sites is 80.35 ppb and 17.72 ppb respectively.

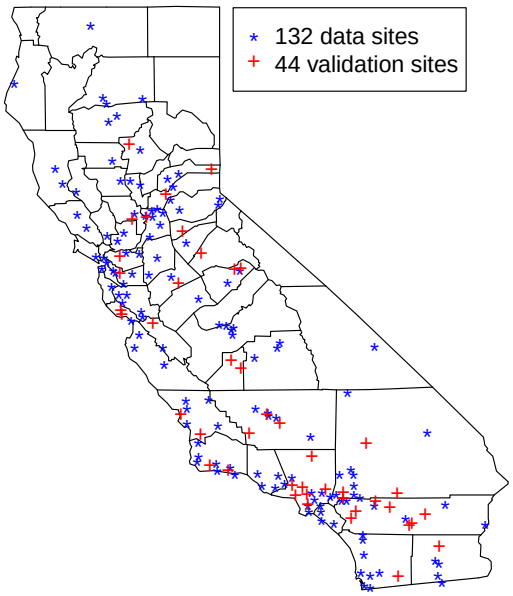

Figure 4.8: Sampling locations for Californian ozone concentration data.

The spatially varying covariate we use is land use. Sites are categorised as urban or suburban and assigned the value one, or they are categorised as rural, and assigned the value zero. Of the 132 data sites, 89 are urban or suburban with mean concentration 78.71 ppb and standard deviation 18.53 ppb. The remaining 43 rural sites have mean 83.74 ppb

[1] see http://http://www.epa.gov/air/criteria.html

and standard deviation 15.59 ppb. Of the 44 validation sites 28 are urban or suburban and 16 are rural.

Given information about land use at each data location we fit model (4.1) with $p = 2$. Therefore we have two processes, an intercept and slope process and so $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0', \tilde{\boldsymbol{\beta}}_1')'$ for the CP and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0', \boldsymbol{\beta}_1')'$ for the NCP. Each process has a corresponding global parameter and a variance parameter and so $\boldsymbol{\theta} = (\theta_0, \theta_1)'$ and $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)'$. We use an exponential correlation function for both processes and so $\boldsymbol{\phi} = (\phi_0, \phi_1)'$. In addition we have the data variance, σ_ϵ^2, and so we have $2(n+3)+1$ parameters to estimate.

For the prior distribution of $\boldsymbol{\theta}$ we let $\boldsymbol{m} = (0,0)'$ and $v_0 = v_1 = 10^4$. We let $a_0 = a_1 = a_\epsilon = 2$ and $b_0 = b_1 = b_\epsilon = 1$, so that each variance parameter is assigned an $IG(2,1)$ prior distribution.

To stabilise the variance and avoid negative predictions, we model the data on the square root scale, as done by Sahu et al. (2007) and Berrocal et al. (2010) when modelling ozone concentrations for the U.S.

4.5.1 Estimating decay parameters

To estimate the spatial decay parameters we perform a grid search over a range of values for ϕ_0 and ϕ_1. A grid search is equivalent to placing a discrete uniform prior distribution upon the decay parameters and is a commonly adopted approach, see Sahu et al. (2011); Berrocal et al. (2010); Sahu et al. (2007). We obtain predictions at the validation sites as described in Section 4.3. The estimates are taken to be the pair of values that minimise the prediction error with respect to the validation data. The criteria used to compute the prediction error are the mean absolute prediction error (MAPE), the root mean squared prediction error (RMSPE) and the continuous ranked probability score (CRPS), defined in Section 2.5.

The greatest distance between any two of the 132 monitoring stations in California is 1190 kilometers (km) and so we select values of ϕ_0 and ϕ_1 corresponding to effective ranges of 50, 100, 250, 500 and 1000 km. For each of the 25 pairs of spatial decay parameters we generate a single chain of 25,000 iterations and discard the first 5,000.

We denote by d_0 and d_1 the effective range implied by ϕ_0 and ϕ_1 respectively. The values of the MAPE, RMSPE and CRPS for the 25 combinations of d_0 and d_1 are given in Table 4.1. We see that the prediction error is minimised for two of the three criteria when $d_0 = 250$ and $d_1 = 500$ and so our estimates for the spatial decay parameters are

$$\hat{\phi}_0 = -\log(0.05)/250, \quad \text{and} \quad \hat{\phi}_1 = -\log(0.05)/500. \tag{4.8}$$

4.5.2 Prior sensitivity

In this section we assess the sensitivity of prediction to changes in the hyperparameters a and b for the $IG(a,b)$ prior placed upon the variance parameters. By the properties of the inverse gamma distribution we have a prior mean and variance for the variance

Table 4.1: Prediction error for different combinations of d_0 and d_1.

d_0	d_1	MAPE	RMSPE	CRPS
50	50	15.84	19.45	11.15
	100	15.83	19.41	11.15
	250	15.87	19.40	11.14
	500	15.91	19.41	11.16
	1000	15.89	19.41	11.15
100	50	14.92	18.58	10.52
	100	14.97	18.58	10.55
	250	14.99	18.53	10.54
	500	14.98	18.50	10.53
	1000	15.01	18.53	10.54
250	50	**14.63**	18.39	10.42
	100	14.69	18.39	10.44
	250	14.70	18.33	10.43
	500	14.65	**18.27**	**10.39**
	1000	14.66	18.28	10.40
500	50	15.37	19.10	11.00
	100	15.36	19.06	10.99
	250	15.29	18.96	10.93
	500	15.30	18.94	10.93
	1000	15.28	18.93	10.93
1000	50	16.17	20.20	11.98
	100	16.24	20.22	11.99
	250	16.17	20.04	11.90
	500	16.19	20.05	11.91
	1000	16.24	20.05	11.94

parameters of $b/(a-1)$ and $b^2/\{(a-1)^2(a-2)\}$ respectively. We assign the same values to the hyperparameters for each of the variance parameters and so $a = a_0 = a_1 = a_\epsilon$ and $b = b_0 = b_1 = b_\epsilon$. Decay parameters are fixed at the values given in (4.8). Table 4.2 gives the prediction error for a range of values for a and b. We see that the predictions are robust to changes in a and b. Only when we impose an extreme prior mean to the variance parameters is the quality of the predictions degraded. In what follows we let $a_0 = a_1 = a_\epsilon = 2$ and $b_0 = b_1 = b_\epsilon = 1$.

4.5.3 The impact of the decay parameters on the performance of the CP and the NCP

In the simulation study of Section 4.4, a model with one process is considered and the effective range d_0 is held at the value that is used to generate the data. There, it is shown that the performance of the sampler is strongly affected by the value of the effective range. For the Californian ozone data we do not know the effective ranges d_0 and d_1 and their values are selected by minimising the prediction error, see Section 4.5.1. Although this is a common approach, we can see from Table 4.1 that the quality of prediction is not sensitive to changes in d_0 and d_1. In this section we look at how the values of d_0 and d_1 affect the performance of the CP and the NCP.

Table 4.2: Prediction error for different hyperparameters of the $IG(a, b)$ prior placed upon the variance parameters.

a	b	MAPE	RMSPE	CRPS
2	1	14.79	18.35	10.44
3	1	14.75	18.33	10.43
4	1	14.76	18.33	10.42
1000	1	14.79	18.35	10.44
2	2	14.79	18.37	10.45
2	3	14.77	18.34	10.43
2	4	14.79	18.36	10.46
2	1000	86.36	90.41	39.76
5	5	14.80	18.36	10.45
10	10	14.84	18.40	10.50

We look at five effective ranges: 50, 100, 250, 500 and 1000 km, hence there are 25 pairs of d_0 and d_1. For each pair we generate five Markov chains of length 25,000 and compute the $\text{MPSRF}_M(1.1)$ and the ESS of θ_0, θ_1, σ_0^2, σ_1^2 and σ_ϵ^2. Results for the CP are given in Table 4.3. We can see that the lowest values for the $\text{MPSRF}_M(1.1)$ are found for the longer effective ranges. It is clear that the ESS for θ_0 increases when d_0 is increased and the ESS for θ_1 and increases when d_1 is increased. The ESS of σ_0^2 reduces with increasing d_0 and is insensitive to changes in d_1, whereas the ESS of σ_1^2 is reduced by increasing d_1 but is not strongly affected by the level of d_0. For a fixed value of d_0 the ESS of σ_ϵ^2 increases in d_1 where the best value for d_0 is 250 km.

Results for the NCP are given in Table 4.4. We can see that the performance degrades significantly for longer effective ranges, and the $\text{MPSRF}_M(1.1)$ is almost 15,000 for $d_0 = d_1 = 1000$. The ESS of θ_0 is reduced by increasing d_0 but stable under changes in d_1, and similarly the ESS of θ_1 is reduced as d_1 is increased but insensitive to changes in d_0. The pattern for the ESS of the variance parameters for the NCP is the same as that of the variance parameters for the CP.

4.5.4 Selecting a fitting method

In this section we compare the performance of the CP and the NCP when fitting model (4.1) to the Californian ozone concentration data. We have an intercept and a slope process and we will compare the performance of each parameterisation when we update all random effects together or in two blocks according to the process of which they are realisations. In Section 3.8.4 it is shown that as long as the covariate $\boldsymbol{x}$ is centred, so that it has zero mean, the convergence rate remains the same regardless of whether we update θ_0 and θ_1 together or separately. As it is more efficient to sample from univariate distributions, we will update the global effects as two separate components. Therefore, we have four fitting strategies; the CP and the NCP with random effects updated in two blocks, which we label CP_2 and NCP_2, and the CP and the NCP where all of the the random effects are updated as one block, which we label CP_1 and NCP_1 respectively. The full conditional distributions needed to block update $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}$ are given in equations (2.10)

Table 4.3: The $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters under different combinations of d_0 and d_1 for the CP.

d_0	d_1	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS θ_1	ESS σ_0^2	ESS σ_1^2	ESS σ_ϵ^2
50	50	675	48559	5486	30511	5109	9264
	100	420	46668	8881	30281	4615	9971
	250	630	48960	16859	31413	4481	10373
	500	570	45185	26310	34031	4166	10281
	1000	410	49989	38392	34099	3583	10322
100	50	465	69147	6237	34693	5985	12481
	100	430	69459	9683	35664	5582	14489
	250	310	64122	18027	37391	5232	15078
	500	600	64652	29680	38646	4626	14606
	1000	440	65409	40612	38285	3635	15420
250	50	715	95206	7455	25203	6315	12477
	100	370	96426	11862	26064	6062	16055
	250	225	94084	21687	26394	5892	16931
	500	170	86589	35370	27553	4887	17209
	1000	470	88235	47911	27048	3883	17432
500	50	560	110988	8061	17153	5976	12187
	100	315	107418	12887	18234	6300	14856
	250	530	105493	23750	19592	5815	15323
	500	180	102489	37490	18078	5005	15672
	1000	180	105549	52678	19026	3974	15509
1000	50	505	113575	8984	13343	6007	11429
	100	590	114256	13903	14006	5753	13368
	250	245	111889	25438	13831	5477	13809
	500	240	114169	39463	14967	4821	14800
	1000	345	111179	54504	14158	3782	14280

and (2.12) respectively.

For each fitting strategy we generate five Markov chains of length 25,000 from the same widely dispersed starting values. The $\text{MPSRF}_M(1.1)$ and the ESS for $\boldsymbol{\theta} = (\theta_0, \theta_1)'$, $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)'$ and σ_ϵ^2 are computed and given in table 4.5. Compare first the CP_2 and NCP_2. The CP_2 requires far fewer iterations for the MPSRF to drop below 1.1 than the NCP_2, 120 versus 1995. There is a significant difference in the ESS of the mean parameters between the parameterisations. The CP_2 yields more than 70 times the number of effect samples for θ_0, and more than 10 times the number of effective samples for θ_1 than the NCP_2. The CP_2 mixes better in the σ_0^2 coordinate and has small advantages over the NCP_2 in terms of the ESS for σ_1^2 and σ_ϵ^2 . Fitting strategies CP_1 and NCP_1 compare similarly. Respectively their $\text{MPSRF}_M(1.1)$'s are 135 and 1405. The ESS is over 80 times greater for θ_0, and over 16 times greater for θ_1 when the CP_1 is used instead of the NCP_1. The ESS for the variance parameters is higher for the CP_1 than the NCP_1, in particular for σ_0^2.

If we compare blocking strategies, CP_2 with CP_1 and NCP_2 with NCP_1, we see that while there is little difference in the results for non-centred parameterisations, updating all random effects at once provides a significant increase in the ESS of the global mean

Table 4.4: The $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters under different combinations of d_0 and d_1 for the NCP.

d_0	d_1	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS θ_1	ESS σ_0^2	ESS σ_1^2	ESS σ_ϵ^2
50	50	575	8031	13135	24580	5081	11480
	100	1150	7752	12079	25118	4700	11781
	250	425	7790	6980	24624	4356	11563
	500	865	7722	3657	26519	4051	11912
	1000	1415	7995	1993	29054	3264	11427
100	50	515	4050	13890	27390	5965	14580
	100	660	4036	11944	27475	5630	15175
	250	1045	4006	6720	27080	4887	14884
	500	790	3820	3502	27862	4495	14708
	1000	2145	4118	1808	27807	3247	14934
250	50	1995	1213	16719	20602	6313	15881
	100	1980	1246	13908	21857	5921	16045
	250	1860	1140	6978	18715	5633	16795
	500	2075	1242	3747	22525	4746	16494
	1000	1405	1191	1693	20419	3214	16053
500	50	1815	428	18699	14663	5974	14874
	100	4425	405	14933	14225	6220	15969
	250	4735	452	7816	16334	5691	16291
	500	3200	445	3755	15572	4703	15705
	1000	4590	451	1997	15758	3724	15278
1000	50	8960	135	20648	11317	5824	13817
	100	4820	145	15944	11616	5675	13897
	250	9800	152	8247	11305	5396	14043
	500	6745	160	4037	12125	4735	14753
	1000	14985	162	2104	12645	3612	14487

parameters for the centred parameterisations. This is consistent with the results of Section 3.8.4 in which the exact convergence rates for this model are investigated.

However, there is a computational cost of jointly updating all random effects. For each iteration of the Gibbs sampler we must construct the Cholesky decomposition of a $2n \times 2n$ matrix, an operation of cubic computational complexity. This means that each iteration takes longer than if we were to update the random effects according to the processes from which they are realised, which requires the decomposition of two, $n \times n$ matrices.

We let

$$\text{MPSRF}_t(1.1) = \text{MPSRF}_M(1.1) \times \text{time per iteration},$$

denote the computation time (in seconds) for the MPSRF to fall below 1.1, and let ESS/s denote the ESS per second. Table 4.6 gives the time adjusted measures for each fitting method. We see that when computation time is considered the approaches that update the random effects in two blocks are more efficient. It is clear that the choice is between the CP_2 and the CP_1. We must decide between the method whose sampler can be run more quickly, the CP_2, and the one that returns samples that have lower autocorrelation, the CP_1. The utility placed upon these two factors will differ between practitioners and it is up to the individual to decide which is favourable.

Table 4.5: MPSRF$_M$(1.1) and ESS of the model parameters.

	MPSRF$_M$(1.1)	ESS θ_0	ESS θ_1	ESS σ_0^2	ESS σ_1^2	ESS σ_ϵ^2
CP$_2$	120	89230	36170	27283	4977	17095
NCP$_2$	1995	1238	3485	20863	4214	16017
CP$_1$	135	103129	56718	29539	5271	16836
NCP$_1$	1405	1252	3242	23371	4333	15797

Table 4.6: MPSRF$_t$(1.1) and ESS/s of the model parameters.

	MPSRF$_t$(1.1)	ESS/s θ_0	ESS/s θ_1	ESS/s σ_0^2	ESS/s σ_1^2	ESS/s σ_ϵ^2
CP$_2$	0.7	119.8	48.6	36.6	6.7	22.9
NCP$_2$	11.9	1.7	4.7	28.0	5.6	21.5
CP$_1$	2.6	42.7	23.5	12.2	2.2	7.0
NCP$_1$	27.2	0.5	1.3	9.7	1.8	6.5

4.5.5 Posterior inference

In this section we use the CP$_2$ to obtain posterior estimates of the model parameters. We run a single long chain of 50,000 iterations and discard the first 10,000. Parameter estimates and their 95% credible intervals are given in Table 4.7. We also include estimates and 95% credible intervals for the variance ratios $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ and $\delta_1 = \sigma_1^2/\sigma_\epsilon^2$.

Table 4.7: Parameter estimates and their 95% credible intervals (CI).

Parameter	Estimate	95% CI
θ_0	8.654	$(8.197, 9.010)$
θ_1	-0.176	$(-0.784, 0.397)$
σ_0^2	0.677	$(0.456, 0.958)$
σ_1^2	0.360	$(0.143, 0.768)$
σ_ϵ^2	0.137	$(0.081, 0.218)$
δ_0	5.329	$(2.428, 9.970)$
δ_1	2.808	$(0.914, 6.716)$

A negative estimate for θ_1 implies that ozone concentrations are higher in rural areas, although we see here that given the spatially correlated random effects, θ_1 is not significantly different from zero. The variances of the spatial processes are estimated to be larger than that of the pure error process as the Gaussian processes capture the spatial variation in the data. The estimates of the variance ratios δ_0 and δ_1 are approximately five and three receptively. Given these results it is not surprising that the CP outperformed the NCP here. The density plots for the model parameters are given in Figure 4.9.

4.6 Summary

In this chapter we have given details of how to construct Gibbs samplers for the CP and NCP of a general spatially varying coefficients model. The sampling efficiency of the parameterisations is compared via the (M)PSRF$_M$(1.1) and the ESS of the unknown

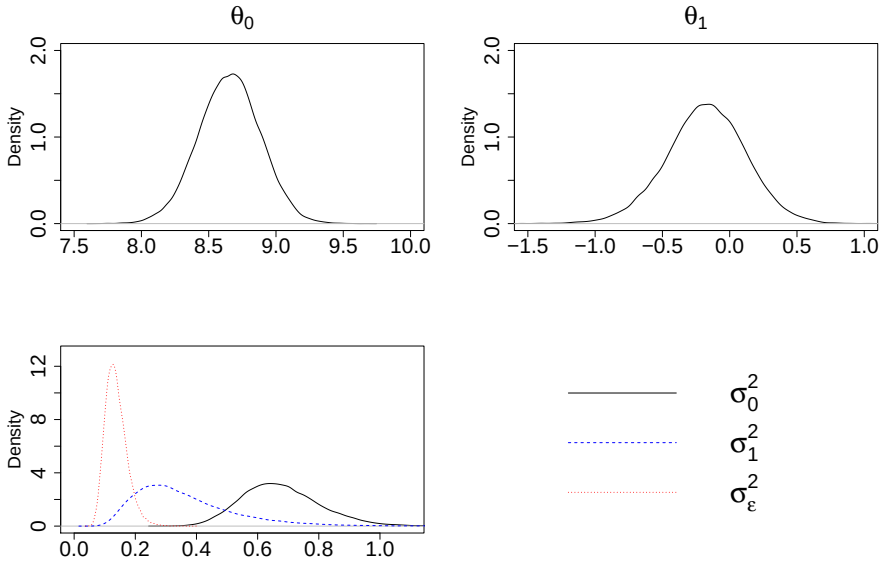

Figure 4.9: Density plots of model parameters for Californian ozone concentration data.

model parameters. Simulation studies suggest that when the covariance parameters are assumed to be known, these measures are in good agreement with the exact convergence rates computed in Chapter 3. Therefore, we use them to judge the efficiency of the Gibbs samplers emitted by the CP and the NCP when we do not have access to the exact convergence rates, i.e. when the variance parameters are unknown.

We have seen that the relationships established in Chapter 3 between the sampling efficiency of the respective parameterisations, and the ratio of the variance parameters and the strength of spatial correlation, still hold for unknown variance parameters. The CP performs better when the data precision is relatively high and when the correlation is strong. Contrary to this, the NCP performs best for when the data is less informative and the correlation is weak.

We have fitted the CP and the NCP of a model with spatially varying intercept and slope to ozone concentration data from California. We find that the performance of the CP is far superior than that of the NCP due to the strong spatial correlation in the data. Block updating all of the random effects together has little effect on the NCP but gives a higher ESS for the mean parameters for the CP. However, this comes with a computational time penalty as we have to invert larger matrices.

Chapter 5

Partially centred parameterisations for spatial models

5.1 Introduction

In Chapters 3 and 4 we have considered the CP and the NCP of spatial models. We find that the performance of the Gibbs samplers under these model parameterisations is dependent on the informativity of the data about the latent surface. Without knowing the values of the covariance parameters we cannot recommend one parameterisation over another *apriori*.

In this chapter we look to construct a parameterisation that has an associated Gibbs sampler whose performance is robust to changes of the values of the covariance parameters. Papaspiliopoulos et al. (2003) consider a family of partially centred parameterisations (PCPs) that lie on a continuum between the CP and the NCP at the extremes. They construct a PCP for a two stage NLHM that has a Gibbs sampler with zero convergence rate. By minimising the posterior covariance between the global and random effects we construct a PCP for the three stage NLHM that has the same property. Furthermore, we show that to achieve immediate convergence we must update all random effects as one block and all global effects as another.

The PCP is constructed conditional on the covariance parameters. When these parameters are unknown we propose a dynamically updated PCP and show that stationarity is only preserved if the parameterisation is updated with each new draw of a covariance parameter. We go on to suggest pilot adaption schemes which limit the number of matrix computations needed when we run the Gibbs sampler under the PCP with unknown variance parameters.

The rest of this chapter is organised as follows: In Section 5.2 we give details of the construction of the PCP for the three stage NLHM. In Section 5.3 we look at how the optimal weights of partial centering vary with the variance parameters and the correlation structure of the random effects. Section 5.4 gives the full conditional distributions needed

for Gibbs sampling under the PCP. We conduct a simulation study to assess the effectiveness of the PCP for the cases when the variance parameters are known and then when they are unknown. Section 5.5 applies the PCP to model Californian ozone concentration data. Section 5.6 contains some summary comments.

5.2 Construction and properties of the PCP

In Section 1.3.1 we consider a simple random effects model to illustrate hierarchical centering. The partially centred form of the model is given by

$$\begin{aligned} Y_i &\sim N((1-w)\theta + U_i^w, \sigma_\epsilon^2) \\ U_i^w &\sim N(w\theta, \sigma_u^2), \end{aligned} \tag{5.1}$$

for $i = 1, \ldots, n$ and $w \in [0, 1]$. This follows the PCP given by Papaspiliopoulos (2003, Chapter 7), but instead they consider $w^* = 1 - w$.

We can show that for a flat prior distribution on θ that the posterior precision matrix of $\boldsymbol{U}^w = (U_1^w, \ldots, U_n^w)'$ and θ is

$$\boldsymbol{Q}^{pc} = \begin{pmatrix} \left(\frac{1}{\sigma_\epsilon^2} + \frac{1}{\sigma_u^2}\right)\boldsymbol{I} & \left(\frac{1-w}{\sigma_\epsilon^2} - \frac{w}{\sigma_u^2}\right)\boldsymbol{1} \\ \left(\frac{1-w}{\sigma_\epsilon^2} - \frac{w}{\sigma_u^2}\right)\boldsymbol{1}' & \frac{(1-w)^2 n}{\sigma_\epsilon^2} + \frac{w^2 n}{\sigma_u^2} \end{pmatrix}.$$

Applying Theorem 2.3.1 we find that the convergence rate for model (5.1) is given by

$$\begin{aligned} \lambda_{pc} &= \frac{((1-w)\sigma_u^2 - w\sigma_\epsilon^2)^2}{((1-w)^2\sigma_u^2 + w^2\sigma_\epsilon^2)(\sigma_\epsilon^2 + \sigma_u^2)} \\ &= \frac{((1-w)\lambda_{nc} - w\lambda_c)^2}{(1-w)^2\lambda_{nc} + w^2\lambda_c}. \end{aligned} \tag{5.2}$$

From equations (2.6) and (2.7) we have that $\lambda_{nc} = \sigma_u^2/(\sigma_\epsilon^2 + \sigma_u^2) = 1 - \lambda_c$. We can see from the numerator of (5.2) that $\lambda_{pc} = 0$ when $w = \lambda_{nc}$. Note that from equation (1.5) we have that $Corr(U_i^w, \theta|\boldsymbol{y}) = 0$ when $w = \lambda_{nc}$.

5.2.1 Constructing the PCP of the spatially varying coefficients model

We can apply the constructive approach of model (5.1) to the spatially varying coefficients model discussed in Chapters 3 and 4. A PCP is found by introducing the partially centred spatially correlated random effects, defined to be

$$\beta_k^w(\boldsymbol{s}_i) = \beta_k(\boldsymbol{s}_i) + w_k\theta_k, \quad k = 0, \ldots, p-1, \quad i = 1, \ldots, n, \tag{5.3}$$

where $\beta_k(\boldsymbol{s}_i)$ is a realisation of a zero mean spatial process and $w_k \in [0, 1]$ is the weight of partial centering for kth process. Substituting equation (5.3) into the general spatial

model given in (4.1) gives us the PCP of the model, which is written as

$$Y(\boldsymbol{s}_i) = \sum_{k=0}^{p-1}\{(1-w_k)\theta_k + \beta_k^w(\boldsymbol{s}_i)\}x_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \tag{5.4}$$

where $x_0(\boldsymbol{s}_i) = 1$ and $i = 1, \ldots, n$.

We let $\boldsymbol{\beta}^w = (\boldsymbol{\beta}_0^{w\prime}, \ldots, \boldsymbol{\beta}_{p-1}^{w\prime})'$, where $\boldsymbol{\beta}_k^w = (\beta_k^w(\boldsymbol{s}_1), \ldots, \beta_k^w(\boldsymbol{s}_n))'$, and then we write the partially centred random effects as

$$\boldsymbol{\beta}^w = \tilde{\boldsymbol{\beta}} - (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \tag{5.5}$$

where we recall that $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0', \ldots, \tilde{\boldsymbol{\beta}}_{p-1}')'$ is the vector of centred random effects of length np, $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ is the $p \times 1$ vector of global regression coefficients and $\boldsymbol{X}_2$ is a $np \times p$ block diagonal matrix with blocks made up of vectors of ones of length n,

$$\boldsymbol{X}_2 = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1} \end{bmatrix}.$$

It follows from (5.4) that $\boldsymbol{W}$ is a $np \times np$ diagonal matrix given by

$$\boldsymbol{W} = \begin{bmatrix} w_0\boldsymbol{I} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & w_1\boldsymbol{I} & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & w_{p-1}\boldsymbol{I} \end{bmatrix}. \tag{5.6}$$

By substituting (5.5) into model (4.2) the partially centred model is written as

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}^w, \boldsymbol{\theta} &\sim N(\boldsymbol{X}_1\boldsymbol{\beta}^w + \boldsymbol{X}_1(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_1) \\ \boldsymbol{\beta}^w|\boldsymbol{\theta} &\sim N(\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3). \end{aligned} \tag{5.7}$$

Note that if $\boldsymbol{W}$ is the identity matrix we recover the CP and where $\boldsymbol{W}$ is the zero matrix we have the NCP. The question is how do we choose the entries of $\boldsymbol{W}$ such that optimal performance of the Gibbs sampler is achieved?

5.2.2 Posterior covariance matrices and convergence rates for the PCP

Now that we have the model written in the three stage hierarchical form given in (5.7) we can apply the approach of Section 2.4.2 and use expression (5.5) to compute the posterior

variance of $\boldsymbol{\beta}^w$. This gives us

$$\begin{aligned} Var(\boldsymbol{\beta}^w|\boldsymbol{y}) &= Cov(\tilde{\boldsymbol{\beta}} - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}} - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}|\boldsymbol{y}) \\ &= Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{y}) - Cov(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}|\boldsymbol{y})\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 Cov(\boldsymbol{\theta}, \tilde{\boldsymbol{\beta}}|\boldsymbol{y}) \\ &\quad +(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 Var(\boldsymbol{\theta}|\boldsymbol{y})\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' \\ &= \boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' \\ &\quad -(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} + (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})', \end{aligned} \tag{5.8}$$

where we recall that

$$\boldsymbol{B} = Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{y}) = \left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1},$$

and

$$\boldsymbol{\Sigma}_{\theta|y} = Var(\boldsymbol{\theta}|\boldsymbol{y}) = \left((\boldsymbol{X}_1\boldsymbol{X}_2)'\boldsymbol{\Sigma}_{Y|\theta}^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1}, \tag{5.9}$$

where $\boldsymbol{\Sigma}_{Y|\theta} = \boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1'$. The posterior covariance of $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$ is given by

$$\begin{aligned} Cov(\boldsymbol{\beta}^w, \boldsymbol{\theta}|\boldsymbol{y}) &= Cov(\tilde{\boldsymbol{\beta}} - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{\theta}|\boldsymbol{y}) \\ &= Cov(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta}|\boldsymbol{y}) - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 Var(\boldsymbol{\theta}|\boldsymbol{y}) \\ &= \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y} - (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y} \\ &= \left(\boldsymbol{B}\boldsymbol{C}_2^{-1} - (\boldsymbol{I}-\boldsymbol{W})\right)\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}. \end{aligned} \tag{5.10}$$

We can see from (5.10) that $Cov(\boldsymbol{\beta}^w, \boldsymbol{\theta}|\boldsymbol{y}) = \boldsymbol{0}$ when $\boldsymbol{B}\boldsymbol{C}_2^{-1} = \boldsymbol{I}-\boldsymbol{W}$, or equivalently when

$$\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1} = \boldsymbol{I} - \left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1}\boldsymbol{C}_2^{-1}. \tag{5.11}$$

Equation (5.11) implies that to minimise the posterior correlation between the random effects and global effects we cannot restrict $\boldsymbol{W}$ to be the diagonal matrix given in (5.6). It then follows that as $\boldsymbol{\beta}^w|\boldsymbol{\theta} \sim N(\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2)$ *apriori*, the prior mean of $\beta_k^w(\boldsymbol{s}_i)$ will be a linear combination of all elements of $\boldsymbol{\theta}$ and not just a proportion of θ_k as expressed in (5.3).

Henceforth we will set $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$, and further explore the consequences of this construction. Substituting this expression into the posterior variance for $\boldsymbol{\beta}^w$ given in equation (5.8) we get

$$\begin{aligned} Var(\boldsymbol{\beta}^w|\boldsymbol{y}) &= \boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' - \\ &\quad -(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} + (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' \\ &= \boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} - \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} - \\ &\quad -\boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B} \\ &= \boldsymbol{B}. \end{aligned}$$

We now look at the implication of (5.11) for the convergence rate of a Gibbs sampler using the PCP. First we need the posterior precision matrix of $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$, which we can

identify by writing

$$\begin{aligned}
\pi(\boldsymbol{\beta}^w, \boldsymbol{\theta}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\beta}^w, \boldsymbol{\theta})\pi(\boldsymbol{\beta}^w|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \\
&\propto \exp\Big\{ -\frac{1}{2}\Big[\big(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}^w - \boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\big)'\boldsymbol{C}_1^{-1}\big(\boldsymbol{Y} - \boldsymbol{X}_1\boldsymbol{\beta}^w \\
&\quad -\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\big) + \big(\boldsymbol{\beta}^w - \boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\big)'\boldsymbol{C}_2^{-1}\big(\boldsymbol{\beta}^w - \boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\big) \\
&\quad +(\boldsymbol{\theta}-\boldsymbol{m})'\boldsymbol{C}_3^{-1}(\boldsymbol{\theta}-\boldsymbol{m})\Big]\Big\} \\
&= \exp\Big\{ -\frac{1}{2}\Big[\ldots + \boldsymbol{\beta}^{w\prime}\big(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\big)\boldsymbol{\beta}^w \\
&\quad +2\boldsymbol{\beta}^{w\prime}\big(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 - \boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2\big)\boldsymbol{\theta} + \boldsymbol{\theta}'\big(\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})' \\
&\quad \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 + \boldsymbol{X}_2'\boldsymbol{W}'\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\big)\boldsymbol{\theta} + \ldots\Big]\Big\}.
\end{aligned}$$

Then we can see that the posterior precision matrix for the PCP is

$$\boldsymbol{Q}^{pc} = \begin{pmatrix} \boldsymbol{Q}^{pc}_{\beta^w} & \boldsymbol{Q}^{pc}_{\beta^w\theta} \\ \boldsymbol{Q}^{pc}_{\theta\beta^w} & \boldsymbol{Q}^{pc}_{\theta} \end{pmatrix}, \tag{5.12}$$

where

$$\begin{aligned}
\boldsymbol{Q}^{pc}_{\beta^w} &= \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}, \\
\boldsymbol{Q}^{pc}_{\beta^w\theta} &= \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 - \boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2, \\
\boldsymbol{Q}^{pc}_{\theta} &= \boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 + \boldsymbol{X}_2'\boldsymbol{W}'\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}.
\end{aligned}$$

If we block update a Gibbs sampler according to the partitioning of the precision matrix (5.12), by corollary 2.3.2, we have that the convergence rate of the PCP is the maximum modulus eigenvalue of the matrix

$$\boldsymbol{F}^{pc}_{22} = (\boldsymbol{Q}^{pc}_{\theta})^{-1}\boldsymbol{Q}^{pc}_{\theta\beta^w}(\boldsymbol{Q}^{pc}_{\beta^w})^{-1}\boldsymbol{Q}^{pc}_{\beta^w\theta}.$$

Consider $\boldsymbol{Q}^{pc}_{\beta^w\theta}$ and substitute $\boldsymbol{W}$ from equation (5.11), then we have

$$\begin{aligned}
\boldsymbol{Q}^{pc}_{\beta^w\theta} &= \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 - \boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2 \\
&= (\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1)\left[(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_2^{-1}\right]\boldsymbol{X}_2 - \boldsymbol{C}_2^{-1}\Big[\boldsymbol{I} - (\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 \\
&\quad +\boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_2^{-1}\Big]\boldsymbol{X}_2 \\
&= \Big[(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1)(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_2^{-1} + \boldsymbol{C}_2^{-1}(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_2^{-1} \\
&\quad -\boldsymbol{C}_2^{-1}\Big]\boldsymbol{X}_2 \\
&= \Big[(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})^{-1}\boldsymbol{C}_2^{-1} - \boldsymbol{C}_2^{-1}\Big]\boldsymbol{X}_2 \\
&= \Big[\boldsymbol{C}_2^{-1} - \boldsymbol{C}_2^{-1}\Big]\boldsymbol{X}_2 \\
&= \boldsymbol{0}.
\end{aligned}$$

Therefore by setting $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$, $\boldsymbol{F}_{22}^{pc}$ becomes the zero matrix and immediate convergence is achieved. A more straightforward way to see this is that by construction we have a 2×2 block diagonal posterior covariance matrix for $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$. Therefore the precision matrix is also block diagonal and $\boldsymbol{F}_{22}^{pc}$ is null.

Suppose now that we have constructed the PCP as before but we partition the partially centred random effects, $\boldsymbol{\beta}^w$, into two disjoint sets, $\boldsymbol{\beta}_1^w$ and $\boldsymbol{\beta}_2^w$, and update them separately in a Gibbs sampler. Partitioned accordingly, the covariance matrix is a 3×3 block matrix given by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\beta_1} & \boldsymbol{\Sigma}_{\beta_{12}} & \mathbf{0} \\ \boldsymbol{\Sigma}_{\beta_{21}} & \boldsymbol{\Sigma}_{\beta_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{\theta} \end{pmatrix}.$$

Using results from Harville (1997, Chapter 8) we find the corresponding precision matrix to be

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{\beta_1} & \boldsymbol{Q}_{\beta_{12}} & \mathbf{0} \\ \boldsymbol{Q}_{\beta_{21}} & \boldsymbol{Q}_{\beta_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_{\theta} \end{pmatrix}, \tag{5.13}$$

where

$$\begin{aligned}
\boldsymbol{Q}_{\beta_1} &= (\boldsymbol{\Sigma}_{\beta_1} - \boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}\boldsymbol{\Sigma}_{\beta_{21}})^{-1}, \\
\boldsymbol{Q}_{\beta_{12}} &= -(\boldsymbol{\Sigma}_{\beta_1} - \boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}\boldsymbol{\Sigma}_{\beta_{21}})^{-1}\boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}, \\
\boldsymbol{Q}_{\beta_{21}} &= -(\boldsymbol{\Sigma}_{\beta_2} - \boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}\boldsymbol{\Sigma}_{\beta_{12}})^{-1}\boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}, \\
\boldsymbol{Q}_{\beta_2} &= (\boldsymbol{\Sigma}_{\beta_2} - \boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}\boldsymbol{\Sigma}_{\beta_{12}})^{-1}, \\
\boldsymbol{Q}_{\theta} &= \boldsymbol{\Sigma}_{\theta}^{-1}.
\end{aligned}$$

It can be shown that a Gibbs sampler with Gaussian target distribution with precision matrix given by

$$\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} & \boldsymbol{Q}_{13} \\ \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} & \boldsymbol{Q}_{23} \\ \boldsymbol{Q}_{31} & \boldsymbol{Q}_{32} & \boldsymbol{Q}_{33} \end{pmatrix},$$

has a convergence rate which is equal to the maximum modulus eigenvalue of

$$\boldsymbol{F} = \begin{pmatrix} \mathbf{0} & -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} & -\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{13} \\ \mathbf{0} & \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12} & \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21}\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{13} - \boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{23} \\ \mathbf{0} & \boldsymbol{F}_{32} & \boldsymbol{F}_{33} \end{pmatrix},$$

where

$$\begin{aligned}
\boldsymbol{F}_{32} &= (\boldsymbol{Q}_{33}^{-1}\boldsymbol{Q}_{31} - \boldsymbol{Q}_{33}^{-1}\boldsymbol{Q}_{32}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21})\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{12}, \\
\boldsymbol{F}_{33} &= (\boldsymbol{Q}_{33}^{-1}\boldsymbol{Q}_{31} - \boldsymbol{Q}_{33}^{-1}\boldsymbol{Q}_{32}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{21})\boldsymbol{Q}_{11}^{-1}\boldsymbol{Q}_{13} + \boldsymbol{Q}_{33}^{-1}\boldsymbol{Q}_{32}\boldsymbol{Q}_{22}^{-1}\boldsymbol{Q}_{23}.
\end{aligned}$$

Therefore, the convergence rate corresponding to the precision matrix given by (5.13) is

the maximum modulus eigenvalue of

$$\boldsymbol{F}^{pc} = \begin{pmatrix} \mathbf{0} & -\boldsymbol{Q}_{\beta_1}^{-1}\boldsymbol{Q}_{\beta_{12}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_{\beta_2}^{-1}\boldsymbol{Q}_{\beta_{21}}\boldsymbol{Q}_{\beta_1}^{-1}\boldsymbol{Q}_{\beta_{12}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}\boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix},$$

which will be zero if the posterior correlation between $\boldsymbol{\beta}_1^w$ and $\boldsymbol{\beta}_2^w$ is zero.

Alternatively, suppose that we update $\boldsymbol{\beta}^w$ as one block but partition $\boldsymbol{\theta}$ into $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$, updating them accordingly. The covariance and precision matrices have the form

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_\beta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\theta_1} & \boldsymbol{\Sigma}_{\theta_{12}} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\theta_{21}} & \boldsymbol{\Sigma}_{\theta_2} \end{pmatrix}, \qquad \boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_\beta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Q}_{\theta_1} & \boldsymbol{Q}_{\theta_{12}} \\ \mathbf{0} & \boldsymbol{Q}_{\theta_{21}} & \boldsymbol{Q}_{\theta_2} \end{pmatrix},$$

where

$$\begin{aligned} \boldsymbol{Q}_\beta &= \boldsymbol{\Sigma}_\beta^{-1}, \\ \boldsymbol{Q}_{\theta_1} &= (\boldsymbol{\Sigma}_{\theta_1} - \boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}\boldsymbol{\Sigma}_{\theta_{21}})^{-1}, \\ \boldsymbol{Q}_{\theta_{12}} &= -(\boldsymbol{\Sigma}_{\theta_1} - \boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}\boldsymbol{\Sigma}_{\theta_{21}})^{-1}\boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}, \\ \boldsymbol{Q}_{\theta_{21}} &= -(\boldsymbol{\Sigma}_{\theta_2} - \boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}\boldsymbol{\Sigma}_{\theta_{12}})^{-1}\boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}, \\ \boldsymbol{Q}_{\theta_2} &= (\boldsymbol{\Sigma}_{\theta_2} - \boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}\boldsymbol{\Sigma}_{\theta_{12}})^{-1}, \end{aligned}$$

and the convergence rate is the maximum modulus eigenvalue of

$$\boldsymbol{F} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\boldsymbol{Q}_{\theta_1}^{-1}\boldsymbol{Q}_{\theta_{12}} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{Q}_{\theta_2}^{-1}\boldsymbol{Q}_{\theta_{21}}\boldsymbol{Q}_{\theta_1}^{-1}\boldsymbol{Q}_{\theta_{12}} \end{pmatrix} = \begin{pmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}\boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1} \end{pmatrix},$$

which will be a null matrix if the two blocks of $\boldsymbol{\theta}$ are uncorrelated *a posteriori*.

It is the relationship between convergence rate and inter-block correlation that we take advantage of when constructing the PCP. For our construction, immediate convergence is only guaranteed if the random effects and global effects are each updated as one complete block. If a greater number of blocks are used we cannot, in general, find a matrix $\boldsymbol{W}$ that will remove all cross covariances and return a convergence rate of zero. To see this first note that the posterior covariance matrix $\boldsymbol{\Sigma}_{\theta|y}$, given in (5.9), is unaffected by hierarchical centering and so partial centering cannot remove and any posterior correlation between subsets of $\boldsymbol{\theta}$, and therefore all of its elements must be updated together. Then suppose that we partition the partially centred random effects into l blocks so that $\boldsymbol{\beta}^w = (\boldsymbol{\beta}_1^{w\prime}, \ldots, \boldsymbol{\beta}_l^{w\prime})'$.

We find the posterior covariance between the ijth block to be

$$\begin{aligned} Cov(\boldsymbol{\beta}_i^w, \boldsymbol{\beta}_j^w|\boldsymbol{y}) &= \boldsymbol{B}_{ij} + \boldsymbol{B}_{i.}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j} - \boldsymbol{B}_{i.}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})_{.j}' \\ &\quad -(\boldsymbol{I}-\boldsymbol{W})_{i.}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'\boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j} + (\boldsymbol{I}-\boldsymbol{W})_{i.}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{I}-\boldsymbol{W})_{.j}' \\ &= \boldsymbol{B}_{ij} + \boldsymbol{B}_{i.}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'(\boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j} - (\boldsymbol{I}-\boldsymbol{W})_{.j}') \\ &\quad +(\boldsymbol{I}-\boldsymbol{W})_{i.}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}\boldsymbol{X}_2'((\boldsymbol{I}-\boldsymbol{W})_{.j}' - \boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j}), \end{aligned}$$

where $\boldsymbol{B}_{ij}$ is the ijth block of $\boldsymbol{B} = Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, y)$. We let $\boldsymbol{B}_{i.}$ denote the rows of $\boldsymbol{B}$ associated with the ith block and let $\boldsymbol{B}_{.j}$ denote the columns of $\boldsymbol{B}$ associated with the jth block, with $(\boldsymbol{I}-\boldsymbol{W})_{i.}$ and $(\boldsymbol{I}-\boldsymbol{W})_{.j}$ having similar interpretations. We see that if $(\boldsymbol{I}-\boldsymbol{W})_{.j}' = \boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j}$ then $Cov(\boldsymbol{\beta}_i^w, \boldsymbol{\beta}_j^w|\boldsymbol{y}) = \boldsymbol{B}_{ij}$, which is generally a non-zero matrix. Therefore we must update $\boldsymbol{\beta}^w$ as one component and $\boldsymbol{\theta}$ as another.

We have seen that letting $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$ gives us the optimum parameterisation in terms of posterior uncorrelatedness and convergence rate of the Gibbs sampler. Using the lemma 3.3.1 we write $\boldsymbol{W}$ as

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1} \\ &\quad \boldsymbol{I} - \left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1}\boldsymbol{C}_2^{-1} \\ &= \boldsymbol{I} - \left[\boldsymbol{C}_2 - \boldsymbol{C}_2\boldsymbol{X}_1'\left(\boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1\boldsymbol{C}_2\right]\boldsymbol{C}_2^{-1} \\ &= \boldsymbol{C}_2\boldsymbol{X}_1'\left(\boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1. \end{aligned} \tag{5.14}$$

which requires the inversion of a matrix of the size $n \times n$ not one of $np \times np$, as is the case for the representation of $\boldsymbol{W}$ given in (5.11). We then re-write model (5.7) as

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}^w, \boldsymbol{\theta} &\sim N(\boldsymbol{X}_1^{opt}(\boldsymbol{\beta}^{w\prime}, \boldsymbol{\theta})', \boldsymbol{C}_1) \\ \boldsymbol{\beta}^w|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2^{opt}\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned}$$

where the optimal design matrix for the latent process at the second of the three stage model is given by the $np \times p$ matrix

$$\boldsymbol{X}_2^{opt} = \boldsymbol{C}_2\boldsymbol{X}_1'\left(\boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1'\right)^{-1}\boldsymbol{X}_1\boldsymbol{X}_2,$$

hence the optimal design matrix for the first stage is given by the augmented $n \times (np+p)$ matrix

$$\begin{aligned} \boldsymbol{X}_1^{opt} &= [\boldsymbol{X}_1, \boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2] \\ &= \left[\boldsymbol{X}_1, \boldsymbol{X}_1(\boldsymbol{X}_2 - \boldsymbol{X}_2^{opt})\right]. \end{aligned}$$

5.3 Spatially varying weights for partial centering

In this section we will be investigating how the weights for the PCP depend on the variance parameters but also the correlation structure of the latent processes. In particular, we will see how the weights vary across the spatial region and see the impact upon the weights of a spatially varying covariate. To focus on these relationships we will consider simplified versions of model (5.7) that have one global parameter and one latent process.

5.3.1 Optimal weights for the equi-correlation model

We begin by looking at the following model,

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}_0^w, \theta_0 &\sim N(\boldsymbol{\beta}_0^w + (\boldsymbol{I} - \boldsymbol{W})\mathbf{1}\theta_0, \sigma_\epsilon^2 \boldsymbol{I}) \\ \boldsymbol{\beta}_0^w|\theta_0 &\sim N(\boldsymbol{W}\mathbf{1}\theta_0, \sigma_0^2 \boldsymbol{R}_0) \\ \theta_0 &\sim N(m_0, \sigma_0^2 v_0). \end{aligned} \tag{5.15}$$

which has one global parameter $\boldsymbol{\theta} = \theta_0$ and one latent spatial process $\boldsymbol{\beta}^w = \boldsymbol{\beta}_0^w$, and hence can be found from (5.7) by letting $\boldsymbol{X}_1 = \boldsymbol{I}$, $\boldsymbol{C}_1 = \sigma_\epsilon^2 \boldsymbol{I}$, $\boldsymbol{X}_2 = \mathbf{1}$, $\boldsymbol{C}_2 = \sigma_0^2 \boldsymbol{R}_0$, $\boldsymbol{m} = m_0$ and $\boldsymbol{C}_3 = \sigma_0^2 v_0$. Therefore, using the representation of $\boldsymbol{W}$ given in equation (5.14) we have

$$\boldsymbol{W} = \sigma_0^2 \boldsymbol{R}_0 (\sigma_\epsilon^2 \boldsymbol{I} + \sigma_0^2 \boldsymbol{R}_0)^{-1}. \tag{5.16}$$

Model (5.15) is the PCP of the model we used to investigate the properties of the CP and the NCP in Section 3.3. Here we look at the equi-correlated model as in Section 3.3.1 which is characterised by the following correlation structure for the random effects

$$(\boldsymbol{R}_0)_{ij} = \begin{cases} \rho & \text{if } i \neq j \\ 1 & \text{if } i = j, \end{cases}$$

for $0 \leq \rho < 1$. We write

$$\sigma_\epsilon^2 \boldsymbol{I} + \sigma_0^2 \boldsymbol{R}_0 = [\sigma_\epsilon^2 + \sigma_0^2(1-\rho)]\boldsymbol{I} + \sigma_0^2 \rho \boldsymbol{J},$$

then using Lemma 3.3.2 to find $(\sigma_\epsilon^2 \boldsymbol{I} + \sigma_0^2 \boldsymbol{R}_0)^{-1}$, we have that

$$\boldsymbol{W} = \frac{1}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho)} \left(\sigma_0^2(1-\rho)\boldsymbol{I} + \frac{\sigma_\epsilon^2 \sigma_0^2 \rho}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2 \rho} \boldsymbol{J} \right).$$

As $\boldsymbol{X}_2 = \mathbf{1}$, the entries of the optimal design matrix are the row sums of $\boldsymbol{W}$. Due to the equi-correlation assumption the weight for each data point is the same, and after some cancellation we get

$$(\boldsymbol{X}_2^{opt})_i = w = \frac{\sigma_0^2(1-\rho) + n\sigma_0^2 \rho}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2 \rho}. \tag{5.17}$$

Note that w is equal to the convergence rate of the NCP for the equi-correlated model for $1/v_0 = 0$, see equation (3.19). This extends the result from Papaspiliopoulos (2003, Chapter 7), who showed it to be true for $\rho = 0$, in which case $w = \sigma_0^2/(\sigma_\epsilon^2 + \sigma_0^2)$

Figure 5.1 illustrates equation (5.17) by plotting w against ρ for $n = 20, 50, 100, 250$, for $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$. We see that for fixed n and ρ that the optimal weight increases with the variance ratio $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$. When n and δ_0 are fixed, w is increases with increasing ρ. Finally we see that when δ_0 and ρ are fixed, increasing n increases w.

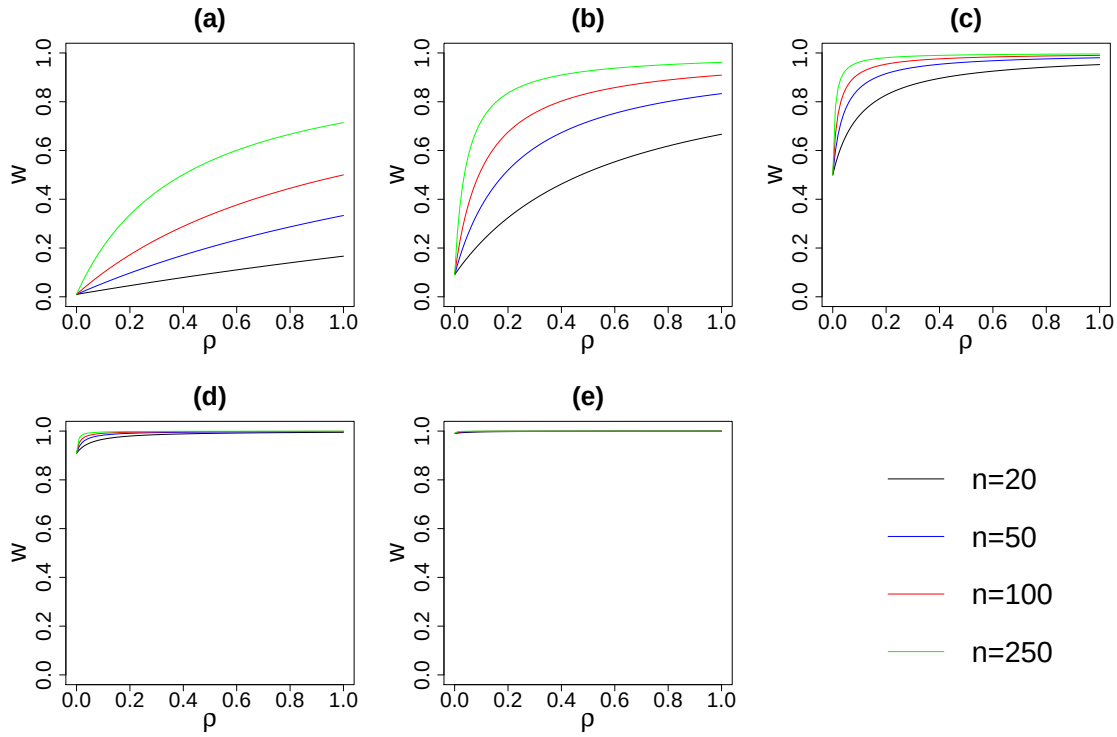

Figure 5.1: Optimal weights against correlation for the PCP of the equi-correlation model for $n = 20, 50, 100, 250$ for different values of δ_0. (a) $\delta_0 = 0.01$, (b) $\delta_0 = 0.1$, (c) $\delta_0 = 1$, (d) $\delta_0 = 10$, (e) $\delta_0 = 100$.

5.3.2 Surfaces of optimal weights for spatially correlated random effects

In this section we consider spatially referenced data. We continue to look at model (5.15) but now, as in Section 3.3.2, we impose a spatial correlation structure upon $\boldsymbol{\beta}_0^w$, such that

$$(\boldsymbol{R}_0)_{ij} = \exp\{-\phi_0 d_{ij}\}, \tag{5.18}$$

where ϕ_0 controls the rate of decay of the correlation between the partially centred random effects at sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$, and d_{ij} denotes the distance between them.

Here we select sampling locations according to a pattern, such that the locations are more densely clustered in some regions of the domain than others. We consider 200 locations in the unit square, see Figure 5.2, which we split into nine sub-squares of equal area. We randomly select 100 points in the top left square and 25 points in the three areas to which it is adjacent. The remaining five sub-squares have five points randomly chosen within them.

We consider five variance ratios: $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$, and three effective ranges: $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$. Note that an effective range of zero, implying independent

random effects, is equivalent to a equi-correlation model with $\rho = 0$, and so the weights are the same at each location. For each of the 15 variance ratio-effective range combinations we compute the spatially varying weights, $w(\boldsymbol{s}_i) = (\boldsymbol{W}\boldsymbol{1})_i$, $i = 1, \ldots, 200$, where $\boldsymbol{W}$ is given in (5.16).

We use the `Tps` (thin plate spline) function in the R package fields (Furrer et al., 2009) to interpolate the weights over the unit square. These interpolated plots of spatially varying weights are given in Figure 5.3. Each row corresponds to a value of δ_0, from 0.01 in top row to 100 in the bottom. For each row going left to right we have increasing effective ranges, $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$. We can see that as the variance ratio increases we favour a greater weight, as we do when the effective range increases. Within each panel, the areas of higher weights are concentrated around the areas of more densely positioned sampling locations. The stronger the correlation, the farther reaching is the influence of these clusters.

We make it clear that although the interpolated plots are informative they do not represent a true surface in the sense that the interpolated values are not estimates of a true value of $w(\boldsymbol{s})$ for an unsampled location. Indeed, if we were to include a new location in our set of points, the values of $w(\boldsymbol{s}_i)$ at the existing locations would be changed.

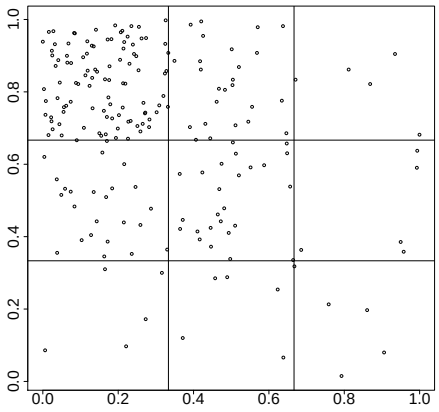

Figure 5.2: Patterned sampling locations. 100 top left; 25 in top middle, middle left and middle middle; five top right, middle right and bottom third.

5.3.3 Covariate surface and optimal weights

In this section we investigate the effect of a covariate upon the spatially varying weights. To do this we look at the PCP of model (3.24) which is used in Section 3.5 to compare the effect of the covariate on the CP and NCP. The model is given by

$$\begin{aligned} \boldsymbol{Y}|\boldsymbol{\beta}_1^w, \theta_1 &\sim N(\boldsymbol{D}\boldsymbol{\beta}_1^w + (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{1}\theta_1, \sigma_\epsilon^2 \boldsymbol{I}) \\ \boldsymbol{\beta}_1^w|\theta_1 &\sim N(\boldsymbol{W}\boldsymbol{1}\theta_1, \sigma_1^2 \boldsymbol{R}_1) \\ \theta_1 &\sim N(m_1, \sigma_1^2 v_1), \end{aligned} \tag{5.19}$$

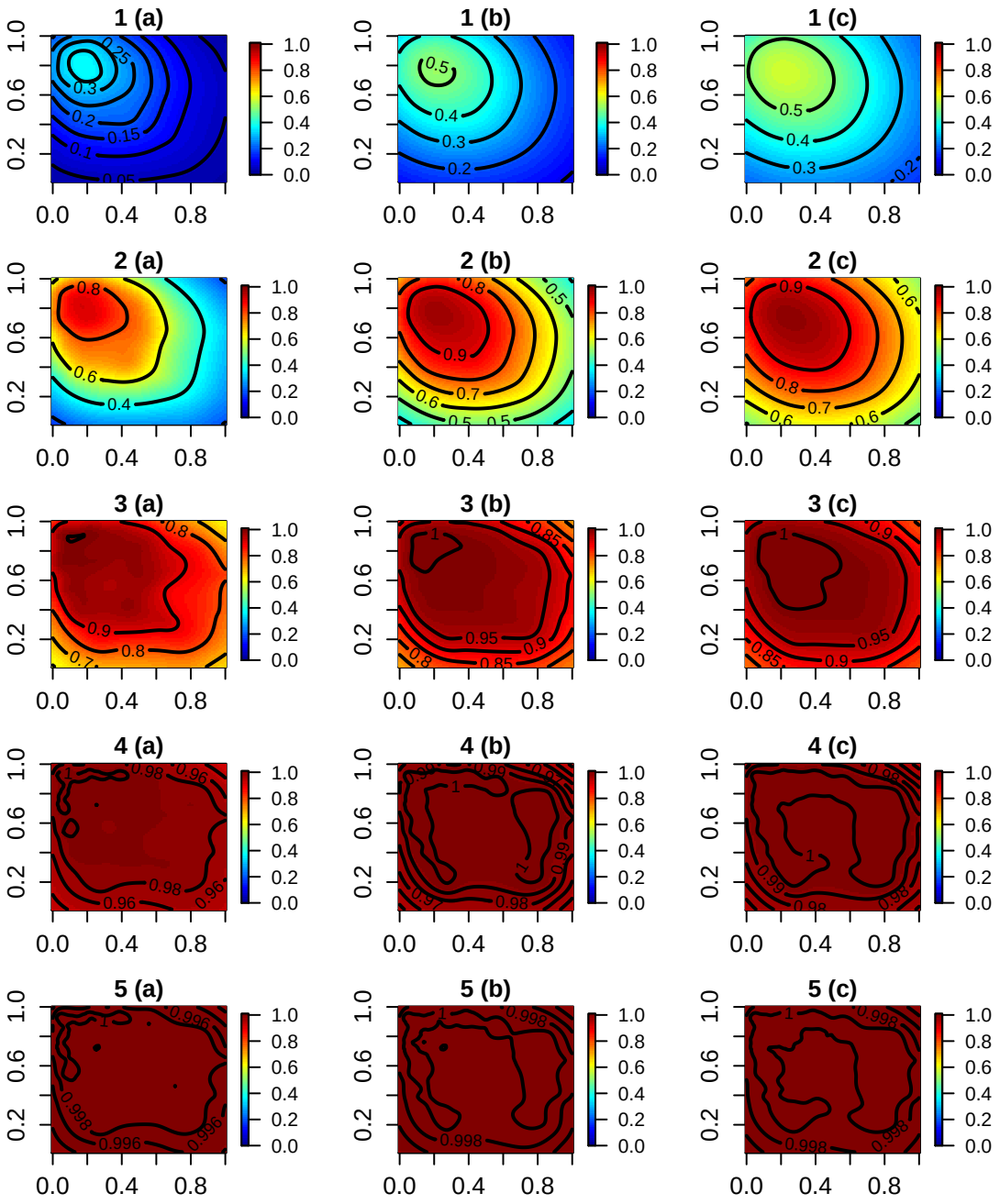

Figure 5.3: Interpolated surfaces of weights for the PCP for 15 combinations of variance ratio δ_0 and effective range d_0. Panels are given an alpha-numeric label. Numbers refer to the five values of $\delta_0 = 0.01, 0.1, 1, 10, 100$. Letters (a), (b) and (c) refer to three values of $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$.

where $\boldsymbol{D} = \text{diag}(\boldsymbol{x})$ and $\boldsymbol{x} = (x(\boldsymbol{s}_1), \ldots, x(\boldsymbol{s}_n))'$ contains the values of a known spatially referenced covariate. We have a global slope, hence $\boldsymbol{\theta} = \theta_1$, and a partially centered spatial process $\boldsymbol{\beta}^w = \boldsymbol{\beta}_1^w$. Model (5.19) can be retrieved from model (5.7) by letting $\boldsymbol{X}_1 = \boldsymbol{D}$, $\boldsymbol{C}_1 = \sigma_\epsilon^2 \boldsymbol{I}$, $\boldsymbol{X}_2 = \boldsymbol{1}$, $\boldsymbol{C}_2 = \sigma_1^2 \boldsymbol{R}_1$, $\boldsymbol{m} = m_1$ and $\boldsymbol{C}_3 = \sigma_1^2 v_1$.

For model (5.19) the $\boldsymbol{W}$ matrix is given by

$$\boldsymbol{W} = \sigma_1^2 \boldsymbol{R}_1 \boldsymbol{D} (\sigma_\epsilon^2 \boldsymbol{I} + \sigma_1^2 \boldsymbol{D} \boldsymbol{R}_1 \boldsymbol{D})^{-1} \boldsymbol{D}.$$

As there is one global parameter in the model, θ_1, $\boldsymbol{X}_2^{opt} = \boldsymbol{W}\mathbf{1}$ is a vector whose ith entry represents the optimum weight of partial centering for $\beta_1^w(\boldsymbol{s}_i)$. To see how these weights vary across the domain we randomly select 200 points uniformly over the unit square, see Figure 5.4. We generate the values of $\boldsymbol{x}$ by selecting a point $\boldsymbol{s}_x$, which we may imagine to

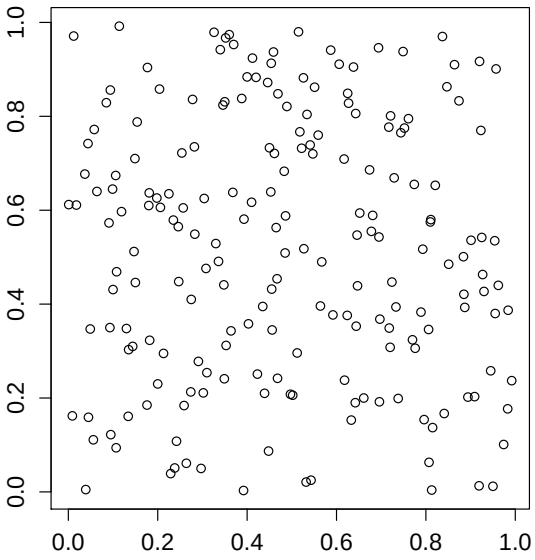

Figure 5.4: 200 randomly selcted locations within the unit square.

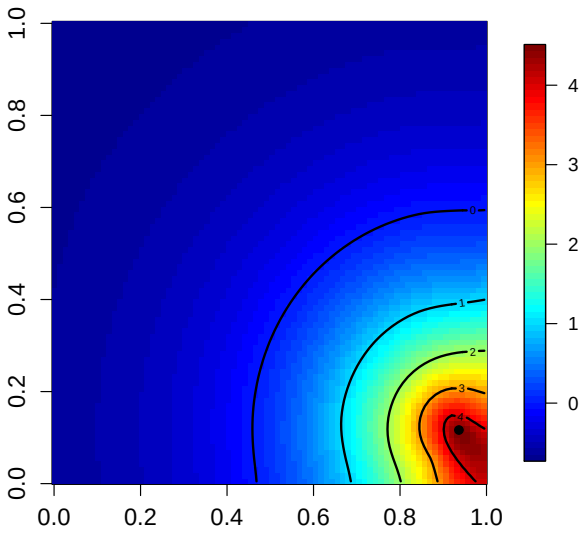

Figure 5.5: Interpolated surface of $\boldsymbol{x}$ for the uniformly sampled data locations given in Figure 5.4.

be the site of a source of pollution. We assume that the value for the observed covariate

at site $\boldsymbol{s}_i$ decays exponentially at rate ϕ_x with increasing separation from $\boldsymbol{s}_x$, so that

$$x(\boldsymbol{s}_i) = \exp\{-\phi_x \|\boldsymbol{s}_i - \boldsymbol{s}_x\|\}, \quad i = 1, \ldots, n.$$

The spatial decay parameter ϕ_x is chosen such that there is an effective spatial range of $\sqrt{2}/2$, i.e. if $\|\boldsymbol{s}_i - \boldsymbol{s}_x\| = \sqrt{2}/2$ then $x(\boldsymbol{s}_i) = 0.05$. The values of $\boldsymbol{x}$ are standardised by subtracting their sample mean and dividing by their sample standard deviation. Figure 5.5 gives the interpolated covariate surface where $\boldsymbol{s}_x = (0.936, 0.117)'$. We can see how the values decay with increased separation from $\boldsymbol{s}_x$. As in Section 5.3.2, interpolation is carried out using the `Tps` (thin plate spline) function in the R package fields (Furrer et al., 2009).

The optimal weights are computed for 15 combinations of variance ratio δ_1 and effective range d_1, where $\delta_1 = \sigma_1^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$ and $d_1 = \sqrt{2}/3$, $2\sqrt{2}/3$,$\sqrt{2}$. The weights are interpolated and plotted in Figure 5.6. The layout is the same as Figure 5.3, with each row corresponding to a value of δ_1, going from 0.01 at the top to 100 at the bottom, and increasing effective range from left to right.

As in Section 5.3.2 we see that the weights increase with increasing δ_1 or d_1. It is also clear that for locations near $\boldsymbol{s}_x$, where the values of the covariate are greatest, the optimum weights are greatest.

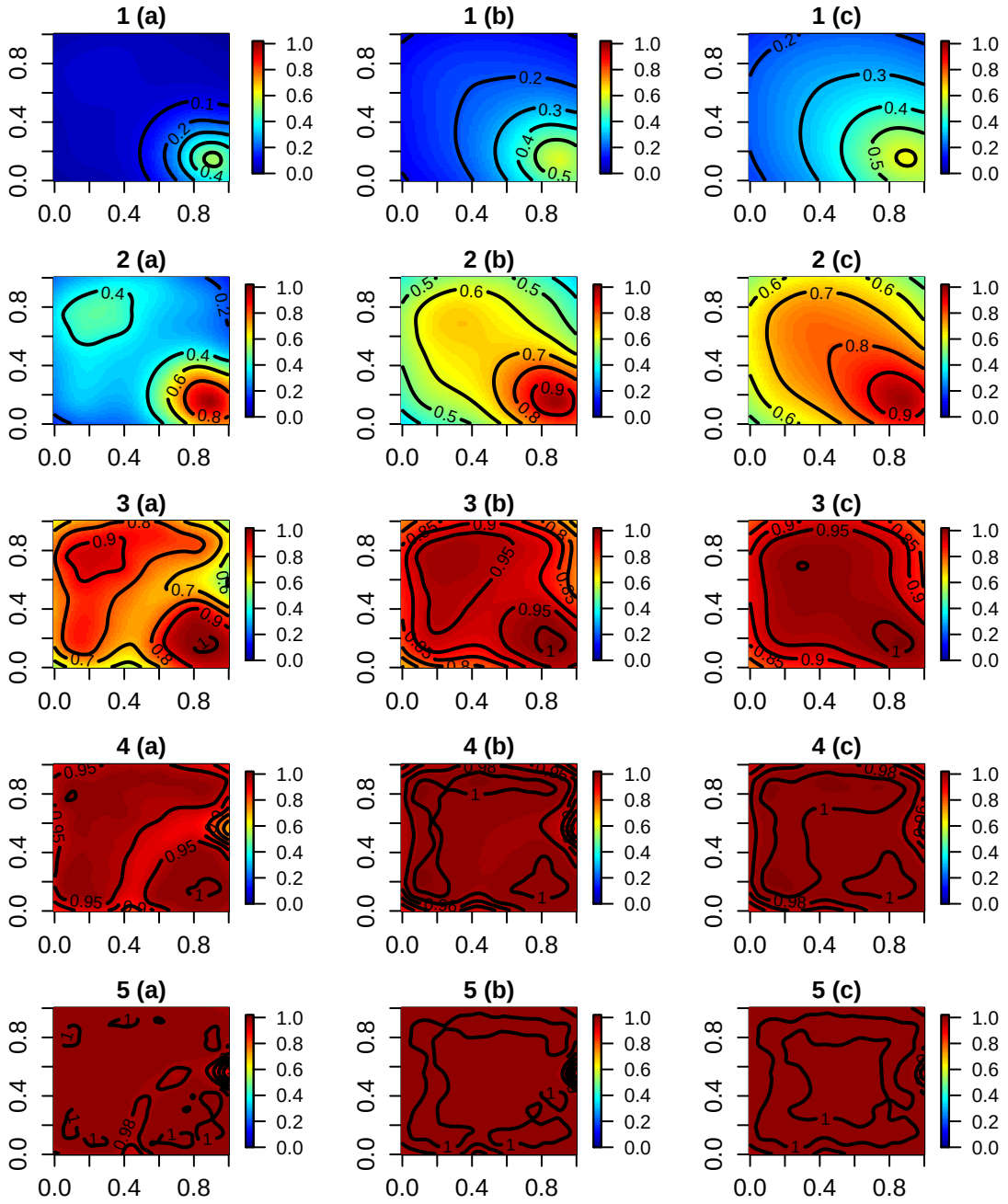

Figure 5.6: Interpolated surfaces of weights for the PCP for 15 combinations of variance ratio δ_0 and effective range d_1. Panels are given an alpha-numeric label. Numbers refer to the five values of $\delta_1 = 0.01, 0.1, 1, 10, 100$. Letters (a), (b) and (c) refer to three values of $d_1 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$.

5.4 Gibbs sampling for the PCP

In this section we investigate the performance of a Gibbs sampler using the PCP constructed as described in Section 5.2. We begin by outlining the joint posterior and full conditional distributions needed for the PCP. We then demonstrate that by dynamically updating $\boldsymbol{W}$ the stationary distribution of the Markov chain is not disturbed. Simulated data is used to investigate the performance of the PCP, first by assuming that the variance parameters are known and then relaxing this assumption. We go on to consider schemes that mitigate the computational cost of the PCP, so called pilot adaption schemes.

5.4.1 Joint posterior and full conditional distributions of the PCP

We begin here by writing down the joint posterior distribution of the parameters in model (5.7). We let $\boldsymbol{\xi} = (\boldsymbol{\beta}^{w\prime}, \boldsymbol{\theta}', \boldsymbol{\sigma^2}', \sigma_\epsilon^2, \boldsymbol{\phi}')'$ be the vector containing all np partially centred random effects, p global effects, p random effect variances, the data variance and p decay parameters for the correlation functions. The joint posterior for $\boldsymbol{\xi}$ is given by

$$
\begin{aligned}
\pi(\boldsymbol{\xi}|\boldsymbol{y}) \quad &\propto \quad \pi(\boldsymbol{Y}|\boldsymbol{\beta}^w, \boldsymbol{\theta}, \sigma_\epsilon^2)\pi(\boldsymbol{\beta}^w|\boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\sigma^2})\pi(\boldsymbol{\sigma^2})\pi(\sigma_\epsilon^2)\pi(\boldsymbol{\phi}) \\
&\propto \quad \prod_{k=0}^{p-1} \left(\sigma_k^2\right)^{-(n/2+1/2+a_k+1)} |\boldsymbol{R}_k|^{-1/2} \left(\sigma_\epsilon^2\right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp\left\{ -\frac{1}{2\sigma_\epsilon^2}\left[\Big(\boldsymbol{Y} - \boldsymbol{X}_1\big(\boldsymbol{\beta}^w + (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\big)\Big)'\Big(\boldsymbol{Y} - \boldsymbol{X}_1\big(\boldsymbol{\beta}^w + \right.\right. \right. \\
&\quad \left.\left.\left. (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\big)\Big) + 2b_\epsilon\right]\right\} \exp\left\{ -\frac{1}{2}\Big(\boldsymbol{\beta}^w - \boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\Big)'\boldsymbol{C}_2^{-1}\Big(\boldsymbol{\beta}^w - \boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\Big)\right\} \\
&\quad \exp\left\{ -\frac{1}{2}\sum_{k=0}^{p-1}\frac{1}{\sigma_k^2}\left(\frac{(\theta_k - m_k)^2}{v_k} + 2b_k\right)\right\} \prod_{k=0}^{p-1}\pi(\phi_k),
\end{aligned}
$$

where a description of the prior distributions $\pi(\boldsymbol{\sigma^2})$, $\pi(\sigma_\epsilon^2)$ and $\pi(\boldsymbol{\phi})$ can be found in Section 4.2.1.

It is argued in Section 5.2 that we must jointly update the $\boldsymbol{\beta}^w$'s and jointly update $\boldsymbol{\theta}$ and this is reflected in the conditional distributions given below.

- The full conditional distribution of $\boldsymbol{\beta}^w$ is

$$\boldsymbol{\beta}^w|\boldsymbol{\theta}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2, \boldsymbol{\phi}, \boldsymbol{y} \sim N\left(\boldsymbol{m}_\beta^*, \boldsymbol{C}_2^*\right),$$

  where

$$
\begin{aligned}
\boldsymbol{C}_2^* &= \left(\sigma_\epsilon^{-2}\boldsymbol{X}_1'\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1}, \\
\boldsymbol{m}_\beta^* &= \boldsymbol{C}_2^*\left(\sigma_\epsilon^{-2}(\boldsymbol{y} - \boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}) + \boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\right).
\end{aligned}
$$

- The full conditional distribution of $\boldsymbol{\theta}$ is

$$\boldsymbol{\theta}|\boldsymbol{\beta}^w, \boldsymbol{\sigma^2}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \sim N\left(\boldsymbol{m}^*_\theta, \boldsymbol{C}^*_3\right),$$

where

$$\begin{aligned}
\boldsymbol{C}^*_3 &= \left(\sigma^{-2}_\epsilon (\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2)'(\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2) + (\boldsymbol{W}\boldsymbol{X}_2)'\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1}, \\
\boldsymbol{m}^*_\theta &= \boldsymbol{C}^*_3\left(\sigma^{-2}_\epsilon (\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2)'(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}^w) + (\boldsymbol{W}\boldsymbol{X}_2)'\boldsymbol{C}_2^{-1}\boldsymbol{\beta}^w + \boldsymbol{C}_3^{-1}\boldsymbol{m}\right).
\end{aligned}$$

- The full conditional distribution of σ^2_k, $k = 0, \ldots, p-1$, is

$$\begin{aligned}
\sigma^2_k|\boldsymbol{\beta}^w, \boldsymbol{\theta}, \boldsymbol{\sigma^2}_{-k}, \sigma^2_\epsilon, \boldsymbol{\phi}, \boldsymbol{y} \quad &\sim \quad IG\left\{\frac{n+1}{2} + a_k, \right. \\
&\left. \frac{1}{2}\left[\left(\boldsymbol{\beta}^w_k - \sum_{m=0}^{p-1}\boldsymbol{W}_{km}\theta_k\mathbf{1}\right)'\boldsymbol{R}_k^{-1}\left(\boldsymbol{\beta}^w_k - \sum_{m=0}^{p-1}\boldsymbol{W}_{km}\theta_k\mathbf{1}\right) + \frac{(\theta_k - m_k)^2}{v_k} + 2b_k\right]\right\},
\end{aligned}$$

where $\boldsymbol{W}_{km}$ denotes the kmth, $n \times n$ block of $\boldsymbol{W}$.

- The full conditional distribution of σ^2_ϵ is

$$\begin{aligned}
\sigma^2_\epsilon|\boldsymbol{\beta}^w, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}, \boldsymbol{y} \quad &\sim \quad IG\left\{\frac{n}{2} + a_\epsilon, \right. \\
&\left. \frac{1}{2}\left[(\boldsymbol{Y} - \boldsymbol{X}_1(\boldsymbol{\beta}^w + (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta})'(\boldsymbol{Y} - \boldsymbol{X}_1(\boldsymbol{\beta}^w + (\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}) + 2b_\epsilon\right]\right\}.
\end{aligned}$$

- The full conditional distribution of ϕ_k is

$$\begin{aligned}
\pi(\phi_k|\boldsymbol{\beta}^w, \boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi}_{-k}, \boldsymbol{y}) \quad &\propto \quad |R_k|^{-1/2} \\
&\exp\left\{-\frac{1}{2\sigma^2_k}\left(\boldsymbol{\beta}^w_k - \sum_{m=0}^{p-1}\boldsymbol{W}_{km}\theta_k\mathbf{1}\right)'\boldsymbol{R}_k^{-1}\left(\boldsymbol{\beta}^w_k - \sum_{m=0}^{p-1}\boldsymbol{W}_{km}\theta_k\mathbf{1}\right)\right\},
\end{aligned}$$

for $l_k < \phi_k < u_k$, zero otherwise.

5.4.2 Dynamically updating the PCP

The PCP relies on the $\boldsymbol{W}$ matrix which, by construction, removes the posterior correlation between $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$. However, the expression for $Cov(\boldsymbol{\beta}^w, \boldsymbol{\theta}|\boldsymbol{y})$ given in (5.10) that leads to the derivation of $\boldsymbol{W}$ is conditional on the covariance matrices, $\boldsymbol{C}_1$, $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$. Therefore when the variance and decay parameters are unknown how do we compute $\boldsymbol{W}$? We propose a dynamically updated parameterisation that uses the most recent values to re-compute $\boldsymbol{W}$ at each move of the Markov chain along a coordinate of which it is a function. As we will see it is essential that $\boldsymbol{W}$ is updated each time a parameter it depends upon is updated, and not just at the end of a complete pass of the sampler through all of the model parameters.

We must ensure that by dynamically updating $\boldsymbol{W}$ we do not disturb the stationary distribution of the Markov chains generated from the Gibbs sampler. To demonstrate that stationarity is preserved we let $p = 1$ in model (5.7), therefore we have the version of the model given in (5.15). If

$$\boldsymbol{\xi}^{(t)} = \left(\boldsymbol{\beta}_0^{w(t)}, \theta_0^{(t)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}\right)'$$

is the state of the Markov chain after t iterations, we must show that

$$\pi(\boldsymbol{\xi}^{(t+1)}|\boldsymbol{y}) = \int P(\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)})\pi(\boldsymbol{\xi}^{(t)}|\boldsymbol{y})d\boldsymbol{\xi}^{(t)}, \tag{5.20}$$

where $P(\cdot|\cdot)$ is the transition kernel of the chain. Given $\boldsymbol{\xi}^{(t)}$, we obtain a new sample, $\boldsymbol{\xi}^{(t+1)}$, from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ as follows:

1. Sample $\boldsymbol{\beta}^{w(t+1)} \sim \pi(\boldsymbol{\beta}^w|\theta_0^{(t)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})$.

2. Sample $\theta_0^{(t+1)} \sim \pi(\theta_0|\boldsymbol{\beta}_0^{w(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})$.

3. Sample $\sigma_0^{2(t+1)} \sim \pi(\sigma_0^2|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})$.

4. Sample $\sigma_\epsilon^{2(t+1)} \sim \pi(\sigma_\epsilon^2|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})$.

5. Sample $\phi^{(t+1)} \sim \pi(\phi_0|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t+1)}, \phi_0^{(t)}, \boldsymbol{y})$.

Note that the full conditional distributions of σ_0^2, σ_ϵ^2 and ϕ_0 are conditional on their respective current values through $\boldsymbol{W}$, i.e. $\sigma_0^{2(t+1)}$ is conditioned on $\sigma_0^{2(t)}$.

The transition kernel of the Markov chain is

$$\begin{aligned} P(\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)}) \;=\; & \pi(\boldsymbol{\beta}^{w(t+1)}|\theta_0^{(t)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})\pi(\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y}) \\ & \pi(\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y}) \\ & \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y}) \\ & \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t+1)}, \phi_0^{(t)}, \boldsymbol{y}), \end{aligned}$$

and it follows that

$$
\begin{aligned}
&\int P(\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)})\pi(\boldsymbol{\xi}^{(t)}|\boldsymbol{y})d\boldsymbol{\xi}^{(t)} \\
&= \int \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \pi(\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y})\pi(\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \underbrace{\pi(\boldsymbol{\beta}^{w(t+1)}|\theta_0^{(t)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y})\left(\int \pi(\boldsymbol{\beta}_0^{w(t)},\theta_0^{(t)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})d\boldsymbol{\beta}_0^{w(t)}\right)}_{=\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})} \\
&\quad d\theta_0^{(t)}d\sigma_0^{2(t)}d\sigma_\epsilon^{2(t)}d\phi_0^{(t)} \\
&= \int \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \pi(\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \underbrace{\pi(\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y})\left(\int \pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})d\theta_0^{(t)}\right)}_{=\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})} \\
&\quad d\sigma_0^{2(t)}d\sigma_\epsilon^{2(t)}d\phi_0^{(t)} \\
&= \int \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \underbrace{\left(\int \pi(\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y})\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})d\sigma_0^{2(t)}\right)}_{=\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})} \\
&\quad d\sigma_\epsilon^{2(t)}d\phi_0^{(t)} \\
&= \int \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)},\boldsymbol{y}) \\
&\quad \underbrace{\left(\int \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)},\boldsymbol{y})\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\phi_0^{(t)}|\boldsymbol{y})d\sigma_\epsilon^{2(t)}\right)}_{=\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)}|\boldsymbol{y})} \\
&\quad d\phi_0^{(t)} \qquad (5.21) \\
&= \int \pi(\phi_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)},\boldsymbol{y})\pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t)}|\boldsymbol{y}) \\
&\quad d\phi_0^{(t)} \\
&= \pi(\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)},\phi_0^{(t+1)}|\boldsymbol{y}), \\
&= \pi(\boldsymbol{\xi}^{(t+1)}|\boldsymbol{y}),
\end{aligned}
$$

and hence stationarity is preserved.

The above argument can easily be extended for $p > 1$, and in which it must be noted that in Section 5.2 it is argued that the desirable properties of the PCP, that of posterior uncorrelatedness and immediate convergence, can only be achieved if $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$ are each updated in one block.

Notice that if we update $\boldsymbol{W}$ and the end of each complete pass of the sampler then the

stationarity condition (5.20) does not hold. For instance, consider σ_ϵ^2, which is conditioned on σ_0^2 through $\boldsymbol{W}$, see Section 5.4.1. If $\boldsymbol{W}$ is not recalculated using $\sigma_0^{2(t+1)}$ then $\sigma_\epsilon^{2(t+1)}$ is conditioned and $\sigma_0^{2(t)}$, and consequently

$$\int \pi(\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}, \boldsymbol{y})\pi(\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t)}, \phi_0^{(t)}|\boldsymbol{y})d\sigma_\epsilon^{2(t)}$$
$$\neq \pi(\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t+1)}, \phi_0^{(t)}|\boldsymbol{y}),$$

but equality is required to complete step (5.21) in the string of equalities proving stationarity.

5.4.3 Performance of the PCP for known variance parameters

In this section we investigate the convergence and mixing properties of a Gibbs sampler employing the PCP. As in the analysis of Chapter 4 we assess performance of the sampler in terms of the number of iterations required for the (multivariate) potential scale reduction factor to fall below 1.1, [(M)PSRF$_M$(1.1)] and the effective sample size (ESS) of the model parameters, see Section 2.3. We focus on how these measures are influenced by the ratio of the variance parameters and the correlation structure between the random effects.

We again look at model (5.15) and assume an exponential correlation function between realisations of the Gaussian process used to model the spatial surface, see (5.18). We generate data from this model by first selecting $n = 40$ points in the unit square. These are given in Figure 4.1. The global mean is fixed at $\theta_0 = 0$ and we let hyperparameters $m_0 = 0$ and $v_0 = 10^4$. We use five levels of variance ratio δ_0, 0.01, 0.1, 1, 10 and 100, and four levels for the effective range d_0, which are 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$, hence we have 20 variance ratio-effective range combinations. For each of these 20 data sets are generated and so we have a total of 400 data sets

Variance parameters, σ_0^2 and σ_ϵ^2, and the decay parameter $\phi_0 = -\log(0.05)/d_0$ are held fixed at their true values and so for each iteration of the Gibbs sampler we generate samples from the full conditional distributions of $\boldsymbol{\beta}_0^w$ and θ_0.

For each data set we generate five chains of length 25,000 using starting values for θ_0 that are outside the range of values described by pilot chains. We then use the output to compute the PSRF$_M$(1.1) and the ESS for θ_0. The results are plotted in Figure 5.7. On the top row we have the PSRF$_M$(1.1) and on the bottom the ESS. Each of the five panels in each row corresponds to a value of δ_0, with 0.01 on the left rising to 100 on the right. Within each panel we have four boxplots, one for each level of the effective range, again rising from left to right. Each boxplot consists of 20 values, for the 20 repetitions of that variance ratio-effective range combination.

The analysis of Section 4.4 is conducted using the same data and starting values and used here. For comparison we can look at the results for the CP and NCP for known variance parameters, given in Figures 4.2 and 4.3 respectively. The CP performs well for higher values of δ_0 and d_0, the opposite being true for the NCP. Figure 5.7 shows that for the PCP with known variance parameters we achieve near immediate convergence and independent samples for θ_0 in all cases, and that it is robust to changes in both variance

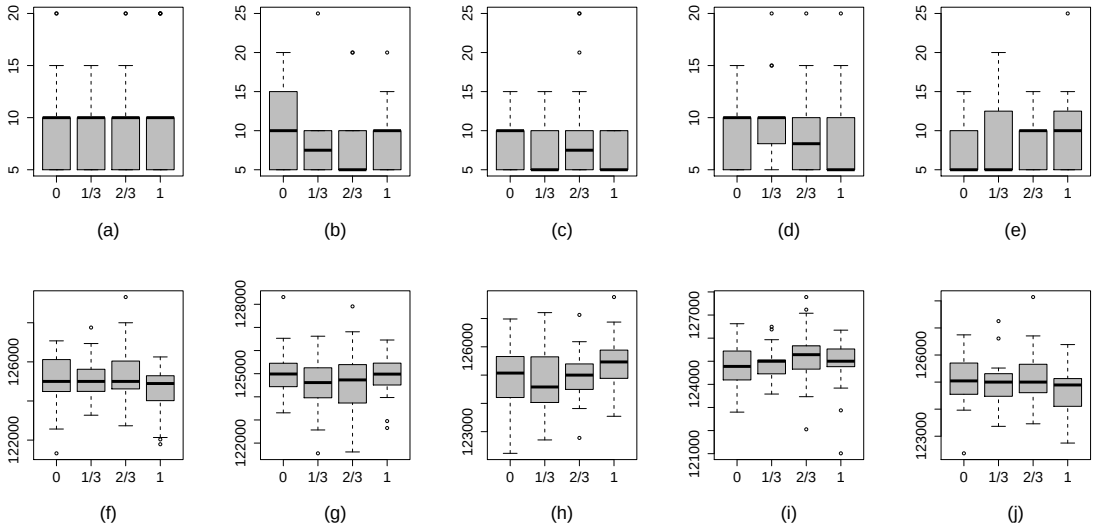

Figure 5.7: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the PCP of the Gaussian model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

ratio and strength of correlation.

5.4.4 Performance of the PCP for unknown variance parameters

In this section we remove the assumption that the variance parameters are known but we still the fix the decay parameter ϕ_0 at its true value. We repeat the analysis of 5.4.3 but now we sample from the full conditional distributions of σ_0^2 and σ_ϵ^2.

Now that we are sampling the variance parameters as well as global mean θ_0, convergence is assessed through the multivariate PSRF. Figure 5.8 shows the $\text{MPSRF}_M(1.1)$ (top row) and the ESS of θ_0 (bottom row) for the 20 combinations of δ_0 and d_0 detailed in Section 4.4.1. There is more variability in the results seen here for the $\text{MPSRF}_M(1.1)$ than we saw for the $\text{PSRF}_M(1.1)$ in Section 5.4.3. When the random effects are independent, weak identifiability of the variance parameters can effect the performance of the sampler. However, the robustness to changes in δ_0 remains and we still see rapid convergence in most cases. The ESS for θ_0 remains high, with a median value above 120,000 for all of the 20 combinations of δ_0 and d_0.

Boxplots of the ESS of the variance parameters are given in Figure 5.9, with the results for σ_0^2 on the top row and σ_ϵ^2 on the bottom row. There is a suggestion that increasing δ_0 increases the ESS of σ_0^2 and decreases the ESS of σ_ϵ^2. For a fixed value of δ_0 we can see that the ESS of both variance parameters increases as the effective range increases. The stronger correlation across the random effects means that the variability seen in the data can be more easily separated between the two components.

We compare the results for the PCP with those obtained for the CP and the NCP by calculating the mean responses of each measure of performance for each of the 20 variance ratio-effective range combinations. Table 5.1 shows the mean $\text{MPSRF}_M(1.1)$ and mean

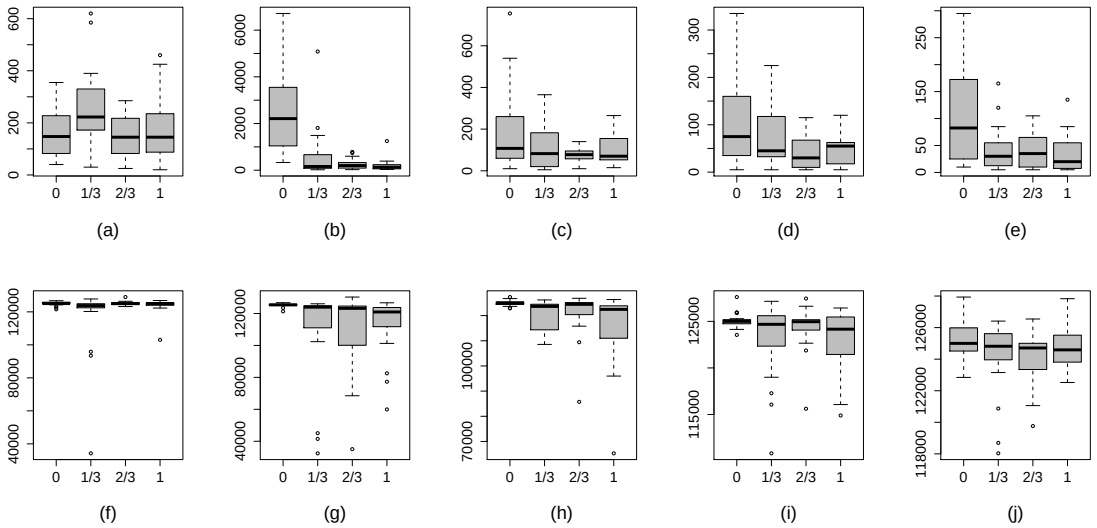

Figure 5.8: MPSRF$_M(1.1)$ and the ESS of θ_0 for the PCP of the Gaussian model with unknown variance parameters Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

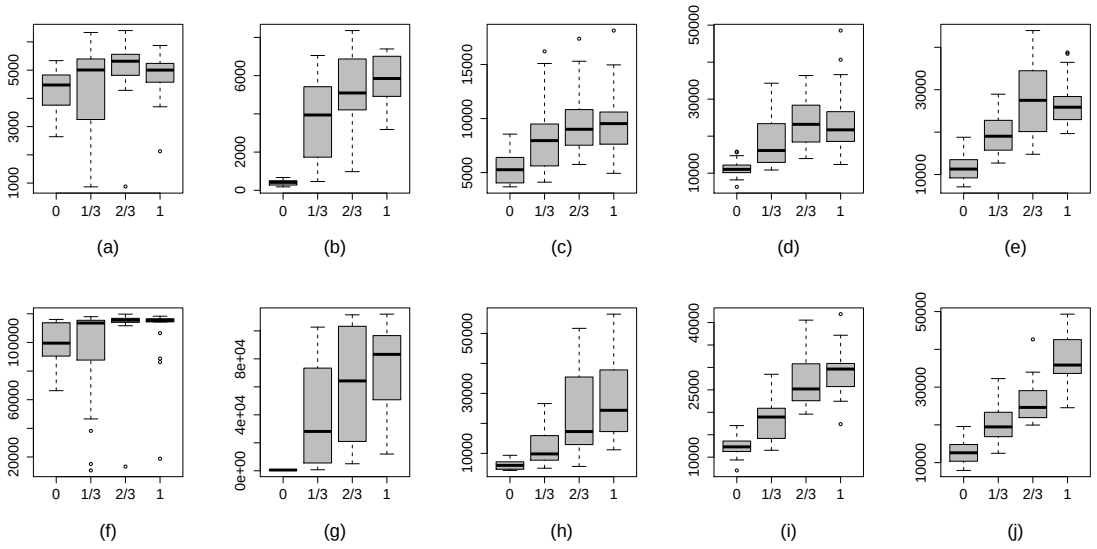

Figure 5.9: ESS of σ_0^2 and σ_ϵ^2 for the PCP of the Gaussian model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

ESS for θ_0. We see that the PCP has a lower average MPSRF$_M(1.1)$ for most cases, and when it does not, the difference is less than 3%. We can also see that, in terms of the ESS of θ_0, it is clear that the PCP is superior to the CP and the NCP in all cases.

A similar comparison for the mean ESS of the variance parameters is given in Table 5.2. The PCP does not always deliver the highest ESS for the variance parameters, but for σ_0^2 it is always within 1% of the ESS for the parameterisation that does return the highest value, and for σ_ϵ^2 it is always within 2%.

Table 5.1: Means of the $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for 20 variance ratio-effective range combinations for the CP, the NCP and the PCP.

		$\text{MPSRF}_M(1.1)$			ESS of θ_0		
δ_0	$d_0/\sqrt{2}$	CP	NCP	PCP	CP	NCP	PCP
0.01	0	3064.50	172.00	**163.25**	463	108819	**124988**
	1/3	1115.75	278.25	**251.25**	1821	103342	**116659**
	2/3	544.50	166.25	**154.00**	3055	105397	**125108**
	1	366.50	184.75	**175.00**	4652	94657	**123730**
0.1	0	3528.25	3455.50	**2464.50**	4707	20420	**125272**
	1/3	**607.00**	1305.50	624.25	13336	33347	**108922**
	2/3	271.25	592.75	**251.50**	25884	28959	**109987**
	1	211.00	518.25	**203.75**	31938	22361	**113191**
1	0	274.50	910.50	**187.50**	25523	7353	**124945**
	1/3	134.50	1639.00	**118.75**	65927	3785	**120325**
	2/3	78.25	2092.00	**76.00**	82100	3148	**121013**
	1	**101.00**	2473.75	103.00	84742	2722	**115700**
10	0	123.50	1140.25	**105.75**	32578	5226	**125091**
	1/3	**79.75**	1556.25	**79.75**	83586	2918	**123055**
	2/3	45.25	2824.25	**44.50**	102306	1734	**124341**
	1	**49.50**	3542.25	50.50	107261	1177	**122774**
100	0	104.75	1124.75	**102.75**	32891	4596	**125388**
	1/3	42.75	1607.50	**42.50**	84755	2772	**124120**
	2/3	41.75	2544.75	**41.50**	104941	1671	**124214**
	1	**32.75**	3427.75	33.00	108050	1186	**124670**

Table 5.2: Means of the ESS of σ_0^2 and ESS of σ_ϵ^2 for 20 variance ratio-effective range combinations for the CP, the NCP and the PCP.

		ESS of σ_0^2			ESS of σ_ϵ^2		
δ_0	$d_0/\sqrt{2}$	CP	NCP	PCP	CP	NCP	PCP
0.01	0	4138	4078	**4244**	27753	95416	**98456**
	1/3	**4312**	4268	4295	56647	**95916**	94741
	2/3	5061	**5071**	5044	81338	**111550**	110636
	1	**4804**	4737	4758	88184	**108623**	107762
0.1	0	383	391	**397**	532	505	**545**
	1/3	3696	3027	**3719**	36659	30190	**39819**
	2/3	5368	4873	**5406**	60338	58930	**62573**
	1	5784	5095	**5790**	72669	70080	**73138**
1	0	5527	3868	**5557**	6188	6208	**6321**
	1/3	**8333**	3557	8297	12400	11488	**12412**
	2/3	**9430**	3945	**9430**	**22846**	18310	22828
	1	**9714**	3712	9711	**28853**	23222	28770
10	0	11270	3945	**11352**	12216	12308	**12495**
	1/3	**18435**	3543	18413	**18226**	18009	18167
	2/3	**23692**	3699	23632	**26945**	26373	26825
	1	**24595**	4015	24572	**29228**	28790	29177
100	0	11423	3652	**11457**	12427	12495	**12690**
	1/3	**19708**	3559	19668	**20118**	19930	20085
	2/3	**27706**	3929	27649	**26197**	25924	26120
	1	**26937**	3012	26936	**37121**	36499	37087

5.4.5 Pilot adaption schemes

The limitation of the dynamically updated PCP is the requirement to compute $\boldsymbol{W}$ multiple times at each iteration of the Gibbs sampler, see Section 5.4.2. In this section we look at a method of mitigating the computational cost. We propose a strategy that runs the sampler until the MPSRF reaches 1.1, and then a further number of iterations M^* to obtain an estimate for $\boldsymbol{W}$, which is fixed thereafter. Therefore, for iterations $t = \text{MPSRF}_M(1.1) + M^* + 1, \ldots, N$, where N is the total length of the chain, we estimate $\boldsymbol{W}$ by, $\widehat{\boldsymbol{W}} = \boldsymbol{W}(\widehat{\boldsymbol{\sigma^2}}, \widehat{\sigma}^2_\epsilon, \widehat{\boldsymbol{\phi}})$, where

$$\widehat{\sigma}^2_\epsilon = \frac{1}{M^*} \sum_{t=\text{MPSRF}_M(1.1)+1}^{\text{MPSRF}_M(1.1)+M^*} \sigma^{2(t)}_\epsilon,$$

with equivalent definitions for σ^2_k and ϕ_k for $k = 0, \ldots, p-1$. We refer to these fitting strategies as *pilot adapted partially centred parameterisations* (PAPCPs).

Table 5.3: Means of ESS/s of θ_0, σ^2_0 and σ^2_ϵ for 20 variance ratio-effective range combinations for the PAPCP with $M^* = 1000, 2000, 3000, 4000, 5000$.

		ESS/s θ_0					ESS/s σ^2_0					ESS/s σ^2_ϵ				
δ_0	$d_0/\sqrt{2}$	1000	2000	3000	4000	5000	1000	2000	3000	4000	5000	1000	2000	3000	4000	5000
0.01	0	**9069**	8033	7244	6593	6022	**338**	299	270	247	226	**7902**	6995	6307	5790	5279
	1/3	**8917**	7964	7142	6501	5903	**354**	314	283	257	235	**7929**	7025	6322	5747	5229
	2/3	**9408**	8411	7596	6902	6328	**421**	374	337	306	280	**9224**	8181	7372	6687	6120
	1	**8799**	7816	7004	6409	5980	**397**	352	317	288	263	**8965**	7942	7177	6509	5949
0.1	0	**1996**	1564	1346	1328	1130	**55**	50	45	39	35	395	**421**	406	414	333
	1/3	**5221**	4765	4322	4062	3694	**328**	283	258	233	211	**3268**	2990	2756	2536	2301
	2/3	**4066**	3736	3487	3244	2951	**483**	427	383	343	314	**4098**	3558	3556	3292	3158
	1	**3757**	3460	3020	2768	2605	**511**	453	405	367	337	**4582**	4260	3556	3415	3147
1	0	**1879**	1727	1470	1392	1344	**647**	543	476	437	380	**961**	849	748	684	605
	1/3	**3906**	3357	2977	2655	2380	**809**	731	632	566	523	**1124**	1007	888	804	733
	2/3	**5528**	4819	4238	3660	3370	**937**	801	715	642	589	**2005**	1760	1576	1410	1302
	1	**5965**	5098	4465	4133	3536	**961**	825	734	664	595	**2577**	2262	2025	1860	1670
10	0	**3433**	2880	2776	2346	2385	**1012**	947	812	755	662	**1123**	1024	895	836	740
	1/3	**5751**	4882	4374	4085	3467	**1724**	1522	1367	1227	1126	**1562**	1381	1248	1134	1038
	2/3	**5864**	4766	4040	3721	3086	**2341**	1940	1787	1552	1400	**2357**	2058	1847	1670	1533
	1	**6744**	5299	4481	4153	3676	**2369**	2078	1826	1670	1451	**2563**	2267	2040	1857	1682
100	0	**3554**	3292	2622	2233	2519	**1037**	915	820	748	669	**1111**	997	905	827	743
	1/3	**6065**	5308	4681	4298	3851	**1887**	1671	1476	1300	1191	**1723**	1533	1371	1242	1135
	2/3	**6985**	5873	5176	4463	3994	**2641**	2274	2023	1859	1650	**2271**	1996	1799	1639	1499
	1	**7794**	6455	5693	5218	4720	**2499**	2197	1921	1743	1600	**3136**	2780	2494	2271	2071

Here we look at a simulation example to investigate the impact of different values of M^*. We wish to investigate the trade-off between the additional computation time of larger values of M^* and the improvement in performance gained by obtaining a more accurate estimate of $\boldsymbol{W}$.

We let $p = 1$ in model (5.7) and so again we reduce the model to the form given in (5.15). We use the same data as in Sections 5.4.3 and 5.4.4. We fix ϕ_0 and sample from θ_0, σ_0^2 and σ_ϵ^2.

We consider five values of M^*: 1000, 2000, 3000, 4000 and 5000. We run five chains of length 25000 for each value of M^*. The same random seed is used each time and so the $\text{MPRSF}_M(1.1)$ is identical across the schemes. We compute the ESS and divide it by the total run time (in seconds) to obtain values for the ESS/s of θ_0, σ_0^2 and σ_ϵ^2. We then take the mean of the 20 values returned for data generated under the same values of δ_0 and d_0.

The results given in Table 5.3 show clearly that there is no advantage in terms of ESS/s for the model parameters of using a value of M^* larger than 1000. The ESS is little improved as M^* is increased and so the reduction in ESS/s is due almost entirely to the increased run times. Therefore in the following section, we take the PAPCP to be the scheme with $M^* = 1000$.

5.5 Californian ozone concentration data

In this section we apply the PCP and the PAPCP of model (5.7) with $p = 2$, to the Californian ozone concentration data introduced in Section 4.5. Given the results of Section 5.4.5 we take the PAPCP to be the scheme that updates $\boldsymbol{W}$ for $\text{MPSRF}_M(1.1)$+1000 iterations, i.e. $M^* = 1000$.

We compute the $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters for both the PCP and the PAPCP fitting strategies. Results are given in Table 5.4. For comparison we also include the results for the CP_1 and the CP_2 from Section 4.5.4, as these were the best of the centred and non-centred parameterisations. The subscript indicates the number of blocks used to update $\tilde{\boldsymbol{\beta}}$. For all four fitting strategies convergence is swift

Table 5.4: $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters.

	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS θ_1	ESS σ_0^2	ESS σ_1^2	ESS σ_ϵ^2
CP_1	135	103129	56718	29539	5271	16836
CP_2	120	89230	36170	27283	4977	17095
PCP	160	125000	121995	28348	5082	15028
PAPCP	160	124248	111495	29515	5221	15661

with the $\text{MPSRF}_M(1.1)$'s between 120−160 iterations. There is also little difference in the ESS of the variance parameters across the methods. The difference lies in the ESS of the global mean parameters. The PCP returns independent samples from the marginal posterior distribution of θ_0 and near independent samples for θ_1. The ESS's are only slightly reduced by using the pilot adaption scheme. In particular we see the partially centred methods are superior to the centred methods in the θ_1 coordinate.

Just as there is a penalty in terms of computational time for using CP_1 instead of CP_2 (see Section 4.5.4) the superior mixing of PCP comes at a cost. We have to update $\boldsymbol{\beta}^w$ as one block and also we must update $\boldsymbol{W}$ repeatedly for each iteration of the Gibbs sampler, consequently increasing run times. This increase is mitigated by the use of the PAPCP,

Table 5.5: $\text{MPSRF}_t(1.1)$ and ESS/s of the model parameters.

	$\text{MPSRF}_t(1.1)$	ESS/s θ_0	ESS/s θ_1	ESS/s σ_0^2	ESS/s σ_1^2	ESS/s σ_ϵ^2
CP_1	2.6	42.7	23.5	12.2	2.2	7.0
CP_2	0.7	119.8	48.6	36.6	6.7	22.9
PCP	5.4	29.6	28.9	6.7	1.2	3.6
PAPCP	3.1	52.1	46.8	12.4	2.2	6.6

but we still have to update all random effects at once.

Table 5.5 gives the time adjusted measures for each fitting strategy. The relatively short run times for the CP_2 give it the advantage over the other methods in terms of $\text{MPSRF}_t(1.1)$ and ESS/s. However, in order to retain the same number of effective samples the CP_2 will need a longer chain than the PAPCP, for example. This means that more data must be stored and handled. Ultimately the user must decide between an algorithm that can be run quickly and one that returns samples with lower autocorrelation.

5.6 Summary

In this chapter we have investigated the performance of a PCP for the spatially varying coefficients model. By minimising the posterior covariance of the random and global effects, we are able to parameterise the model in such a way that the convergence rate for the associated Gibbs sampler is zero. The construction is conditioned on the covariance matrices in the model. We have shown that the parameterisation can be updated dynamically within the Gibbs sampler for the case when these matrices are known only up to a set of covariance parameters which must be estimated.

The optimal weights of partial centering are shown to vary over the spatial domain, with higher weights given to locations where the data is more informative about the latent surface. Therefore, higher weights are found when the data precision is relatively high, or there is some clustering of locations. We also saw higher weights for locations where the value of the covariate was larger.

Our investigations show that the unlike the CP and the NCP, the performance of the PCP is robust to changes in the relative informativity of the data and the strength of correlation. Swift convergence and independent, or near independent samples from the posterior distributions of the mean parameters are achieved for all of the data sets we considered, whether it was simulated or real data.

The PCP requires us to update all of the random effects in one block and all of the global effects in another, hence it is a computationally intensive strategy. The pilot adaption schemes are shown to reduce the computational burden associated with the PCP while inheriting its desirable properties of fast convergence and good mixing.

Chapter 6

Different parameterisations of non-Gaussian spatial models

6.1 Introduction

In previous chapters we have confined ourselves to a Gaussian error structure for the data. In this chapter we investigate the effect of parameterisation on the efficiency of Gibbs samplers for non-Gaussian spatial models. These models are referred to as spatial GLMMs, see for example Diggle et al. (1998) and Christensen et al. (2006). We consider models of the form

$$Y(\boldsymbol{s}) \sim f(\cdot|Z(\boldsymbol{s})) \quad \text{where} \quad E[Y(\boldsymbol{s})|Z(\boldsymbol{s})] = l^{-1}(Z(\boldsymbol{s})), \tag{6.1}$$

for some link function l. We assume that $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n))'$ are conditionally independent given $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))'$ and that

$$\begin{aligned} \boldsymbol{Z}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned} \tag{6.2}$$

hence the latent processes are Gaussian, but the model for the data is not Gaussian in general.

Gelfand et al. (1996) find that for GLMMs, the centred parameterisation (CP) reduces the posterior correlation between the parameters describing the mean function when compared to the non-centred parameterisation (NCP), a result that extends their work on normal linear mixed models (Gelfand et al., 1995). Here we are concerned with models with latent spatial processes, and so we wish to discover how the presence of spatial correlation effects the efficiency of the Gibbs sampler under different model parameterisations.

Together, (6.1) and (6.2) describe the CP of the model, with the NCP found by letting $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}} - \boldsymbol{X}_2\boldsymbol{\theta}$. The partially centred parameterisation (PCP) is analogous to the one constructed in Chapter 5 for Gaussian likelihoods and is found by estimating the conditional posterior covariance matrix of the random effects.

Through simulated and real data examples we compare the efficiency of the model parameterisations. In the case of Gaussian likelihoods we have expressions for the exact convergence rate when the variance parameters are known. For non-Gaussian likelihoods we do not have an equivalent result, but we feel justified in comparing parameterisations in terms of the (M)PSRF and the ESS of the model parameters, given the close agreement between these measures and the exact convergence rate when the convergence rate is available.

We look at two widely employed models for non-Gaussian spatial data. We begin with a Tobit model for the data (Tobin, 1958). Developed to model non-negative economic data, Tobit models have been generalised to model data that is constrained in some way, see Amemiya (1984) for a detailed review. Here we apply the Tobit model to create a predictive map of the probability of positive precipitation over New York for the week beginning July 30, 2001. The second model we look at is the probit model for binary data. Binary spatial data arises in many disciplines, i.e. ecology, politics and biology. Here we use it to indicate whether the observed of ozone concentration exceeds a predetermined air pollution standard, and then create a map of the probability of exceedance.

The rest of this chapter is organised as follows: Sections 6.2 and 6.3 contain the full conditional distributions and compare the performance of the CP and the NCP with simulated and real data sets for the spatial Tobit and the spatial probit models respectively. Section 6.4 looks at the construction and performance of PCPs for non-Gaussian models and the chapter is concluded in Section 6.5 with some summary remarks.

6.2 Spatial Tobit model

Tobit regression models are usually applied to data that are truncated at zero. Reich et al. (2010) use a spatial Tobit model with spatially varying coefficients to model the activity of pregnant women across North Carolina. Berrocal et al. (2008) use a similar model to predict precipitation over the Pacific Northwest where they use the output of a numerical weather prediction model as a covariate. Spatial Tobit models are also widely used in econometrics, where it is common to model the latent spatial process as a Markov random field (Pace and LeSage, 2009).

We consider the following Tobit model with spatially varying coefficients

$$\begin{aligned} Y(\boldsymbol{s}_i) &= \max\{Z(\boldsymbol{s}_i), 0\} \\ Z(\boldsymbol{s}_i) &= \sum_{k=0}^{p-1} \{\theta_k + \beta_k(\boldsymbol{s}_i)\} x_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \end{aligned} \tag{6.3}$$

where $x_0(\boldsymbol{s}_i) = 1$ and $\epsilon(\boldsymbol{s}_i) \sim N(0, \sigma_\epsilon^2)$ independently for all $i = 1, \ldots, n$. We only observe $Z(\boldsymbol{s}_i)$ when it is non-negative and so we have a truncated normal distribution with a point mass at zero for observations $Y(\boldsymbol{s}_i)$. Covariate information for site $\boldsymbol{s}_i$ is contained within the $p \times 1$ vector $\boldsymbol{x}(\boldsymbol{s}_i) = (x_0(\boldsymbol{s}_i), \ldots, x_{p-1}(\boldsymbol{s}_i))'$ and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ is a $p \times 1$ vector of global regression coefficients. Elements of $\boldsymbol{\theta}$ are assumed to be independent *a priori* and each θ_k is assigned a normal prior distribution with mean m_k and variance $\sigma_k^2 v_k$, so that

$\theta_k \sim N(m_k, \sigma_k^2 v_k)$.

We denote by $\beta_k(\boldsymbol{s}_i)$ the realisation of the kth spatial process at site $\boldsymbol{s}_i$, which we model as a zero mean Gaussian process. Therefore, we take

$$\boldsymbol{\beta}_k = (\beta_k(\boldsymbol{s}_1), \ldots, \beta_k(\boldsymbol{s}_n))' \sim N(0, \sigma_k^2 \boldsymbol{R}_k),$$

The $\boldsymbol{\beta}_k$ are centred on zero and so model (6.3) is referred to as the NCP for the spatial Tobit model. Introducing the variable $\tilde{\beta}_k(\boldsymbol{s}_i) = \beta_k(\boldsymbol{s}_i) + \theta_k$ gives us the CP and allows us to write the model in its hierarchically centred form as

$$\begin{aligned} Y(\boldsymbol{s}) &= \max\{Z(\boldsymbol{s}), 0\} \\ \boldsymbol{Z}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \sigma_\epsilon^2 \boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned} \tag{6.4}$$

where $\boldsymbol{Z} = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))'$ and $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0', \ldots, \tilde{\boldsymbol{\beta}}_{p-1}')'$ is the $np \times 1$ vector of centred spatially correlated random effects where $\tilde{\boldsymbol{\beta}}_k = (\tilde{\beta}_k(\boldsymbol{s}_1), \ldots, \tilde{\beta}_k(\boldsymbol{s}_n))'$.

The distributional specification for $\boldsymbol{Z}$, $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{\theta}$ given in (6.4) is used to model Gaussian data in Chapters 3 and 4, where $\boldsymbol{Z}$ is the top level of the hierarchy. For a description of design matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ and covariance matrices $\boldsymbol{C}_2$ and $\boldsymbol{C}_3$ see Section 3.2.

We consider exponential correlation functions so that

$$(\boldsymbol{R}_k)_{ij} = \exp\{-\chi_k d_{ij}\},$$

where $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ and χ_k is the spatial decay parameter for the kth process.[1] To complete the model specification we assign inverse gamma prior distributions to the variance parameters, such that

$$\sigma_k^2 \sim IG(a_k, b_k), \quad \text{and} \quad \sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon),$$

and uniform prior distributions to each decay parameter

$$\chi_k \sim U(l_k, u_k),$$

where l_k and u_k are the lower and upper bounds of the support of the uniform distribution.

6.2.1 Gibbs sampling for the Tobit model

In this section we demonstrate how to perform Gibbs sampling for the Tobit model. Let $\mathscr{C} = \{i : Y(\boldsymbol{s}_i) = 0\}$ and define $\boldsymbol{Z}^- = \{Z(\boldsymbol{s}_i) : i \in \mathscr{C}\}$ so that $Z^-(\boldsymbol{s}_i) < 0$ for $i \in \mathscr{C}$. We treat $\boldsymbol{Z}^-$ as missing data and following Chib (1992) we combine the data augmentation (DA) algorithm (Tanner and Wong, 1987) with the Gibbs sampler (Gelfand and Smith, 1990) to obtain draws from the posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. We consider only the CP

[1] In other chapters the spatial decay parameters are represented by ϕ. Later in this chapter we take ϕ to be the density of a standard normal distribution, hence the change of notation.

for the time being and so the vector of model parameters is $\boldsymbol{\xi} = (\tilde{\boldsymbol{\beta}}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2}}', \sigma^2_\epsilon, \boldsymbol{\chi}')'$, where $\boldsymbol{\sigma^{2}} = (\sigma^2_0, \ldots, \sigma^2_{p-1})'$ and $\boldsymbol{\chi} = (\chi_0, \ldots, \chi_{p-1})'$.

The DA algorithm uses the following equalities

$$\pi(\boldsymbol{\xi}|\boldsymbol{y}) = \int \pi(\boldsymbol{\xi}|\boldsymbol{Z}^-, \boldsymbol{y})\pi(\boldsymbol{Z}^-|\boldsymbol{y})d\boldsymbol{Z}^-,$$

and

$$\pi(\boldsymbol{Z}^-|\boldsymbol{y}) = \int \pi(\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y})\pi(\boldsymbol{\xi}|\boldsymbol{y})d\boldsymbol{\xi},$$

and is performed as follows. Given $\boldsymbol{\xi}^{(0)}$, we alternate between the following two steps. For $t = 1, \ldots, T$,

1. Sample $\boldsymbol{Z}^{-(t)} \sim \pi(\boldsymbol{Z}^-|\boldsymbol{\xi}^{(t-1)}, \boldsymbol{y})$.
2. Sample $\boldsymbol{\xi}^{(t)} \sim \pi(\boldsymbol{\xi}|\boldsymbol{Z}^{-(t)}, \boldsymbol{y})$.

The draws $\{\boldsymbol{\xi}^{(t)}\}_{t=1}^T$ are samples from the marginal posterior distribution $\pi(\boldsymbol{\xi}|\boldsymbol{y})$. The conditional distribution of $\pi(\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y})$ is

$$\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(-\infty,0]}(\boldsymbol{X}_1^-\tilde{\boldsymbol{\beta}}, \sigma^2_\epsilon \boldsymbol{I}),$$

where $\boldsymbol{X}_1^-$ contains only the rows of $\boldsymbol{X}_1$ that correspond to locations at which $Y(\boldsymbol{s}_i) = 0$. We denote by $N_A(\cdot, \cdot)$ the normal distribution truncated to the set $A \in \mathbb{R}$.

To generate values from a truncated normal distribution we use a one-to-one inversion method, see (Devroye, 1986, Chapter 2). Let $u \sim U(0, 1)$ be a random draw from a uniform distribution over the unit interval. Then

$$v = \mu + \sigma\Phi^{-1}(\Phi(\alpha) + [\Phi(\gamma) - \Phi(\alpha)]u),$$

is a draw from a truncated normal $N_{[a,b]}(\mu, \sigma^2)$, where $\alpha = (a - \mu)/\sigma$, $\gamma = (b - \mu)/\sigma$ and $\Phi(\cdot)$ is the cdf of a standard normal distribution.

We use the value of $\boldsymbol{Z}^-$ to impute the missing values in $\boldsymbol{y}$ and then we can use the full conditional distributions given in Section (4.2.2) to sample from $\pi(\boldsymbol{\xi}|\boldsymbol{Z}^-, \boldsymbol{y})$.

For the NCP the set of all model parameters is $\boldsymbol{\xi} = (\boldsymbol{\beta}', \boldsymbol{\theta}', \boldsymbol{\sigma^{2}}', \sigma^2_\epsilon, \boldsymbol{\chi}')'$, and the conditional distribution $\pi(\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y})$ is

$$\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(-\infty,0]}(\boldsymbol{X}_1^-(\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\theta}), \sigma^2_\epsilon \boldsymbol{I}).$$

Details of how to sample from $\pi(\boldsymbol{\xi}|\boldsymbol{Z}^-, \boldsymbol{y})$ for the NCP are given in Section (4.2.3).

6.2.2 CP versus NCP for the Tobit model

In this section we investigate the performance of the CP and the NCP of the Tobit model. We simulate data and compare the performance of each parameterisation in terms of the (M)PSRF$_M$(1.1) and ESS of the model parameters. The measures are not adjusted for computation time as there is a negligible difference between the CP and the NCP in this regard.

Data is generated from model (6.3) with $p = 1$ and so we have a global mean parameter θ_0 which is locally adjusted by the realisations of a spatial process, $\tilde{\boldsymbol{\beta}}_0$ (or $\boldsymbol{\beta}_0$ in the non-centred case). We take the unit square to be the spatial domain and randomly select n=40 sampling locations, see Figure 3.3, and fix $\theta_0 = 0$. We let $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ be the ratio of the random effects variance to the data variance. We let $\sigma_0^2 = 1$ and vary σ_ϵ^2 so that we have five variance ratios of $\delta_0 = 0.01, 0.1, 1, 10, 100$. For each variance ratio we have four levels of the decay parameter, χ_0, corresponding to effective ranges of $d_0 = 0, \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$. We generate 20 data sets for each of the 20 variance ratio-effective range combinations.

We begin by assuming that σ_0^2 and σ_ϵ^2 are known, and hence they are fixed at their true values within the sampler. Therefore, we only sample from $\tilde{\boldsymbol{\beta}}_0$ (or $\boldsymbol{\beta}_0$) and θ_0. For the prior distribution of θ_0 we take hyperparameters $m_0 = 0$ and $v_0 = 10^4$. We run five chains of length 25,000 with from widely dispersed starting values, with the same values used for both the CP and the NCP. From the output we compute the $\text{PSRF}_M(1.1)$ and the ESS of θ_0. The results for the CP and the NCP are given in Figures 6.1 and 6.2 respectively. The top rows give the results for the $\text{PSRF}_M(1.1)$ and the bottom rows give the results for the ESS of θ_0. Each panel corresponds to a value of δ_0, rising from 0.01 on the left to 100 on the right. Within each panel of the four boxplots represent the results for each variance ratio-effective range pair, with each boxplot formed of 20 values.

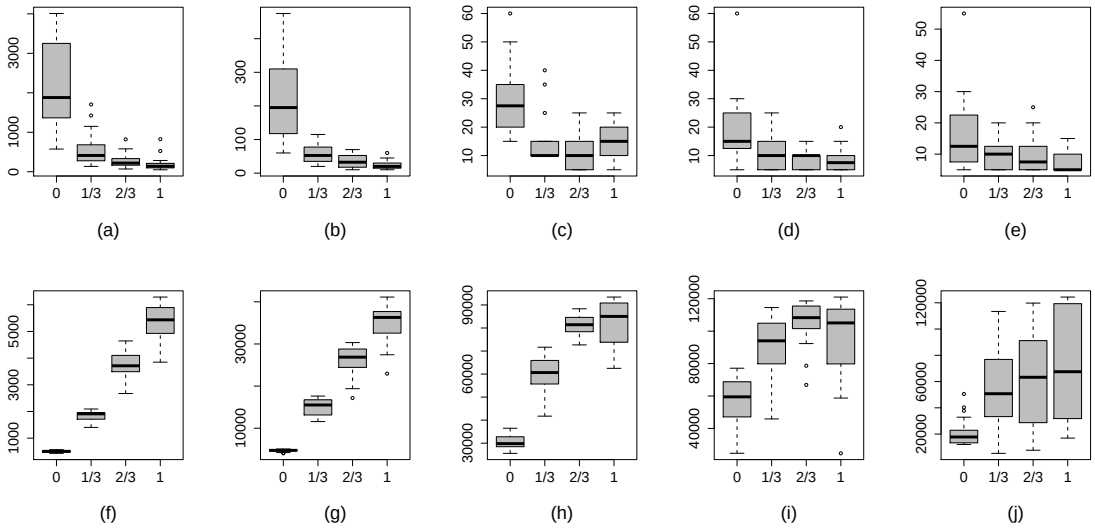

Figure 6.1: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the CP of the Tobit model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

We see that the effect of that varying δ_0 and χ_0 on the performance of the different parameterisations of the Tobit model is similar to that seen for the Gaussian model, investigated in Chapter 4. As the relative size of the data variance decreases, the performance of the CP is improved. Furthermore, for a fixed variance ratio, increasing the strength of correlation also improves the CP's performance. The opposite is seen for the NCP, where a relative increase in the data variance or a reduction in the strength of correlation yields

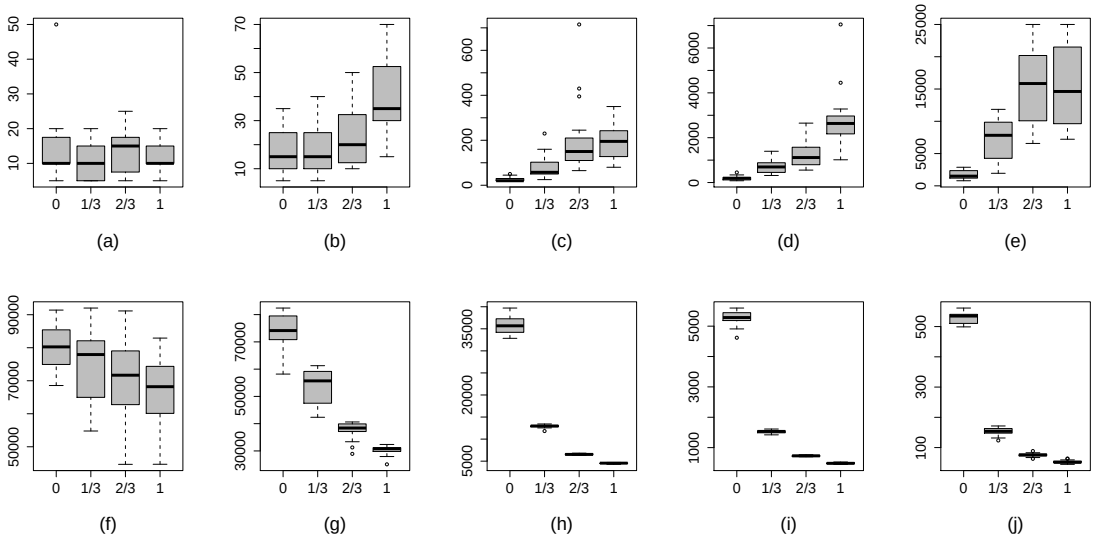

Figure 6.2: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the Tobit model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

an improved performance.

We now relax the assumption that σ_0^2 and σ_ϵ^2 are known and sample from their full conditional distributions. Following Gelfand et al. (2000) we take the prior distributions for the variance parameters to be $\pi(\sigma_0^2) = \pi(\sigma_\epsilon^2) = IG(2, 1)$. As we are now sampling the variance parameters, the CP and the NCP are compared by their $\text{MPSRF}_M(1.1)$'s and the ESS of θ_0, σ_0^2 and σ_ϵ^2.

Figure 6.3 gives the $\text{MPSRF}_M(1.1)$ and the ESS of θ for the CP. We see a similar picture here as when the variance parameters are assumed to be known. As the ratio of δ_0 increases, the $\text{MPSRF}_M(1.1)$ decreases and the ESS of θ_0 increases. We also see improved performance with increasing strength of correlation for a fixed variance ratio. It should be noted that for independent random effects, marginally $Var(Z(\boldsymbol{s}_i)) = \sigma_0^2 + \sigma_\epsilon^2$, which explains the poor performance for zero effective range.

The $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the NCP are given in Figure 6.4. The NCP also displays the trends here that were apparent when the variance parameters were fixed at their true values. The performance worsens as δ_0 increases or as the strength of correlation increases.

The ESS of the variance parameters is shown for the CP and the NCP in Figures 6.5 and 6.6 respectively. The pattern of results for the different parameterisations is similar. Both CP and NCP display poor mixing in the σ_0^2 and σ_ϵ^2 coordinates for the case when the random effects are independent. However, the performance is much improved in the presence of spatial correlation. The ESS of the variance parameters shows a slight downward trend as δ_0 increases, but not significant enough to claim that any relationship exists.

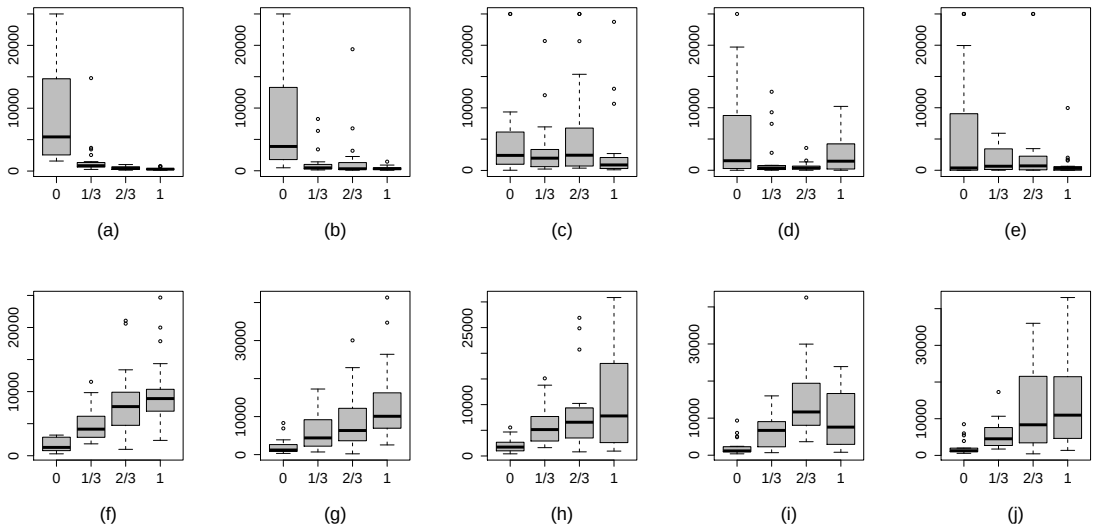

Figure 6.3: MPSRF$_M(1.1)$ and the ESS of θ_0 for the CP of the Tobit model with unknown variance parameters. Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

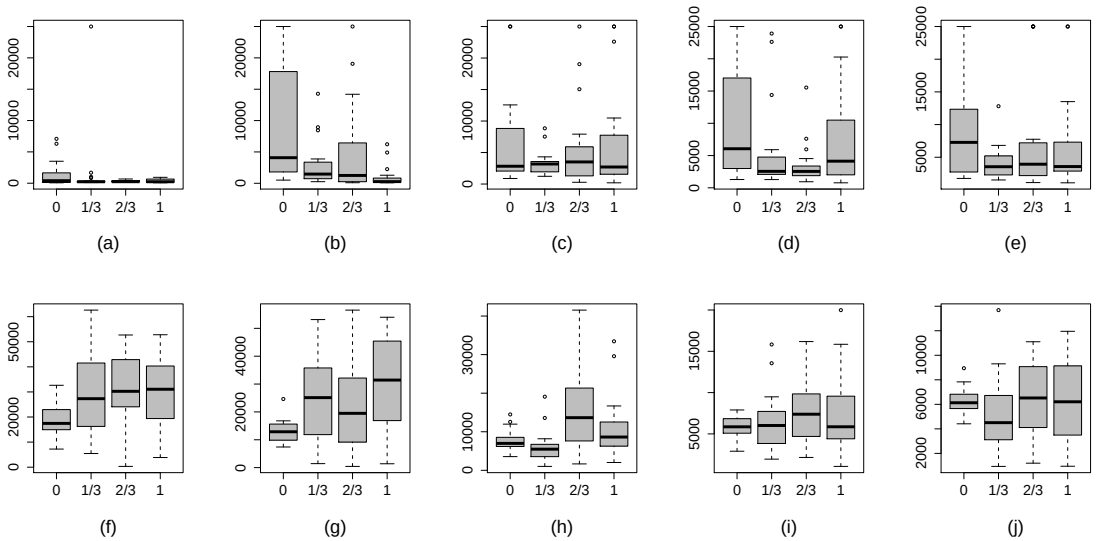

Figure 6.4: MPSRF$_M(1.1)$ and the ESS of θ_0 for the NCP of the Tobit model with unknown variance parameters. Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

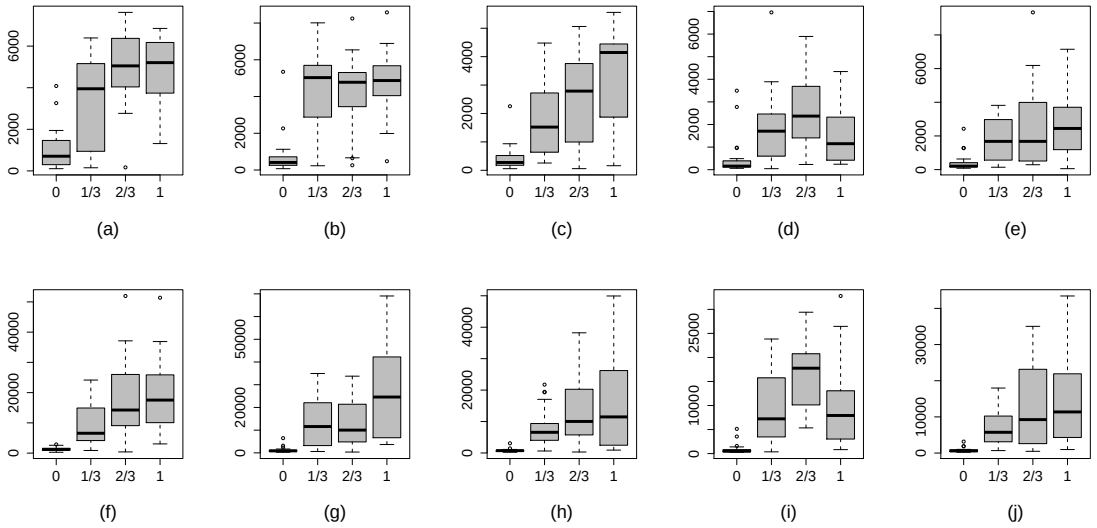

Figure 6.5: ESS of σ_0^2 and σ_ϵ^2 for the CP of the Tobit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

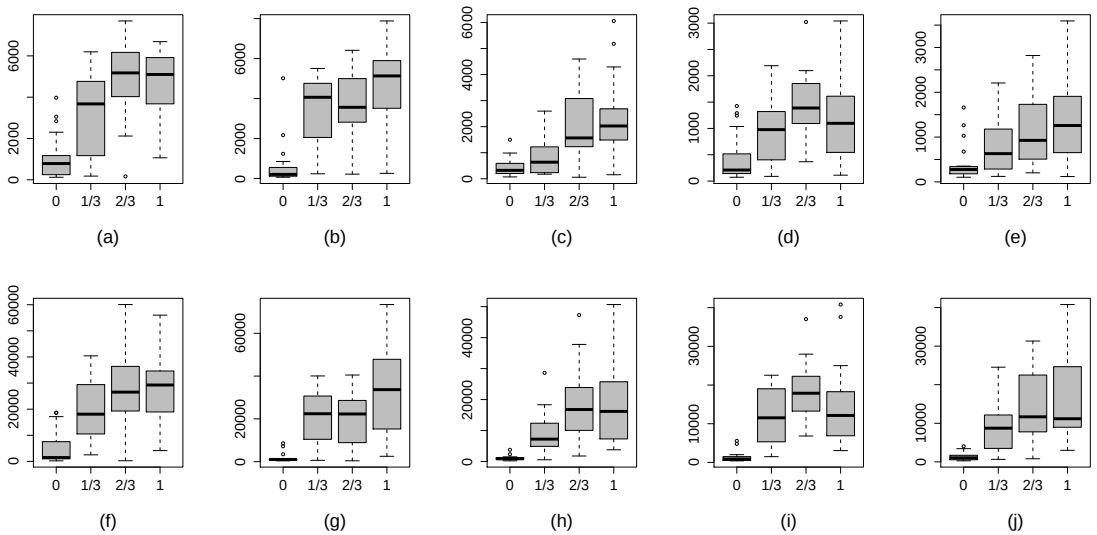

Figure 6.6: ESS of σ_0^2 and σ_ϵ^2 for the NCP of the Tobit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

6.2.3 Global mean and parameterisation for Tobit data

In this section we investigate how the performance of the CP and the NCP is affected by the global mean θ_0 and whether its value need inform us about how we choose to parameterise model (6.4).

The probability that $Y(\boldsymbol{s})$ is equal to zero is given by

$$Pr\left(Y(\boldsymbol{s})=0\right)=Pr\left(Z(\boldsymbol{s})<0\right)=Pr\left(\frac{Z(\boldsymbol{s})-\theta_0}{\sqrt{\sigma_0^2+\sigma_\epsilon^2}}<\frac{-\theta_0}{\sqrt{\sigma_0^2+\sigma_\epsilon^2}}\right)=\Phi\left(\frac{-\theta_0}{\sqrt{\sigma_0^2+\sigma_\epsilon^2}}\right),$$

and hence in the above simulation study by fixing $\theta_0=0$ we fix probability of not observing $Z(\boldsymbol{s})$ at 0.5. We now look at how varying this probability affects the performance of the CP and the NCP. To generate the data we fix $\sigma_0^2=\sigma_\epsilon^2=1$ and vary θ_0 such that the $Pr(Z(\boldsymbol{s})<0)=0.9,0.7,0.5,0.3,0.1$. For each level of $Pr(Z(\boldsymbol{s})<0)$ we have four effective ranges 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$, and we generate 20 data sets for each of the 20 pairs of $Pr(Z(\boldsymbol{s})<0)$ and effective range.

We begin by fixing the variance parameters and looking at the $\text{PSRF}_M(1.1)$ and the ESS of θ_0. Results for the CP and the NCP are given in Figures 6.7 and 6.8 respectively. The $\text{PSRF}_M(1.1)$ is given in the top row and ESS of θ_0 given in the bottom row. Each panel relates to a $Pr(Z(\boldsymbol{s})<0)$, ranging from 0.9 on the left to 0.1 on the right. Within each panel the four boxplots correspond to the four effective ranges, increasing from left to right. Each boxplot is made of the results of the 20 data sets generated for the respective pairing of $Pr(Z(\boldsymbol{s})<0)$ and effective range.

We see that for a given θ_0 increasing the strength of correlation improves the performance of the CP and hinders that of the NCP, as expected given the results of Section 6.2.2. When the probability of observing $Z(\boldsymbol{s})$ is increased, we see a slight improvement in the performance of both the CP and the NCP, but nothing to suggest that one parameterisation should be favoured over the other for a given level of censoring.

We now drop the assumption that the variance parameters are known and sample from their full conditional distributions. Figures 6.9 and 6.10 give the $\text{MPSRF}_M(1.1)$ and ESS of θ_0 for the CP and the NCP. Again we can see an improvement in performance for both the CP and the NCP as the $Pr(Z(\boldsymbol{s})<0)$ decreases. The ESS for the variance parameters for the CP and the NCP is given in Figures 6.11 and 6.12 respectively. The results are very similar for both parameterisations, each displaying an increase in the ESS of both of the variance parameters as $Pr(Z(\boldsymbol{s})<0)$ decreases. We conclude that the value of the global mean θ_0 should not be a factor in choosing between the CP and the NCP.

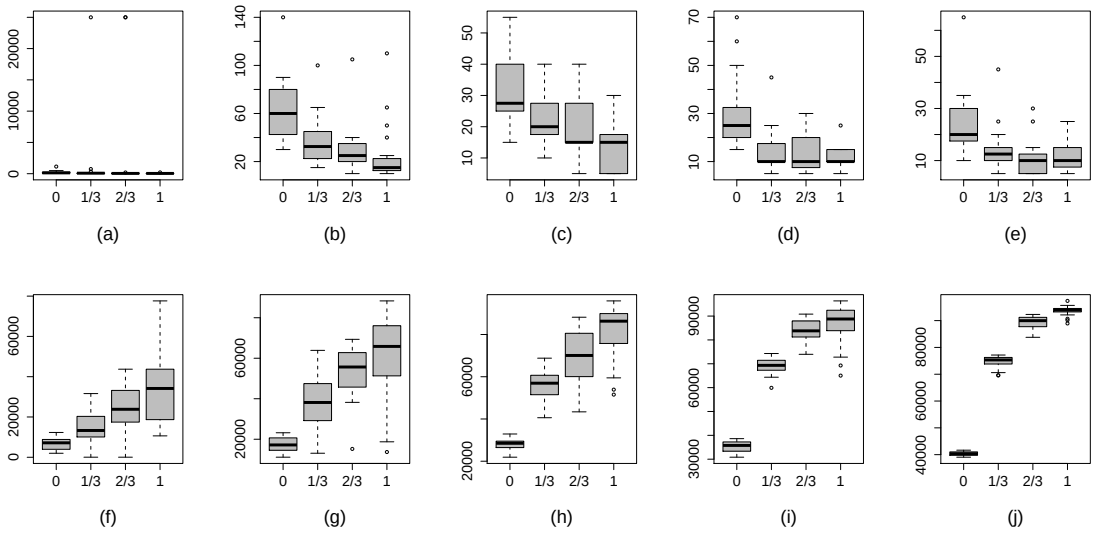

Figure 6.7: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 of the CP of the Tobit model with different global mean with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

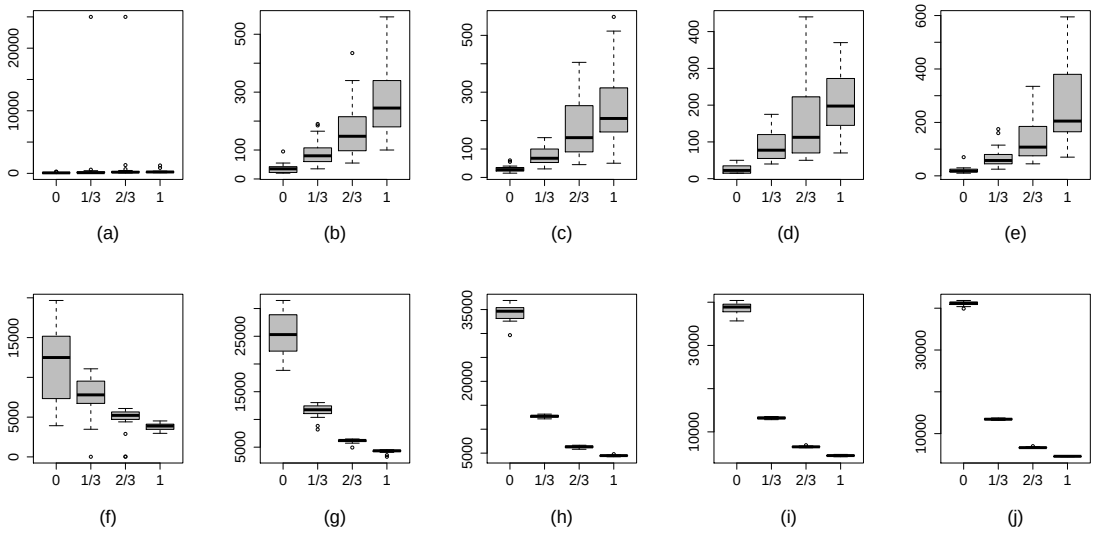

Figure 6.8: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the Tobit model with different global mean with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

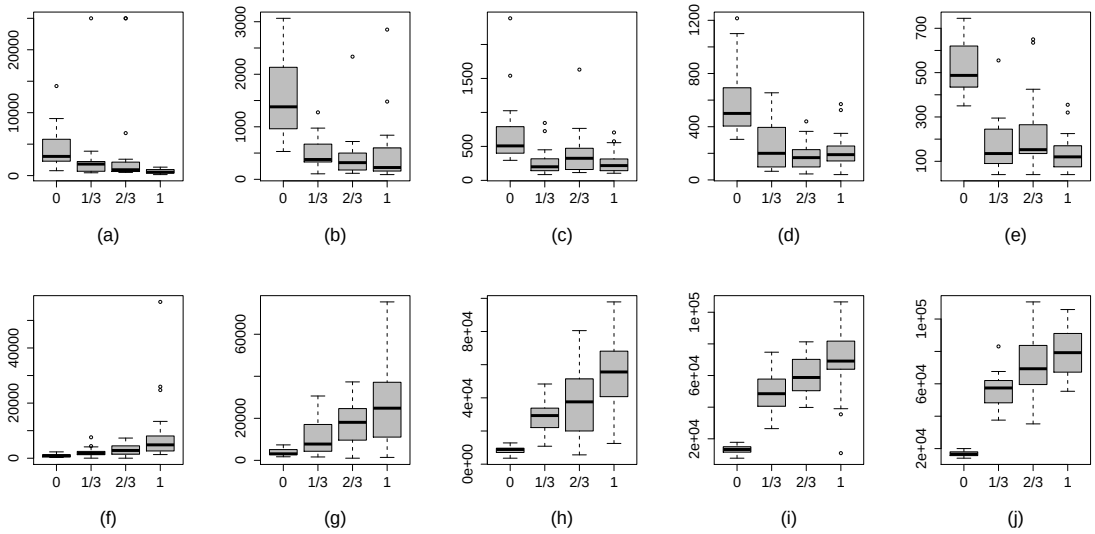

Figure 6.9: MPSRF$_M(1.1)$ and the ESS of θ_0 for the CP for the Tobit model with different global mean with unknown variance parameters. Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

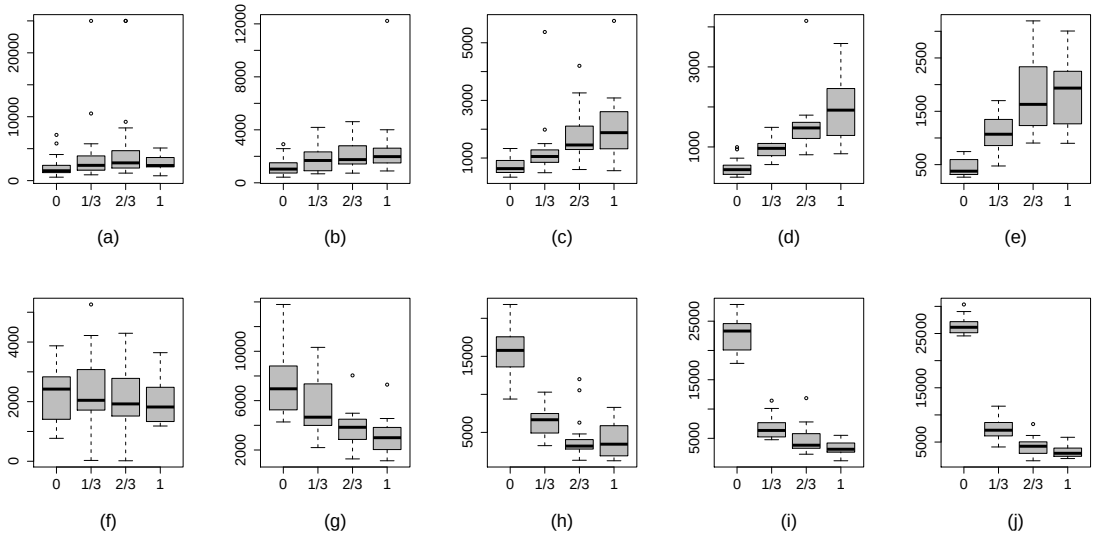

Figure 6.10: MPSRF$_M(1.1)$ and the ESS of θ_0 for the NCP for the Tobit model with different global mean with unknown variances. Plots (a)–(e) give the MPSRF$_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

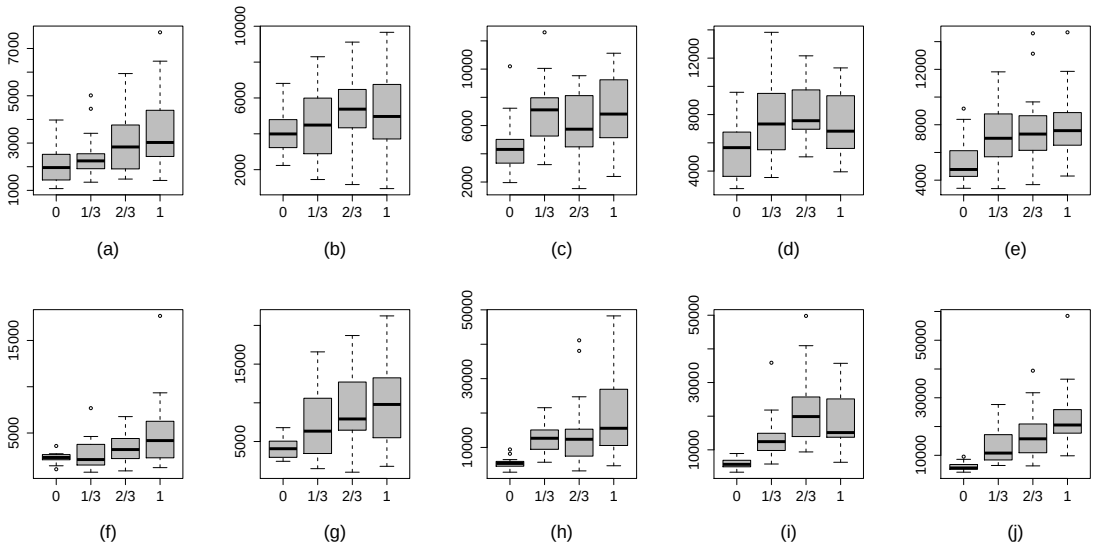

Figure 6.11: ESS of σ_0^2 and σ_ϵ^2 for the CP of the Tobit model with different global mean with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2 plots (f)–(j) the ESS of σ_ϵ^2. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

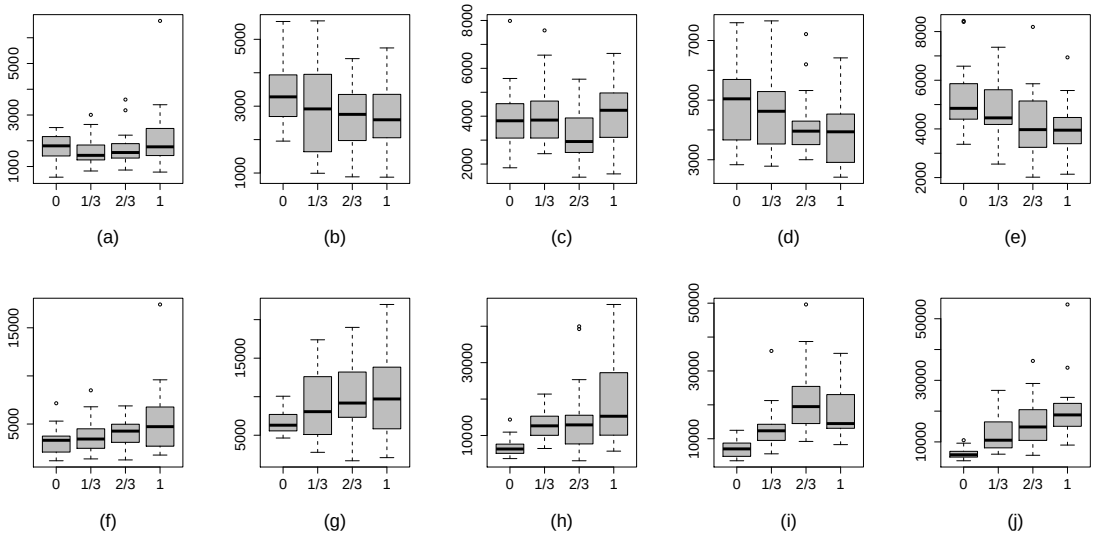

Figure 6.12: ESS of σ_0^2 and σ_ϵ^2 for the NCP of the Tobit model with different global mean with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $Pr(Z(\boldsymbol{s}) < 0) = 0.9, 0.7, 0.5, 0.3, 0.1$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

6.2.4 Tobit model applied to New York precipitation data

We now apply the CP and the NCP to a real data set. We have precipitation data from New York for the week July 30–August 5, 2001. Observed are the total weekly precipitation at 130 sampling locations. We fit the model to data from 104 locations having randomly selected 26 points to leave out for validation. These are shown in Figure 6.13. Of the 104 data sites there was no precipitation that week at 29 of them, and of the 26 validation sites there was no precipitation at seven. The dry locations are shown in Figure 6.14

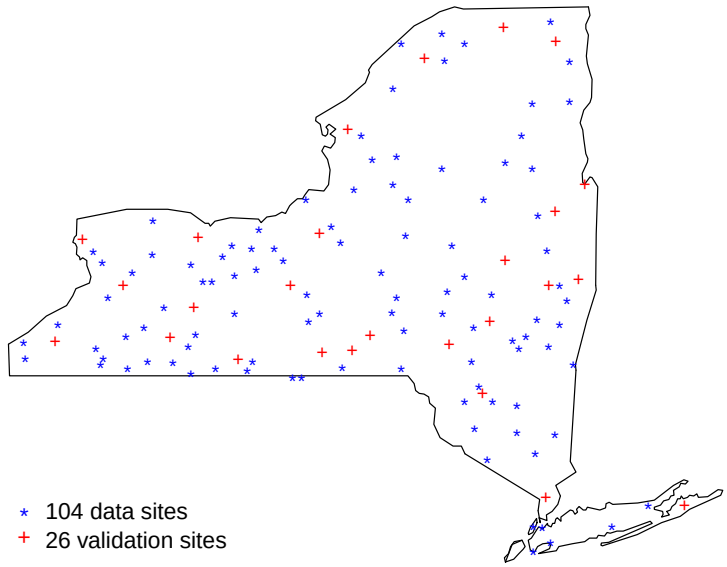

Figure 6.13: Locations of precipitation monitoring stations in New York.

We do not have any covariate information and therefore we set $p = 1$ in model (6.3), and so $\boldsymbol{\theta} = \theta_0$, $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_0$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. To estimate χ_0 we fit the model with five different values corresponding to effective ranges, d_0, of 50, 100, 250, 500 and 1000 km. Predictions are made at the validations sites and prediction errors are computed. Table 6.1 gives the mean absolute prediction error (MAPE), the root mean squared prediction error (RMSPE) and the continuous ranked probability score (CRPS) for each effective range. We see that an effective range of 100 km yields the lowest value for each criterion and so we set $\hat{\chi}_0 = -\log(0.05)/100 \approx 0.03$.

Table 6.1: Prediction error for different values of d_0 under the Tobit model.

d_0	MAPE	RMSPE	CRPS
50	0.791	0.883	0.470
100	**0.767**	**0.858**	**0.456**
250	0.812	0.885	0.472
500	0.876	0.933	0.497
1000	1.011	1.072	0.555

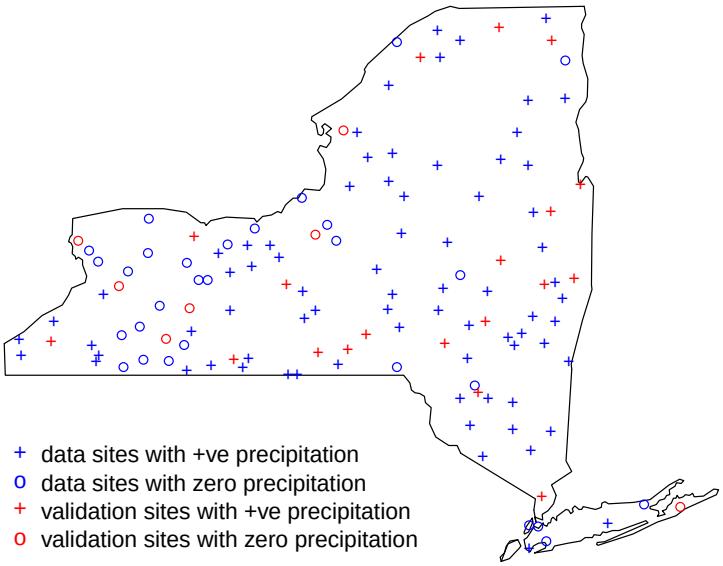

Figure 6.14: Locations of precipitation monitoring stations in New York indicating which measured positive precipitation.

Having fixed the decay parameter we now compare the CP and the NCP. We run five chains each of length 25,000. Table 6.2 gives the $\text{MPSRF}_M(1.1)$ and ESS of each of the model parameters for the CP and the NCP. We see that the CP has an $\text{MPSRF}_M(1.1)$

Table 6.2: $\text{MPSRF}_M(1.1)$ and the ESS of the Tobit model parameters.

	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS σ_0^2	ESS σ_ϵ^2
CP	500	52982	6662	5207
NCP	1875	2001	3878	4270

that is nearly four times lower, and an ESS for θ_0 that is over 26 times greater than that of the NCP.

Due to its superior performance we use the CP to obtain parameter estimates. A single chain of 50,000 iterations is run and the first 10,000 are discarded. The arithmetic mean and the 2.5 and 97.5 percentiles for the model parameters are given in Table 6.3 along with density plots given in Figure 6.15.

Table 6.3: Parameter estimates and their 95% credible intervals (CI) for the Tobit model.

Parameter	Estimate	95% CI
θ_0	0.686	$(-0.011, 1.358)$
σ_0^2	2.730	$(1.453, 4.243)$
σ_ϵ^2	0.589	$(0.184, 1.377)$

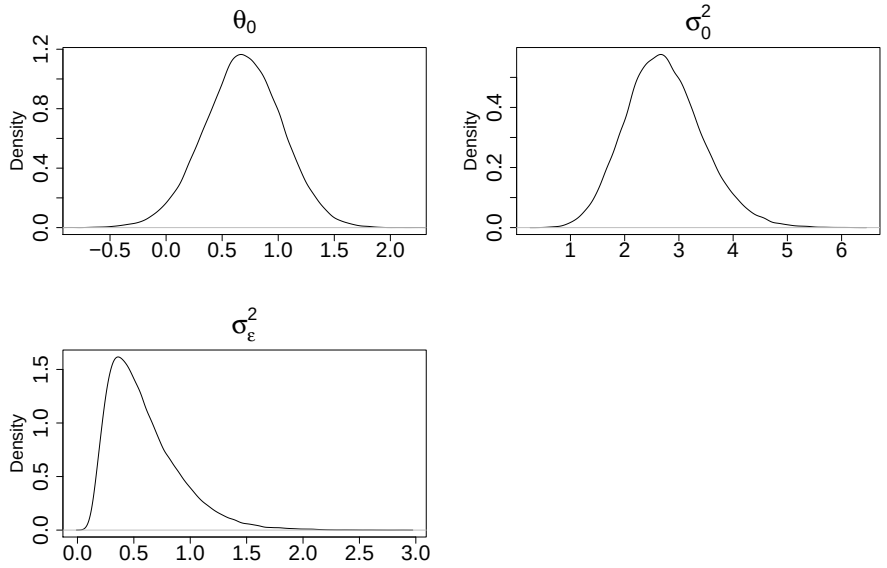

Figure 6.15: Density plots of the Tobit model parameters for New York precipitation data.

By evaluating $\delta_0^{(t)} = \sigma_0^{2(t)}/\sigma_\epsilon^{2(t)}$ at the tth iteration of the sampler, we obtain an estimate of $\hat{\delta}_0 = 6.531$. This explains why the CP outperforms the NCP for this data set.

We now look to create a predictive map of the probability of precipitation across New York during the week that the data were recorded. We draw samples from the posterior predictive distribution of Z($\boldsymbol{s}$) at knot locations on a 18×12 grid given in Figure 6.16. For a given knot location $\boldsymbol{s}^*$ and model parameters $\boldsymbol{\xi} = (\tilde{\boldsymbol{\beta}}_0', \theta_0, \sigma_0^2, \sigma_\epsilon^2)'$, the posterior predictive distribution is given by

$$\pi(Z(\boldsymbol{s}^*)|\boldsymbol{y}) = \int \pi(Z(\boldsymbol{s}^*)|\tilde{\beta}_0(\boldsymbol{s}^*), \boldsymbol{\xi}, \boldsymbol{Z}^-, \boldsymbol{y})\pi(\beta_0(\boldsymbol{s}^*)|\boldsymbol{\xi}, \boldsymbol{Z}^-, \boldsymbol{y})\pi(\boldsymbol{\xi}, \boldsymbol{Z}^-|\boldsymbol{y})d\tilde{\beta}_0(\boldsymbol{s}^*)d\boldsymbol{\xi}d\boldsymbol{Z}^-.$$

For each post-burn in sample $(\boldsymbol{\xi}^{(t)}, \boldsymbol{Z}^{-(t)})' \sim \pi(\boldsymbol{\xi}, \boldsymbol{Z}^-|\boldsymbol{y})$ obtained from the Gibbs sampler we draw $\tilde{\beta}_0^{(t)}(\boldsymbol{s}^*) \sim \pi(\beta_0(\boldsymbol{s}^*)|\boldsymbol{\xi}^{(t)}, \boldsymbol{Z}^{-(t)}, \boldsymbol{y})$ and then

$$Z^{(t)}(\boldsymbol{s}^*) \sim \pi(Z(\boldsymbol{s}^*)|\tilde{\beta}_0^{(t)}(\boldsymbol{s}^*), \boldsymbol{\xi}^{(t)}, \boldsymbol{Z}^{-(t)}, \boldsymbol{y}),$$

is a draw from the posterior predictive distribution $\pi(Z(\boldsymbol{s}^*)|\boldsymbol{y})$. The relevant conditional distributions are

$$Z(\boldsymbol{s}^*)|\tilde{\beta}_0(\boldsymbol{s}^*), \boldsymbol{\xi}, \boldsymbol{Z}^-, \boldsymbol{y} \sim N(\tilde{\beta}_0(\boldsymbol{s}^*), \sigma_\epsilon^2),$$

and

$$\tilde{\beta}_0(\boldsymbol{s}^*)|\boldsymbol{\xi}, \boldsymbol{Z}^-, \boldsymbol{y} \sim N\left(\theta_0 + \boldsymbol{c}_0'\boldsymbol{R}_0^{-1}(\tilde{\boldsymbol{\beta}}_0 - \theta_0\mathbf{1}), \sigma_0^2(1 - \boldsymbol{c}_0'\boldsymbol{R}_0^{-1}\boldsymbol{c}_0)\right),$$

where $\boldsymbol{c}_0$ is a n-dimensional vector whose elements are given by $\exp\{-\chi_0\|\boldsymbol{s}_i - \boldsymbol{s}^*\|\}$.

By computing the proportion draws from $Z^{(t)}(\boldsymbol{s}^*) \sim \pi(Z(\boldsymbol{s}^*)|\boldsymbol{y})$ that are greater than zero, we compute the probability of positive precipitation at $\boldsymbol{s}^*$ during the week that the data was recorded. This is done for all knot locations. The probabilities are then interpolated to produce the predictive probability map given in Figure 6.17.

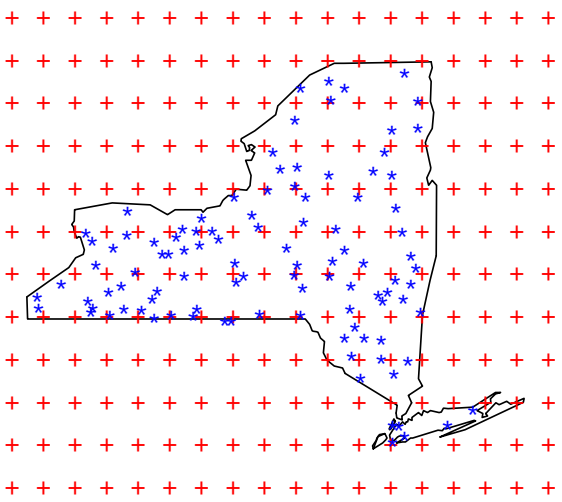

Figure 6.16: Data locations and predictive grid for New York precipitation data.

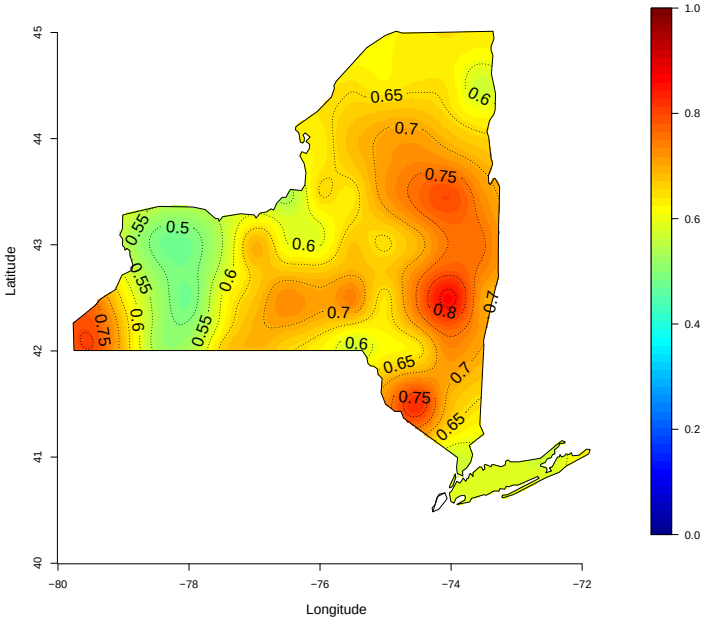

Figure 6.17: Predictive map of the probabilty of positive precipitation across New York for the week July 30–August 5, 2001.

6.3 Spatial probit model

In this section we look at the spatial probit model. The probit model has been widely applied to model binary data due to the easy implementation of Gibbs samplers for making inference. The spatial probit model is widely used in ecology for presence-absence data, (Johnson et al., 2013; Rathbun and Fei, 2006; Musio et al., 2008), and is applied to model voting behaviour (Salazar et al., 2013) and real estate markets (Gelfand et al., 2000).

We consider the following model

$$\begin{aligned} Y(\boldsymbol{s}_i) &= \begin{cases} 1 & \text{if } Z(\boldsymbol{s}_i) > z^* \\ 0 & \text{if } Z(\boldsymbol{s}_i) \leq z^* \end{cases} \\ Z(\boldsymbol{s}_i) &= \sum_{k=0}^{p-1} \{\theta_k + \beta_k(\boldsymbol{s}_i)\} x_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \end{aligned} \tag{6.5}$$

for $i = 1, \ldots, n$, where $Z(\boldsymbol{s}_i)$ is the realisation at $\boldsymbol{s}_i$ of the process driving the observable $Y(\boldsymbol{s}_i)$ and z^* is some threshold value that dichotomises $Z(\boldsymbol{s}_i)$. Errors $\epsilon(\boldsymbol{s}_i)$ are modelled as independent, normally distributed random variates with mean zero and variance σ_ϵ^2. The rest of the model is identical to that of the Tobit model, given in Section 6.2.

Setting $\beta_1(\boldsymbol{s}_i) = \ldots = \beta_{p-1}(\boldsymbol{s}_i) = 0$, so that

$$Z(\boldsymbol{s}_i) = \boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\theta} + \beta_0(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i),$$

where $\boldsymbol{x}'(\boldsymbol{s}_i) = (1, x_1(\boldsymbol{s}_i), \ldots, x_{p-1}(\boldsymbol{s}_i))$, gives us the model considered by Gelfand et al. (2000). They have $z^* = 0$ and note that

$$\begin{aligned} Pr(Y(\boldsymbol{s}_i) = 1 | \boldsymbol{\theta}, \boldsymbol{\beta}_0, \sigma_\epsilon^2) &= Pr(Z(\boldsymbol{s}) > 0 | \boldsymbol{\theta}, \boldsymbol{\beta}_0, \sigma_\epsilon^2) \\ &= 1 - \Phi\left(-\frac{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\theta} + \beta_0(\boldsymbol{s}_i)}{\sigma_\epsilon} \right) \\ &= \Phi\left(\frac{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\theta} + \sigma_0 \gamma(\boldsymbol{s}_i)}{\sigma_\epsilon} \right), \end{aligned}$$

where $\Phi(\cdot)$ is the cdf of a standard normal distribution, and $\boldsymbol{\gamma} = (\gamma(\boldsymbol{s}_1), \ldots, \gamma(\boldsymbol{s}_n))'$ with $\boldsymbol{\beta}_0 = \sigma_0 \boldsymbol{\gamma}$, so that $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{R}_0)$. Therefore $Pr(Y(\boldsymbol{s}) = 1 | \boldsymbol{\theta}, \boldsymbol{\beta}_0, \sigma_\epsilon^2)$ is unchanged if σ_ϵ, σ_0 and $\boldsymbol{\theta}$ are multiplied by a constant. To combat this indentifiability problem, De Oliveira (2000) and omit the nugget term and have

$$\boldsymbol{Z} \sim N(\boldsymbol{X}\boldsymbol{\theta}, \sigma_0^2 \boldsymbol{R}_0),$$

where the ith row of $\boldsymbol{X}$ is $\boldsymbol{x}'(\boldsymbol{s}_i)$. However, Gelfand et al. (2000) point out that without including $\epsilon(\boldsymbol{s}_i)$, there is no conditional independence for the $Y(\boldsymbol{s}_i)$, and so they set $\sigma_\epsilon = 1$.

These issues only arise if $z^* = 0$, as in general

$$Pr(Y(\boldsymbol{s}_i) = 1 | \boldsymbol{\theta}, \boldsymbol{\beta}_0, \sigma_\epsilon^2) = \Phi\left(\frac{\boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{\theta} + \sigma_0 \gamma(\boldsymbol{s}_i) - z^*}{\sigma_\epsilon} \right),$$

and we have $z^* \neq 0$ for the simulated and real data examples analysed in Sections 6.3.2 and 6.3.3.

6.3.1 Gibbs sampling for the probit model

Gibbs sampling for the probit model is performed similarly to the Tobit model, but for binary data we must obtain samples from all of $\boldsymbol{Z}$, and not just the negative part. We have that $\mathscr{C} = \{i : Y(\boldsymbol{s}_i) = 0\}$ and $\boldsymbol{Z}^- = \{Z(\boldsymbol{s}_i) : i \in \mathscr{C}\}$, see Section 6.2.1. Now we equivalently define $\mathscr{E} = \{i : Y(\boldsymbol{s}_i) = 1\}$ and $\boldsymbol{Z}^+ = \{Z(\boldsymbol{s}_i) : i \in \mathscr{E}\}$. Therefore we partition $\boldsymbol{Z} = (\boldsymbol{Z}^{-\prime}, \boldsymbol{Z}^{+\prime})'$ where $\boldsymbol{Z}^-$ contains the unobserved values of $\boldsymbol{Z}$ at locations where $Y(\boldsymbol{s}_i) = 0$, and $\boldsymbol{Z}^+$ contains the unobserved values of $\boldsymbol{Z}$ at locations where $Y(\boldsymbol{s}) = 1$.

For the CP the full conditional distributions for $\boldsymbol{Z}^-$ and $\boldsymbol{Z}^+$ are

$$\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(-\infty, z^*]}(\boldsymbol{X}^-\tilde{\boldsymbol{\beta}}, \sigma_\epsilon^2 \boldsymbol{I}) \quad \text{and} \quad \boldsymbol{Z}^+|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(z^*, \infty)}(\boldsymbol{X}^+\tilde{\boldsymbol{\beta}}_1, \sigma_\epsilon^2 \boldsymbol{I}),$$

where $\boldsymbol{X}^-$ and $\boldsymbol{X}^+$ are the rows of $\boldsymbol{X}_1$ corresponding to locations where $Y(\boldsymbol{s}_i) = 0$ and $Y(\boldsymbol{s}_i) = 1$, respectively. For the NCP we have

$$\boldsymbol{Z}^-|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(-\infty, z^*]}(\boldsymbol{X}_1^-(\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\theta}), \sigma_\epsilon^2 \boldsymbol{I}) \quad \boldsymbol{Z}^+|\boldsymbol{\xi}, \boldsymbol{y} \sim N_{(z^*, \infty)}(\boldsymbol{X}_1^+(\boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\theta}), \sigma_\epsilon^2 \boldsymbol{I}),$$

The remaining full conditional distributions for $\tilde{\boldsymbol{\beta}}$ (for the CP), $\boldsymbol{\beta}$ (for the NCP), $\boldsymbol{\theta}$, $\boldsymbol{\sigma^2}$, σ_ϵ^2 and $\boldsymbol{\chi}$ are the same as those given for the Gaussian model, see Section 4.2.

6.3.2 CP versus NCP for the probit model

We generate data from the probit model (6.5) where we set $p = 1$ and $\theta_0 = 1$ and $z^* = 1$. We use the unit square as the spatial domain and the same sampling locations as were used for the Tobit model in Section 6.3.2. We use 20 variance ratio-effective range combinations. There are five levels for $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ which are 0.01, 0.1, 1, 10 and 100. We use effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$, which correspond to fractions of the largest possible separation of two points in the unit square. We have 20 data sets for each combination, and hence 400 data sets in total. We run five chains of length 25,000 for the CP and the NCP, using the same set of widely dispersed starting values for each parameterisation.

We begin with case when σ_0^2 and σ_ϵ^2 are assumed to be known. The $\text{PSRF}_M(1.1)$ and ESS of θ_0 are given for the CP and the NCP in Figures 6.18 and 6.19 respectively. We see a now familiar pattern. The performance of the CP is improved with increasing δ_0 with the opposite observed for the NCP. For a fixed δ_0 increasing the effective range favours the CP but hinders the NCP.

When we sample from the variance components the picture remains largely unchanged. The $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the CP is given in Figure 6.20, with the equivalent results for the NCP given in Figure 6.21. Again, we can see that as δ_0 is increased the CP becomes more efficient, and the NCP less so. It is not as clear here as it is when the variance parameters are fixed, but increasing the effective range has a positive effect on the ESS of θ_0 for the CP, and a negative one for the NCP.

The results for the variance parameters σ_0^2 and σ_ϵ^2 are given in Figures 6.22 and 6.23. The pattern of results for each parameterisation is the same. We see that the ESS of σ_0^2 falls with δ_0. For a fixed δ_0 the ESS increases with effective range, d_0, especially for $\delta_0 \leq 1$.

For both the CP and the NCP, the results for σ_ϵ^2 are fairly constant across the different values of δ_0. For a fixed δ_0, we see that the ESS of σ_ϵ^2 increases with d_0.

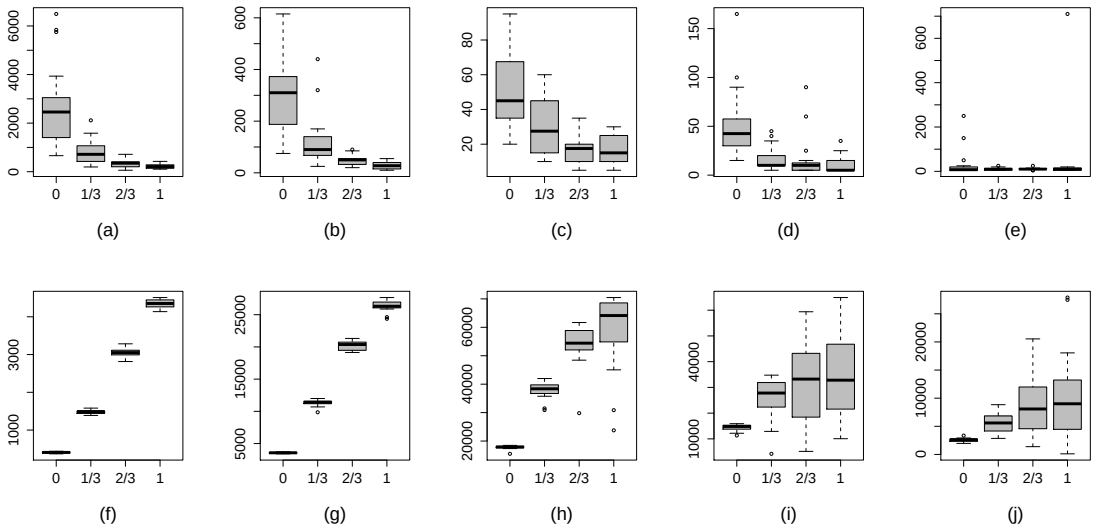

Figure 6.18: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the CP of the probit model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

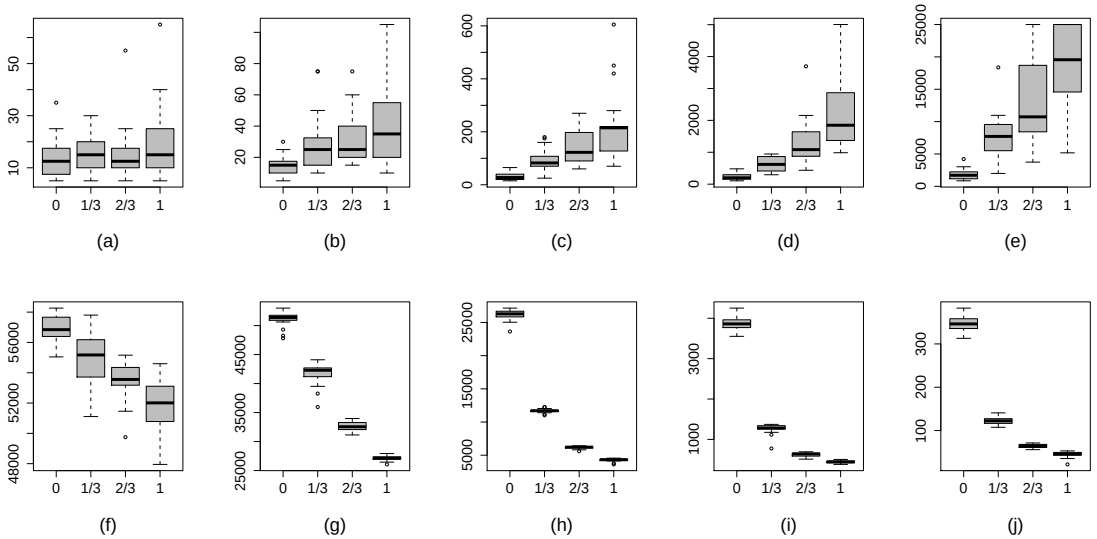

Figure 6.19: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the probit model with known variance parameters. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

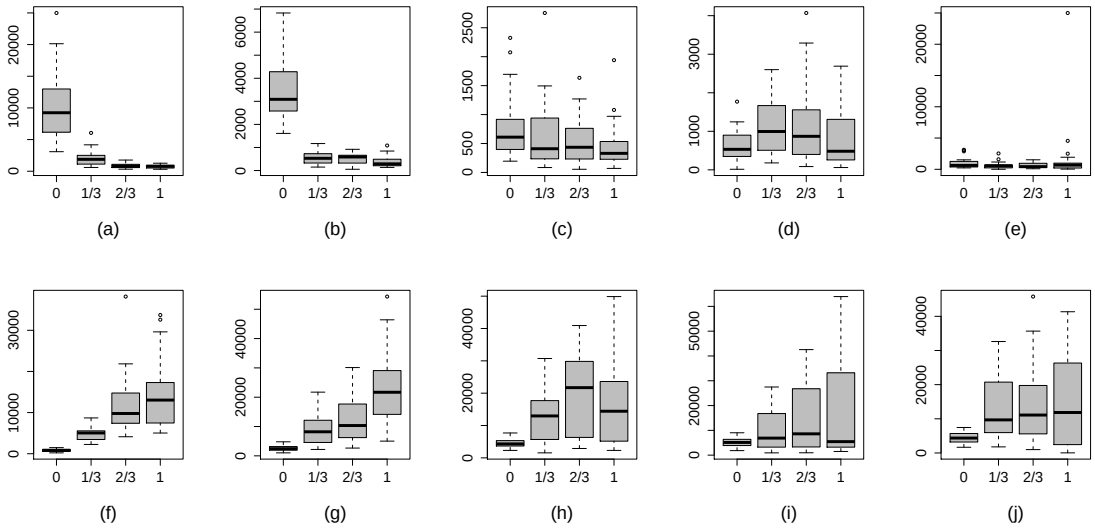

Figure 6.20: $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the CP of the probit model with unknown variance parameters. Plots (a)–(e) give the $\text{MPSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

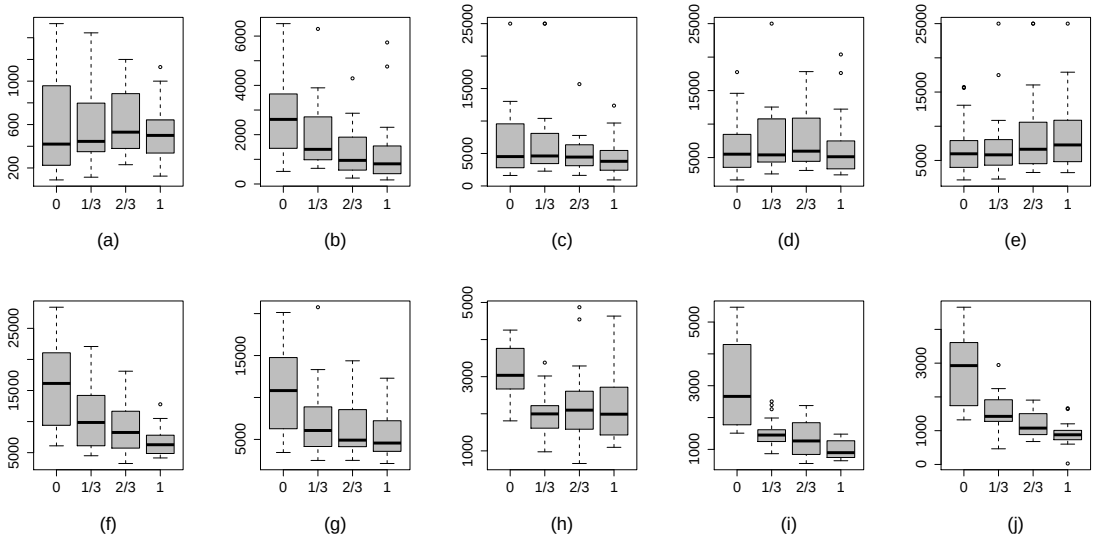

Figure 6.21: $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the NCP of the probit model with unknown variance parameters. Plots (a)–(e) give the $\text{MPSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

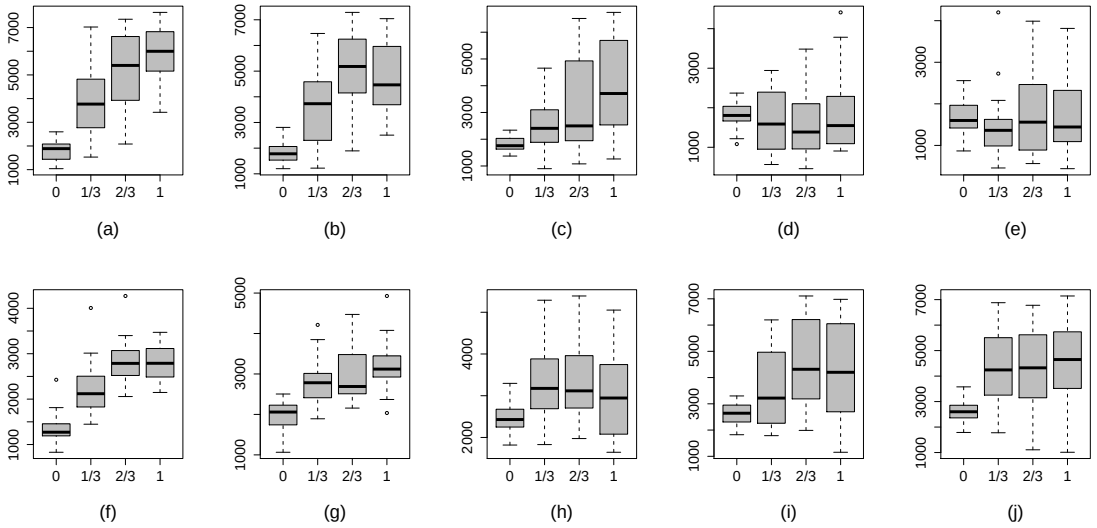

Figure 6.22: ESS of σ_0^2 and σ_ϵ^2 for the CP of the probit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

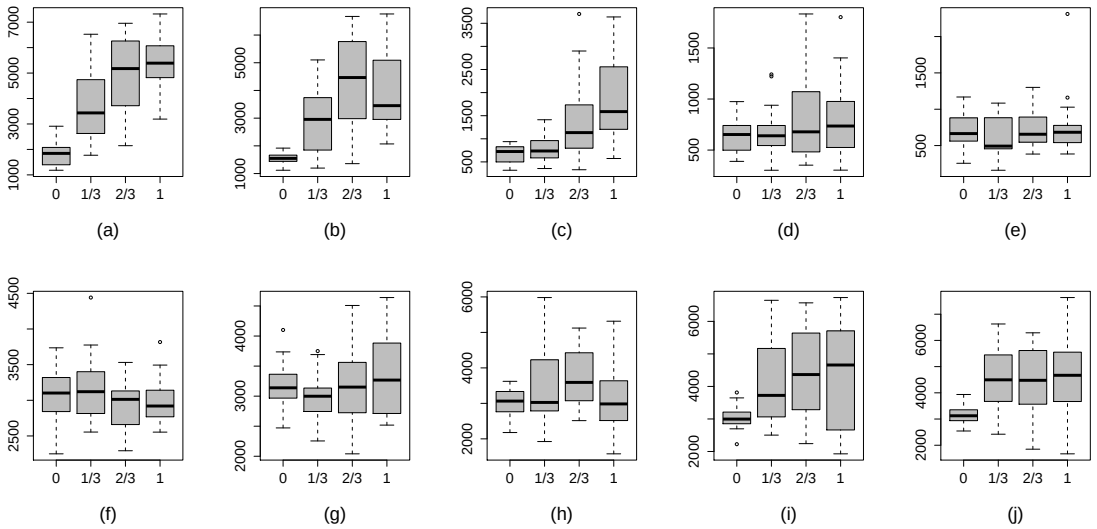

Figure 6.23: ESS of σ_0^2 and σ_ϵ^2 for the NCP of the probit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

6.3.3 Probit model applied to Californian ozone concentration data

In this section we apply the CP and the NCP of the probit model to the Californian ozone data set analysed in Section 4.5. Here, we are interested in the probability that the ozone concentration exceeds 75ppb, the limit set by the US Environmental Protection Agency. We model data on the square root scale to stabilise the variance and so $z^* = \sqrt{75}$ in model (6.5). As in the simulation study in Section 6.3.2 we have $\boldsymbol{x}(\boldsymbol{s}_i) = 1$ and hence $\boldsymbol{\theta} = \theta_0$, $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}_0$ and $\boldsymbol{\beta} = \boldsymbol{\beta}_0$.

We begin by estimating the decay parameter, χ_0. We consider the same effective ranges that were considered in Section 4.5: 50, 100, 250, 500 and 1000 km. The true values of $Z(\boldsymbol{s})$ are the observed ozone concentrations, and in this case they are available. Therefore, we compare the predictions for $Z(\boldsymbol{s})$ at the validation sites with the ozone concentration data. Table 6.4 shows the MAPE, the RMSPE and the CRPS for the five different effective ranges. We see that judged by the RMSPE and the CRPS, an effective range of 250 km gives the best performance, and by the MAPE criterion it comes a close second. Therefore we let $\hat{\chi}_0 = -\log(0.05)/250$.

Table 6.4: Prediction error for different values of d_0 under the probit model.

d_0	MAPE	RMSPE	CRPS
50	13.99	17.96	10.91
100	12.65	16.81	10.21
250	11.72	**15.96**	**10.05**
500	**11.69**	16.16	10.78
1000	13.83	18.36	11.81

With χ_0 fixed, we run five chains with widely dispersed starting values of length 25,000 each for both the CP and the NCP. The $\text{MPSRF}_M(1.1)$ and the ESS of θ_0, σ_0^2 and σ_ϵ^2 are given in Table 6.5. There is not much between the CP and the NCP in terms of the $\text{MPSRF}_M(1.1)$, both taking around 10,000 iterations to converge by this measure. The ESS of σ_0^2 is very low for both parameterisations and this contributes to their slow convergence. However, the ESS of θ_0 is over 23 times greater for the CP than the NCP.

Table 6.5: $\text{MPSRF}_M(1.1)$ and the ESS of the probit model parameters.

	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS σ_0^2	ESS σ_ϵ^2
CP	9925	19245	268	1887
NCP	10335	740	271	1899

We now use the CP to obtain estimates of the model parameters. A single chain of length 50,000 is generated. We discard the first 10,000 iterations, performing inference on what remains. Parameter estimates and their 95% credible intervals are given in Table 6.6 and density plots of model parameters given in Figure 6.24.

We obtain an estimate for $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ of $\hat{\delta}_0 = 10.498$, which is why we have far better mixing for θ_0 for the CP than the NCP. Note the uncertainty in the estimate of σ_0^2. High correlation between successive iterates means that after excursions to the tails the sampler is slow to return to areas of high posterior density.

Table 6.6: Parameter estimates and their 95% credible intervals (CI) for the probit model.

Parameter	Estimate	95% CI
θ_0	8.885	$(8.581, 10.180)$
σ_0^2	3.824	$(0.456, 17.475)$
σ_ϵ^2	0.406	$(0.133, 1.137)$

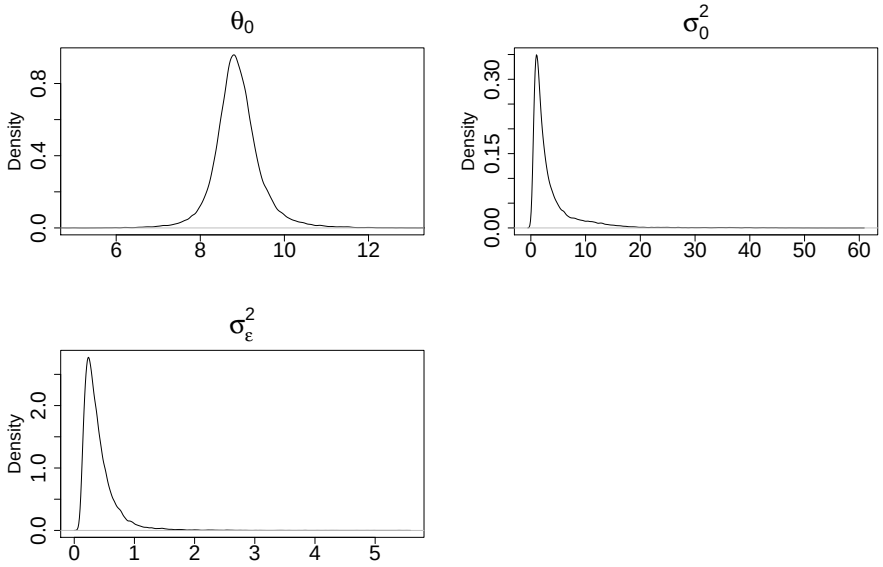

Figure 6.24: Density plots of the probit model parameters for Californian ozone concentration data.

To create a predictive map of the probability of exceedance, predictions are made at a collection of points described by the 12×12 grid given in Figure 6.25 and are then smoothed to create the map given in Figure 6.26. We can see that the probability of exceeding the 75 ppb threshold is greatly reduced for areas near the coast.

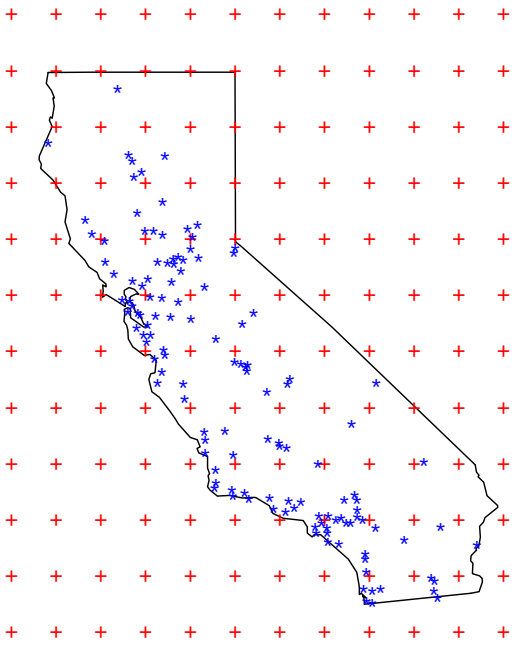

Figure 6.25: Data locations and predictive grid for Californian ozone concentration data.

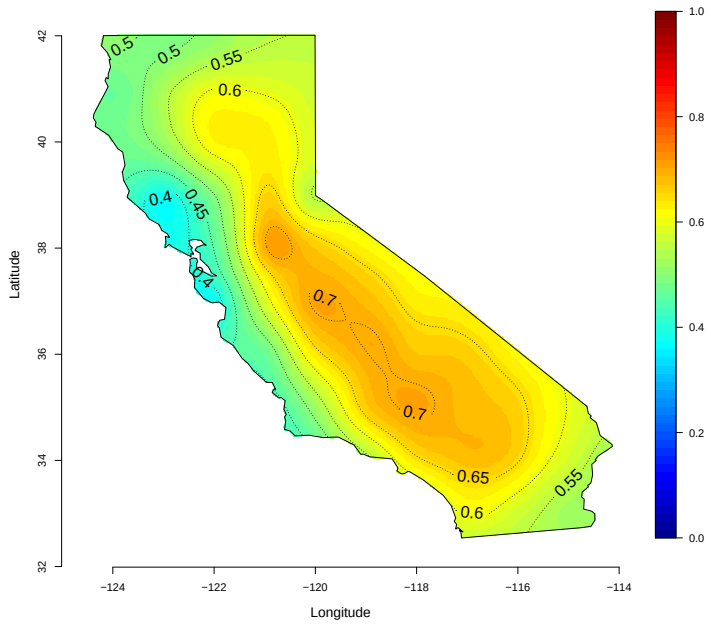

Figure 6.26: Predictive map of the probabilty of that ozone concentrations in California exceed 75 ppb.

6.4 Partial centering for non-Gaussian spatial models

In this section we look at the construction and performance of the partially centred parameterisation (PCP) for non-Gaussian spatial models. Recall that in the case of Gaussian likelihoods the PCP is induced by constructing a weight matrix $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$, where $\boldsymbol{B} = Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{y})$ and $\boldsymbol{C}_2 = Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta})$ are the conditional posterior and prior variances of $\tilde{\boldsymbol{\beta}}$ respectively.

In the Gaussian case we can compute $\boldsymbol{B}$ (see Lemma 2.4.1), but we do not have an equivalent expression in the non-Gaussian case, and so it must be estimated. We approximate the $Var(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{y})$ by $\widehat{\boldsymbol{B}}$, where

$$\widehat{\boldsymbol{B}}^{-1} = -\frac{\partial^2}{\partial \tilde{\boldsymbol{\beta}}^2} \log \pi(\tilde{\boldsymbol{\beta}}|\theta, \boldsymbol{y}) \Big|_{\tilde{\boldsymbol{\beta}}=\widehat{\boldsymbol{\beta}}}, \tag{6.6}$$

and $\widehat{\boldsymbol{\beta}}$ is the MLE of $\tilde{\boldsymbol{\beta}}$. Justification of this approximation can be found in (Bernardo and Smith, 1994, Chapter 5.3).

This approach is used by Papaspiliopoulos et al. (2003) to find a PCP for the Poisson log-normal model often used for count data. Christensen et al. (2006) suggest transforming $\tilde{\boldsymbol{\beta}}$ to obtain approximately independent random effects, so that instead of updating $\tilde{\boldsymbol{\beta}}$, they update $\tilde{\boldsymbol{\beta}}^*$ where $\tilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{B}}^{1/2}\tilde{\boldsymbol{\beta}}^*$, and $\widehat{\boldsymbol{B}}^{1/2}$ is the Cholesky decomposition of $\widehat{\boldsymbol{B}}$. We use $\widehat{\boldsymbol{B}}$ to find $\widehat{\boldsymbol{W}} = \boldsymbol{I} - \widehat{\boldsymbol{B}}\boldsymbol{C}_2^{-1}$, and hence construct a PCP for the Tobit and probit models.

6.4.1 Partial centering for the Tobit model

In Section 6.2 we analysed the performance of the CP and the NCP for the spatial Tobit model. In this section we revisit the Tobit model and investigate the construction and performance of its PCP. We consider model (6.3) with $p = 1$. Initially we must compute $\widehat{\boldsymbol{B}}$, given in expression (6.6). To begin, let e_i be an indicator variable such that

$$e_i = \begin{cases} 1 & \text{if } Y(\boldsymbol{s}_i) > 0 \\ 0 & \text{if } Y(\boldsymbol{s}_i) = 0. \end{cases}$$

Now we write $\log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y})$ as

$$\begin{aligned} \log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y}) &\propto \log \pi(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0) + \log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0) \\ &= \sum_{i=1}^{n} (1 - e_i) \log(1 - \Phi_i) \\ &\quad + e_i \left(-\frac{1}{2} \log 2\pi - \log \sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2} (Y(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i))^2 \right) \\ &\quad - \frac{n}{2} \log 2\pi - \frac{n}{2} \log \sigma_0^2 - \frac{1}{2\sigma_0^2} \left(\tilde{\boldsymbol{\beta}}_0 - \theta_0 \mathbf{1} \right)' \boldsymbol{R}_0^{-1} \left(\tilde{\boldsymbol{\beta}}_0 - \theta_0 \mathbf{1} \right), \end{aligned} \tag{6.7}$$

where $\Phi_i = \Phi(\eta_i)$, the cdf of a standard normal distribution taking argument η_i, where $\eta_i = \tilde{\beta}(\boldsymbol{s}_i)/\sigma_\epsilon$. Twice differentiating (6.7) with respect to $\tilde{\boldsymbol{\beta}}_0$ and negating gives

$$-\frac{\partial^2}{\partial \tilde{\boldsymbol{\beta}}_0^2} \log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y}) \quad = \quad \text{diag}\left\{ \frac{1-e_i}{\sigma_\epsilon^2} \left[\left(\frac{\phi_i}{1-\Phi_i} \right)^2 - \frac{\eta_i \phi_i}{1-\Phi_i} \right] + \frac{e_i}{\sigma_\epsilon^2} \right\} + \frac{1}{\sigma_0^2} \boldsymbol{R}_0^{-1} \quad (6.8)$$

where $\phi_i = \phi(\eta_i)$, the pdf of a standard normal distribution taking argument η_i. Evaluating expression (6.8) at the $\widehat{\boldsymbol{\beta}}_0$ gives $\widehat{\boldsymbol{B}}^{-1}$ from which $\widehat{\boldsymbol{W}}$ follows.

Properties of the $\widehat{W}$ for the Tobit model

Note that if $Z(\boldsymbol{s}_i)$ is observed for all $i = 1, \ldots, n$, then we get $\widehat{\boldsymbol{B}} = (\sigma_\epsilon^{-2}\boldsymbol{I} + \sigma_0^{-2}\boldsymbol{R}_0^{-1})^{-1}$, which is the exact expression for the conditional posterior variance of $\tilde{\boldsymbol{\beta}}_0$ for the equivalent hierarchical model with a Gaussian first stage.

When $Z(\boldsymbol{s}_i)$ is unobserved the data precision, $1/\sigma_\epsilon^2$, is multiplied by h_i, where

$$h_i = \left(\frac{\phi_i}{1-\Phi_i} \right)^2 - \frac{\eta_i \phi_i}{1-\Phi_i}.$$

By Mills' Ratio inequality (Gordon, 1941) we have that for a real constant $x > 0$,

$$\frac{\phi(x)}{1-\Phi(x)} - x > 0, \quad (6.9)$$

and it is clear that inequality (6.9) holds for $x \leq 0$ and hence $h_i > 0$. Furthermore, by the properties of the truncated normal distribution (Barr and Sherrill, 1999), the variance of $Z^-(\boldsymbol{s}_i)$ is given by

$$Var(Z^-(\boldsymbol{s}_i)) = \sigma_\epsilon^2 \left(1 + \frac{\eta_i \phi_i}{1-\Phi_i} - \left(\frac{\phi_i}{1-\Phi_i} \right)^2 \right) < \sigma_\epsilon^2. \quad (6.10)$$

By combining (6.9) and (6.10) we get the following bounds on h_i

$$0 < h_i = 1 - \frac{Var(Z^-(\boldsymbol{s}_i))}{\sigma_\epsilon^2} < 1,$$

for $\sigma_\epsilon^2 \neq 0$. Therefore, the effect of not observing $Z(\boldsymbol{s}_i)$ is to reduce the data precision at location $\boldsymbol{s}_i$ by a factor h_i. For independent random effects, where $\boldsymbol{R}_0$ equals the identity matrix, $\widehat{\boldsymbol{W}}$ is a diagonal matrix with ith diagonal element equal to w_i, where

$$w_i = \begin{cases} \dfrac{\sigma_0^2}{\sigma_0^2 + \sigma_\epsilon^2} & \text{if } Y(\boldsymbol{s}_i) > 0 \\ \dfrac{\sigma_0^2}{\sigma_0^2 + \sigma_\epsilon^2/h_i} & \text{if } Y(\boldsymbol{s}_i) = 0, \end{cases}$$

hence if $Z(\boldsymbol{s}_i)$ is unobserved the weight associated with $\boldsymbol{s}_i$ is shifted downwards.

We can write $\widehat{\boldsymbol{B}}$ as

$$\widehat{\boldsymbol{B}} = \sigma_0^2 \left(\frac{\sigma_0^2}{\sigma_\epsilon^2} \text{diag}\left\{(1-e_i)h_i + e_i\right\} + \boldsymbol{R}_0^{-1} \right)^{-1},$$

and we can see that as $\sigma_0^2/\sigma_\epsilon^2 \to 0$, $\widehat{\boldsymbol{B}} \to \sigma_0^2 \boldsymbol{R}$ and $\widehat{\boldsymbol{W}}$ tends towards zero matrix. Alternatively, Lemma 3.3.1 lets us write $\widehat{\boldsymbol{B}}$ as

$$\widehat{\boldsymbol{B}} = \sigma_0^2 \boldsymbol{R}_0 - \sigma_0^2 \boldsymbol{R}_0 \left(\sigma_\epsilon^2 \text{diag}\left\{ [(1-e_i)h_i + e_i]^{-1} \right\} + \sigma_0^2 \boldsymbol{R}_0 \right)^{-1} \sigma_0^2 \boldsymbol{R}_0,$$

from which we get the following expression for $\widehat{\boldsymbol{W}}$:

$$\widehat{\boldsymbol{W}} = \boldsymbol{R}_0 \left(\frac{\sigma_\epsilon^2}{\sigma_0^2} \text{diag}\left\{ [(1-e_i)h_i + e_i]^{-1} \right\} + \boldsymbol{R}_0 \right)^{-1},$$

and we see that as $\sigma_\epsilon^2/\sigma_0^2 \to 0$, $\widehat{\boldsymbol{W}} \to \boldsymbol{I}$. Therefore, a PCP constructed in this way has the desired property that as the data variance overwhelms that of the random effects we move toward the NCP and opposingly, as the variance of the random effects overwhelms that of the data we move toward the CP.

Maximum likelihood estimates for the Tobit model

To obtain an estimate for $\boldsymbol{B}$ we must evaluate expression (6.8) at the maximum likelihood estimates for $\tilde{\boldsymbol{\beta}}_0$. Closed form solutions for the maximum likelihood estimates are not available for the Tobit model, and hence iterative methods must be employed. We use the EM algorithm (Dempster et al., 1977), details are taken from Amemiya (1984). Given current values $\boldsymbol{\xi}^{(t)} = (\tilde{\boldsymbol{\beta}}_0^{(t)}, \sigma_\epsilon^{2(t)})'$, the EM algorithm finds $\boldsymbol{\xi}^{(t+1)}$ as follows:

1. E-step : Compute $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) = E[\log \pi(\boldsymbol{Z}|\boldsymbol{\xi})|\boldsymbol{\xi}^{(t)}, \boldsymbol{y}]$.

2. M-step : Let $\boldsymbol{\xi}^{(t+1)}$ be the value of $\boldsymbol{\xi}$ that maximises $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$.

Recalling that $e_i = 1$ if $Z(\boldsymbol{s}_i)$ is observed, 0 otherwise, $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$ is given by

$$\begin{aligned}
K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) &= -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n} e_i \left(y(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i) \right)^2 \\
&\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(1-e_i) E\left[\left(Z^-(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i) \right)^2 \mid \boldsymbol{\xi}^{(t)} \right] \\
&= -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n} e_i \left(y(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i) \right)^2 \\
&\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(1-e_i) \left(E\left[Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)} \right] - \tilde{\beta}_0(\boldsymbol{s}_i) \right)^2 \\
&\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(1-e_i) Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)} \right),
\end{aligned}$$

where

$$E[Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}] = \tilde{\beta}_0^{(t)}(\boldsymbol{s}_i) - \sigma_\epsilon^{(t)} \frac{\phi_i^{(t)}}{1-\Phi_i^{(t)}} = z^-(\boldsymbol{s}_i), \tag{6.11}$$

and

$$Var(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}) = \sigma_\epsilon^{2(t)} \left(1 + \frac{\eta_i^{(t)}\phi_i^{(t)}}{1-\Phi_i^{(t)}} - \left(\frac{\phi_i^{(t)}}{1-\Phi_i^{(t)}}\right)^2\right),$$

where $\eta_i^{(t)} = \tilde{\beta}_0^{(t)}(\boldsymbol{s}_i)/\sigma_\epsilon^{(t)}$, $\phi_i^{(t)} = \phi(\eta_i^{(t)})$ and $\Phi_i^{(t)} = \Phi(\eta_i^{(t)})$.

Let $\boldsymbol{y}_z$ be the vector of observations with zeros replaced by $z^-(\boldsymbol{s}_i)$, as defined in (6.11). Then we have

$$K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) = -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\left(y_z(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(1-e_i)Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right). \tag{6.12}$$

Differentiating (6.12) with respect to $\tilde{\beta}_0(\boldsymbol{s}_i)$ and then σ_ϵ^2, it is clear that to maximise $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$ we should set

$$\tilde{\beta}_0^{(t+1)}(\boldsymbol{s}_i) = y_z(\boldsymbol{s}_i),$$

and

$$\sigma_\epsilon^{2(t+1)} = \frac{1}{n}\left[\sum_{i=1}^{n}\left(y_z(\boldsymbol{s}_i) - \tilde{\beta}_0^{(t+1)}(\boldsymbol{s}_i)\right)^2 + (1-e_i)Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right)\right].$$

Once the algorithm has converged we have estimates for $\tilde{\boldsymbol{\beta}}_0$ which can be substituted into (6.8) to find $\widehat{\boldsymbol{B}}$.

Performance of the PCP for the Tobit model

With the data generated for use in Section 6.2.2 we investigate the performance of the PCP. We run two simulation studies. For the first we assume σ_0^2 and σ_ϵ^2 are known. In addition, the true values of $\tilde{\beta}_0(\boldsymbol{s}_i)$ are substituted into (6.6) to compute $\widehat{\boldsymbol{B}}$. The $\text{PSRF}_M(1.1)$ and ESS of θ_0 are plotted in Figure 6.27. We can see rapid convergence and a high ESS for the three cases when $\delta_0 \leq 1$. The performance is not so good for larger relative values of σ_0^2. In particular, the ESS is at its lowest when $\delta_0 = 100$ and there is no spatial correlation between the random effects. However, across the range of variance ratio-effective range pairs, there is far less variability in the performance of the PCP than we saw from either the CP or the NCP, whose results are given in Figures 6.1 and 6.2 respectively.

Figure 6.28 compares the performance of the PCP with the CP and the NCP for the same 400 data sets, under the assumption that the variance parameters are known and the true values of $\boldsymbol{\beta}_0$ have been used to compute $\widehat{\boldsymbol{B}}$. On the top left panel we have the $\text{PSRF}_M(1.1)$ for the CP minus the $\text{PSRF}_M(1.1)$ for the PCP for each of the 400 data sets. The bottom left panel gives the ESS of θ_0 for the PCP minus the ESS of θ_0 for the CP. Therefore, values above the horizontal line indicate superiority of the PCP. We see that for lower values of δ_0, the PCP out performs the CP but our construction of the PCP fails to match the CP for higher values of δ_0.

The equivalent comparisons between the PCP and the NCP are given in the right two

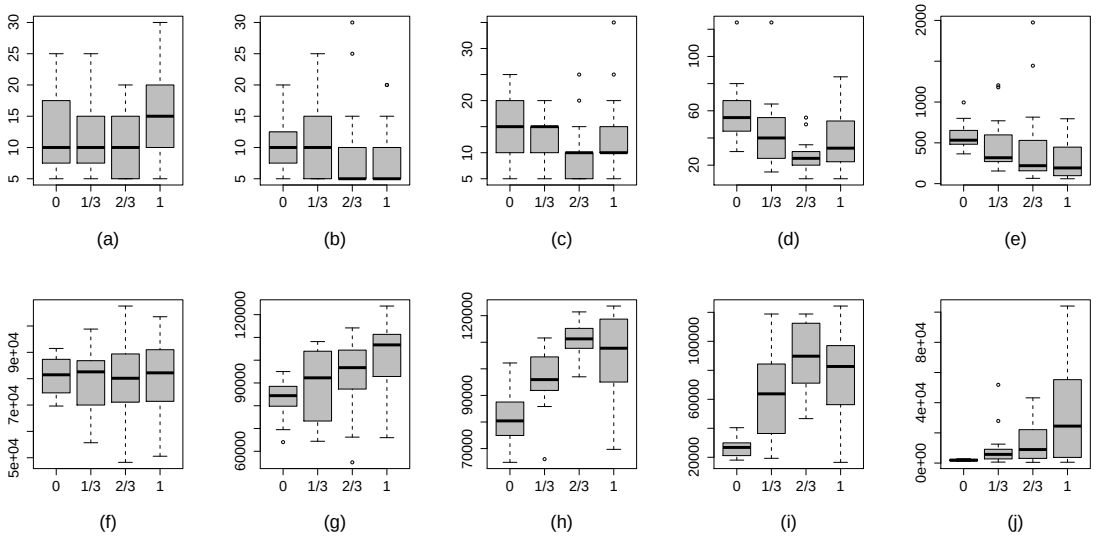

Figure 6.27: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the PCP for the Tobit model for known variance parameters using the known values of $\tilde{\boldsymbol{\beta}}_0$ to compute $\widehat{\boldsymbol{B}}$. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$ of θ_0, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

panels of Figure 6.28. We can see that the PCP outperforms the NCP for almost all data sets both in terms of $\text{PSRF}_M(1.1)$ and ESS for θ_0.

For the second simulation study, the analysis is repeated but now we use the EM algorithm to obtain maximum likelihood estimates for $\boldsymbol{\beta}_0$ and sample from the variance parameters, dynamically updating $\widehat{\boldsymbol{W}}$ within the sampler.

We see from Figures 6.29 and 6.30 that the PCP is robust to changes in the variance ratio and effective range. The differences of the $\text{MPSRF}_M(1.1)$ and ESS of θ_0 between the PCP and each of the CP and the NCP are plotted in Figure 6.31. The PCP clearly outperforms the CP and the NCP in terms of ESS of θ_0. Of the 400 data sets, the PCP returns a greater ESS for θ_0 than the CP for 393 data sets, and one greater than the NCP for 384 data sets. The comparison of the ESS of the variance parameters is given in 6.32. The PCP performs well against the CP in the σ^2_ϵ coordinate and well against the NCP for both variance parameters.

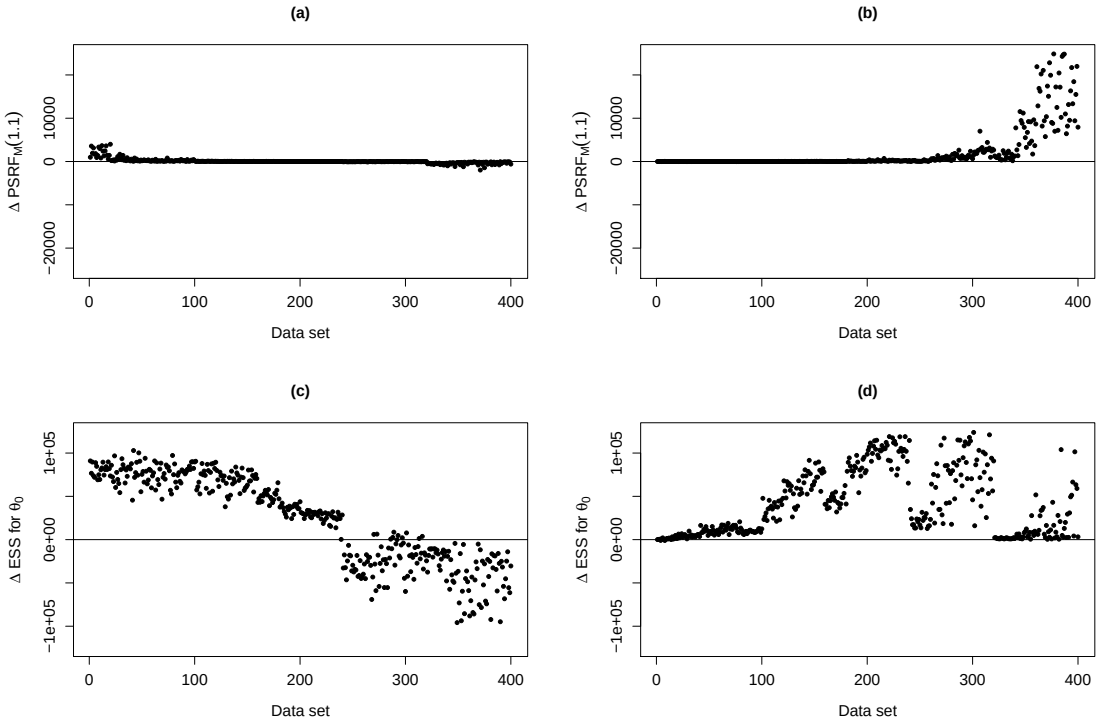

Figure 6.28: Comparison of the $\text{PSRF}_M(1.1)$ and ESS of θ_0 for the PCP with the CP and the NCP of the Tobit model with known variance parameters using the known values of $\tilde{\boldsymbol{\beta}}_0$ to compute $\widehat{\boldsymbol{B}}$. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

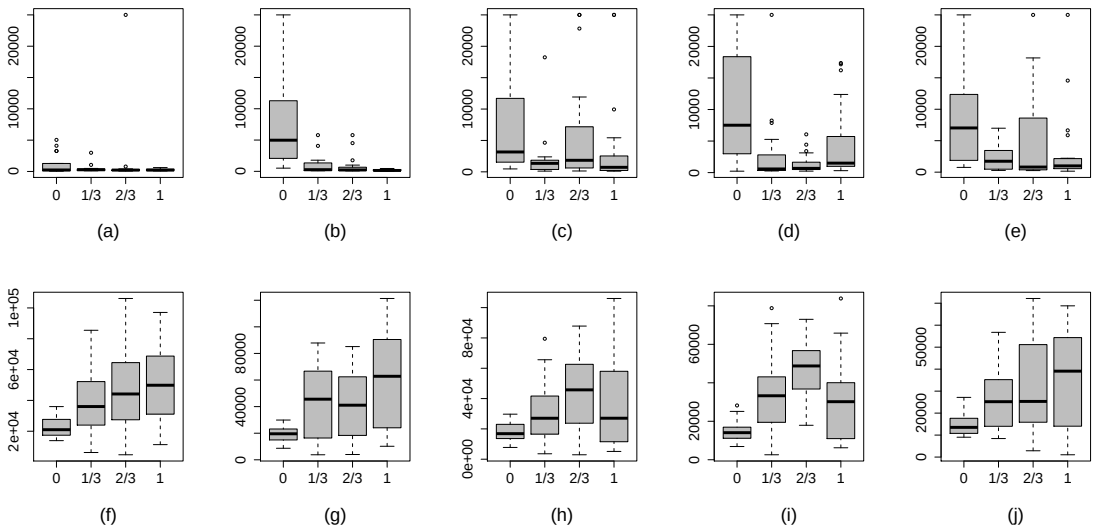

Figure 6.29: $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the PCP of the Tobit model with unknown variance parameters Plots (a)–(e) give the $\text{MPSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

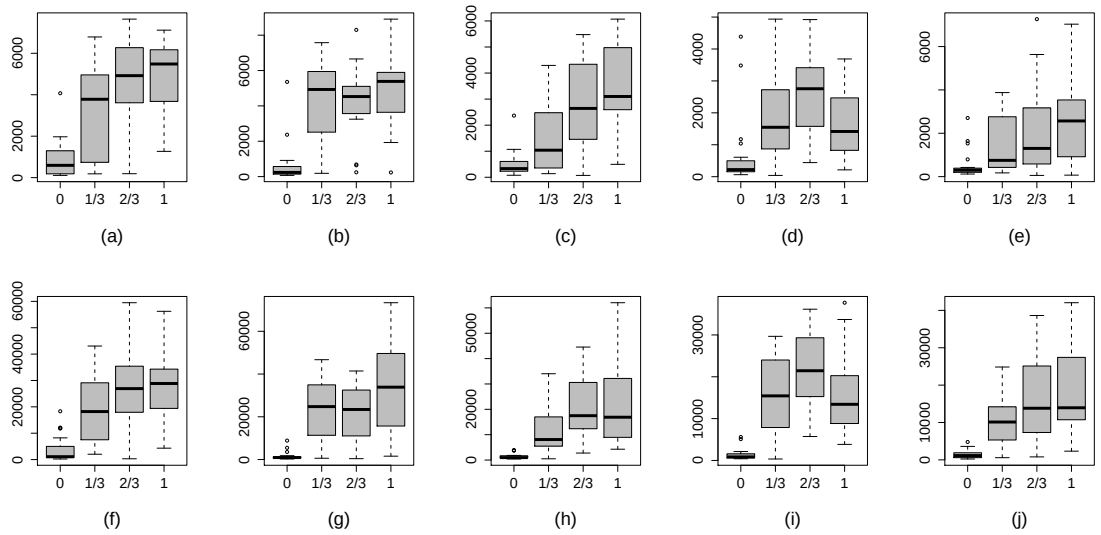

Figure 6.30: ESS of σ_0^2 and σ_ϵ^2 for the PCP of the Tobit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

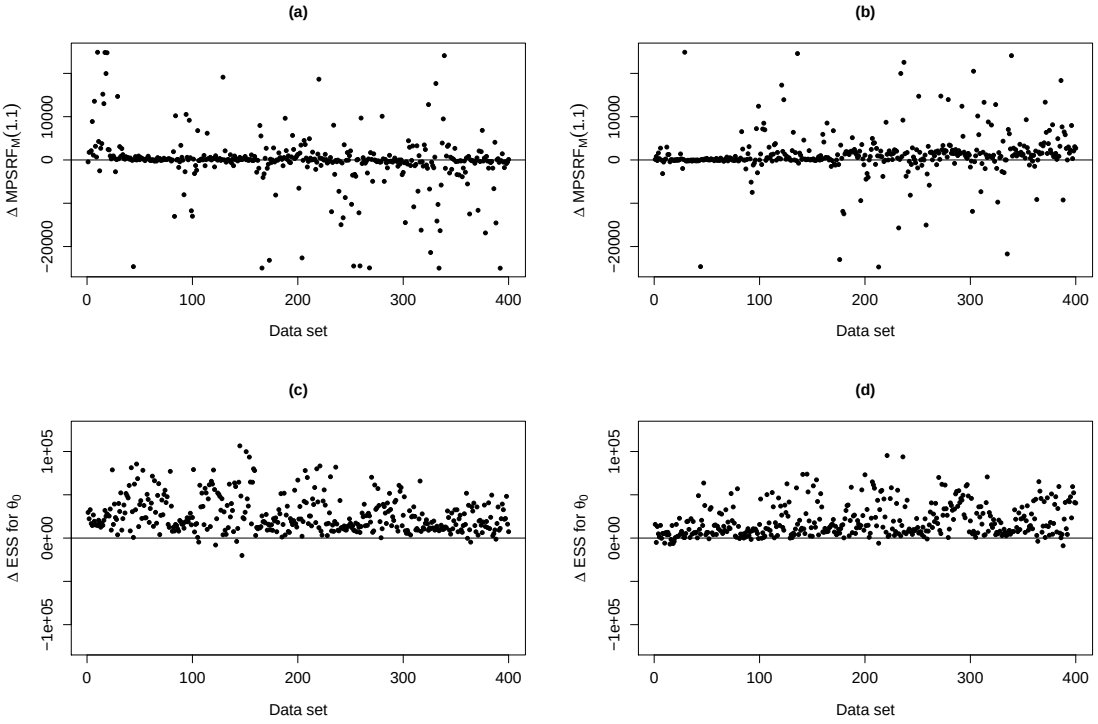

Figure 6.31: Comparison of the MPSRF$_M$(1.1) and ESS of θ_0 for the PCP with the CP and the NCP of the Tobit model with unknown variance parameters. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

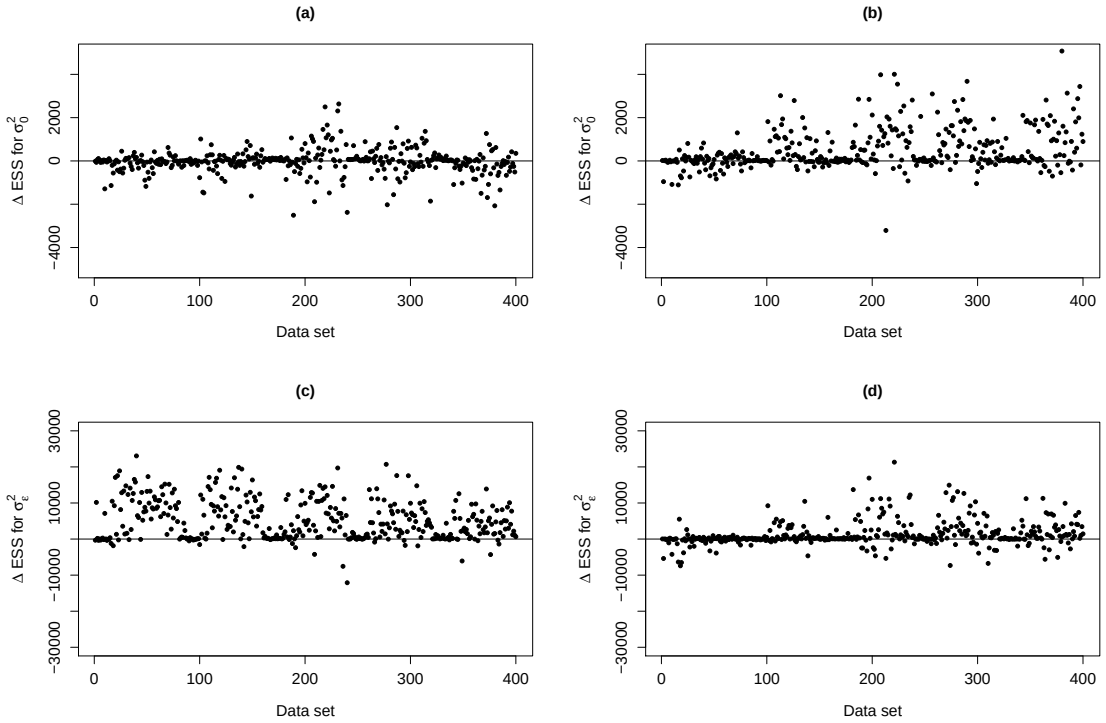

Figure 6.32: Comparison of the ESS of σ_0^2 and σ_ϵ^2 for the PCP with the CP and the NCP of the Tobit model with unknown variance parameters. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

PCP for New York precipitation data

We now apply the PCP to the New York precipitation data analysed in Section 6.2.4. We run five chains of length 25,000 and compute the $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters. We also run the pilot adapted PCP (PAPCP) introduced in Section 5.4.5. Recall that the PAPCP adapts $\boldsymbol{W}$, or in this case $\widehat{\boldsymbol{W}}$, for an additional 1000 iterations after the MPSRF dips below 1.1, and then fixes it for subsequent iterations.

Table 6.7 shows the $\text{MPSRF}_M(1.1)$ and ESS of the model parameters for each of the different fitting methods. Recall from Section 6.2.4 that we got an estimate for the ratio of the variance parameters of $\hat{\delta}_0 = 6.531$. We can see that the PCP is comparable to the CP in terms of $\text{MPSRF}_M(1.1)$ and the ESS of the variance parameters, but returns a lower ESS for θ_0. We saw in the first simulation study earlier in this section that where $\delta_0 > 1$ the respective $\text{PSRF}_M(1.1)$'s for the CP and the PCP are similar but the CP has a higher ESS for θ_0, see Figure 6.28. This may be due to the difficulty in estimating $Var(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y})$ for relatively high values of σ_0^2.

Table 6.7: $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters for the Tobit model.

	$\text{MPSRF}_M(1.1)$	ESS θ_0	ESS σ_0^2	ESS σ_ϵ^2
CP	500	52982	6662	5207
NCP	1875	2001	3878	4270
PCP	515	38316	7055	4297
PAPCP	515	36268	7026	4303

Table 6.8: $\text{MPSRF}_t(1.1)$ and the ESS/s of the model parameters for the Tobit model.

	$\text{MPSRF}_t(1.1)$	ESS/s θ_0	ESS/s σ_0^2	ESS/s σ_ϵ^2
CP	6.76	313.97	39.42	30.81
NCP	25.35	11.84	22.95	25.27
PCP	49.44	31.93	5.88	3.58
PAPCP	24.8	60.25	11.67	7.15

Table 6.8 gives a comparison of the fitting strategies with the measures adjusted for computation time. Here the relative computational efficiency of the CP is clear.

6.4.2 Partial centering for the probit model

We now turn our attention to the PCP for the probit model. As in Section 6.4.1 we are without access to the exact conditional posterior covariance matrix of $\tilde{\boldsymbol{\beta}}_0$ and therefore we evaluate the negated Hessian matrix of the log full conditional distribution at the MLE, $\widehat{\boldsymbol{\beta}}_0$. This gives $\widehat{\boldsymbol{B}}^{-1}$ and in turn allows us to compute $\widehat{\boldsymbol{W}}$.

We begin by writing the log-conditional posterior distribution of $\tilde{\boldsymbol{\beta}}_0$ as

$$\begin{aligned}\log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y}) &\propto \log \pi(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0) + \log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0) \\ &= \sum_{i=1}^{n}(1-y_i)\log(1-\Phi_i) + y_i \log(\Phi_i) \\ &\quad -\frac{n}{2}\log 2\pi - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2\sigma_0^2}(\tilde{\boldsymbol{\beta}}_0 - \theta_0\mathbf{1})'\boldsymbol{R}_0^{-1}(\tilde{\boldsymbol{\beta}}_0 - \theta_0\mathbf{1}),\end{aligned}$$

where $\Phi_i = \Phi(\eta_i)$, the cdf of a standard normal distribution taking argument η_i, where $\eta_i = (\tilde{\beta}_0(\boldsymbol{s}_i) - z^*)/\sigma_\epsilon$. Twice differentiating with respect to $\tilde{\boldsymbol{\beta}}_0$ and negating gives

$$-\frac{\partial^2}{\partial \tilde{\boldsymbol{\beta}}_0^2}\log \pi(\tilde{\boldsymbol{\beta}}_0|\theta_0, \boldsymbol{y}) = \text{diag}\left\{\frac{1}{\sigma_\epsilon^2}\left[(1-y_i)h_i + y_i g_i\right]\right\} + \frac{1}{\sigma_0^2}\boldsymbol{R}_0^{-1} \tag{6.13}$$

where

$$h_i = \left(\frac{\phi_i}{1-\Phi_i}\right)^2 - \frac{\eta_i\phi_i}{1-\Phi_i}, \quad g_i = \left(\frac{\phi_i}{\Phi_i}\right)^2 + \frac{\eta_i\phi_i}{\Phi_i},$$

and $\phi_i = \phi(\eta_i)$, the pdf of a standard normal distribution taking argument η_i. Evaluating expression (6.13) at the $\widehat{\boldsymbol{\beta}}_0$ gives $\widehat{\boldsymbol{B}}^{-1}$.

Properties of $\widehat{\boldsymbol{W}}$ for the probit model

We saw in Section 6.4.1 that $0 < h_i < 1$. By similar argument it is easy to show that $0 < g_i < 1$. Therefore, by only being able to observe the sign of $Z(\boldsymbol{s}_i)$ we reduce the data precision by either h_i, if $Z(\boldsymbol{s}_i)$ is negative, or by g_i if $Z(\boldsymbol{s}_i)$ is positive.

Recall that $\widehat{\boldsymbol{W}} = \boldsymbol{I} - \widehat{\boldsymbol{B}}\boldsymbol{C}_2^{-1}$, where here $\boldsymbol{C}_2 = \sigma_0^2\boldsymbol{R}_0$ is the prior covariance matrix of $\tilde{\boldsymbol{\beta}}_0$. Writing $\widehat{\boldsymbol{B}}$ as

$$\widehat{\boldsymbol{B}} = \sigma_0^2\left[\frac{\sigma_0^2}{\sigma_\epsilon^2}\text{diag}\left\{(1-y_i)h_i + y_i g_i\right\} + \boldsymbol{R}_0^{-1}\right]^{-1},$$

we see that as $\sigma_0^2/\sigma_\epsilon^2 \to 0$ then $\widehat{\boldsymbol{B}} \to \boldsymbol{R}_0$ and $\widehat{\boldsymbol{W}} \to \mathbf{0}$. Hence as the data variance increases relative to that of the random effects the parameterisation tends to the NCP. Alternatively, by Lemma 3.3.1 we can write

$$\widehat{\boldsymbol{B}} = \sigma_0^2\boldsymbol{R} - \sigma_0^2\boldsymbol{R}_0\left(\sigma_\epsilon^2\text{diag}\left\{(1-y_i)h_i + y_i g_i\right\}^{-1} + \sigma_0^2\boldsymbol{R}_0\right)^{-1}\sigma_0^2\boldsymbol{R}_0$$

which implies that

$$\begin{aligned}\widehat{\boldsymbol{W}} &= \sigma_0^2 \boldsymbol{R}_0 \left(\sigma_\epsilon^2 \text{diag}\left\{(1-y_i)h_i + y_i g_i\right\}^{-1} + \sigma_0^2 \boldsymbol{R}_0\right)^{-1} \\ &= \boldsymbol{R}_0 \left(\frac{\sigma_\epsilon^2}{\sigma_0^2}\text{diag}\left\{(1-y_i)h_i + y_i g_i\right\}^{-1} + \boldsymbol{R}_0\right)^{-1}. \end{aligned} \tag{6.14}$$

We see from equation (6.14) that as the variance of the random effects grows large relative to that of the data and $\sigma_\epsilon^2/\sigma_0^2 \to 0$, $\widehat{\boldsymbol{W}} \to \boldsymbol{I}$ and we recover the CP.

Maximum likelihood estimates for the probit model

Closed form solutions for the maximum likelihood estimates are not available for the probit model, and hence iterative methods must be employed.

Given current values $\boldsymbol{\xi}^{(t)} = (\tilde{\boldsymbol{\beta}}_0^{(t)}, \sigma_\epsilon^{2(t)})'$, the EM algorithm finds $\boldsymbol{\xi}^{(t+1)}$ as follows:

1. E-step : Compute $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) = E[\log \pi(\boldsymbol{Z}|\boldsymbol{\xi})|\boldsymbol{\xi}^{(t)}, \boldsymbol{y}]$.
2. M-step : Let $\boldsymbol{\xi}^{(t+1)}$ be the value of $\boldsymbol{\xi}$ that maximises $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$.

$$\begin{aligned} K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) &= -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n (1-y_i) E\left[\left(Z^-(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 \mid \boldsymbol{\xi}^{(t)}\right] \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n y_i E\left[\left(Z^+(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 \mid \boldsymbol{\xi}^{(t)}\right] \\ &= -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n (1-y_i)\left(E\left[Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right] - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n (1-y_i) Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right) \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n y_i \left(E\left[Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right] - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^n y_i Var\left(Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right), \end{aligned}$$

where

$$E[Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}] = \tilde{\beta}_0^{(t)}(\boldsymbol{s}_i) - \sigma_\epsilon^{(t)}\frac{\phi_i^{(t)}}{1-\Phi_i^{(t)}} = z^-(\boldsymbol{s}_i), \tag{6.15}$$

$$E[Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}] = \tilde{\beta}_0^{(t)}(\boldsymbol{s}_i) + \sigma_\epsilon^{(t)}\frac{\phi_i^{(t)}}{\Phi_i^{(t)}} = z^+(\boldsymbol{s}_i), \tag{6.16}$$

and

$$Var(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}) = \sigma_\epsilon^{2(t)}\left(1 + \frac{\eta_i^{(t)}\phi_i^{(t)}}{1-\Phi_i^{(t)}} - \left(\frac{\phi_i^{(t)}}{1-\Phi_i^{(t)}}\right)^2\right),$$

$$Var(Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}) = \sigma_\epsilon^{2(t)}\left(1 - \frac{\eta_i^{(t)}\phi_i^{(t)}}{\Phi_i^{(t)}} - \left(\frac{\phi_i^{(t)}}{\Phi_i^{(t)}}\right)^2\right),$$

where $\eta_i^{(t)} = (\tilde{\beta}_0^{(t)}(\boldsymbol{s}_i) - z^*)/\sigma_\epsilon^{2(t)}$, $\phi_i^{(t)} = \phi(\eta_i^{(t)})$ and $\Phi_i^{(t)} = \Phi(\eta_i^{(t)})$.

Let $\boldsymbol{y}_z$ be the vector of observations with zeros replaced by $z^-(\boldsymbol{s}_i)$ and ones replaced with $z^+(\boldsymbol{s}_i)$, as defined in (6.15) and (6.16). Then we have

$$\begin{aligned} K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)}) &= -\frac{n}{2}\log\sigma_\epsilon^2 - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}\left(y_z(\boldsymbol{s}_i) - \tilde{\beta}_0(\boldsymbol{s}_i)\right)^2 \\ &\quad -\frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(1-y_i)Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right) + y_i Var\left(Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right). \quad (6.17) \end{aligned}$$

Differentiating (6.17) with respect to $\tilde{\beta}_0(\boldsymbol{s}_i)$ and then σ_ϵ^2, it is clear that to maximise $K(\boldsymbol{\xi}|\boldsymbol{\xi}^{(t)})$ we should set

$$\tilde{\beta}_0^{(t+1)}(\boldsymbol{s}_i) = y_z(\boldsymbol{s}_i),$$

and

$$\begin{aligned} \sigma_\epsilon^{2(t+1)} &= \frac{1}{n}\left[\sum_{i=1}^{n}\left(y_z(\boldsymbol{s}_i) - \tilde{\beta}_0^{(t+1)}(\boldsymbol{s}_i)\right)^2\right. \\ &\quad \left. +(1-y_i)\,Var\left(Z^-(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right) + y_i Var\left(Z^+(\boldsymbol{s}_i)|\boldsymbol{\xi}^{(t)}\right)\right]. \end{aligned}$$

Once the algorithm has converged we have estimates for $\tilde{\boldsymbol{\beta}}_0$ which can be substituted into (6.13) to find $\widehat{\boldsymbol{B}}$.

Performance of the PCP of the probit model

We fit the PCP for the probit model to the 400 data sets that were generated to compare the CP and the NCP for probit model in Section 6.3.2. We begin by fixing the variance parameters at their true values and using the true values of $\tilde{\boldsymbol{\beta}}_0$ to generate the matrix $\widehat{\boldsymbol{B}}$. Results are given in Figure 6.33. We can see for values of $\delta_0 \leq 1$ that we have rapid convergence, but the performance deteriorates as σ_0^2 becomes relatively large. The ESS of θ_0 reduces accordingly. For a fixed δ_0, mixing is poorest when there is no correlation between the random effects.

Figure 6.34 compares the results of the PCP with those obtained for the CP and NCP (plotted in Figures 6.18 and 6.19 respectively). Positive values indicate a superior performance for the PCP. The left hand panels shows that for low values of δ_0 the PCP outperforms the CP, but as the variance ratio increases and the CP improves, the PCP fails to deliver the optimal matrix $\widehat{\boldsymbol{W}}$. The right hand panels shows that the PCP increasingly outperforms the NCP as δ_0 increases.

We now go on to consider the case of unknown variance parameters. The $\text{MPSRF}_M(1.1)$ and ESS of θ_0 are plotted in Figure 6.35. We can see that the performance of the sampler is more consistent across the range of variance ratio-effective range pairs than was the case when we fixed the variance parameters at their true values, see Figure 6.33.

Looking at Figure 6.36 we see that the performance of the variance parameters for the PCP is similar to that of the CP and the NCP, given in Figures 6.22 and 6.23 respectively. The ESS of σ_0^2 falls with increasing δ_0 and increases with increasing effective range, d_0.

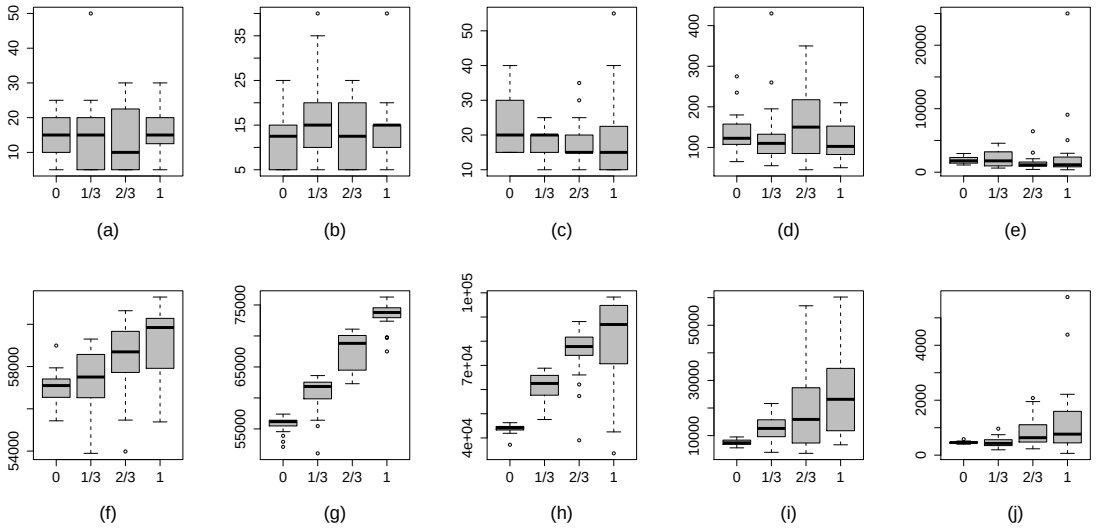

Figure 6.33: $\text{PSRF}_M(1.1)$ and the ESS of θ_0 for the PCP of the probit model with known variance parameters using the known values of $\tilde{\boldsymbol{\beta}}_0$ to compute $\widehat{\boldsymbol{B}}$. Plots (a)–(e) give the $\text{PSRF}_M(1.1)$ of θ_0, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

The ESS of σ_ϵ^2 is fairly constant across the different values of δ_0 and d_0.

Figure 6.37 compares the $\text{MPSRF}_M(1.1)$ and ESS of θ_0 for the PCP with those obtained for the CP and the NCP, which are plotted in Figures 6.20 and 6.21 respectively. Values above the horizontal line indicate a superior performance for the PCP. From the top left panel of Figure 6.37 we can see that the PCP does better than the CP for the lower values of δ_0, but this is reversed as the variance ratio is increased. The bottom left panel compares the ESS of θ_0 for the PCP and the CP. The wave like pattern indicates that the PCP does better than the CP when there is a shorter effective range. This is because the PCP is robust to changes in effective range whereas the CP improves with increasing effective range. Looking at the panels on the right we see that an increase in δ_0 adversely affects the NCP much more heavily than the PCP and so the advantage of the PCP over the NCP grows with δ_0.

Figure 6.38 compares the ESS of the variance parameters obtained with the PCP to those obtained with either the CP or the NCP. From the top left panel we can see that the ESS for σ_0^2 is regularly higher for the CP than the PCP. The bottom left panel shows us that the PCP returns a higher ESS for σ_ϵ^2 than the CP, especially for smaller values of δ_0 and for shorter effective ranges. We can see on the right that only for low values of δ_0 does the NCP have a higher ESS for σ_0^2 than the PCP. There is little difference in the ESS of σ_ϵ^2 between the PCP and the NCP, with points evenly spread either side of the horizontal line.

PCP for Californian ozone data

We now fit the PCP of the probit model to the Californian ozone data analysed in Section 6.3.3. In that section we got an estimate for the variance ratio, $\hat{\delta}_0 = 10.498$. Table 6.9 gives

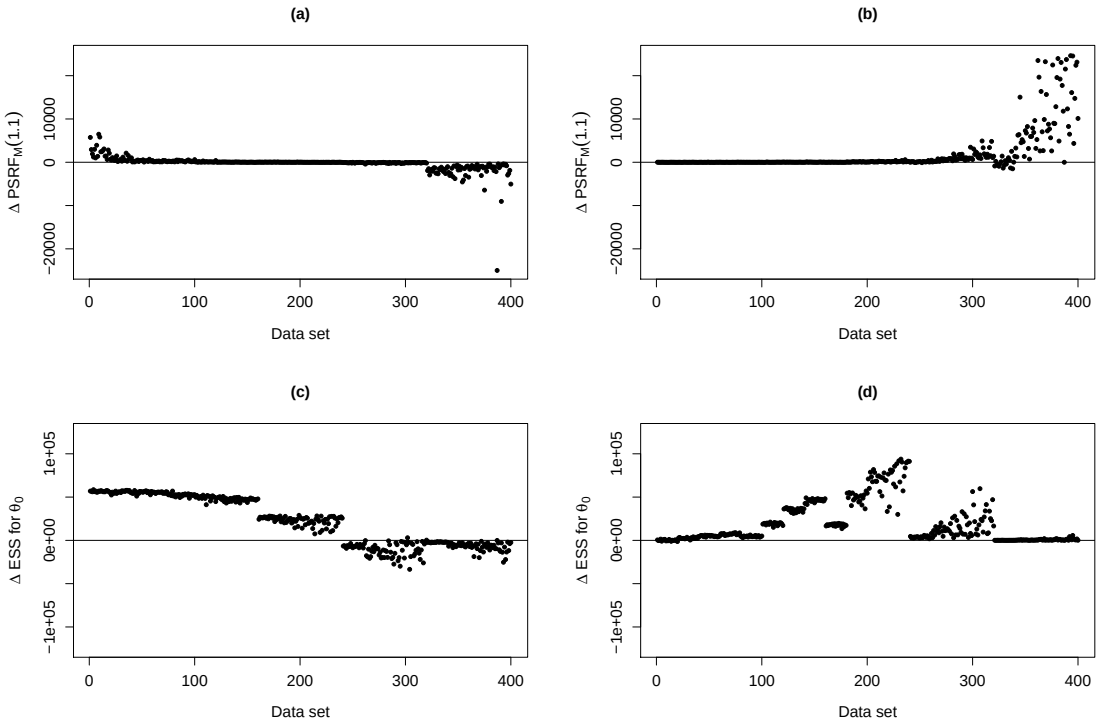

Figure 6.34: Comparison of the $\text{PSRF}_M(1.1)$ and ESS of θ_0 for the PCP with the CP and the NCP of the probit model with known variance parameters using the known values of $\tilde{\boldsymbol{\beta}}_0$ to compute $\widehat{\boldsymbol{B}}$. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

the $\text{MPSRF}_M(1.1)$ and ESS of the model parameters for different fitting strategies. We can see that convergence is slow for all methods, each requiring around 10,000 iterations before the MPSRF falls below 1.1. This is due to the poor mixing in the σ_0^2 coordinate, we achieve an ESS of less than 300 out of a total 125,000 MCMC samples for all of the parameterisations. We do see that the ESS for θ_0 is the greatest for the CP, this is not surprising given the value of $\hat{\delta}_0$ that we observed.

Table 6.9: $\text{MPSRF}_M(1.1)$ and the ESS of the model parameters for the probit model.

	$\text{MPSRF}_M(1.1)$	ESS θ	ESS σ_0^2	ESS σ_ϵ^2
CP	9925	19245	268	1887
NCP	10335	740	271	1899
PCP	9705	4887	219	1888
PAPCP	9705	5318	254	1908

Table 6.10 compares the fitting strategies in terms of the time adjusted measures $\text{MPSRF}_t(1.1)$ and ESS/s. Here we can see that the CP is the preferred fitting method.

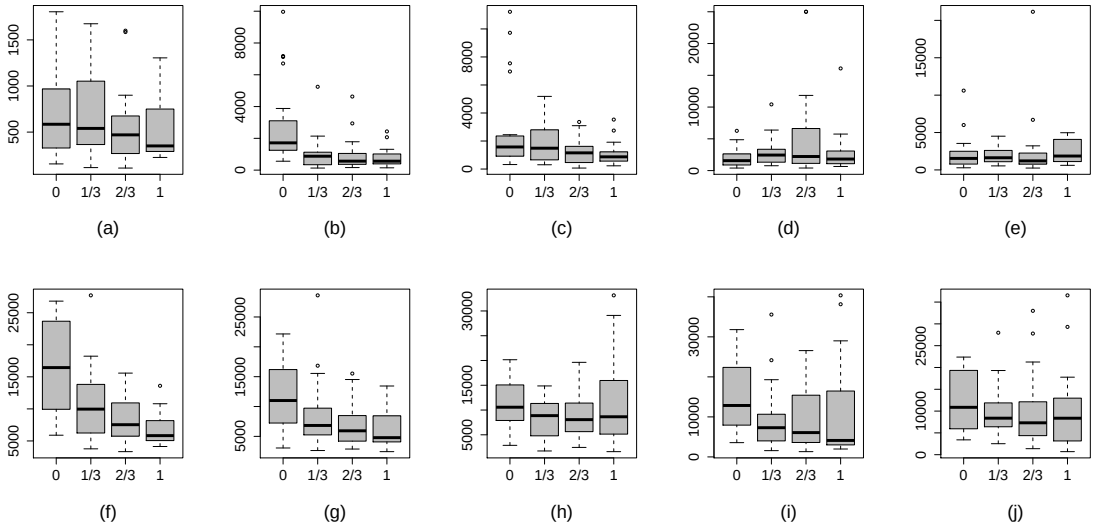

Figure 6.35: $\text{MPSRF}_M(1.1)$ and the ESS of θ_0 for the PCP of the probit model with unknown variance parameters. Plots (a)–(e) give the $\text{MPSRF}_M(1.1)$, plots (f)–(j) the ESS of θ_0. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

Table 6.10: $\text{MPSRF}_t(1.1)$ and the ESS/s of the model parameters for the probit model.

	$\text{MPSRF}_t(1.1)$	ESS/s θ_0	ESS/s σ_0^2	ESS/s σ_ϵ^2
CP	29.85	51.18	0.71	5.02
NCP	31.09	1.97	0.72	5.05
PCP	188.67	2.01	0.09	0.78
PAPCP	99.07	4.17	0.20	1.50

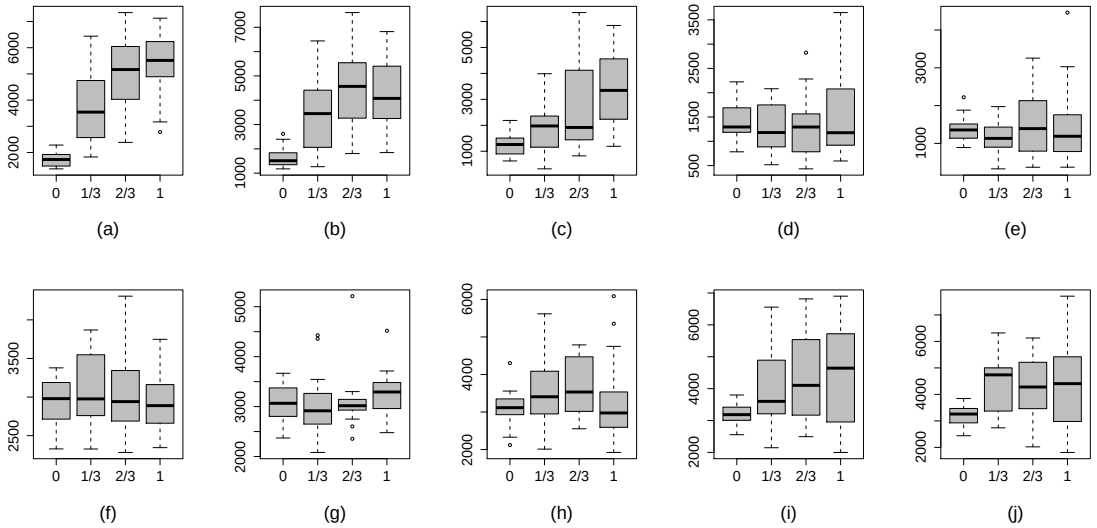

Figure 6.36: ESS of σ_0^2 and σ_ϵ^2 for the PCP of the probit model with unknown variance parameters. Plots (a)–(e) give the ESS of σ_0^2, plots (f)–(j) the ESS of σ_ϵ^2. L–R $\delta_0 = 0.01, 0.1, 1, 10, 100$. Within each plot effective ranges of 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$ are used.

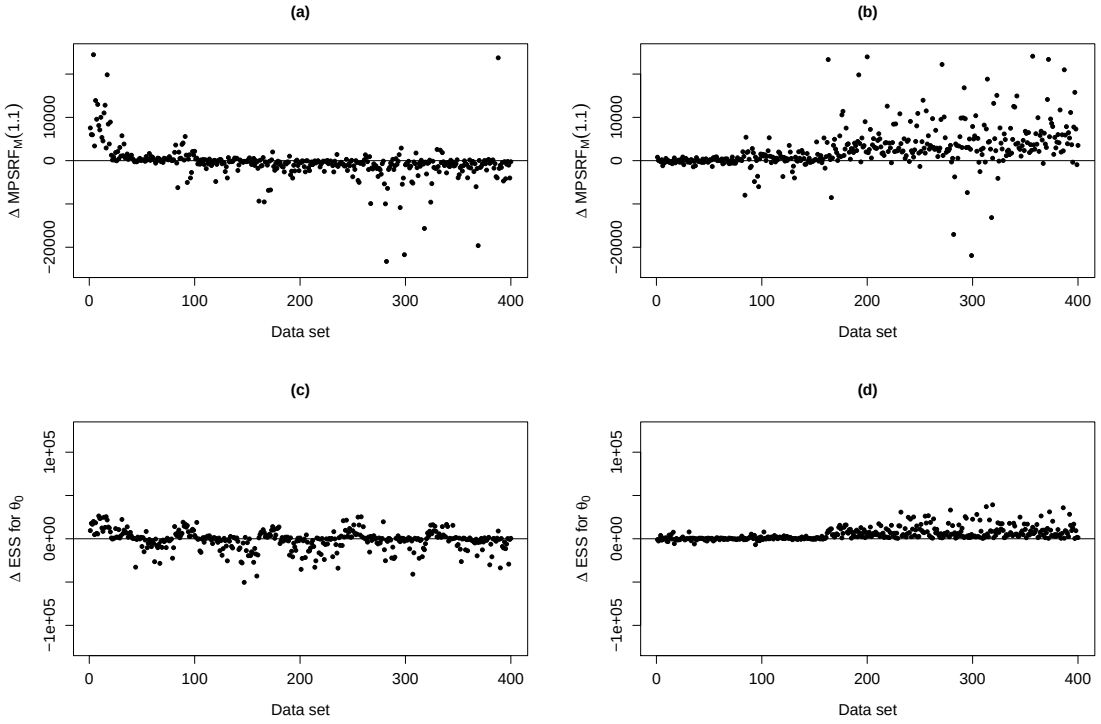

Figure 6.37: Comparison of the MPSRF$_M$(1.1) and ESS for θ_0 for the PCP with the CP and the NCP of the probit model with unknown variance parameters. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

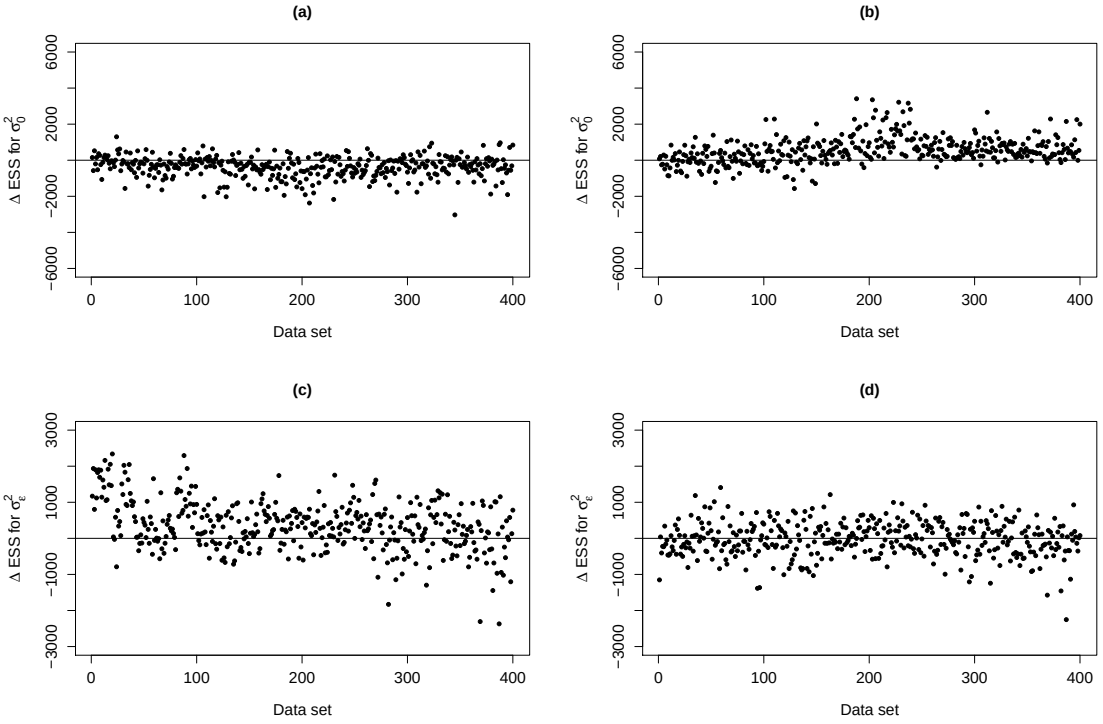

Figure 6.38: Comparison of the ESS of σ_0^2 and σ_ϵ^2 for the PCP with the CP and the NCP of the probit model with unknown variance parameters. Panels (a) and (c) compare PCP with CP. Panels (b) and (d) compare PCP with NCP.

6.5 Summary

In this chapter we have investigated the efficiency of Gibbs samplers produced under different model parameterisations for non-Gaussian spatial models. Parameterisations are compared in terms of the $\text{MPSRF}_M(1.1)$ and the ESS of model parameters using simulated and real data sets. We have focused on the spatial Tobit and spatial probit models. For both models, we see that the CP is preferable to the NCP when the nugget variance is low relative to the marginal variance of the spatially correlated random effects. Furthermore, strengthening spatial correlation improves the performance of the CP, but impairs that of the NCP.

We showed how to construct a PCP for non-Gaussian data which is analogous to the one employed for Gaussian data. The construction used here relies on an estimate of the conditional posterior covariance matrix of the random effects, denoted by $\widehat{\boldsymbol{B}}$. We saw that in many cases the PCP outperforms the CP and the NCP, but when the random effects variance is relatively high, the accuracy of $\widehat{\boldsymbol{B}}$ can be diminished and with it the efficiency of the PCP. In these cases, in which the CP is at its most effective, the PCP may not perform as well as the CP. However, we saw that the efficiency of the PCP is robust to changes in the ratio of the variance components and the effective range, and so it becomes a useful strategy when these quantities are unknown.

Chapter 7

Conclusions and future work

In this final chapter we summarise the work in this thesis and give details of two possible extensions to the analysis presented in the preceding chapters. In Section 7.1 we highlight the important results and detail the limitations of the approaches undertaken. In Section 7.2 we give brief details of how to extend our work to spatio-temporal and multivariate spatial models, and state some of questions that emerge as a consequence of these extensions.

7.1 Conclusions

The goal of this thesis is to investigate the impact of the correlation structure of latent random processes upon the efficiency of the Gibbs sampler used for inference. It is known that for a normal linear hierarchical model (NLHM) with independent random effects that if the data precision is high relative to that of the random effects, then the centred parameterisation (CP) gives us a more efficient sampler than the non-centred parameterisation (NCP). This work shows how the strength of spatial correlation between realisations of a latent Gaussian process impacts upon the convergence rates associated with the CP and the NCP.

In Chapter 3 we looked at a spatially varying coefficients models with Gaussian errors at the first stage. We compared the CP and the NCP in terms of the exact convergence rates of their associated Gibbs samplers, where the rates are computable for known posterior precision matrices. We show that with an exponential correlation structure for the random effects, that as we strengthen the spatial correlation the convergence rate is quickened for the CP but slowed for the NCP.

The notion that the CP becomes more efficient with stronger correlation and the NCP less so, is borne out when we look at covariance tapering and geometric anisotropic exponential correlation functions. Removing long range correlation through tapering favours the NCP but degrades the performance of the CP. Whereas when we increase the correlation in one direction by inducing geometric anisotropy, the performance of the CP is improved and the performance of the NCP is worsened.

In Chapter 3 it is also demonstrated that if there is any correlation between the random effects that they should be updated together to achieve faster convergence. This does

not, however, take into account the additional computational burden of jointly updating variables of higher dimension within a Gibbs sampler. These practical issues are addressed in Chapter 4 where we look at the implementation of Gibbs samplers for the CP and the NCP of spatially varying coefficient models. We assume that covariance matrices are known only up to set of parameters. Therefore the posterior precision matrices are unknown and we cannot compute the exact convergence rate as we did in Chapter 3. Instead we compare the CP and the NCP in terms of the potential scale reduction factor (PSRF) and the effective sample size (ESS) of the model parameters. It is shown for both simulated and real data examples that the CP performs best when the variance of the random effects is much greater than that of the data, and when there is strong spatial correlation. The NCP performs best when the random effects variance is relatively low and when there is weak spatial correlation.

The work of Chapters 3 and 4 adds weight to the notion that the CP and the NCP are complimentary pairs, where one does well the other does badly. The performance of either parameterisation depends on the data through the covariance parameters in the model, which are typically unknown. Therefore, we cannot know *a priori* which of the CP or the NCP to implement. In Chapter 5 we tackle this problem by constructing a parameterisation whose performance is robust to the data. By computing the conditional posterior correlation between the random and global effects we are able to produce a partially centred parameterisation (PCP) for a three stage NLHM. The PCP is determined by a weight matrix $\boldsymbol{W}$ and returns a Gibbs sampler that converges immediately and produces independent samples from the marginal posterior distribution of the global effects. Therefore we can optimise the performance of any model that can be written as a three stage NLHM.

The derivation of the PCP is conditional on the covariance matrices. When these are known only up to a set of covariance parameters we show that the PCP can be updated dynamically within the Gibbs sampler without disturbing the stationary distribution of the Markov chain. The PCP is shown to converge more quickly and to return samples from the posterior distributions of the global effects with lower autocorrelation than either the CP or the NCP.

However, the PCP is a computationally demanding fitting strategy. We have to update all random effects as one block and all global effects as another. Also, when the covariance parameters are unknown we have to repeatedly update the associated weight matrix $\boldsymbol{W}$. Pilot adapted PCPs offer some reduction in computation time for little reduction in performance but are still slower to run than the CP or the NCP.

In Chapter 6 we consider models for non-Gaussian data for which we cannot compute the exact convergence rate of their associated Gibbs samplers. Therefore we again compare parameterisations in terms of the PSRF and the ESS. We looked at spatial Tobit and spatial probit models. The simulated and real data examples show that, just as for Gaussian data, the CP is best when the ratio of the random effects variance to the nugget variance is high, and when there is strong spatial correlation, with the opposite holding for the NCP.

We demonstrated how to construct a PCP for non-Gaussian models with a method analogous to the one used to construct the PCP for Gaussian data in Chapter 5. This requires the estimation of the conditional posterior covariance matrix of the random effects, $Var(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y})$. The performance is good for relatively low ratios of the random effects variance to the nugget variance, but the performance is degraded as the variance ratio increases as this makes estimating $Var(\boldsymbol{\beta}|\boldsymbol{\theta}, \boldsymbol{y})$ becomes more problematic.

7.2 Future work

In this thesis we have only considered univariate spatial models. In this section we outline how we might extend our models to spatio-temporal or multivariate models. For discussions of spatio-temporal modelling see, for example, Gelfand et al. (2010, Chapter 23) or Banerjee et al. (2003, Chapter 8), and for mutlivariate modelling see Gelfand et al. (2010, Chapter 28) or Banerjee et al. (2003, Chapter 7). We are concerned with how the increasingly complicated covariance structures resulting from these extensions affects the efficiency of the Gibbs samplers for the different model parameterisations.

7.2.1 Spatio-temporal models

It is natural to extend the spatial models discussed in this thesis to model spatial data that is collected over time at equally spaced intervals. We let $Y(\boldsymbol{s}_i, t)$ be the response at site $\boldsymbol{s}_i$ and at time t. We model $Y(\boldsymbol{s}_i, t)$ as

$$Y(\boldsymbol{s}_i,t) = \theta_0+\beta_0(\boldsymbol{s}_i,t)+\{\theta_1+\beta_1(\boldsymbol{s}_i,t)\}x_1(\boldsymbol{s}_i,t)+\ldots+\{\theta_{p-1}+\beta_{p-1}(\boldsymbol{s}_i,t)\}x_{p-1}(\boldsymbol{s}_i,t)+\epsilon(\boldsymbol{s}_i,t), \tag{7.1}$$

for $i = 1, \ldots, n$, and $t = 1, \ldots, T$, where $\epsilon(\boldsymbol{s}_i, t) \overset{ind}{\sim} N(0, \sigma^2_\epsilon)$, and $x_k(\boldsymbol{s}_i, t)$ is the value of the kth covariate at location $\boldsymbol{s}_i$ and at time t, for $k = 1, \ldots, p-1$. Global regression coefficients θ_k are perturbed by realisations of space-time processes $\beta_k(\boldsymbol{s}_i, t)$, and hence we have a model with spatio-temporally varying coefficients.

It is straightforward to see that we can write model (7.1) in the form of the three stage NLHM given in (2.8). We collect the responses into a vector $\boldsymbol{Y} = (\boldsymbol{Y}'(\boldsymbol{s}_1), \ldots, \boldsymbol{Y}'(\boldsymbol{s}_n))'$, where $\boldsymbol{Y}(\boldsymbol{s}_i) = (Y(\boldsymbol{s}_i, 1), \ldots, Y(\boldsymbol{s}_i, T))'$, and gather measurements for the kth covariate into the vector $\boldsymbol{x}_k = (\boldsymbol{x}'_k(\boldsymbol{s}_1), \ldots, \boldsymbol{x}'_k(\boldsymbol{s}_n))'$ where $\boldsymbol{x}_k(\boldsymbol{s}_i) = (x_k(\boldsymbol{s}_i, 1), \ldots, x_k(\boldsymbol{s}_i, T))'$. The vector $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})'$ contains the global regression coefficients and has a Gaussian prior distribution with mean $\boldsymbol{m}$ and covariance $\boldsymbol{C}_3$. We let $\boldsymbol{\beta}_k = (\boldsymbol{\beta}'_k(\boldsymbol{s}_1), \ldots, \boldsymbol{\beta}'_k(\boldsymbol{s}_n))'$ where $\boldsymbol{\beta}_k(\boldsymbol{s}_i) = (\beta_k(\boldsymbol{s}_i, 1), \ldots, \beta_k(\boldsymbol{s}_i, T))'$. Centering the random effects on their associated global parameter introduces $\tilde{\boldsymbol{\beta}}_k = \boldsymbol{\beta}_k + \theta_k \mathbf{1}$, where $\mathbf{1}$ is a $nT \times 1$ vector of ones. We now write model (7.1) as

$$\begin{aligned} \boldsymbol{Y}|\tilde{\boldsymbol{\beta}} &\sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \sigma^2_\epsilon \boldsymbol{I}) \\ \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} &\sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2) \\ \boldsymbol{\theta} &\sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned}$$

where $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0', \ldots, \tilde{\boldsymbol{\beta}}_{p-1}')'$ and $\boldsymbol{X}_1 = (\boldsymbol{I}, \boldsymbol{D}_1, \ldots, \boldsymbol{D}_{p-1})$ is a $nT \times nTp$ matrix with $\boldsymbol{D}_k = \text{diag}(\boldsymbol{x}_k)$. The matrix $\boldsymbol{X}_2$ is a $nTp \times p$ block diagonal matrix, with blocks made up of vectors of ones of length nT. Using the results of Chapter 2 we can compute the conditional posterior covariances of $\tilde{\boldsymbol{\beta}}$, $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, and the calculate exact convergence rates for the CP and the NCP of model (7.1). Moreover, we can compute the $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$ to find the optimal PCP, where $\boldsymbol{B} = (\sigma_\epsilon^{-2}\boldsymbol{X}_1'\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1})^{-1}$.

We have yet to specify the covariance structure of the space-time processes, $\boldsymbol{C}_2$. Suppose that we follow Sahu et al. (2011) and model the $\beta_k(\boldsymbol{s}_i, t)'s$ as a zero mean Gaussian process with separable covariance structure, such that

$$Cov(\beta_k(\boldsymbol{s}_i, t_l), \beta_k(\boldsymbol{s}_j, t_m) = \sigma_k^2 \rho_k^s(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi}_k^s)\rho_k^t(t_l, t_m; \boldsymbol{\phi}_k^t),$$

for $i, j = 1, \ldots, n$ and $l, m = 1, \ldots, T$. We take $\rho_k^s(\cdot; \boldsymbol{\phi}_k^s)$ and $\rho_k^t(\cdot; \boldsymbol{\phi}_k^t)$ to be valid correlation functions, possibly non-stationary and anisotropic (Rasmussen and Williams, 2006), with parameters $\boldsymbol{\phi}_k^s$ and $\boldsymbol{\phi}_k^t$ respectively.

We assume *apriori* independence across the p space-time processes and so we have

$$\boldsymbol{C}_2 = \begin{bmatrix} \sigma_0^2 \boldsymbol{R}_0^s \otimes \boldsymbol{R}_0^t & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \sigma_1^2 \boldsymbol{R}_1^s \otimes \boldsymbol{R}_1^t & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \sigma_{p-1}^2 \boldsymbol{R}_{p-1}^s \otimes \boldsymbol{R}_{p-1}^t \end{bmatrix},$$

where $\otimes$ denotes the Kronecker product and

$$(\boldsymbol{R}_k^s)_{ij} = \rho_k^s(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi}_k^s), \quad (\boldsymbol{R}_k^t)_{lm} = \rho_k^t(t_l, t_m; \boldsymbol{\phi}_k^t).$$

Note that the separable covariance structure means that to invert the $nT \times nT$ matrix $\boldsymbol{R}_k^s \otimes \boldsymbol{R}_k^t$ we have only to invert one $n \times n$ matrix and one $T \times T$ matrix, as by the properties of the Kronecker product we have

$$(\boldsymbol{R}_k^s \otimes \boldsymbol{R}_k^t)^{-1} = (\boldsymbol{R}_k^s)^{-1} \otimes (\boldsymbol{R}_k^t)^{-1}.$$

It is also useful to note that for determinants we have $|\boldsymbol{R}_k^s \otimes \boldsymbol{R}_k^t| = |\boldsymbol{R}_k^s|^T |\boldsymbol{R}_k^t|^n$.

Given this model set up we may ask how the spatial and temporal dependencies interact to determine the convergence rate for the CP and the NCP. If, for example, there is strong spatial correlation but weak temporal correlation what effect might that have on the convergence rate, and does it depend on the relative sizes of n and T?

Given the results of this thesis we might expect to see that strengthening either spatial or temporal correlation, or indeed both, will result in increased efficiency for the CP, but a reduction in efficiency for the NCP. If $T >> n$ say, it is likely that the temporal dependence will be more important in determining the efficiency of the CP and the NCP than the spatial dependence.

7.2.2 Multivariate spatial models

In this thesis we have only considered univariate responses. Consider the multivariate extension of the standard Gaussian process model, given in (3.1). Suppose we collect data $\boldsymbol{Y}(\boldsymbol{s}_i) = (Y_1(\boldsymbol{s}_i), \ldots, Y_V(\boldsymbol{s}_i))'$ on V variables at each location $\boldsymbol{s}_i$, for $i = 1, \ldots, n$. We model $\boldsymbol{Y}(\boldsymbol{s}_i)$ as

$$\boldsymbol{Y}(\boldsymbol{s}_i) = \boldsymbol{X}(\boldsymbol{s}_i)\boldsymbol{\theta} + \boldsymbol{\beta}(\boldsymbol{s}_i) + \boldsymbol{\epsilon}(\boldsymbol{s}_i),$$

where $\boldsymbol{X}(\boldsymbol{s}_i)$ is a $V \times Vp$ block diagonal matrix with vth block given by the $1 \times p$ row vector of covariate information for the vth quantity of interest at location $\boldsymbol{s}_i$, denoted $\boldsymbol{x}'_v(\boldsymbol{s}_i)$. The concatenated vector of all Vp global regression coefficients is denoted by $\boldsymbol{\theta}$, $\boldsymbol{\beta}(\boldsymbol{s}_i) = (\beta_1(\boldsymbol{s}_i), \ldots, \beta_V(\boldsymbol{s}_i))'$ is a realisation of a zero mean multivariate spatial process at location $\boldsymbol{s}_i$, and non-spatial errors $\boldsymbol{\epsilon}(\boldsymbol{s}_i) = (\epsilon_1(\boldsymbol{s}_i), \ldots, \epsilon_V(\boldsymbol{s}_i))' \sim N(\boldsymbol{0}, \sigma^2_\epsilon \boldsymbol{I}_V)$, for all $i = 1, \ldots, n$, where $\boldsymbol{I}_V$ is the identity matrix of order V.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}'(\boldsymbol{s}_1), \ldots, \boldsymbol{\beta}'(\boldsymbol{s}_n))'$ be the vector of all spatially correlated random effects. We must take care when choosing $Cov(\beta_u(\boldsymbol{s}_i), \beta_v(\boldsymbol{s}_j))$, $u, v = 1, \ldots, V$, $i, j = 1, \ldots, n$, to ensure that $Var(\boldsymbol{\beta}) = \boldsymbol{C}_2$ is positive definite. Suppose we follow Banerjee et al. (2003, Chapter 7.1) and have a seperable covariance structure such that

$$Cov(\beta_u(\boldsymbol{s}_i), \beta_v(\boldsymbol{s}_j)) = \rho(\boldsymbol{s}_i, \boldsymbol{s}_j) Cov(\beta_u(\boldsymbol{s}), \beta_v(\boldsymbol{s})), \tag{7.2}$$

where $Cov(\beta_u(\boldsymbol{s}), \beta_v(\boldsymbol{s}))$ is the covariance between variables u and v at location $\boldsymbol{s}$, for all $\boldsymbol{s}$, and $\rho(\cdot, \cdot)$ is a valid correlation function for a univariate spatial process. Then we can write

$$\boldsymbol{C}_2 = \boldsymbol{R} \otimes \boldsymbol{S},$$

where $\boldsymbol{R}_{ij} = \rho(\boldsymbol{s}_i, \boldsymbol{s}_j)$ and $\boldsymbol{S}_{uv} = Cov(\beta_u(\boldsymbol{s}), \beta_v(\boldsymbol{s}))$ with $\boldsymbol{S}$ the $V \times V$ covariance matrix assiciated with $\boldsymbol{\beta}(\boldsymbol{s})$. Therefore, if $\boldsymbol{S}$ is positive definite, then so is $\boldsymbol{C}_2$.

With the foregoing model specification the following questions arise: How do the entries of $\boldsymbol{S}$ affect the convergence rate for the CP and the NCP? Furthermore, how does $\boldsymbol{S}$ affect the spatial pattern of the optimal weights of partial centering for the different variables?

Given this thesis we might speculate that strengthening the correlation between variables will will make the CP more efficient and the NCP less so. For a seperable covariance structure, like that given in (7.2), it seems likely that the optimal weights of partial centering will be increased with increasing correlation across the variables. We would expect that when $\boldsymbol{S}$ is the identity matrix, the spatial pattern of weights for the individual variables will be the same as if we had fitted a series of univariate models. As the correlation between the variables strengthens it seems plausible that the weights will increase, possibly by the same amount at each location.

Bibliography

Abramowitz, M. and Stegun, I. A. (1972) *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. No. 55. Courier Dover Publications.

Amemiya, T. (1984) Tobit models: A survey. *Journal of Econometrics*, **24**, 3–61.

Amit, Y. (1991) On rates of convergence of stochastic relaxation for Gaussian and non-Gaussian distributions. *Journal of Multivariate Analysis*, **38**, 82–99.

Apputhurai, P. and Stephenson, A. G. (2013) Spatiotemporal hierarchical modelling of extreme precipitation in Western Australia using anisotropic Gaussian random fields. *Environmental and ecological statistics*, **20**, 667–677.

Bakar, K. S. and Sahu, S. K. (2015) spTimer: Spatio-temporal Bayesian modeling using R. *Journal of Statistical Software*, **63**, 1–32.

Banerjee, A., Dunson, D. B. and Tokdar, S. T. (2012) Efficient Gaussian process regression for large datasets. *Biometrika*, **100**, 75–89.

Banerjee, S., Finley, A. O., Waldmann, P. and Ericsson, T. (2010) Hierarchical spatial process models for multiple traits in large genetic trials. *Journal of the American Statistical Association*, **105**, 506–521.

Banerjee, S., Gelfand, A. E. and Carlin, B. P. (2003) *Hierarchical Modeling and Analysis for Spatial Data*. CRC Press.

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

Barndorff-Nielsen, O. E. and Shephard, N. (2001) Non-Gaussian Ornstein–Uhlenbeck-based models and some of their uses in financial economics. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **63**, 167–241.

Barr, D. R. and Sherrill, E. T. (1999) Mean and variance of truncated normal distributions. *The American Statistician*, **53**, 357–361.

Berliner, L. M. (1996) Hierarchical Bayesian time series models. In *Maximum entropy and Bayesian methods*, 15–22. Springer.

Bernardo, J. M. and Smith, A. (1994) *Bayesian Theory*. Chichester: John Wiley and Sons, Ltd.

Berrocal, V. J., Gelfand, A. E. and Holland, D. M. (2010) A spatio-temporal downscaler for output from numerical models. *Journal of Agricultural, Biological, and Environmental statistics*, **15**, 176–197.

Berrocal, V. J., Raftery, A. E. and Gneiting, T. (2008) Probabilistic quantitative precipitation field forecasting using a two-stage spatial model. *The Annals of Applied Statistics*, **2**, 1170–1193.

Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, **36**, 192–236.

Breslow, N. E. and Clayton, D. G. (1993) Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, **88**, 9–25.

Brooks, S. P. and Gelman, A. (1998) General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

Chib, S. (1992) Bayes inference in the Tobit censored regression model. *Journal of Econometrics*, **51**, 79–99.

Chib, S. and Greenberg, E. (1995) Understanding the Metropolis-Hastings algorithm. *The American Statistician*, **49**, 327–335.

Christensen, O. F., Roberts, G. O. and Sköld, M. (2006) Robust Markov chain Monte Carlo methods for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, **15**, 1–17.

Christensen, O. F. and Waagepetersen, R. (2002) Bayesian prediction of spatial count data using generalized linear mixed models. *Biometrics*, **58**, 280–286.

Cowles, M. K. and Carlin, B. P. (1996) Markov chain Monte Carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, **91**, 883–904.

Cowles, M. K., Roberts, G. O. and Rosenthal, J. S. (1999) Possible biases induced by MCMC convergence diagnostics. *Journal of Statistical Computation and Simulation*, **64**, 87–104.

Cressie, N. (1993) *Statistics for Spatial Data*. John Wiley and Sons, Inc.

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

Cressie, N. and Wikle, C. K. (2011) *Statistics for Spatio-temporal Data*. Wiley.com.

De Oliveira, V. (2000) Bayesian prediction of clipped Gaussian random fields. *Computational Statistics and Data Analysis*, **34**, 299–314.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977) Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, **39**, 1–38.

Devroye, L. (1986) *Non-uniform Random Variate Generation*. Springer Verlag, New York.

Dey, D. K., Ghosh, S. K. and Mallick, B. K. (2000) *Generalized Linear Models: A Bayesian Perspective*. Marcel Dekker Inc, New York.

Diggle, P. J. and Ribeiro Jr, P. J. (2002) Bayesian inference in Gaussian model-based geostatistics. *Geographical and Environmental Modelling*, **6**, 129–146.

Diggle, P. J., Tawn, J. and Moyeed, R. (1998) Model-based geostatistics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **47**, 299–350.

Ecker, M. D. and Gelfand, A. E. (1999) Bayesian modeling and inference for geometrically anisotropic spatial data. *Mathematical Geology*, **31**, 67–83.

Finley, A. O., Banerjee, S. and Carlin, B. P. (2007) spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of Statistical Software*, **19**, 1–24.

Finley, A. O., Banerjee, S. and Gelfand, A. E. (2015) spBayes for large univariate and multivariate point-referenced spatio-temporal data models. *Journal of Statistical Software*, **63**, 1–28.

Finley, A. O., Banerjee, S. and MacFarlane, D. W. (2011) A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, **106**, 31–48.

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

Furrer, R., Nychka, D. and Sain, S. (2009) fields: Tools for spatial data. *R package version*, **6**.

Gelfand, A. E. (2012) Hierarchical modeling for spatial data problems. *Spatial Statistics*, **1**, 30–39.

Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010) *Handbook of Spatial Statistics*. CRC Press.

Gelfand, A. E., Kim, H.-J., Sirmans, C. and Banerjee, S. (2003) Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**, 387–396.

Gelfand, A. E., Sahu, S. K. and Carlin, B. P. (1995) Efficient parameterisations for normal linear mixed models. *Biometrika*, **82**, 479–488.

— (1996) Efficient parameterizations for generalized linear mixed models, (with discussion). In *Bayesian Statistics 5, J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith*, 165–180. Oxford University Press.

Gelfand, A. E. and Smith, A. F. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**, 398–409.

Gelfand, A. J., Ravishanker, N. and Ecker, M. D. (2000) Point-referenced binary spatial data. In *Generalized Linear Models: A Bayesian Perspective* (eds. D. K. Dey, S. K. Ghosh and B. K. Mallick). Marcel Dekker Inc, New York.

Gelman, A. (1992) Iterative and non-iterative simulation algorithms. *Computing Science and Statistics*, **24**, 433–438.

Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (2004) *Bayesian Data Analysis*. CRC press.

Gelman, A., Roberts, G. and Gilks, W. (1996) Efficient Metropolis jumping rules. *Bayesian Statistics*, **5**, 599–608.

Gelman, A. and Rubin, D. B. (1992) Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**, 457–472.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **6**, 721–741.

Gilks, W. R., Richardson, S. and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*. CRC press.

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

Gordon, R. D. (1941) Values of Mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, **12**, 364–366.

Gray, R. M. (2005) Toeplitz and circulant matrices: A review. *Communications and Information Theory*, **2**, 155–239.

Guhaniyogi, R., Finley, A. O., Banerjee, S. and Gelfand, A. E. (2011) Adaptive Gaussian predictive process models for large spatial datasets. *Environmetrics*, **22**, 997–1007.

Hamm, N., Finley, A., Schaap, M. and Stein, A. (2015) A spatially varying coefficient model for mapping PM10 air quality at the European scale. *Atmospheric Environment*, **102**, 393–405.

Handcock, M. S. and Stein, M. L. (1993) A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Handcock, M. S. and Wallis, J. R. (1994) An approach to statistical spatial-temporal modeling of meteorological fields. *Journal of the American Statistical Association*, **89**, 368–378.

Harville, D. A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Horn, R. A. and Johnson, C. R. (2012) *Matrix Analysis*. Cambridge University Press.

Huerta, G., Sansó, B. and Stroud, J. R. (2004) A spatiotemporal model for Mexico City ozone levels. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **53**, 231–248.

Johnson, D. S., Conn, P. B., Hooten, M. B., Ray, J. C. and Pond, B. A. (2013) Spatial occupancy models for large data sets. *Ecology*, **94**, 801–808.

Johnson, V. E. (1996) Studying convergence of Markov chain Monte Carlo algorithms using coupled sample paths. *Journal of the American Statistical Association*, **91**, 154–166.

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

Koenker, R. and Ng, P. (2003) Sparsem: A sparse matrix package for R. *Journal of Statistical Software*, **8**, 1–9.

Laird, N. M. and Ware, J. H. (1982) Random-effects models for longitudinal data. *Biometrics*, **38**, 963–974.

Lindgren, F., Rue, H. and Lindström, J. (2011) An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **73**, 423–498.

Lindley, D. V. and Smith, A. F. (1972) Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 1–41.

Liu, C., Liu, J. and Rubin, D. B. (1992) A variational control variable for assessing the convergence of the Gibbs sampler. *1992 Proceedings of Statistical Computing Section of American Statistical Association*, 74–78.

MacEachern, S. N. and Berliner, L. M. (1994) Subsampling the Gibbs sampler. *The American Statistician*, **48**, 188–190.

Matérn, B. (1986) *Spatial Variation*. Springer Verlag, Berlin, 2nd. edn.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21**, 1087.

Meyn, S. P. and Tweedie, R. L. (1993) *Markov Chains and Stochastic Stability*. Springer Verlag, London.

Musio, M., Augustin, N. H. and von Wilpert, K. (2008) Geoadditive Bayesian models for forestry defoliation data: a case study. *Environmetrics*, **19**, 630–642.

Neal, P. and Roberts, G. (2005) A case study in non-centering for data augmentation: stochastic epidemics. *Statistics and Computing*, **15**, 315–327.

— (2006) Optimal scaling for partially updating MCMC algorithms. *The Annals of Applied Probability*, **16**, 475–515.

Neal, R. M. (2003) Slice sampling. *Annals of Statistics*, **31**, 705–741.

Pace, R. K. and LeSage, J. (2009) *Introduction to Spatial Econometrics*. Boca Raton, FL: Chapman &Hall/CRC.

Papaspiliopoulos, O. (2003) *Non-centered parameterisations for data augmentation and hierarchical models with applications to inference for Lévy-based stochastic volatility models*. Ph.D. thesis, University of Lancaster.

Papaspiliopoulos, O. and Roberts, G. (2008) Stability of the Gibbs sampler for Bayesian hierarchical models. *The Annals of Statistics*, **36**, 95–117.

Papaspiliopoulos, O., Roberts, G. O. and Sköld, M. (2003) Non-centered parameterisations for hierarchical models and data augmentation (with discussion). In *Bayesian Statistics 7 (Bernardo, JM and Bayarri, MJ and Berger, JO and Dawid, AP and Heckerman, D and Smith, AFM and West, M): Proceedings of the Seventh Valencia International Meeting*, 307–326. Oxford University Press, USA.

— (2007) A general framework for the parameterization of hierarchical models. *Statistical Science*, **22**, 59–73.

Plummer, M., Best, N., Cowles, K. and Vines, K. (2006) CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**, 7–11.

Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. MIT press.

Rathbun, S. L. and Fei, S. (2006) A spatial zero-inflated Poisson regression model for oak regeneration. *Environmental and Ecological Statistics*, **13**, 409–426.

Reich, B. J., Fuentes, M., Herring, A. H. and Evenson, K. R. (2010) Bayesian variable selection for multivariate spatially varying coefficient regression. *Biometrics*, **66**, 772–782.

Ribeiro Jr, P. J. and Diggle, P. J. (2001) geoR: A package for geostatistical analysis. *R news*, **1**, 14–18.

Ripley, B. and Kirkland, M. (1990) Iterative simulation methods. *Journal of Computational and Applied Mathematics*, **31**, 165–172.

Ripley, B. D. (1987) *Stochastic Simulation*. Wiley, New York.

Robert, C. O. and Casella, G. (2004) *Monte Carlo Statistical Methods*. Springer-Verlag New York.

Roberts, G. O. (1996) Methods for estimating L^2 convergence of Markov chain Monte Carlo. In *Bayesian Analysis in Statistics and Econometrics: Essays in Honor of Arnold Zellner*, 373–384. Amsterdam: North-Holland.

Roberts, G. O., Gelman, A. and Gilks, W. R. (1997) Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, **7**, 110–120.

Roberts, G. O., Papaspiliopoulos, O. and Dellaportas, P. (2004) Bayesian inference for non-Gaussian Ornstein–Uhlenbeck stochastic volatility processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**, 369–393.

Roberts, G. O. and Rosenthal, J. S. (1998) Markov chain Monte Carlo: Some practical implications of theoretical results. *Canadian Journal of Statistics*, **26**, 5–20.

Roberts, G. O. and Sahu, S. K. (1997) Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **59**, 291–317.

— (2001) Approximate predetermined convergence properties of the Gibbs sampler. *Journal of Computational and Graphical Statistics*, **10**, 216–229.

Rue, H., Martino, S. and Chopin, N. (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series B (statistical methodology)*, **71**, 319–392.

Sahu, S. K. and Bakar, K. S. (2012) Hierarchical Bayesian autoregressive models for large spacetime data with applications to ozone concentration modelling. *Applied Stochastic Models in Business and Industry*, **28**, 395–415.

Sahu, S. K., Gelfand, A. E. and Holland, D. M. (2007) High resolution space–time ozone modeling for assessing trends. *Journal of the American Statistical Association*, **102**, 1221–1234.

— (2010) Fusing point and areal level space–time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **59**, 77–103.

Sahu, S. K., Yip, S. and Holland, D. M. (2011) A fast Bayesian method for updating and forecasting hourly ozone levels. *Environmental and Ecological Statistics*, **18**, 185–207.

Salazar, E., Dunson, D. B. and Carin, L. (2013) Analysis of space–time relational data with application to legislative voting. *Computational Statistics and Data Analysis*, **68**, 141–154.

Sang, H. and Huang, J. Z. (2012) A full scale approximation of covariance functions for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **74**, 111–132.

Schabenberger, O. and Gotway, C. A. (2004) *Statistical Methods for Spatial Data Analysis*. CRC Press.

Smith, A. F. and Roberts, G. O. (1993) Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, **55**, 3–23.

Smith, B. J., Yan, J. and Cowles, M. K. (2008) Unified geostatistical modeling for data fusion and spatial heteroskedasticity with R package ramps. *Journal of Statistical Software*, **25**, 1–21.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993) Bayesian analysis in expert systems. *Statistical Science*, **8**, 219–247.

Sun, Y., Li, B. and Genton, M. G. (2012) Geostatistics for large datasets. In *Advances and Challenges in Space-Time Modelling of Natural Events*, 55–77. Springer.

Tanner, M. A. and Wong, W. H. (1987) The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, **82**, 528–540.

Tobin, J. (1958) Estimation of relationships for limited dependent variables. *Econometrica: Journal of the Econometric Society*, **12**, 24–36.

Wheeler, D. C., Páez, A., Spinney, J. and Waller, L. A. (2014) A Bayesian approach to hedonic price analysis. *Papers in Regional Science*, **93**, 663–683.

Woodbury, M. A. (1950) Inverting modified matrices. *Memorandum report 42*, Princeton University, Statistical Research Group.

Yan, J., Cowles, M. K., Wang, S. and Armstrong, M. P. (2007) Parallelizing MCMC for Bayesian spatiotemporal geostatistical models. *Statistics and Computing*, **17**, 323–335.

Yu, Y. and Meng, X.-L. (2011) To center or not to center: That is not the question: an Ancillarity–Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. *Journal of Computational and Graphical Statistics*, **20**, 531–570.

Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

Zimmerman, D. L. (1993) Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**, 453–470.