Dynamically Updated Spatially Varying Parameterizations of Hierarchical Bayesian Models for Spatial Data

Mark R. Bass and Sujit K. Sahu*

Abstract

Fitting hierarchical Bayesian models to spatially correlated data sets using Markov chain Monte Carlo (MCMC) techniques is computationally expensive. Complicated covariance structures of the underlying spatial processes, together with high dimensional parameter space, mean that the number of calculations required grows cubically with the number of spatial locations at each MCMC iteration. This necessitates the need for efficient model parameterisations that hasten the convergence and improve the mixing of the associated algorithms. We consider partially centred parameterisations (PCPs) which lie on a continuum between what are known as the centred (CP) and noncentered parameterisations (NCP). By introducing a weight matrix we remove the conditional posterior correlation between the fixed and the random effects, and hence construct a PCP which achieves immediate convergence for a three stage model, based on multiple Gaussian processes with known covariance parameters. When the covariance parameters are unknown we dynamically update the parameterisation within the sampler. The PCP outperforms both the CP and the NCP and leads to a fully automated algorithm which has been demonstrated in two simulation examples. The effectiveness of the spatially varying PCP is illustrated with a practical data set of nitrogen dioxide concentration levels. Supplemental materials consisting of appendices, data sets and computer code to reproduce the results are available online.

Keywords: Gibbs sampling; Parameterisation; Rate of convergence; Spatially varying coefficient model.

1 Introduction

There is a growing interest among researchers in spatially varying coefficient (SVC) models (Hamm *et al.*, 2015; Wheeler *et al.*, 2014; Finley *et al.*, 2011; Berrocal *et al.*, 2010;

*Mark R. Bass is statistician, Barclays Services Limited, City of London, United Kingdom. Sujit K. Sahu is Professor of Statistics, University of Southampton, Southampton, SO17 1BJ, UK. Email: *S.K.Sahu@southampton.ac.uk.* Much of this work was completed when the first author was a PhD student in the University of Southampton.

Gelfand *et al.*, 2003). 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 fitting these models is challenging and computationally expensive. 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.

Parameterisation of a hierarchical model is known to affect the performance of the Markov chain Monte Carlo (MCMC) method used for inference. For normal linear hierarchical models (NLHMs) the centred parameterisation (CP) yields an efficient Gibbs sampler when the variance of the data model is low relative to that of the random effects, and the noncentred parameterisation (NCP) yields an efficient Gibbs sampler when the variance of the data model is relatively high (Papaspiliopoulos *et al.*, 2003; Gelfand *et al.*, 1995). Where the latent variables are realisations of a spatial process with an exponential correlation function, Bass and Sahu (2017) show that increasing the strength of correlation improves the efficiency of the CP but degrades that of the NCP. Hence the sampling efficiency of the CP and the NCP is dependent upon the typically unknown variance and spatial correlation parameters and will therefore differ across different data sets. Consequently, deciding which parameterisation to employ can be problematic.

With the aim of developing a robust parameterisation for NLHMs, Papaspiliopoulos *et al.* (2003) consider the CP and the NCP as extremes of a family of partially centred parameterisations (PCPs). They find the optimal PCP which results in a Gibbs sampler that produces independent samples from the posterior distributions of the mean parameters, but again this is conditioned on the covariance parameters. The question we look to answer in this paper is can we create a parameterisation for spatial models that is robust to the data and does not require *a priori* knowledge of the model parameters, and hence can be routinely implemented?

To address this question we write a general SVC model as a three stage NLHM. The PCP is constructed by introducing a weight matrix that allows us to eliminate the conditional posterior correlation between the global and random effects. This in turn implies immediate convergence of the associated Gibbs sampler for known variance and correlation parameters, which collectively we will call the covariance parameters. When these parameters are unknown we propose dynamically updating the weight matrix, which leads to a parameterisation that is both spatially varying and dynamically updated within the Gibbs sampler. As it is necessary to invert an $n \times n$ matrix to compute the weight matrix, implementing the PCP results in an algorithm with longer run times than those associated

with the CP and NCP. Consequently, we recommend this method for modest sized spatial datasets of hundreds, but not thousands, of observations.

For an exponential correlation function we demonstrate how the weights of partial centering depend on the covariance parameters and the sampling locations, and hence vary over the spatial domain. We show that weights are higher when the data variance is relatively low and when the spatial correlation is high. Also, higher weights are given to more densely clustered sampling locations and where the covariate values are higher. In order to judge sampling efficiency of the PCP we use well known convergence diagnostics. The performance of the PCP is shown to be robust to changes in the covariance parameters of the data generating mechanism. Moreover, the PCP converges more quickly and produces posterior samples with lower autocorrelation than either the CP or the NCP. Note that Bass and Sahu (2017) do not consider the PCP at all.

A related approach is the interweaving algorithm proposed by Yu and Meng (2011). The algorithm results in a Gibbs sampler that is more efficient than the worst of the CP and NCP. However, the interweaving algorithm does not guarantee immediate convergence for known covariance parameters, whereas the PCP does, see Section A.3. Another approach is to marginalise over the random effects, thus reducing the dimension of the posterior distribution. This method can be employed when the error structures of the data and the random effects are both assumed to be Gaussian. Marginalised likelihoods are used by Gelfand *et al.* (2003) for fitting SVC regression models. However, marginalisation results in a loss of conditional conjugacy of the variance parameters and means that they have to be updated by using Metropolis-type steps, which require difficult and time consuming tuning. The scheme proposed here is fully automated.

The rest of this paper is laid out as follows: Section 2 gives details of the general spatial model and the construction and properties of its PCP. Section 3 illustrates how the weights of partial centering are influenced by the model parameters and the sampling locations. Section 4 demonstrates the sampling efficiency of the PCP by applying it to simulated spatial data sets and compares its performance to the CP and the NCP. Section 5 applies the different model parameterisations to a real air pollution data set on annnual mean concentration levels of NO_2 observed in Greater London in 2011. Section 6 contains some concluding remarks.

MCMC samplers are coded in the C programming language and the output is analysed in R (R Core Team, 2016).

2 Partial centering of space-varying coefficient models

2.1 Model specification

For data observed at a set of locations $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$ we consider the normal linear model with spatially varying regression coefficients (Gelfand *et al.*, 2003):

$$Y(\boldsymbol{s}_i) = \theta_0 + \beta_0(\boldsymbol{s}_i) + \sum_{k=1}^{p-1} \{\theta_k + \beta_k(\boldsymbol{s}_i)\} x_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i), \quad i = 1, \ldots, n. \tag{1}$$

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

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

where $\boldsymbol{x}(\boldsymbol{s}_i) = \{1, x_1(\boldsymbol{s}_i), \ldots, x_{p-1}(\boldsymbol{s}_i)\}^{\mathrm{T}}$ is a vector containing covariate information for site $\boldsymbol{s}_i$ and $\boldsymbol{\theta} = (\theta_0, \ldots, \theta_{p-1})^{\mathrm{T}}$ 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 $\boldsymbol{\beta}(\boldsymbol{s}_i) = \{\beta_0(\boldsymbol{s}_i), \ldots, \beta_{p-1}(\boldsymbol{s}_i)\}^{\mathrm{T}}$. The n realisations of the Gaussian process associated with the kth covariate are given by $\boldsymbol{\beta}_k = \{\beta_k(\boldsymbol{s}_1), \ldots, \beta_k(\boldsymbol{s}_n)\}^{\mathrm{T}} \sim N(0, \boldsymbol{\Sigma}_k), \quad k = 0, \ldots, p-1$, where $\boldsymbol{\Sigma}_k = \sigma^2_k \boldsymbol{R}_k$, and $(\boldsymbol{R}_k)_{ij} = corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\}$. The form of the model given in (1) is the NCP. The CP is found by introducing the variables $\tilde{\beta}_k(\boldsymbol{s}_i) = \theta_k + \beta_k(\boldsymbol{s}_i)$. Therefore $\tilde{\boldsymbol{\beta}}_k = \{\tilde{\beta}_k(\boldsymbol{s}_1), \ldots, \tilde{\beta}_k(\boldsymbol{s}_n)\}^{\mathrm{T}} \sim N(\theta_k \mathbf{1}, \boldsymbol{\Sigma}_k)$.

Global effects $\boldsymbol{\theta}$ are assumed to be multivariate normal *a priori* and so we write model (1) in its hierarchically centred form as

$$\boldsymbol{Y}|\tilde{\boldsymbol{\beta}} \sim N(\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}, \boldsymbol{C}_1), \quad \tilde{\boldsymbol{\beta}}|\boldsymbol{\theta} \sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2), \quad \boldsymbol{\theta} \sim N(\boldsymbol{m}, \boldsymbol{C}_3), \tag{2}$$

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 where $\boldsymbol{D}_k$ is a diagonal matrix with entries $\boldsymbol{x}_k = \{x_k(\boldsymbol{s}_1), \ldots, x_k(\boldsymbol{s}_n)\}^{\mathrm{T}}$. We denote by $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0^{\mathrm{T}}, \ldots, \tilde{\boldsymbol{\beta}}_{p-1}^{\mathrm{T}})^{\mathrm{T}}$ 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. The p processes are assumed independent *a priori* and so $\boldsymbol{C}_2$ is block diagonal where the kth block is $\boldsymbol{\Sigma}_k$ for $k = 0, 1, \ldots, p-1$.

The global effects $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_{p-1})^{\mathrm{T}}$ are assumed to be independent *a priori* with the kth element assigned a Gaussian prior distribution with mean m_k and variance $\sigma^2_k v_k$, hence we write $\theta_k \sim N(m_k, \sigma^2_k v_k)$. Therefore $\boldsymbol{m} = (m_0, \ldots, m_{p-1})^{\mathrm{T}}$ and $\boldsymbol{C}_3$ is a diagonal matrix with diagonal entries $\sigma^2_k v_k$ for $k = 0, \ldots, p-1$.

We complete the model specification by assigning prior distributions to the covariance parameters. The realisations of the kth zero mean Gaussian process, $\boldsymbol{\beta}_k$, have a prior covariance matrix given by $\boldsymbol{\Sigma}_k = \sigma^2_k \boldsymbol{R}_k$. 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 $\sigma_k^2 \sim IG(a_k, b_k), \quad k = 0, \ldots, p-1, \quad \sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$, where we write $X \sim IG(a, b)$ if X has a density proportional to $x^{-(a+1)}e^{-b/x}$.

In this paper we consider only the exponential correlation function, which is a member of the Matérn family (Handcock and Stein, 1993; Matérn, 1986) and is widely applied to spatial processes (Sahu *et al.*, 2010; Berrocal *et al.*, 2010; Sahu *et al.*, 2007; Huerta *et al.*, 2004). Therefore entries of the $\boldsymbol{R}_k$ are $(\boldsymbol{R}_k)_{ij} = corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\} = \exp(-\phi_k d_{ij})$, where $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ denotes the distance between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ and ϕ_k controls the rate of decay of the correlation. Here the strength of correlation is characterised by the effective range, d_k, which is the distance such that $corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\} = 0.05$. For an exponential correlation function we have that $d_k = -\log(0.05)/\phi_k \approx 3/\phi_k$.

It is not possible to consistently estimate all of the variance and decay parameters under vague prior distributions (Zhang, 2004), and so we do not sample from the posterior distributions of the decay parameters. For the simulation studies in Section 4 the decay parameters are fixed, thus helping us to examine their impact upon the sampling efficiency of the PCP. For the real data example in Section 5 we perform a grid search over a range of values, selecting those that offer the best out-of-sample predictions. A grid search is equivalent to placing a discrete uniform prior distribution upon the decay parameters and is a commonly adopted approach (Berrocal *et al.*, 2010; Sahu *et al.*, 2007).

2.2 Construction of the PCP

In Section 2.1 we consider two parameterisations, non-centred and centred, that differ by the prior mean of the spatial processes. We have $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \boldsymbol{C}_2)$ for the NCP and $\tilde{\boldsymbol{\beta}} \sim N(\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2)$ for the CP where a linear shift relates the two processes such that $\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta} + \boldsymbol{X}_2\boldsymbol{\theta}$. The PCP is formed by a partial shift, which is defined by

$$\boldsymbol{\beta}^w = \tilde{\boldsymbol{\beta}} - (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \tag{3}$$

and therefore the partially centred model is written as

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{\beta}^w, \boldsymbol{\theta} &\sim N\{\boldsymbol{X}_1\boldsymbol{\beta}^w + \boldsymbol{X}_1(\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_1\}, \\ \boldsymbol{\beta}^w \mid \boldsymbol{\theta} &\sim N(\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2), \quad \boldsymbol{\theta} \sim N(\boldsymbol{m}, \boldsymbol{C}_3), \end{aligned} \tag{4}$$

where $\boldsymbol{\beta}^w = (\boldsymbol{\beta}_0^{w\mathrm{T}}, \ldots, \boldsymbol{\beta}_{p-1}^{w\mathrm{T}})^\mathrm{T}$, and $\boldsymbol{\beta}_k^w = \{\beta_k^w(\boldsymbol{s}_1), \ldots, \beta_k^w(\boldsymbol{s}_n)\}^\mathrm{T}$.

Defining the PCP this way gives us tremendous flexibility in terms of parameterisation. If $\boldsymbol{W}$ is the identity matrix we recover the CP and where $\boldsymbol{W}$ is the zero matrix we have the NCP. If we let $\boldsymbol{W} = \mathrm{diag}(w_0\boldsymbol{I}, w_1\boldsymbol{I}, \ldots, w_{p-1}\boldsymbol{I})$ then we have the PCP investigated by Papaspiliopoulos *et al.* (2003, Section 4) and see also Papaspiliopoulos (2003) for an independent random effect Poisson count model.

The question is how do we choose the entries of $\boldsymbol{W}$ such that optimal performance of the Gibbs sampler is achieved? We answer this question by analysing the posterior correlation

between global and random effects. Returning to the CP, if we apply the calculations given in Gelfand *et al.* (1995, Section 2) to our model set up it can be shown that

$$cov(\tilde{\boldsymbol{\beta}}, \boldsymbol{\theta} \mid \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{y}) = \boldsymbol{B}\boldsymbol{C}_2^{-1}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y} \tag{5}$$

where

$$\boldsymbol{B} = var(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\theta}, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{y}) = \left(\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1},$$

and

$$\boldsymbol{\Sigma}_{\theta|y} = var(\boldsymbol{\theta} \mid \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{y}) = \left\{(\boldsymbol{X}_1\boldsymbol{X}_2)^{\mathrm{T}}\boldsymbol{\Sigma}_{Y|\theta}^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right\}^{-1}. \tag{6}$$

Therefore by substituting equation (3) into equation (5) we have that the posterior covariance of $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$ is

$$\begin{aligned} cov(\boldsymbol{\beta}^w, \boldsymbol{\theta} \mid \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{y}) &= cov\{\tilde{\boldsymbol{\beta}} - (\boldsymbol{I} - \boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{\theta} \mid \boldsymbol{y}\} \\ &= \left\{\boldsymbol{B}\boldsymbol{C}_2^{-1} - (\boldsymbol{I} - \boldsymbol{W})\right\}\boldsymbol{X}_2\boldsymbol{\Sigma}_{\theta|y}. \end{aligned} \tag{7}$$

We can see from (7) that $cov(\boldsymbol{\beta}^w, \boldsymbol{\theta} \mid \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{y}) = \boldsymbol{0}$ when $\boldsymbol{B}\boldsymbol{C}_2^{-1} = \boldsymbol{I} - \boldsymbol{W}$. Therefore we define the optimal $\boldsymbol{W}$to be

$$\boldsymbol{W}^{opt} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}. \tag{8}$$

For all that follows we drop the superscript from $\boldsymbol{W}^{opt}$ and any time we refer to $\boldsymbol{W}$ it will be the matrix defined in (8). By the Sherman-Woodbury-Morrison identity (Harville, 1997, Chapter 18) we can write $\boldsymbol{W}$ as

$$\boldsymbol{W} = \boldsymbol{C}_2\boldsymbol{X}_1^{\mathrm{T}}\left(\boldsymbol{C}_1 + \boldsymbol{X}_1\boldsymbol{C}_2\boldsymbol{X}_1^{\mathrm{T}}\right)^{-1}\boldsymbol{X}_1, \tag{9}$$

which requires the inversion of matrix of order n and not of order np.

Equation (8) implies that to minimise the posterior correlation between the random effects and global effects we cannot, in general, restrict $\boldsymbol{W}$ to be a diagonal matrix with entries w_i, for $i = 0, \ldots, p-1$. It then follows that as $\boldsymbol{\beta}^w|\boldsymbol{\theta} \sim N(\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}, \boldsymbol{C}_2)$ *a priori*, 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. For example, suppose we have $p = 2$ processes, $\boldsymbol{\beta}_0^w$ and $\boldsymbol{\beta}_1^w$, and we partition $\boldsymbol{W}$ into four $n \times n$ blocks. Then we have

$$\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{W}_{00} & \boldsymbol{W}_{01} \\ \boldsymbol{W}_{10} & \boldsymbol{W}_{11} \end{pmatrix}\begin{pmatrix} \boldsymbol{1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{1} \end{pmatrix}\begin{pmatrix} \theta_0 \\ \theta_0 \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{00}\boldsymbol{1}\theta_0 + \boldsymbol{W}_{01}\boldsymbol{1}\theta_1 \\ \boldsymbol{W}_{10}\boldsymbol{1}\theta_0 + \boldsymbol{W}_{11}\boldsymbol{1}\theta_1 \end{pmatrix},$$

and the ith row of $\boldsymbol{W}_{00}\boldsymbol{1}$ is the weight assigned to θ_0 for $\beta_0^w(\boldsymbol{s}_i)$. More generally the ith row of $\boldsymbol{W}_{kj}\boldsymbol{1}$ $(k, j = 0, \ldots, p-1)$, is the weight assigned to θ_j for $\beta_k^w(\boldsymbol{s}_i)$. This is illustrated in Section 5, Figure 9. Equation (8) also implies that when $\boldsymbol{X}$ contains the values of spatially referenced covariates or when $\boldsymbol{C}_2$ is the covariance of a spatial process, then the weights, $\boldsymbol{W}\boldsymbol{X}_2$, will vary over space, as demonstrated in Section 3.

2.3 Exploring block updating for the PCP

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, $\boldsymbol{\Sigma}$, of the joint posterior distribution of $(\boldsymbol{\beta}_1^w, \boldsymbol{\beta}_2^w, \boldsymbol{\theta})$ is a 3×3 block matrix and $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$, see e.g. Harville (1997, Chapter 8), is also 3×3 block matrix. These are given by:

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

where $\boldsymbol{Q}_{\beta_1} = (\boldsymbol{\Sigma}_{\beta_1} - \boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}\boldsymbol{\Sigma}_{\beta_{21}})^{-1}$, $\boldsymbol{Q}_{\beta_{12}} = -(\boldsymbol{\Sigma}_{\beta_1} - \boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}\boldsymbol{\Sigma}_{\beta_{21}})^{-1}\boldsymbol{\Sigma}_{\beta_{12}}\boldsymbol{\Sigma}_{\beta_2}^{-1}$, $\boldsymbol{Q}_{\beta_{21}} = -(\boldsymbol{\Sigma}_{\beta_2} - \boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}\boldsymbol{\Sigma}_{\beta_{12}})^{-1}\boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}$, $\boldsymbol{Q}_{\beta_2} = (\boldsymbol{\Sigma}_{\beta_2} - \boldsymbol{\Sigma}_{\beta_{21}}\boldsymbol{\Sigma}_{\beta_1}^{-1}\boldsymbol{\Sigma}_{\beta_{12}})^{-1}$, $\boldsymbol{Q}_{\theta} = \boldsymbol{\Sigma}_{\theta}^{-1}$. Using the intermediate calculations given in Appendix A.4 in Supplementary materials it can be shown that the convergence rate corresponding to the above precision matrix $\boldsymbol{Q}$ 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 posterior 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 $\boldsymbol{Q}_{\beta} = \boldsymbol{\Sigma}_{\beta}^{-1}$, $\boldsymbol{Q}_{\theta_1} = (\boldsymbol{\Sigma}_{\theta_1} - \boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}\boldsymbol{\Sigma}_{\theta_{21}})^{-1}$, $\boldsymbol{Q}_{\theta_{12}} = -(\boldsymbol{\Sigma}_{\theta_1} - \boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}\boldsymbol{\Sigma}_{\theta_{21}})^{-1}\boldsymbol{\Sigma}_{\theta_{12}}\boldsymbol{\Sigma}_{\theta_2}^{-1}$, $\boldsymbol{Q}_{\theta_{21}} = -(\boldsymbol{\Sigma}_{\theta_2} - \boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}\boldsymbol{\Sigma}_{\theta_{12}})^{-1}\boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}$, $\boldsymbol{Q}_{\theta_2} = (\boldsymbol{\Sigma}_{\theta_2} - \boldsymbol{\Sigma}_{\theta_{21}}\boldsymbol{\Sigma}_{\theta_1}^{-1}\boldsymbol{\Sigma}_{\theta_{12}})^{-1}$, 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 (6), is unaffected by hierarchical centering. Therefore partial centering cannot remove any posterior correlation between subsets of $\boldsymbol{\theta}$, and so all of its elements must be updated together. Then suppose that we partition the partially centred random effects into l blocks so that $\boldsymbol{\beta}^w = (\boldsymbol{\beta}_1^{w\mathrm{T}}, \ldots, \boldsymbol{\beta}_l^{w\mathrm{T}})^{\mathrm{T}}$. We find the posterior covariance between the ijth block to be

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

where $\boldsymbol{B}_{ij}$ is the ijth block of $\boldsymbol{B} = var(\tilde{\boldsymbol{\beta}} \mid \boldsymbol{\theta}, y)$. 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 $(\boldsymbol{I}-\boldsymbol{W})_{.j}^{\mathrm{T}} = \boldsymbol{C}_2^{-1}\boldsymbol{B}_{.j}$ then $cov(\boldsymbol{\beta}_i^w, \boldsymbol{\beta}_j^w \mid \boldsymbol{y}) = \boldsymbol{B}_{ij}$, which is generally a non-zero matrix. Therefore we must update $\boldsymbol{\beta}^w$ as one component and $\boldsymbol{\theta}$ as another.

2.4 PCP for unknown variance parameters

The PCP relies on the $\boldsymbol{W}$ matrix which, by construction, removes the posterior correlation between $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$. However, 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 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 MCMC iteration.

Let $\boldsymbol{\xi}^{(t)} = \{\boldsymbol{\beta}^{w(t)}, \boldsymbol{\theta}^{(t)}, \boldsymbol{\sigma}^{\boldsymbol{2}(t)}, \sigma_\epsilon^{2(t)}\}^{\mathrm{T}}$ be the current state of the Markov chain, where $\boldsymbol{\sigma}^{\boldsymbol{2}(t)} = \{\sigma_0^{2(t)}, \ldots, \sigma_{p-1}^{2(t)}\}^{\mathrm{T}}$. Also let $\boldsymbol{\sigma}_{-k}^{\boldsymbol{2}(t+1)} = \{\sigma_0^{2(t+1)}, \ldots, \sigma_{k-1}^{2(t+1)}, \sigma_{k+1}^{2(t)}, \ldots, \sigma_{p-1}^{2(t)}\}^{\mathrm{T}}$ be the partially updated vector without σ_k^2. We write $\boldsymbol{W} = \boldsymbol{W}(\boldsymbol{\sigma}^{\boldsymbol{2}}, \sigma_\epsilon^2)$ to highlight the dependency of $\boldsymbol{W}$ on the variance parameters. Recall that in Section 2.3 we show that $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$ should each be updated as one block. We obtain a new sample, $\boldsymbol{\xi}^{(t+1)} \sim \pi(\boldsymbol{\xi} \mid \boldsymbol{y})$ as follows:

1. Sample $\boldsymbol{\beta}^{w(t+1)} \sim \pi\{\boldsymbol{\beta}^w \mid \boldsymbol{\theta}_0^{(t)}, \boldsymbol{\sigma}^{\boldsymbol{2}(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{W}(\boldsymbol{\sigma}^{\boldsymbol{2}(t)}, \sigma_\epsilon^{2(t)}), \boldsymbol{y}\}$.
2. Sample $\boldsymbol{\theta}^{(t+1)} \sim \pi\{\boldsymbol{\theta} \mid \boldsymbol{\beta}^{w(t+1)}, \boldsymbol{\sigma}^{\boldsymbol{2}(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{W}(\boldsymbol{\sigma}^{\boldsymbol{2}(t)}, \sigma_\epsilon^{2(t)}), \boldsymbol{y}\}$.
3. For $k = 0, \ldots, p-1$,
   Sample $\sigma_k^{2(t+1)} \sim \pi\{\sigma_k^2 \mid \boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \boldsymbol{\sigma}_{-k}^{\boldsymbol{2}(t+1)}, \sigma_\epsilon^{2(t)}, \boldsymbol{W}(\boldsymbol{\sigma}_{-k}^{\boldsymbol{2}(t+1)}, \sigma_k^{2(t)}, \sigma_\epsilon^{2(t)}), \boldsymbol{y}\}$.
4. Sample $\sigma_\epsilon^{2(t+1)} \sim \pi\{\sigma_\epsilon^2 | \boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \boldsymbol{\sigma}^{\boldsymbol{2}(t+1)}, \boldsymbol{W}(\boldsymbol{\sigma}^{\boldsymbol{2}(t+1)}, \sigma_\epsilon^{2(t)}), \boldsymbol{y}\}$.

The distributions of σ_k^2 and σ_ϵ^2 are conditioned on their respective current values through $\boldsymbol{W}$, i.e. $\sigma_0^{2(t+1)}$ is conditioned on $\sigma_0^{2(t)}$. There is no stochastic relationship between the parameters and $\boldsymbol{W}$; given the parameters $\boldsymbol{W}$ is completely determined. However, 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. We need to show that

$$\pi\{\boldsymbol{\xi}^{(t+1)}|\boldsymbol{y}\} = \int P\{\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)}\}\pi\{\boldsymbol{\xi}^{(t)}|\boldsymbol{y}\}d\boldsymbol{\xi}^{(t)}, \tag{10}$$

where $P\{\cdot|\cdot\}$ is the transition kernel of the chain. The definition of $P\{\cdot|\cdot\}$ and the proof of that (10) holds is provided in Appendix A.5 in Supplementary materials.

3 Spatially varying weights of partial centering

3.1 A spatially varying intercept model

In this section we show that the weights of the PCP may vary over space and how these weights are influenced by the spatial distribution of the sampling locations. We also illustrate how the weights of partial centering depend upon 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 the impact of a spatially varying covariate. To focus on these relationships we will consider simplified versions of model (4) that have one global parameter and one latent process. We do not need to simulate data as given a set of sampling locations we can investigate the weights through the variance and decay parameters.

We begin by looking at the following model,

$$\begin{aligned} \boldsymbol{Y} \mid \boldsymbol{\beta}_0^w, \theta_0 &\sim N\{\boldsymbol{\beta}_0^w + (\boldsymbol{I} - \boldsymbol{W})\mathbf{1}\theta_0, \sigma_\epsilon^2 \boldsymbol{I}\}, \\ \boldsymbol{\beta}_0^w \mid \theta_0 &\sim N(\boldsymbol{W}\mathbf{1}\theta_0, \sigma_0^2 \boldsymbol{R}_0), \quad \theta_0 \sim N(m_0, \sigma_0^2 v_0). \end{aligned} \tag{11}$$

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 (4) 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 (9) we have

$$\boldsymbol{W} = \sigma_0^2 \boldsymbol{R}_0 (\sigma_\epsilon^2 \boldsymbol{I} + \sigma_0^2 \boldsymbol{R}_0)^{-1}, \tag{12}$$

and so the entries of $\boldsymbol{W}$ depend on variance parameters σ_0^2, σ_ϵ^2 and, through correlation matrix $\boldsymbol{R}$, the spatial decay parameter and the set of sampling locations, $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n$.

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 which we split into nine sub-squares of equal area, see Figure 1. 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.

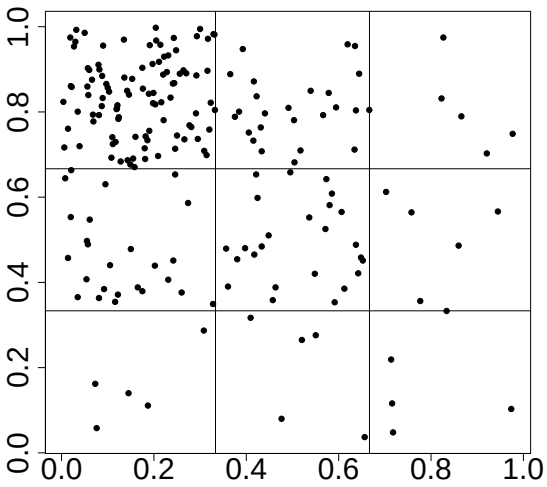

Figure 1: Patterned sampling locations. 100 top left; 25 in top middle, middle left and middle middle; five top right, middle right and bottom third. We use the points shown here to analyse the effect of the sampling locations on the spatially varying weights, c.f. Figure 2.

We consider five variance ratios: $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$, and three effective ranges: $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$, which are chosen with respect to the maximum separation of two points in the unit square, $\sqrt{2}$. An effective range of zero, implying independent random effects, returns weights that are the same at each location. For each of the 15 variance ratio-effective range combinations we compute $\boldsymbol{W}\boldsymbol{X}_2$, whose ith value is the weight assigned to θ_0 at each $\boldsymbol{s}_i$ $i = 1, \ldots, 200$.

We use the 'interp()' function in the R package akima (Akima and Gebhardt, 2016) to interpolate the weights over the unit square. These interpolated plots of spatially varying weights are given in Figure 2 and are created in ggplot2 (Wickham, 2009), making use of the 'gather()' function in tidyr (Wickham, 2017) to set up the data frame. 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 the weights are higher, as they are 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.

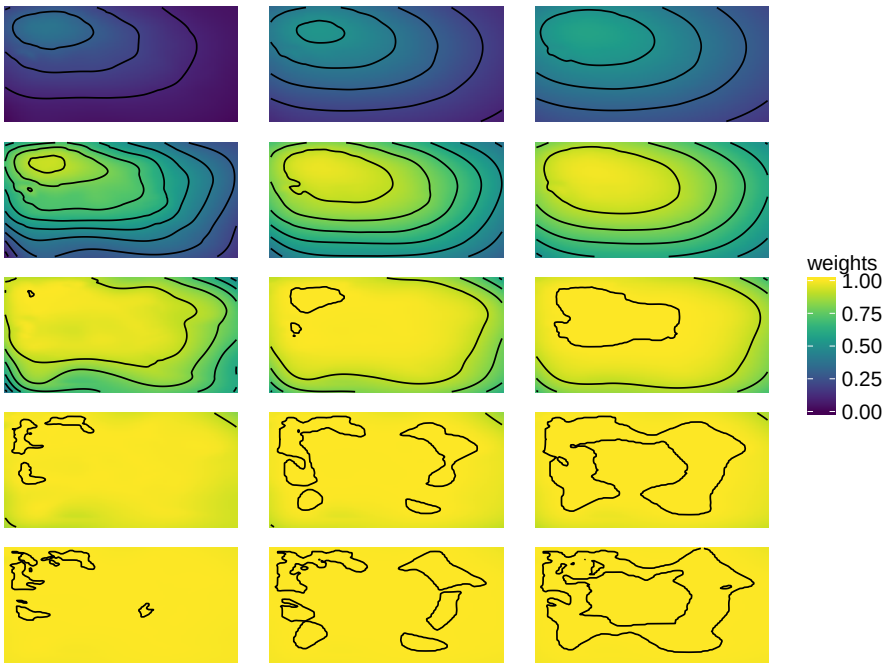

Figure 2: Interpolated surfaces of weights for the PCP for 15 combinations of variance ratio δ_0 and effective range d_0. Rows correspond to five values of $\delta_0 = 0.01, 0.1, 1, 10, 100$, increasing top to bottom. Columns correspond to three values of $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$, increasing left to right. The plot shows that the optimal parameterisation gives greater weight to the CP where sampling locations are more densely clustered.

3.2 A spatially varying slope model

We now investigate the effect of a covariate upon the spatially varying weights. To do this we look at the following model

$$\boldsymbol{Y} \mid \boldsymbol{\beta}_1^w, \theta_1 \sim N\{\boldsymbol{D}\boldsymbol{\beta}_1^w + (\boldsymbol{I} - \boldsymbol{W})\mathbf{1}\theta_1, \sigma_\epsilon^2 \boldsymbol{I}\}, \quad \boldsymbol{\beta}_1^w \mid \theta_1 \sim N(\boldsymbol{W}\mathbf{1}\theta_1, \sigma_1^2 \boldsymbol{R}_1), \tag{13}$$

$\theta_1 \sim N(m_1, \sigma_1^2 v_1)$, where $\boldsymbol{D} = \text{diag}(\boldsymbol{x})$ and $\boldsymbol{x} = \{x(\boldsymbol{s}_1), \ldots, x(\boldsymbol{s}_n)\}^{\mathrm{T}}$ 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 (13) can be retrieved from model (4) by letting $\boldsymbol{X}_1 = \boldsymbol{D}$, $\boldsymbol{C}_1 = \sigma_\epsilon^2 \boldsymbol{I}$, $\boldsymbol{X}_2 = \mathbf{1}$, $\boldsymbol{C}_2 = \sigma_1^2 \boldsymbol{R}_1$, $\boldsymbol{m} = m_1$ and $\boldsymbol{C}_3 = \sigma_1^2 v_1$.

For model (13) the $\boldsymbol{W}$ matrix is $\boldsymbol{W} = \sigma_1^2 \boldsymbol{R}_1 \boldsymbol{D}(\sigma_\epsilon^2 \boldsymbol{I} + \sigma_1^2 \boldsymbol{D}\boldsymbol{R}_1\boldsymbol{D})^{-1}\boldsymbol{D}$. To investigate how these weights vary across the domain we randomly select 200 points uniformly over the unit square. We generate the values of $\boldsymbol{x}$ by selecting 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 $\boldsymbol{s}_i$ decays exponentially at rate ϕ_x with increasing separation from $\boldsymbol{s}_x$, so that $x(\boldsymbol{s}_i) = \exp(-\phi_x \|\boldsymbol{s}_i - \boldsymbol{s}_x\|)$ for $i = 1, \ldots, n$. 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 3 gives the interpolated covariate surface where $\boldsymbol{s}_x = (0.936, 0.117)^{\mathrm{T}}$. We can see how the values decay with increased separation from $\boldsymbol{s}_x$.

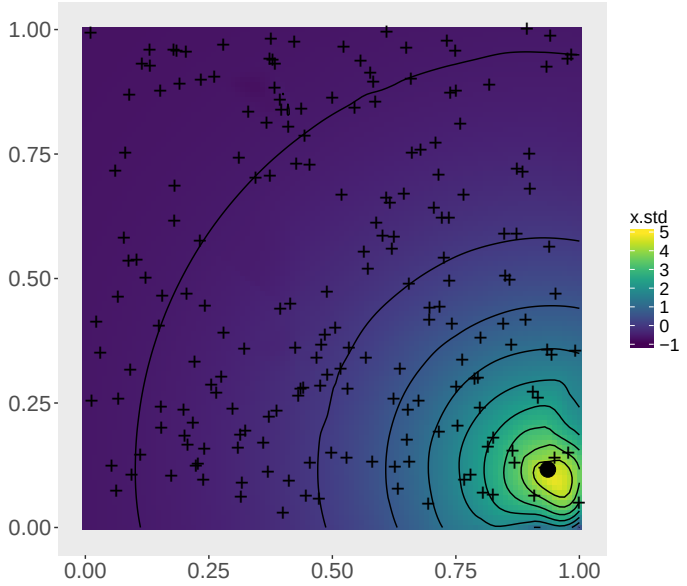

Figure 3: Interpolated surface of $\boldsymbol{x}$ from the 200 uniformly sampled data locations, given by '+'. The values of $\boldsymbol{x}$ decay with distance from the source, represented by a black dot. We use $\boldsymbol{x}$ to analyse the effect of covariates on the spatially varying weights for the PCP, see Figure 4.

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 illustrated in Figure 4. The weights increase with increasing δ_1 or d_1. It is also evident that for locations near $\boldsymbol{s}_x$, where the values of the covariate are greatest, the weights are higher.

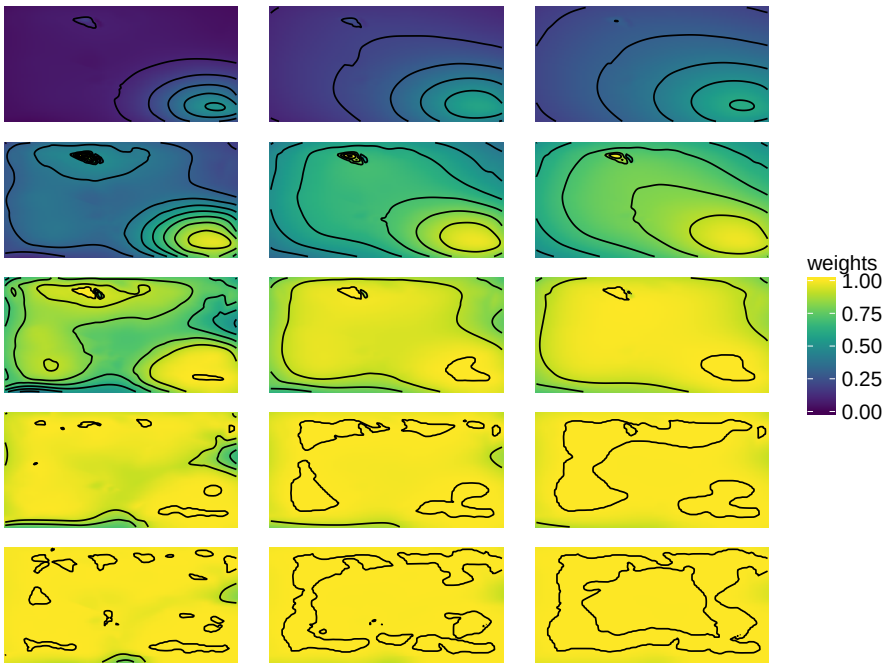

Figure 4: Interpolated weights for the PCP for 15 combinations of δ_1 and d_1. Rows correspond to five values of $\delta_0 = 0.01, 0.1, 1, 10, 100$, increasing top to bottom. Columns correspond to three values of $d_0 = \sqrt{2}/3, 2\sqrt{2}/3, \sqrt{2}$, increasing left to right. Weights are greatest where the value of $\boldsymbol{x}$ is at its greatest, c.f. Figure 3.

4 Simulation studies

4.1 Simulation example 1

In this section we use simulated data to investigate the performance of the Gibbs sampler associated with the PCP. All of the relevant posterior distributions can be found in Appendix A.6 in Supplementary materials. We simulate data from model (11) for $n = 40$ randomly chosen locations across the unit square. We let hyperparameters $m_0 = 0$ and $v_0 = 10^4$.

We set $\theta_0 = 0$ 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, 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.

To begin the 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.

The efficiency of the sampler is judged by two measures, both of which are calulated using functions within the R package coda Plummer *et al.* (2006). The first statistic we use is based on the potential scale reduction factor (PSRF) (Gelman and Rubin, 1992). We define the $\text{PSRF}_M(1.1)$ to be the number of iterations required for the upper limit of the 95 % confidence interval of the PSRF to fall below 1.1 for the first time.

To compute the $\text{PSRF}_M(1.1)$ we simulate multiple chains from widely dispersed starting values. In particular, we take values that are outside of the intervals described by pilot chains. At every fifth iteration the PSRF is calculated and the number of iterations for its value to drop below 1.1 is the value that we record. The second statistic we use is the effective sample size (ESS) of θ_0, (Robert and Casella, 2004, Chapter 12).

For each data set we generate five chains of length 25,000 and compute the $\text{PSRF}_M(1.1)$ and the ESS of θ_0 out of a total of 125,000 samples. The results are plotted in Figure 5. On the top row we have the $\text{PSRF}_M(1.1)$ and on the bottom the ESS. The five panels in each row correspond to values of $\delta_0 = 0.01, 0.1, 1, 10, 100$ from left to right and within each panel we have four boxplots, one for each level of the effective range 0, $\sqrt{2}/3$, $2\sqrt{2}/3$ and $\sqrt{2}$. Each boxplot consists of 20 values, for the 20 repetitions of that variance ratio-effective range combination. The figure 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 ratio and strength of correlation. In some instances we observe an ESS greater than the total number of samples. This occurs when the estimate of the integrated autocorrelation time is less than zero, which is theoretically possible, see e.g. Besag and Green (1993).

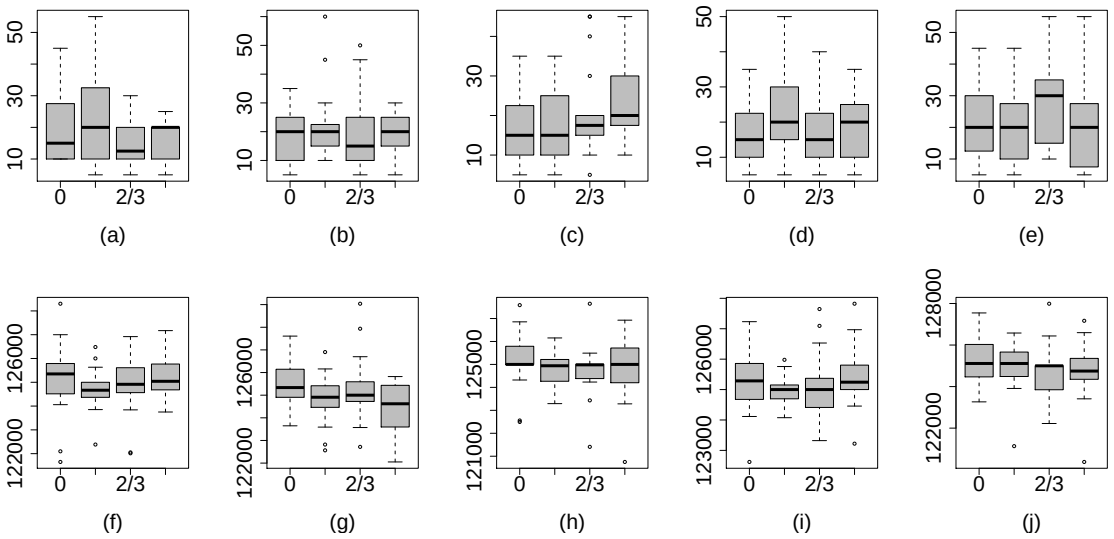

Figure 5: $\text{PSRF}_M(1.1)$, panels (a)–(e), and the ESS, panels (f)–(j), of θ_0 for the PCP with known variance parameters. The plots show that the performance of the PCP is insensitive to values of the variance and spatial decay parameters.

4.2 Simulation example 2

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 iterations needed for the multivariate PSRF (Brooks and Gelman, 1998) to fall below 1.1 for the first time. Note that the MPSRF is approximate upper bound to the maximum of the univariate PRSF's over all monitored variables, see (Brooks and Gelman, 1998), Lemma 3. In addition, we record the ESS of θ_0, σ_0^2 and σ_ϵ^2.

Recall that 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).

Figure 6 shows the $\text{MPSRF}_M(1.1)$ on the top row and the ESS of θ_0 on the bottom row for the 20 combinations of δ_0 and d_0 as in the case of Figure 5 . There is more variability in the results seen here for the $\text{MPSRF}_M(1.1)$ than we saw for the $\text{PSRF}_M(1.1)$ when the variance parameters were fixed, Figure 5. When the random effects are independent, weak identifiability of the variance parameters can effect the performance of the sampler as marginally $var\{Y(\boldsymbol{s}_i)\} = \sigma_\epsilon^2 + \sigma_0^2$. 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.

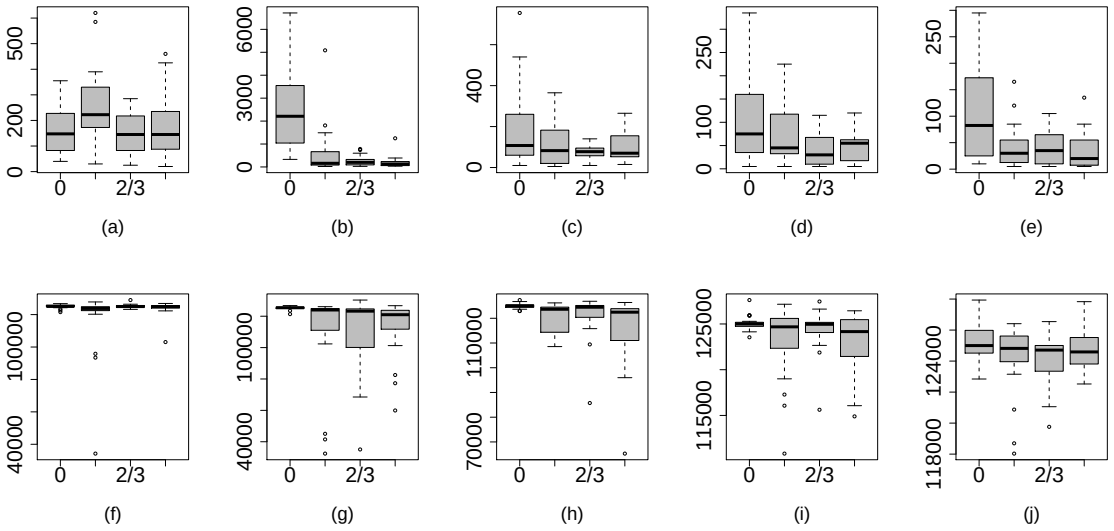

Figure 6: MPSRF$_M(1.1)$, panels (a)–(e), and the ESS panels (f)–(j), of θ_0 for the PCP with unknown variance parameters. Now we are updating the parameterisation within the sampler, but performance is still good and insensitive to the values of the variance and spatial decay parameters.

Boxplots of the ESS of the variance parameters are given in Figure 7, 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.

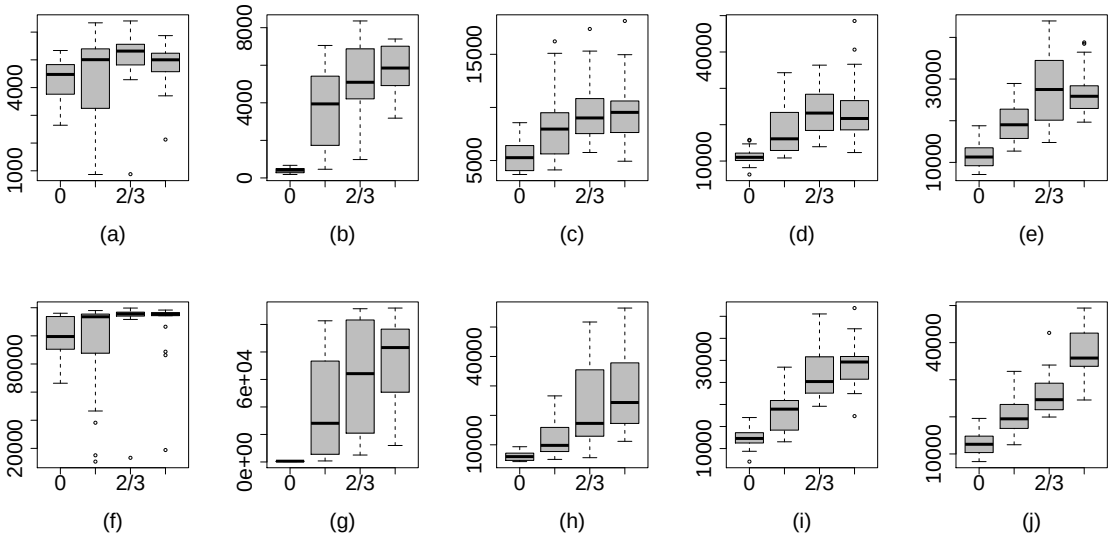

Figure 7: ESS of σ_0^2, panels (a)–(e), and σ_ϵ^2, panels (f)–(j) for the PCP with unknown variance parameters. The ESS of the variance parameters is not strongly effected by their ratio but does improve with stronger spatial correlation.

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 1 shows the mean $\text{MPSRF}_M(1.1)$ and mean ESS for θ_0. The sampling efficiency of the CP and the NCP is dependent on the covariance parameters. The CP performs best for higher values of δ_0 and longer effective ranges, and the NCP performs best for lower values of δ_0 and shorter effective ranges. In contrast, the PCP is robust to changes in δ_0 and d_0. Moreover, we see that the PCP has a lower average $\text{MPSRF}_M(1.1)$ for most cases, and when it does not, the difference is less than 3%. It is clear that, in terms of the ESS of θ_0, the PCP is superior to the CP and the NCP in all cases.

The ESS of the variance parameters is not strongly effected by the fitting methods we consider here as a reparameterisation only acts upon the mean structure of the model. However, in cases where a particular parameterisation is very inefficient, for example for the NCP when $\delta_0 > 1$, poor mixing can occur for the variance parameters.

5 London air pollution modelling example

In this section we compare the efficiency of the CP, the NCP and the PCP when fitting model (1) to a data set of NO_2 concentration levels, downloaded from *www.londonair.org.uk*, for the city of Greater London. It is a spatial data set with values, in microgram per cubic metre ($\mu g/m^3$), of the annual mean concentration for the year 2011. Data have been

Table 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**

collected at 63 irregularly spaced locations across London. We fit the model using data from 47 sites, leaving out 16 sites for validation (see Figure 8). The mean and standard deviation for the 47 data sites is 81.8 $\mu g/m^3$ and 38.7 $\mu g/m^3$ respectively. This annual mean value is quite high compared to the annual limit value of $40\mu g/m^3$, which has been stated in the national air quality objectives [1]. This objective has been in place since 2005 and so was in effect prior to the collection of these data. Also note that there is considerable variability in the data as is evident by the high value of the standard deviation.

Figure 8: Sampling locations for NO_2 concentration data.

The spatially varying covariate we use is output from the Air Quality Unified Model (AQUM), a numerical model giving air pollution predictions at 1-km grid cell resolution (Savage *et al.*, 2013). The AQUM is used as a covariate in the model, where $x(\boldsymbol{s})$ is the AQUM output for the grid cell containing s. Therefore, we use a downscaler model as employed by Berrocal *et al.* (2010).

As we have a spatially varying coefficient we fit the different parameterisations of model (1) with $p = 2$. Therefore we have two spatial processes, an intercept and a slope process and so $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_0^{\mathrm{T}}, \tilde{\boldsymbol{\beta}}_1^{\mathrm{T}})^{\mathrm{T}}$ for the CP, $\boldsymbol{\beta} = (\boldsymbol{\beta}_0^{\mathrm{T}}, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ for the NCP and $\boldsymbol{\beta}^w = (\boldsymbol{\beta}_0^{w\mathrm{T}}, \boldsymbol{\beta}_1^{w\mathrm{T}})^{\mathrm{T}}$ for the PCP. Each process has a corresponding global parameter and a variance parameter and so $\boldsymbol{\theta} = (\theta_0, \theta_1)^{\mathrm{T}}$ and $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)^{\mathrm{T}}$. We use an exponential correlation function for both processes and so $\boldsymbol{\phi} = (\phi_0, \phi_1)^{\mathrm{T}}$. In addition we have the data variance, σ_ϵ^2, and so we have $2(n+3)+1$ parameters to estimate for each parameterisation.

For the prior distribution of $\boldsymbol{\theta}$ we let $\boldsymbol{m} = (0, 0)^{\mathrm{T}}$ and $v_0 = v_1 = 10^4$. We let $a_0 = a_1 =$

[1] https://uk-air.defra.gov.uk/air-pollution/uk-eu-limits, last modified March 3, 2017

$a_\epsilon = 2$ 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.

As mentioned in Section 2.1 we estimate the spatial decay parameters by performing a grid search over a range of values for ϕ_0 and ϕ_1. 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 square prediction error (RMSPE) and the continuous ranked probability score (CRPS), see Gneiting *et al.* (2007) for example.

We select values of ϕ_0 and ϕ_1 corresponding to effective ranges of 5, 10, 25, 50, 100 and 250 km. For each of the 36 pairs of spatial decay parameters we generate a single chain of 25,000 iterations and discard the first 5,000. In this way the estimates for the spatial decay parameters are $\hat{\phi}_0 = -\log(0.05)/5 \approx 0.6$, and $\hat{\phi}_1 = -\log(0.05)/100 \approx 0.03$.

For each parameterisation we generate five Markov chains of length 25,000 from the same set widely dispersed starting values. The $\text{MPSRF}_M(1.1)$ and the ESS for $\boldsymbol{\theta} = (\theta_0, \theta_1)^{\mathrm{T}}$, $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)^{\mathrm{T}}$ and σ_ϵ^2 are computed and given in Table 2.

In this example, and also previously in the simulation examples, we evaluate the efficiency of the sampling algorithms for the fixed effects and global variance parameters which are of interests for inferential purposes. However, we have also monitored the convergence for a random sample of the spatially varying random effects, $\beta(\boldsymbol{s})$. We did not encounter any problem in their convergence behaviour.

Table 2: $\text{MPSRF}_M(1.1)$ 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
PCP	360	120956	121092	5996	7499	20925
CP	460	20548	57177	2279	7578	2452
NCP	1230	15132	4421	2273	5038	2392

We see that the CP requires far fewer iterations for the MPSRF to drop below 1.1 than the NCP, 460 versus 1230. The short effective range for the intercept process means that the ESS for θ_0 is only slightly greater for the CP than the NCP, but due to the stronger spatial association in the regression process the ESS for θ_1 is nearly 13 times greater for the CP than the NCP. However, the PCP, whose performance insensitive to the values of the model parameters, demonstrates extremely low autocorrelation between successive samples with an ESS greater than 120,000 for both θ_0 and θ_1.

The run times for the CP and the NCP are almost the same, but updating $\boldsymbol{W}$ within the sampler means that the PCP is more computationally demanding. Table 3 gives the

same measures as Table 2 but adjusted for computation time. 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.

Even accounting for the shorter run times of the CP, the PCP is comparable in terms of $\text{MPSRF}_t(1.1)$ and ESS/s of the variance components, but most importantly it is superior in terms of ESS/s for the mean parameters.

Table 3: $\text{MPSRF}_t(1.1)$ 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
PCP	6.96	50.06	50.12	2.48	3.10	8.66
CP	4.89	15.47	43.05	1.72	5.71	1.85
NCP	13.07	11.39	3.33	1.71	3.79	1.80

Figure 9 shows the spatially varying parameterisation that results from the PCP. Contained are interpolated maps of the average weight of partial centering given by each process to each global parameter for all 63 sampling locations. The top row displays the weights given by $\boldsymbol{\beta}_0^w$ to θ_0 on the left, and the weights given by $\boldsymbol{\beta}_1^w$ to θ_1 on the right. To restrict the interpolated weight values within the unit interval (0, 1), we first logit-transform the weights and then use the inverse logit transformation on the interpolated values.

The weights for θ_0 are low; the estimated weights at the data locations (not the interpolated weights) are between 0.11 and 0.38. This reflects the short effective range of the intercept process, which is 5 kilometres (km). The weights for θ_1 are much higher, between 0.50 and 0.88, and this is due to the longer effective range of the slope process, 100 km. The weights vary over the spatial locations and are neither exactly zero or one, which would correspond to the NCP and CP respectively. They are higher where there are clusters of monitoring sites. We see this behaviour in Figure 2. It is more clearly demonstrated when there is a shorter effective range as the influence of the clustering of data points is not so far reaching, thus the pattern can be discerned from the plot.

The bottom row displays the weights given by $\boldsymbol{\beta}_0^w$ to θ_1 on the left, and the weight given by $\boldsymbol{\beta}_1^w$ to θ_0 on the right. These take values in (–1, 1) and so we transform them to the unit interval before taking the logit transformation, then the interpolated weights are back-transformed.

When producing the plots, in order to have Greater London fully contained within the convex hull of the sampling locations, we move the most northerly observation site further north by adding 0.05 to its latitude. In addition to the packages needed to produce Figure 2, we use functions within maptools (Bivand and Lewin-Koh, 2016) to handle the shape files and the 'pointsInPoly' function in spBayes to remove those points on the interpolated

grid that lie outside of the polygon described by the Greater London Boundary.

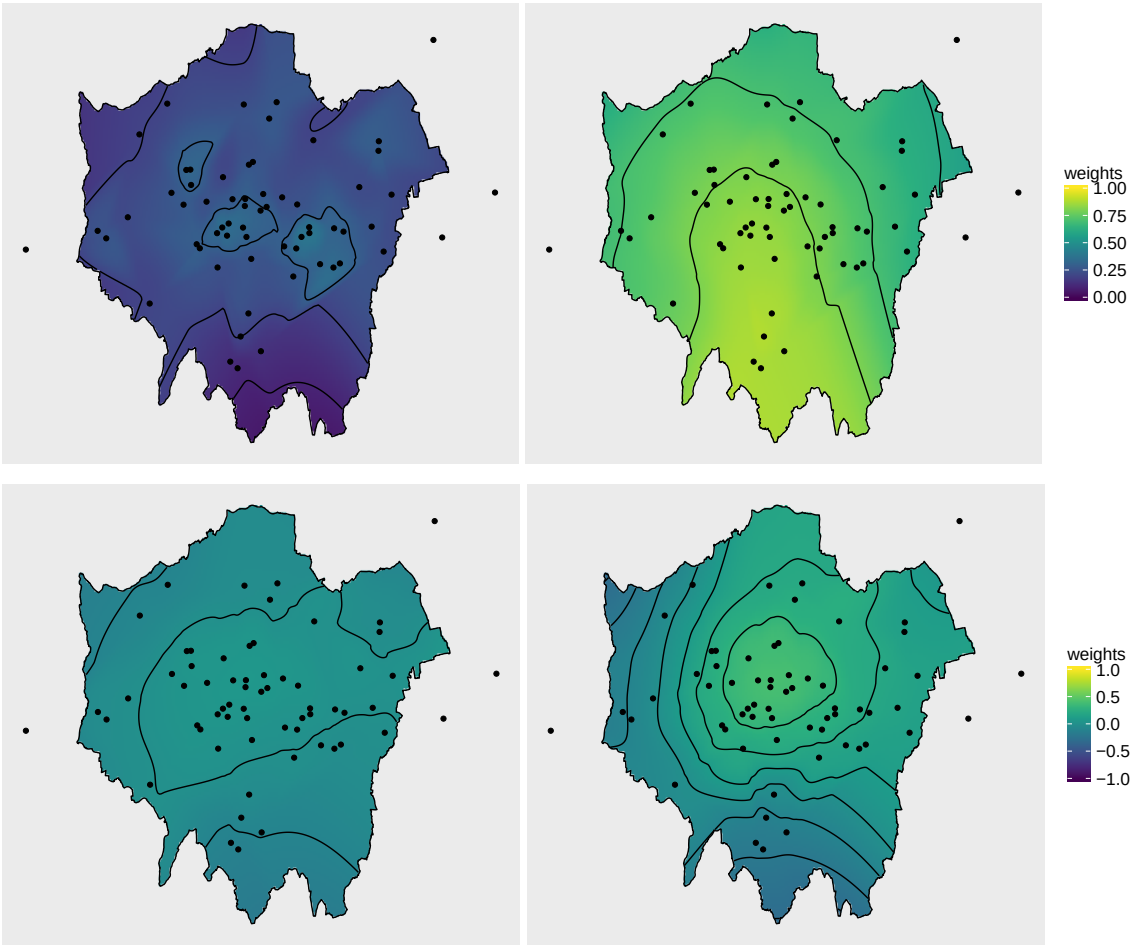

Figure 9: Spatially varying weights for London NO_2 concentration data. On the top row are the weights assigned to the global parameter corresponding to each process. On the left, the weight of θ_0 in the mean of $\boldsymbol{\beta}_0^w$, and on the right the weight of θ_1 in the mean of $\boldsymbol{\beta}_1^w$. The bottom row gives the weights for the alternate global parameter. On the left, the weight of θ_1 in the mean of $\boldsymbol{\beta}_0^w$, and on the right the weight of θ_0 in the mean of $\boldsymbol{\beta}_1^w$.

6 Discussion

We have investigated the performance of a PCP for the spatially varying coefficients model. We are able to parameterise the model in such a way that we remove the posterior covariance between the random and global effects and produce a Gibbs sampler which converges immediately. 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 higher.

We make it clear that although the interpolated plots given in Sections 3 and 5 are informative they do not represent a true surface in the sense that the interpolated values are not estimates of a true value of the weights at an unsampled location. Indeed, if we were to obtain a further measurement at a new location then the values of the weights at the existing locations would change.

Our investigations show that the unlike the CP and the NCP, the performance of the PCP is robust to changes in the variance and correlation parameters. 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. It is a computationally intensive strategy which we recommend for modest sized spatial datasets. However, for larger datasets many practitioners turn to Gaussian predictive process approximations, as implemented in R software packages spBayes (Finley *et al.*, 2015) and spTimer (Bakar and Sahu, 2015). As the method proposed in this paper can be applied to any model that can be written as a three stage NLHM, we believe that the PCP could be used in conjunction with GPP models thus broadening its applicability. We also note that PCP can be used for the Generalised linear models incorporating spatially varying random effects.

Acknowledgement

This research was supported in part by the EPSRC research grant EP/J017485/1 awarded to the correspinding author. The authors thank the referees for helpful comments.

Supplementary materials

Appendices: A.1 to A.6 containing proofs and further details as referenced in the main paper.

C code: All of the C code used to generate the MCMC output which is analysed to produce the results given in this article are contained within the folder called c_files. Full details are provided within the README.txt file contained within the folder.

Data: All data files can be found in the folder labelled data_files. A full description is given in the README.txt file contained therein

R code: The R code used to analyse the MCMC output produced by the C code can be found in the folder labelled R_scripts. A full description of the files is given in README.txt.

Figures: The figures produced by the R code included in this article are given in the folder labelled figures.

References

Akima, H. and Gebhardt, A. (2016). *akima: Interpolation of Irregularly and Regularly Spaced Data*. R package version 0.6-2.

Bakar, K. S. and Sahu, S. K. (2015). spTimer: Spatio-temporal Bayesian modeling using R. *Journal of Statistical Software*, **63**(15), 1–32.

Bass, M. R. and Sahu, S. K. (2017). A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC. *Statistics and Computing*, **27**, 1491–1512.

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**(2), 176–197.

Besag, J. and Green, Peter, J. (1993). Spatial statistics and Bayesian computation (with discussion). *Journal of the Royal Statistical Society. Series B*, **55**, 25–37.

Bivand, R. and Lewin-Koh, N. (2016). *maptools: Tools for Reading and Handling Spatial Objects*. R package version 0.8-39.

Brooks, S. P. and Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**(4), 434–455.

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

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**(493), 31–48.

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**(13), 1–28.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American statistical association*, **85**(410), 398–409.

Gelfand, A. E., Sahu, S. K., and Carlin, B. P. (1995). Efficient parameterisations for normal linear mixed models. *Biometrika*, **82**(3), 479–488.

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**(462), 387–396.

Gelman, A. and Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, **7**(4), 457–472.

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B*, **69**(2), 243–268.

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**(4), 403–410.

Harville, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York.

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*, **53**(2), 231–248.

Matérn, B. (1986). *Spatial Variation*. Springer Verlag, Berlin, 2nd. edition.

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., 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*, pages 307–326. Oxford University Press, USA.

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6**(1), 7–11.

R Core Team (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Robert, C. O. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer-Verlag New York, 2nd. edition.

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*, **59**(2), 291–317.

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**(480), 1221–1234.

Sahu, S. K., Gelfand, A. E., and Holland, D. M. (2010). Fusing point and areal level space–

time data with application to wet deposition. *Journal of the Royal Statistical Society: Series C*, **59**(1), 77–103.

Savage, N., Agnew, P., Davis, L., Ordóñez, C., Thorpe, R., Johnson, C., O'Connor, F., and Dalvi, M. (2013). Air quality modelling using the met office unified model (aqum os24-26): model description and initial evaluation. *Geoscientific Model Development*, **6**(2), 353–372.

Wheeler, D. C., Páez, A., Spinney, J., and Waller, L. A. (2014). A Bayesian approach to hedonic price analysis. *Papers in Regional Science*, **93**(3), 663–683.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wickham, H. (2017). *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.1.

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**(3), 531–570.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**(465), 250–261.

Supplementary materials

Appendices

A.1 Identifying the posterior precision matrix

To identify the precision matrix of the joint posterior distribution of $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$ for the PCP we write:

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

The entries of the precision matrix can then be read off of the final expression.

A.2 Convergence rate of the PCP

Consider $\boldsymbol{Q}^{pc}_{\beta^w\theta}$ and substitute $\boldsymbol{W}$ from equation (8), then we have

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

Therefore by setting $\boldsymbol{W} = \boldsymbol{I} - \boldsymbol{B}\boldsymbol{C}_2^{-1}$, $\boldsymbol{F}^{pc}_{22}$ becomes the null matrix and immediate convergence follows.

A.3 Convergence rate of the PCP

We now look at the implication of setting $\boldsymbol{W}$ according to (8) for the convergence rate of a Gibbs sampler using the PCP. 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.

Suppose that $\boldsymbol{\xi} \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ denote the posterior precision matrix. To compute the convergence rate first partition $\boldsymbol{Q}$ according to a number of blocks, denoted by l, that are used for updating $\boldsymbol{\xi}$, i.e.,

$$(\boldsymbol{Q})_{ij} = \boldsymbol{Q}_{ij}, \text{ for } i,j = 1,\ldots,l. \tag{14}$$

Let $\boldsymbol{A} = \boldsymbol{I} - \mathrm{diag}(\boldsymbol{Q}_{11}^{-1},\ldots,\boldsymbol{Q}_{ll}^{-1})\boldsymbol{Q}$ 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 (14), 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 $\boldsymbol{\Sigma} - \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}$. Their observation leads to the following:

Theorem A.1 *(Roberts and Sahu, 1997) A Markov chain with transition density*

$$N\{\boldsymbol{F}\boldsymbol{\xi}^{(t)} + \boldsymbol{f}, \boldsymbol{\Sigma} - \boldsymbol{F}\boldsymbol{\Sigma}\boldsymbol{F}^{\mathrm{T}}\},$$

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

Corollary A.2 *If we update $\boldsymbol{\xi}$ in two blocks so that $l = 2$ then*

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

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

To compute the convergence rate of the PCP we first need the posterior precision matrix of $\boldsymbol{\beta}^w$ and $\boldsymbol{\theta}$, which we can identify by writing down $\pi(\boldsymbol{\beta}^w, \boldsymbol{\theta}|\boldsymbol{y})$ explicitly (more details are provided in Appendix A.1 in Supplementary materials). The posterior precision matrix for the PCP is

$$\boldsymbol{Q}^{pc} = \begin{pmatrix} \boldsymbol{Q}^{pc}_{\beta^w} & \boldsymbol{Q}^{pc}_{\beta^w\theta} \\ \boldsymbol{Q}^{pc}_{\theta\beta^w} & \boldsymbol{Q}^{pc}_{\theta} \end{pmatrix}, \tag{15}$$

where $\boldsymbol{Q}^{pc}_{\beta^w} = \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}$, $\boldsymbol{Q}^{pc}_{\beta^w\theta} = \boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 - \boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2$, and $\boldsymbol{Q}^{pc}_{\theta} = \boldsymbol{X}_2^{\mathrm{T}}(\boldsymbol{I}-\boldsymbol{W})^{\mathrm{T}}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{C}_1^{-1}\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2 + \boldsymbol{X}_2^{\mathrm{T}}\boldsymbol{W}^{\mathrm{T}}\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}$. If we block update a Gibbs sampler according to the partitioning of the precision matrix (15), by Corollary A.2, we have that the convergence rate of the PCP is the maximum modulus eigenvalue of the matrix $\boldsymbol{F}^{pc}_{22} = (\boldsymbol{Q}^{pc}_{\theta})^{-1}\boldsymbol{Q}^{pc}_{\theta\beta^w}(\boldsymbol{Q}^{pc}_{\beta^w})^{-1}\boldsymbol{Q}^{pc}_{\beta^w\theta}$. 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 $\boldsymbol{F}^{pc}_{22}$ is null and immediate convergence is achieved.

A.4 Convergence rate of a three component Gibbs sampler

It can be shown that a Gibbs sampler with Gaussian target distribution with precision matrix given by $\boldsymbol{Q}$ having elements $(\boldsymbol{Q})_{ij} = \boldsymbol{Q}_{ij}$ for $i, j = 1, 2, 3$ has a convergence rate which is equal to the maximum modulus eigenvalue of

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

where

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

A.5 Proof of stationarity of the PCP

To demonstrate that stationarity is preserved we let $p = 1$ in model (4). The transition kernel of the Markov chain is:

$$\begin{aligned} P\{\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)}) &= \pi\{\boldsymbol{\beta}^{w(t+1)}|\theta_0^{(t)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{y}\}\pi\{\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{y}\} \\ &\quad \pi\{\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t)}, \sigma_\epsilon^{2(t)}, \boldsymbol{y}\} \\ &\quad \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)}, \theta_0^{(t+1)}, \sigma_0^{2(t+1)}, \sigma_\epsilon^{2(t)}, \boldsymbol{y}\}. \end{aligned}$$

We have dropped the $\boldsymbol{W}$'s to save space, conditioning the variance parameters on their current values where necessary. It follows that

$$\begin{aligned}
&\int P\{\boldsymbol{\xi}^{(t+1)}|\boldsymbol{\xi}^{(t)}\}\pi(\boldsymbol{\xi}^{(t)}|\boldsymbol{y}\}d\boldsymbol{\xi}^{(t)} \\
&= \int \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\} \\
&\quad \pi\{\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\pi\{\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}d\boldsymbol{\xi}^{(t)}
\end{aligned}$$

$$\begin{aligned}
&= \int \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\pi\{\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\} \\
&\quad \underbrace{\pi\{\theta_0^{(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\left[\int \pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}d\theta_0^{(t)}\right]}_{=\pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}} d\sigma_0^{2(t)}d\sigma_\epsilon^{2(t)} \\
&= \int \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\} \\
&\quad \underbrace{\left[\int \pi\{\sigma_0^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}d\sigma_0^{2(t)}\right]}_{=\pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}} d\sigma_\epsilon^{2(t)} \\
&= \int \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}d\sigma_\epsilon^{2(t)} \\
&= \pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)}|\boldsymbol{y}\} \\
&= \pi\{\boldsymbol{\xi}^{(t+1)}|\boldsymbol{y}\},
\end{aligned}$$

and hence stationarity is preserved. The above argument can easily be extended for $p > 1$ or to include other correlation parameters if they are being modelled.

If we update $\boldsymbol{W}$ and the end of each complete pass of the sampler then the stationarity condition (10) does not hold. For instance, consider σ_ϵ^2, which is conditioned on σ_0^2 through $\boldsymbol{W}$. 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

$$\begin{aligned}
&\int \pi\{\sigma_\epsilon^{2(t+1)}|\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t)},\sigma_\epsilon^{2(t)},\boldsymbol{y}\}\pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t)}|\boldsymbol{y}\}d\sigma_\epsilon^{2(t)} \\
&\neq \pi\{\boldsymbol{\beta}_0^{w(t+1)},\theta_0^{(t+1)},\sigma_0^{2(t+1)},\sigma_\epsilon^{2(t+1)}|\boldsymbol{y}\},
\end{aligned}$$

but equality is required to complete step (16) in the string of equalities proving stationarity.

A.6 Joint posterior and full conditional distributions

We begin here by writing down the joint posterior distribution of the parameters in model (4). We let $\boldsymbol{\xi} = (\boldsymbol{\beta}^{w\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\sigma^2}^{\mathrm{T}}, \sigma_\epsilon^2)^{\mathrm{T}}$ 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

$$\begin{aligned}
\pi(\boldsymbol{\xi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\boldsymbol{\beta}^w,\boldsymbol{\theta},\sigma_\epsilon^2)\pi(\boldsymbol{\beta}^w|\boldsymbol{\theta},\boldsymbol{\sigma^2})\pi(\boldsymbol{\theta}|\boldsymbol{\sigma^2})\pi(\boldsymbol{\sigma^2})\pi(\sigma_\epsilon^2) \\
&\propto \prod_{k=0}^{p-1}\left(\sigma_k^2\right)^{-(n/2+1/2+a_k+1)}|\boldsymbol{R}_k|^{-1/2}\left(\sigma_\epsilon^2\right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp\left\{-\frac{1}{2\sigma_\epsilon^2}\Big(\Big[\boldsymbol{Y}-\boldsymbol{X}_1\{\boldsymbol{\beta}^w+(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\}\Big]^{\mathrm{T}}\Big[\boldsymbol{Y}-\boldsymbol{X}_1\{\boldsymbol{\beta}^w+\right.\right. \\
&\quad \left.\left.(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\}\Big]+2b_\epsilon\Big)\right\}\exp\left\{-\frac{1}{2}\Big(\boldsymbol{\beta}^w-\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\Big)^{\mathrm{T}}\boldsymbol{C}_2^{-1}\Big(\boldsymbol{\beta}^w-\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\Big)\right\} \\
&\quad \exp\left[-\frac{1}{2}\sum_{k=0}^{p-1}\frac{1}{\sigma_k^2}\left\{\frac{(\theta_k-m_k)^2}{v_k}+2b_k\right\}\right],
\end{aligned}$$

where a description of the prior distributions $\pi(\boldsymbol{\sigma^2})$ and $\pi(\sigma_\epsilon^2)$ can be found in Section 2.1.

It is argued in Section A.3 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 $\boldsymbol{\beta}^w|\boldsymbol{\theta},\boldsymbol{\sigma^2},\sigma_\epsilon^2,\boldsymbol{y}\sim N\left(\boldsymbol{m}_\beta^*,\boldsymbol{C}_2^*\right)$, where $\boldsymbol{C}_2^*=\left(\sigma_\epsilon^{-2}\boldsymbol{X}_1^{\mathrm{T}}\boldsymbol{X}_1+\boldsymbol{C}_2^{-1}\right)^{-1}$ and $\boldsymbol{m}_\beta^*=\boldsymbol{C}_2^*\left[\sigma_\epsilon^{-2}\{\boldsymbol{y}-\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\}+\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2\boldsymbol{\theta}\right]$.

- The full conditional distribution of $\boldsymbol{\theta}$ is $\boldsymbol{\theta}|\boldsymbol{\beta}^w,\boldsymbol{\sigma^2},\sigma_\epsilon^2,\boldsymbol{y}\sim N\left(\boldsymbol{m}_\theta^*,\boldsymbol{C}_3^*\right)$, where

$$\begin{aligned}
\boldsymbol{C}_3^* &= \left[\sigma_\epsilon^{-2}\{\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\}^{\mathrm{T}}\{\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\}+(\boldsymbol{W}\boldsymbol{X}_2)^{\mathrm{T}}\boldsymbol{C}_2^{-1}\boldsymbol{W}\boldsymbol{X}_2+\boldsymbol{C}_3^{-1}\right]^{-1}, \\
\boldsymbol{m}_\theta^* &= \boldsymbol{C}_3^*\left[\sigma_\epsilon^{-2}\{\boldsymbol{X}_1(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\}^{\mathrm{T}}(\boldsymbol{y}-\boldsymbol{X}_1\boldsymbol{\beta}^w)+(\boldsymbol{W}\boldsymbol{X}_2)^{\mathrm{T}}\boldsymbol{C}_2^{-1}\boldsymbol{\beta}^w+\boldsymbol{C}_3^{-1}\boldsymbol{m}\right].
\end{aligned}$$

- The full conditional distribution of σ_k^2 is $\sigma_k^2|\boldsymbol{\beta}^w,\boldsymbol{\theta},\boldsymbol{\sigma^2}_{-k},\sigma_\epsilon^2,\boldsymbol{y}$

$$IG\left[\frac{n+1}{2}+a_k,\ b_k+\frac{1}{2}\left\{\left(\boldsymbol{\beta}_k^w-\boldsymbol{\eta}_k\right)^{\mathrm{T}}\boldsymbol{R}_k^{-1}\left(\boldsymbol{\beta}_k^w-\boldsymbol{\eta}_k\right)+\frac{(\theta_k-m_k)^2}{v_k}\right\}\right],$$

  for $k=0,\ldots,p-1$, where $\boldsymbol{W}_{km}$ denotes the kmth, $n\times n$ block of $\boldsymbol{W}$ and $\boldsymbol{\eta}_k=\theta_k\sum_{m=0}^{p-1}\boldsymbol{W}_{km}\boldsymbol{1}$.

- The full conditional distribution of σ_ϵ^2 given $\boldsymbol{\beta}^w,\boldsymbol{\theta},\boldsymbol{\sigma^2},\boldsymbol{y}$ is

$$IG\left[\frac{n}{2}+a_\epsilon,\quad b_\epsilon+\frac{1}{2}\left\{\left(\boldsymbol{Y}-\boldsymbol{Z}\right)^{\mathrm{T}}\left(\boldsymbol{Y}-\boldsymbol{Z}\right)\right\}\right],$$

  where $\boldsymbol{Z}=\boldsymbol{X}_1\left(\boldsymbol{\beta}^w+(\boldsymbol{I}-\boldsymbol{W})\boldsymbol{X}_2\boldsymbol{\theta}\right)$.