Stat Comput
DOI 10.1007/s11222-016-9700-z

A comparison of centring parameterisations of Gaussian process-based models for Bayesian computation using MCMC

Mark R. Bass[1] · **Sujit K. Sahu**[1]

Received: 18 February 2016 / Accepted: 30 August 2016

Abstract Markov chain Monte Carlo (MCMC) algorithms for Bayesian computation for Gaussian process-based models under default parameterisations are slow to converge due to the presence of spatial- and other-induced dependence structures. The main focus of this paper is to study the effect of the assumed spatial correlation structure on the convergence properties of the Gibbs sampler under the default non-centred parameterisation and a rival centred parameterisation (CP), for the mean structure of a general multi-process Gaussian spatial model. Our investigation finds answers to many pertinent, but as yet unanswered, questions on the choice between the two. Assuming the covariance parameters to be known, we compare the exact rates of convergence of the two by varying the strength of the spatial correlation, the level of covariance tapering, the scale of the spatially varying covariates, the number of data points, the number and the structure of block updating of the spatial effects and the amount of smoothness assumed in a Matérn covariance function. We also study the effects of introducing differing levels of geometric anisotropy in the spatial model. The case of unknown variance parameters is investigated using well-known MCMC convergence diagnostics. A simulation study and a real-data example on modelling air pollution levels in London are used for illustrations. A generic pattern emerges that the CP is preferable in the presence of more spatial correlation or more information obtained through, for example, additional data points or by increased covariate variability.

Keywords Bayesian inference · Gibbs sampler · Hierarchical models · Rate of convergence · Spatial data

✉ Sujit K. Sahu
S.K.Sahu@soton.ac.uk

[1] Mathematical Sciences, University of Southampton, Southampton SO17 1BJ, UK

1 Introduction

Spatially correlated data are 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, in particular Gibbs sampling type algorithms (Gelfand and Smith 1990). There is a growing interest among researchers in regression models with spatially varying coefficients (Gelfand et al. 2003). Fitting these highly overparameterised and nonstationary models is challenging and computationally expensive. Latent processes correlated across space produce dense covariance matrices that require calculations of order $O(n^3)$ to invert, for n spatial locations (Cressie and Johannesson 2008).

For normal linear hierarchical models with independent random effects, it is known that the ratio of the variance parameters determines the convergence rates of the Gibbs samplers (Papaspiliopoulos et al. 2003; Gelfand et al. 1995). When the data precision is relatively high, the centred parameterisation (CP) will yield an efficient Gibbs sampler, and when the data precision is relatively low, the non-centred parameterisation (NCP) is more efficient. 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 quicker when the data model has lighter tails than that of the latent variables, with the opposite scenario favouring the NCP.

Published online: 14 September 2016

There has been little investigation into the influence of correlation across the random effects on the rate of convergence of the Gibbs sampler. Simulation studies conducted by Papaspiliopoulos et al. (2003) on the spatial Poisson-log-Normal model suggest that stronger spatial correlation improves the sampling efficiency of the CP relative to that of the NCP. However, there are several unresolved questions regarding the choice of the CP versus NCP for the mean structure of a general multi-process Gaussian spatial model. Which of the two parameterisations will converge faster when spatial correlation is increased? What happens to the rates of convergence when tapering (Furrer et al. 2006; Kaufman et al. 2008) is introduced? How does the smoothness parameter in an assumed Matérn covariance function influence the rates? In addition, there are other unexplored issues regarding the choice and the number of blocks for the random effects, the influence of the scale of spatially varying covariates and the introduction of different levels of geometric anisotropy.

In this paper, we cast the general spatial model with multiple spatially varying covariates as a three stage normal linear hierarchical model. This model formulation allows us to compute the exact rates of convergence for both CP and NCP for known prior covariance matrices by following (Roberts and Sahu 1997). These exact rates of convergence facilitate comparison between the two rival parameterisations, CP and NCP. For an exponential correlation function, convergence for the CP is hastened when spatial correlation is stronger, the opposite being true for the NCP. This is also demonstrated in the context of tapered covariance matrices, geometric anisotropic correlation functions and the regression process associated with a spatially varying covariate. The exponential correlation function is a member of the broader Matérn family (Matérn 1986). When we increase the smoothness parameter, the effect is to slow convergence for both the CP and NCP. In the limiting case of the Gaussian correlation, when the smoothness parameter tends to infinity, both CP and NCP may fail to converge when the sample size is large enough due to the associated singularity of the covariance matrices, and Sect. 3.3 invesigates the issues.

When the prior covariance matrices are unknown, the exact convergence is intractable and so the CP and NCP are compared by statistics based on well-known convergence diagnostics on the potential scale reduction factor, see e.g., Gelman and Rubin (1992). We use a simulation and a real-data example to show that increasing the effective range for an exponential correlation function improves the sampling efficiency of the CP, whereas shortening the effective range helps the NCP.

The following remarks are in order. First, a related approach is to marginalise over the random effects, thus reducing the dimension of the posterior distribution. This approach 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 spatially varying coefficient models and by Banerjee et al. (2008) to implement Gaussian predictive process 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. On the other hand, the Gibbs sampler for the full model can potentially be completely automated and run without the need for any tuning.

Secondly, it is possible to generate intermediate partially centred parameterisations by considering CP and the NCP as extremes of a family of parameterisations. Indeed, this has been followed up by Bass and Sahu (2016) in a companion paper. Interweaving of the CP and NCP as proposed by Yu and Meng (2011) is particularly useful when the practitioner has little knowledge of the convergence properties of either parameterisation. These authors obtain an upper bound on the convergence rate of the interweaving algorithm based on an intractable maximal correlation between the latent variables under the two parameterisations and to our knowledge the exact convergence rate for the interweaving algorithm has not yet been computed. Lastly, both CP- and NCP-based computation methods are similar in spirit to various data augmentation (DA) schemes (Liu and Wu 1999; van Dyk and Meng 2001; Imai and van Dyk 2005; Filippone et al. 2013). A direct theoretical comparison between the exact convergence rates of the centring parameterisations for Gaussian process (GP)-based models and the DA algorithms is desirable, as has been done by Sahu and Roberts (1999) for the EM algorithm and the Gibbs sampler. However, this requires further methodological development of DA algorithms for the GP models and evaluation of their exact rates of convergence.

The rest of this paper is organized as follows. In Sect. 2, we give details of a general spatial model and obtain expressions for the rates of convergence. A simple example here illustrates the rates and brings out the rivalry between the two parameterisations. Section 3 is devoted to comparison of the rates of convergence under different settings of correlation structures and introduction of geometric anisotropy. It also studies the effect of tapering and the scale of the spatially varying covariates on the rates of convergence. In Sect. 4, we drop the assumption of known precision matrices and use two convergence diagnostics to judge the sampling efficiencies of the CP and NCP methods. Analyses are carried out on simulated data and PM10 concentration data taken from Greater London in 2011. Section 5 contains some concluding remarks. Appendices A and B, respectively, contain the technical details for calculating the rates of convergence and the full conditional distributions needed for Gibbs sampling.

2 General spatial model

2.1 Model specification

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

$$Y(\boldsymbol{s}_i) = \sum_{k=0}^{p-1} \{\theta_k + \beta_k(\boldsymbol{s}_i)\} x_k(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i) \quad (i = 1, \ldots, n). \quad (1)$$

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

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

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 independent 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 $\boldsymbol{\beta}_k = \{\beta_k(\boldsymbol{s}_1), \ldots, \beta_k(\boldsymbol{s}_n)\}^{\mathrm{T}} \sim N(0, \boldsymbol{\Sigma}_k)$ $(k = 0, \ldots, p-1)$, where $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{R}_k$, 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 known as 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

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

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 $\boldsymbol{x}_k = \{x_k(\boldsymbol{s}_1), \ldots, x_k(\boldsymbol{s}_n)\}^{\mathrm{T}}$. 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 $\boldsymbol{\Sigma}_k$.

2.2 Prior distributions

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 an independent Gaussian prior distribution with mean m_k and variance $\sigma_k^2 v_k$, and hence, we write $\theta_k \sim N(m_k, \sigma_k^2 v_k)$ for $k = 0, \ldots, p-1$. Therefore, $\boldsymbol{m} = (m_0, \ldots, m_{p-1})^{\mathrm{T}}$ and $\boldsymbol{C}_3$ is a diagonal matrix with diagonal entries $\sigma_k^2 v_k$.

The realisations of the kth non-centred Gaussian process, $\boldsymbol{\beta}_k$, have a prior covariance matrix $\boldsymbol{\Sigma}_k = \sigma_k^2 \boldsymbol{R}_k$. This prior covariance matrix is shared by the kth centred Gaussian process, $\tilde{\boldsymbol{\beta}}_k$. The prior distributions for the variance parameters are $\sigma_k^2 \sim IG(a_k, b_k)$ $(k = 0, \ldots, p-1)$, $\sigma_\epsilon^2 \sim IG(a_\epsilon, b_\epsilon)$, 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 $(\boldsymbol{R}_k)_{ij} = corr\{\beta_k(\boldsymbol{s}_i), \beta_k(\boldsymbol{s}_j)\} = \rho_k(d_{ij}; \phi_k, \nu_k)$ 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 (Handcock and Stein 1993; Matérn 1986).

The Matérn correlation function for a pair of random variables at sites $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ is

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

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, Sect. 9.6). The parameter ϕ controls the rate of decay of 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. A number of parameterisations of the Matérn correlation function exist, for examples see Schabenberger and Gotway (2004, Sect. 4.7.2). The form given in (2) is taken from Rasmussen and Williams (2006, Sect. 4.2.1) and its special cases for different values of ν are discussed in Sect. 3.2.

2.3 Exact rates of convergence

For a Gibbs sampler with Gaussian target distribution with a known precision matrix, we can compute the exact rate of convergence (Roberts and Sahu 1997). The convergence rate λ is bounded in the interval $[0, 1]$, with $\lambda = 0$ indicating immediate convergence and $\lambda = 1$ indicating sub-geometric convergence (Meyn and Tweedie 1993).

Suppose we block update all random effects $\boldsymbol{\beta}$, or $\tilde{\boldsymbol{\beta}}$ in the case of the CP, and block update all global effects $\boldsymbol{\theta}$. Using the results given in Appendix 1 we can show that the respective rates of convergence for the CP and the NCP of model (1) are given by the maximum modulus eigenvalue of

$$\boldsymbol{F}^c = \left(\boldsymbol{X}_2' \boldsymbol{C}_2^{-1} \boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1} \boldsymbol{X}_2' \boldsymbol{C}_2^{-1} \times \left(\boldsymbol{X}_1' \boldsymbol{C}_1^{-1} \boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1} \boldsymbol{C}_2^{-1} \boldsymbol{X}_2, \quad (3)$$

and the maximum modulus eigenvalue of

$$\boldsymbol{F}^{nc} = \left(\boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2 + \boldsymbol{C}_3^{-1}\right)^{-1} \times \boldsymbol{X}_2'\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\left(\boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1 + \boldsymbol{C}_2^{-1}\right)^{-1} \times \boldsymbol{X}_1'\boldsymbol{C}_1^{-1}\boldsymbol{X}_1\boldsymbol{X}_2. \quad (4)$$

In Sect. 3, we use this result to investigate how the entries of $\boldsymbol{X}_1$, $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ determine the rate of convergence for the CP and the NCP.

2.4 A simple example

To illustrate how parameterisation effects the posterior correlation of the model parameters and the rate of convergence of the Gibbs sampler, consider the following simple model, taken from Gelfand et al. (1996, Sect. 2). Let

$$Y_i = \theta + \beta_i + \epsilon_i, \quad (5)$$

with $\beta_i \sim N(0, \sigma_\beta^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ independently distributed for all $i = 1, \ldots, n$. The form of the model given by (5) is the NCP. The CP is found by replacing β_i with $\tilde{\beta}_i = \beta_i + \theta$, and so $Y_i = \tilde{\beta}_i + \epsilon_i$, and $\tilde{\beta}_i \sim N(\theta, \sigma_\beta^2)$.

Assuming a locally uniform prior distribution for θ, and that σ_β^2 and σ_ϵ^2 are known, Papaspiliopoulos et al. (2003) show that the exact convergence rates of the CP and the NCP of model (5) are

$$\lambda_c = \frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_\beta^2},$$

and $\lambda_{nc} = 1 - \lambda_c$.

The rates of convergence highlight two important features of the sampling efficiency 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.

3 CP versus NCP

In this section, we investigate how the rates of convergence are affected by the variance parameters and the correlation structure of the spatial processes. To focus on these relationships, we let $p = 1$ in model (1), giving us the following hierarchically centred model:

$$\begin{aligned} &\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}_0 \sim N(\tilde{\boldsymbol{\beta}}_0, \sigma_\epsilon^2\boldsymbol{I}) \\ &\tilde{\boldsymbol{\beta}}_0|\theta_0 \sim N(\theta_0\mathbf{1}, \sigma_0^2\boldsymbol{R}_0) \\ &\theta_0 \sim N(m_0, v_0). \end{aligned} \quad (6)$$

It follows from Eqs. (3) and (4) that the respective rates of convergence for the CP and the NCP of model (6) are

$$\lambda_c = \left(1/\sigma_0^2\mathbf{1}^{\mathrm{T}}\boldsymbol{R}_0^{-1}\mathbf{1} + 1/(\sigma_0^2 v_0)\right)^{-1} 1/\sigma_0^2\mathbf{1}^{\mathrm{T}}\boldsymbol{R}_0^{-1} \times \left(1/\sigma_\epsilon^2\boldsymbol{I} + 1/\sigma_0^2\boldsymbol{R}_0^{-1}\right)^{-1} 1/\sigma_0^2\boldsymbol{R}_0^{-1}\mathbf{1}, \quad (7)$$

and

$$\lambda_{nc} = \left(n/\sigma_\epsilon^2 + 1/(\sigma_0^2 v_0)\right)^{-1} 1/\sigma_\epsilon^2\mathbf{1}^{\mathrm{T}} \times \left(1/\sigma_\epsilon^2\boldsymbol{I} + 1/\sigma_0^2\boldsymbol{R}_0^{-1}\right)^{-1} 1/\sigma_\epsilon^2\mathbf{1}. \quad (8)$$

For independent random effects, the ratio of variance parameters is important in determining the rates of convergence and so we introduce the quantity: $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$. In Sects. 3.1–3.5, we use expressions (7) and (8) to compare the convergence rates for the CP and the NCP for different values of $\delta_0 = \sigma_0^2/\sigma_\epsilon^2$ and forms of $\boldsymbol{R}_0$. In Sect. 3.6, we alter the model to include a covariate.

In Sects. 3.2–3.5, we confine ourselves to the case $1/v_0 = 0$, such that we have an improper prior distribution for θ_0. This serves to clarify the effects of the other parameters on the convergence rates. To see the effect that the prior precision of θ_0 has on the convergence rate consider two different values for v_0; $v_{0,1}$ and $v_{0,2}$, with corresponding rates of convergence $\lambda_{c,1}$, $\lambda_{nc,1}$, $\lambda_{c,2}$ and $\lambda_{nc,2}$. Comparing the ratio of the convergence rates for the two different priors, we have

$$\frac{\lambda_{c,1}}{\lambda_{c,2}} = \frac{\mathbf{1}^{\mathrm{T}}\boldsymbol{R}_0^{-1}\mathbf{1} + 1/v_{0,2}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{R}_0^{-1}\mathbf{1} + 1/v_{0,1}},$$

and clearly if $v_{0,1} < v_{0,2}$ then $\lambda_{c,1} < \lambda_{c,2}$. The same result can be seen for the NCP where

$$\frac{\lambda_{nc,1}}{\lambda_{nc,2}} = \frac{\sigma_0^2 n + \sigma_\epsilon^2/v_{0,2}}{\sigma_0^2 n + \sigma_\epsilon^2/v_{0,1}}.$$

Therefore, a more precise prior distribution for θ_0 will hasten convergence for both the CP and the NCP.

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

$$(\boldsymbol{R}_0)_{ij} = \begin{cases} \rho \text{ if } i \neq j \\ 1 \text{ if } i = j, \end{cases} \quad (9)$$

for $0 \leq \rho < 1$. We restrict ρ to take only non-negative values, as is usual in spatial data modelling. Roberts and Sahu (1997) consider a similar structure for the dispersion matrix of a Gaussian target distribution but do not include a global mean parameter as we do here.

To assist in the computation of convergence rates λ_c and λ_{nc} we make use of the following two matrix inversion identities. The first is the Sherman-Morrison-Woodbury formula, see for example Harville (1997, p. 423). 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

$$(\boldsymbol{N} + \boldsymbol{U}\boldsymbol{M}\boldsymbol{V})^{-1} = \boldsymbol{N}^{-1} - \boldsymbol{N}^{-1}\boldsymbol{U}(\boldsymbol{M}^{-1} + \boldsymbol{V}\boldsymbol{N}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}\boldsymbol{N}^{-1}. \quad (10)$$

The second result takes $\boldsymbol{I}$ to be the $n \times n$ identity matrix and $\boldsymbol{J}$ the $n \times n$ matrix of ones, then

$$(a\boldsymbol{I} + b\boldsymbol{J})^{-1} = \frac{1}{a}\boldsymbol{I} - \frac{b}{a(a+nb)}\boldsymbol{J}, \quad (11)$$

for constants $a > 0, b \neq -(a/n)$. This can be easily checked by direct multiplication and noting that

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

Note also that identity (11) follows from (10) 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 (7) and (8) 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 Eq. (10) we see that

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

For $\boldsymbol{R}_0$ defined by (9), we write

$$\sigma_0^2\boldsymbol{R}_0 = \sigma_0^2(1-\rho)\boldsymbol{I} + \sigma_0^2\rho\boldsymbol{J}, \quad (12)$$

and

$$\sigma_\epsilon^2\boldsymbol{I} + \sigma_0^2\boldsymbol{R}_0 = \left(\sigma_\epsilon^2 + \sigma_0^2(1-\rho)\right)\boldsymbol{I} + \sigma_0^2\rho\boldsymbol{J}, \quad (13)$$

and we can be invert the matrices given in (12) and (13) using Eq. (11).

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

$$\lambda_c = \frac{nv_0}{\sigma_0^2(1-\rho) + n\sigma_0^2\rho + nv_0}\left(\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho}\right), \quad (14)$$

and

$$\lambda_{nc} = \frac{n\sigma_0^2 v_0}{\sigma_\epsilon^2 + n\sigma_0^2 v_0}\left(\frac{\sigma_0^2(1-\rho) + n\sigma_0^2\rho}{\sigma_\epsilon^2 + \sigma_0^2(1-\rho) + n\sigma_0^2\rho}\right). \quad (15)$$

If we assume an improper prior distribution, achieved by letting $1/v_0 = 0$, then $\lambda_{nc} = 1 - \lambda_c$, but otherwise equality does not hold. Note also that when $\rho = 0$, we recover the rates for the independent random effects model, see Sect. 2.4.

It is useful to re-write Eqs. (14) and (15) as

$$\lambda_c^{-1} = \left\{1 + v_0^{-1}\left[\sigma_0^2(1-\rho)/n + \sigma_0^2\rho\right]\right\}\{1 + \delta_0[1 + (n-1)\rho]\}, \quad (16)$$

and

$$\lambda_{nc}^{-1} = [1 + (n\delta_0 v_0)^{-1}](1 + \{\delta_0[1 + (n-1)\rho]\}^{-1}), \quad (17)$$

respectively. Consider first the case $1/v_0 = 0$ and recall that a lower rate indicates faster convergence. For the CP, increasing either δ_0, ρ or n speeds up convergence. Increasing any one of these quantities has the opposing effect on the NCP. When $1/v_0 \neq 0$, λ_{nc} behaves as in the improper case. This is true of λ_c with respect to δ_0 and ρ, but it is no longer monotonic in n.

3.2 Effect of spatial correlation

In spatial modelling, the correlation between two realisations of a latent process is usually assumed to be a function of their separation. Here, we consider exponential correlation functions, which are used widely in applications (Sahu et al. 2010; Berrocal et al. 2010; Sahu et al. 2007; Huerta et al. 2004). We have that

$$(\boldsymbol{R}_0)_{ij} = \exp(-\phi_0 d_{ij}),$$

where ϕ_0 is the spatial decay parameter. The exponential correlation function belongs to the Matérn family and is found by letting $\nu = 0.5$ in Eq. (2). To see this we can use the following results taken from Schabenberger and Gotway (2004, Sect. 4.3.2)

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

We characterise the strength of correlation in terms of the effective range, which we define as the distance, d_0, such that $corr\{\beta_0(\boldsymbol{s}_i), \beta_0(\boldsymbol{s}_j)\} = 0.05$. For an exponential correlation function, we have that

$$d_0 = -\log(0.05)/\phi_0 \approx 3/\phi_0.$$

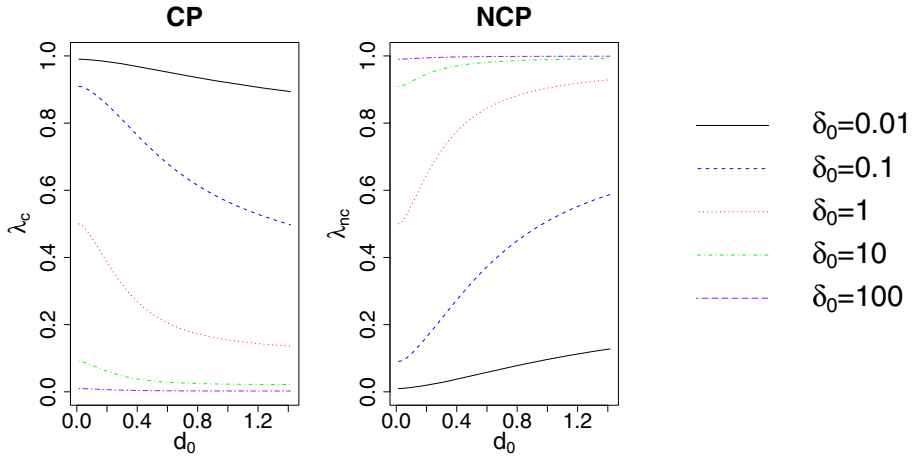

Fig. 1 Convergence rate against effective range for the CP and the NCP at different levels of δ_0

We cannot compute explicit expressions for the entries of $\boldsymbol{R}_0^{-1}$, and hence, we cannot find expressions for the convergence rate in terms of ϕ_0. Therefore, we use a simulation approach. We take the unit square to be the spatial domain, randomly selecting $n = 40$ locations which will be used throughout the rest of this section.

We consider five values of the variance ratio δ_0 and vary the strength of spatial correlation by controlling the effective range $d_0 = -\log(0.05)/\phi_0$. For each value of δ_0, we compute the convergence rates given in Eqs. (7) and (8) for effective ranges between zero, implying no spatial correlation, and $\sqrt{2}$, the maximum possible separation of two points in the domain.

Convergence rates are plotted against the effective range for the CP and the NCP in Fig. 1, where again a lower rate indicates faster convergence. For a fixed d_0, we can see that increasing the δ_0 decreases the convergence rate for the CP, but increases it for the NCP, as we might expect given the results for independent random effects. 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. Hence, the complimentary nature of the CP and NCP is maintained when we vary the strength of exponential correlation across the random effects. However, the two rates do not add to 1 unlike in the simpler case of just below-mentioned Eq. (15).

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

3.3 Effect of the smoothness parameter in the Matérn correlation function

In this section, we consider different correlation functions from the Matérn family, see Eq. (2) for the general form. ν is a half integer, such that $\nu = b + 0.5$ for $b = 0, 1, 2, \ldots$, the correlation function takes on a simpler form. Taken from Rasmussen and Williams (2006, Sect. 4.2), we have that

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

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

$$\rho(d_{ij}, \phi) = \left(1 + \sqrt{3}\phi d_{ij}\right) \exp\left(-\sqrt{3}\phi d_{ij}\right), \tag{18}$$

and $\nu = 2.5$, it becomes

$$\rho(d_{ij}, \phi) = \left(1 + \sqrt{5}\phi d_{ij} + \frac{5\phi^2 d_{ij}^2}{3}\right) \exp(-\sqrt{5}\phi d_{ij}). \tag{19}$$

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

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

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

We consider model (6) and compare the convergence rates for the CP and the NCP for the exponential, $\nu = 1.5$, $\nu = 2.5$ and Gaussian correlation functions. In Sect. 3.2, the strength of correlation is considered in terms of the effective range, which for the exponential correlation function is $-\log(0.05)/\phi_0$. In terms of ϕ_0, the effective range for the Gaussian correlation function is given by $\sqrt{-2\log(0.05)}/\phi_0$. For other members of the Matérn class, there is no closed form expression for the effective range. Therefore, for the cases ν is equal to 1.5 and 2.5, we take an effective range d_0

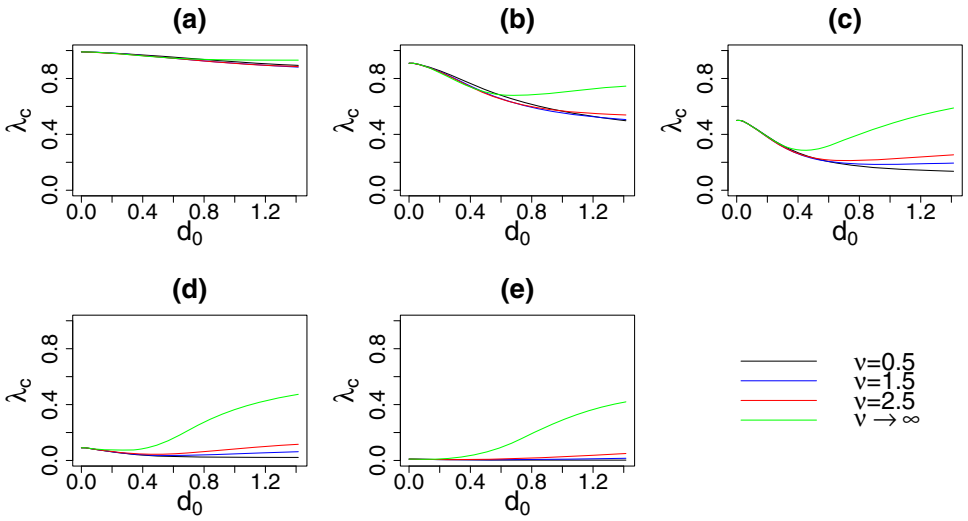

Fig. 2 Convergence rates for the CP of model (6) for different values of ν. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

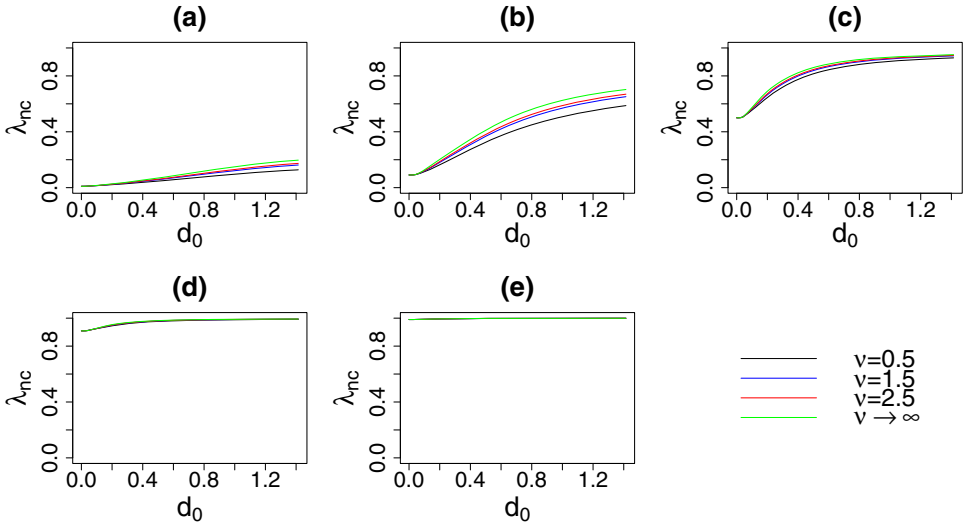

Fig. 3 Convergence rates for the NCP of model (6) for different values of ν. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

and search for the value of ϕ_0 that solves

$$\rho(d_0, \phi_0) - 0.05 = 0, \tag{20}$$

where $\rho(d_0, \phi_0)$ is given by functions (18) and (19) respectively.

Convergence rates are computed for each parameterisation for effective ranges between 0 and $\sqrt{2}$ and for five values of $\delta_0 = 0.01, 0.1, 1, 10, 100$. Recall that as previously in Sect. 3.2 we take the unit square to be the spatial domain and randomly select 40 locations. The results for the CP are given in Fig. 2. We see that for fixed ν and ϕ_0, increasing the δ_0 reduces the convergence rate. Also we see that for fixed ϕ_0 and δ_0, the convergence rate is increased 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 of $\nu = 0.5$, increasing the effective range does not reduce the convergence rate for other values of ν.

The equivalent plot for the NCP is given in Fig. 3. For fixed ν and ϕ_0, the increasing δ_0 increases the convergence rate. For fixed ϕ_0 and δ_0, the increasing ν slows convergence as it does for the CP. The convergence rate is monotonically increasing with the 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.

Figures 2 and 3 show that increasing ν, the smoothness parameter, leads to slower convergence for both the CP and NCP, which contradicts the complimentary behaviour of the rates of convergence seen in the previous two subsections. In order to understand this further, we investigate as follows. The rates of convergence, as obtained in (7) and (8) depend on R_0^{-1} which is guaranteed to be positive definite for any n when a member of the Matérn class of correlation functions is adopted. However, in practical numerical calculations with a large value of n (much greater than 40 used in Figs. 2 and 3) R_0 becomes increasingly singular in the presence of high level of spatial correlation and smoothness. The minimum value of n for which this near singularity is observed depends on the minimum distance between the ${}^nC_2$ pairs of the n

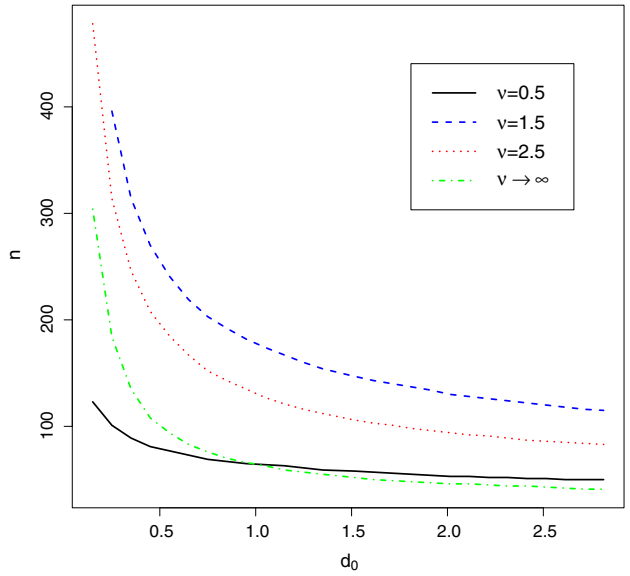

Fig. 4 The minimum value of n for which R_0 is nearly singular for different values of d_0 and ν

locations and the values of the parameters ϕ and ν in the Matérn correlation function. To investigate this minimum value of n, we randomly sample 500 points in the unit square where the minimum Euclidean distance between any two points is greater than a threshold value, which is taken to be 0.01. This ensures that the singularity is not caused by very closely located points in the sample.

Next we find the minimum value of n for which R_0 is approximately singular, viz., the value of the determinant less than 10^{-7}, for a given value of ν and a given value of d_0. The value of ϕ_0 is chosen by (20) $\nu = 1.5$ and 2.5 and $\phi_0 = -\log(0.05)/d_0$ when $\nu = 0.5$, i.e. the case of exponential correlation function and $\phi_0 = \sqrt{-2\log(0.05)}/d_0$ in the case of the Gaussian correlation function, when $\nu \to \infty$. The minimum value is plotted in Fig. 4 against d_0 for all four correlation functions considered here. As expected, for smaller values of d_0, tending to the case of zero spatial correlation, the minimum value of n becomes very large. Increasing spatial correlation, as effected by increasing d_0, leads to a smaller value of n at which near singularity of R_0 is reached. Interestingly, increasing smoothness, through values of ν hasten this except for the case of the exponential correlation function. This is due to the nonlinear effect of the ν and ϕ on the Matérn correlation function.

The near singularity in the limiting Gaussian case is related to the near predictability of the associated spatial processes. In the limiting Gaussian case, it is possible to predict $Y(\boldsymbol{s}')$ for any $\boldsymbol{s}'$ in the same spatial domain upon observing the same spatial process $Y(\boldsymbol{s})$ at any location $\boldsymbol{s}$, see, e.g. page 62 of Banerjee et al. (2015). Such deterministic behaviour leads to the near singularity of covariance matrices which in turn leads to non-convergence. Indeed, Stein (1999) (page 70) explicitly recommends not to use the Gaussian correlation function to model physical processes. The investigation here confirms this view by pointing to non-convergence of the MCMC fitting algorithms for smooth and highly correlated processes for large values of n. This also allows us to conclude that increasing n in an infill asymptotic sense (Zhang 2004), in the presence of high spatial correlation, will lead to near singularity of the covariance matrices, which in turn will lead to non-convergence of either of the two parameterisations. This asymptotic result, however, does not spell disaster for the CP and NCP in practical problems where the strength of the spatial correlation is such that the resulting effective range (d_0 in the exponential case) is significantly less than the maximum distance ($\sqrt{2}$ in the unit square) between any two points in a closed spatial domain. Such a situation will allow modelling of a reasonably large number of spatial observations (e.g. low 100s) using the centring parameterisations.

3.4 Effect of introducing 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

$$d^*_{ij} = \|\boldsymbol{G}(\boldsymbol{s}_i - \boldsymbol{s}_j)\|, \tag{21}$$

where $\boldsymbol{G}$ is a 2×2 transformation matrix. In Euclidean space, (21) is equivalent to

$$d^*_{ij} = \left\{(\boldsymbol{s}_i - \boldsymbol{s}_j)^{\mathrm{T}} \boldsymbol{H}(\boldsymbol{s}_i - \boldsymbol{s}_j)\right\}^{1/2},$$

where $\boldsymbol{H} = \boldsymbol{G}^{\mathrm{T}}\boldsymbol{G}$. 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, Corollary 14.2.14). By replacing d_{ij} with d^*_{ij} in (2), we have a geometric anisotropic Matérn correlation function with elliptical contours of iso-correlation.

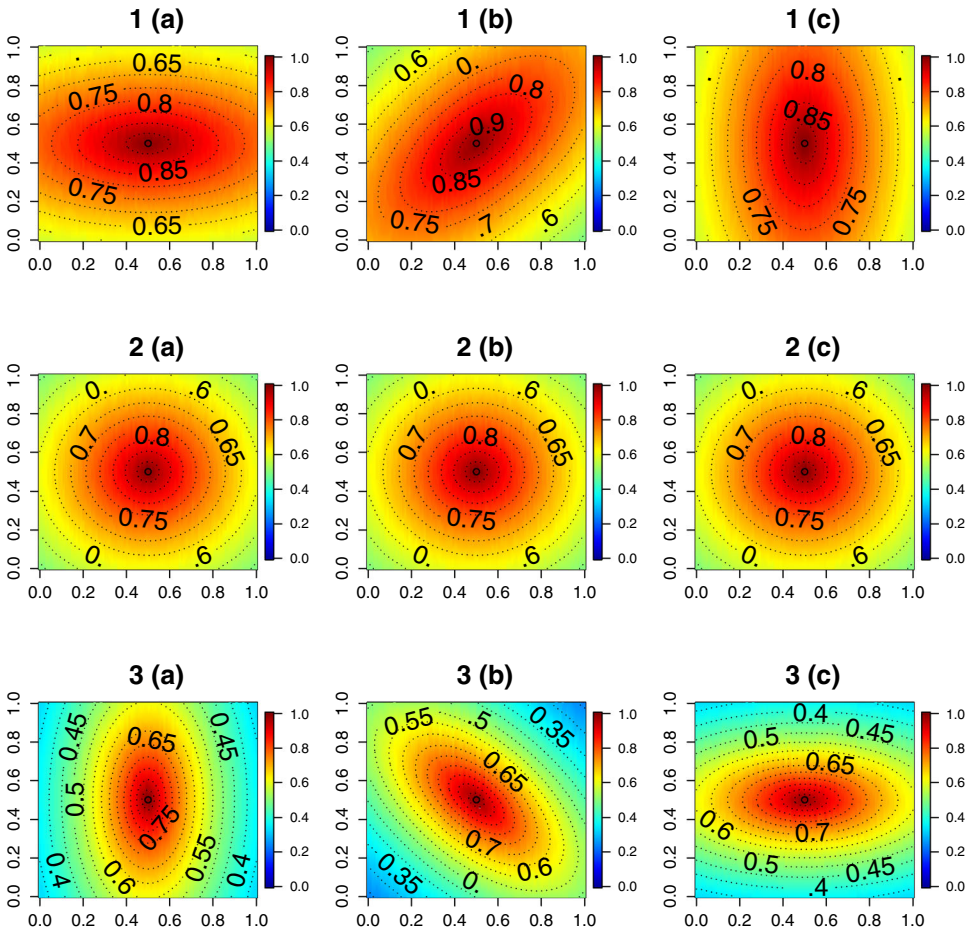

Fig. 5 Correlation surface for $\beta(\boldsymbol{s}^*)$, $\boldsymbol{s}^* = (0.5, 0.5)^{\mathrm{T}}$, for exponential anisotropic correlation functions with transformation matrix $\boldsymbol{G}$ given in (22). Panels are given an *alpha-numeric* label. *Numbers* refer to three values of $\alpha = 0.5, 1, 2$. *Letters* (a), (b), and (c) refer to three values of $\psi = 0, \pi/4, \pi/2$

As an example, we follow Schabenberger and Gotway (2004, Chap. 4) and let

$$\boldsymbol{G} = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\psi & \sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix} = \begin{pmatrix} \alpha\cos\psi & \alpha\sin\psi \\ -\sin\psi & \cos\psi \end{pmatrix}, \quad (22)$$

and hence, the axes 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 $\alpha \neq 0$, and hence, $\boldsymbol{H}$ is positive definite as required. For $\boldsymbol{G}$ given in (22), we have

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

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

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

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

$$\rho(d_{ij}^*, \phi) = \exp(-\phi d_{ij}^*). \quad (23)$$

We take the point $\boldsymbol{s}^* = (0.5, 0.5)^{\mathrm{T}}$ 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 (23). 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 Fig. 5.

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. Clearly, setting $\alpha = 2$ has the effect of weakening correlation in the x-direction.

To assess the impact of anisotropy on the convergence rates for the CP and the NCP we return to model (6). We consider an anisotropic exponential correlation function for the spatial process and so

$$(\boldsymbol{R}_0)_{ij} = \exp\left(-\phi_0 d_{ij}^*\right),$$

where d_{ij}^* is given by Eq. (21). We begin by fixing $\psi = 0$ and letting $\alpha = 0.5, 1, 2$. This corresponds to pan-

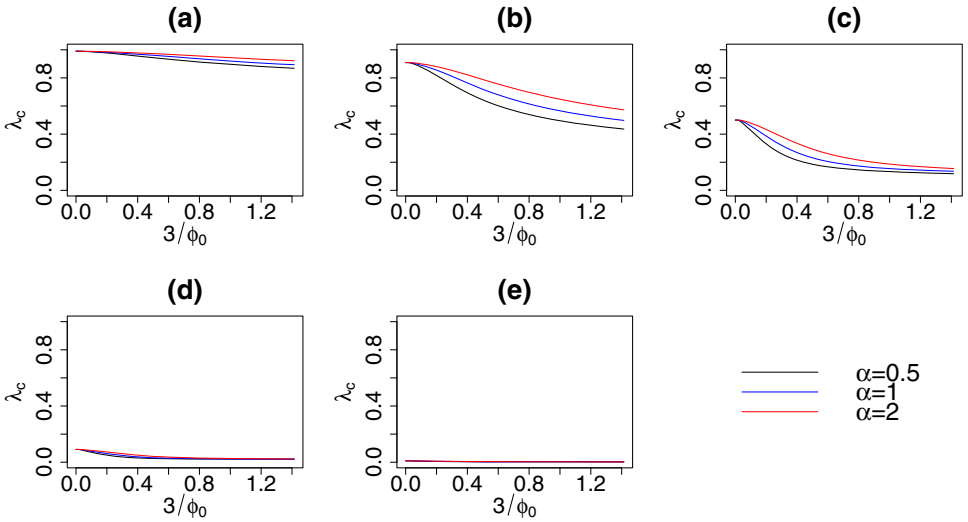

Fig. 6 Convergence rates for the CP of model (6) with an anisotropic exponential correlation function for different values of α. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

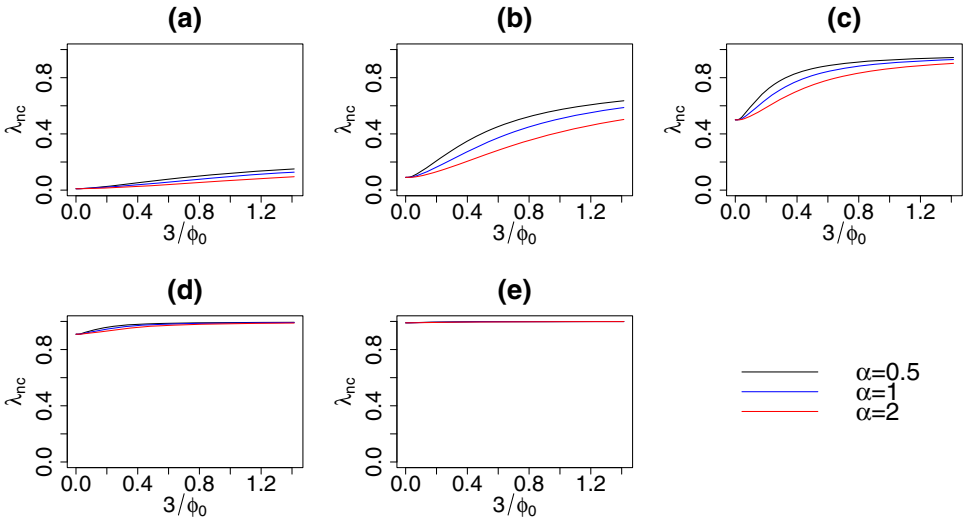

Fig. 7 Convergence rates for the NCP of model (6) with an anisotropic exponential correlation function for different values of α. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

els 1(a), 2(a), and 3(a), in Fig. 5. We use five values for $\delta_0 = \sigma_0^2/\sigma_\epsilon^2 = 0.01, 0.1, 1, 10, 100$ and vary ϕ_0 such that $3/\phi_0 \in (0, \sqrt{2}]$. Here, the effective range is direction dependent so we no longer refer to $3/\phi_0$ as the effective range.

We compute convergence rates for the CP and the NCP and plot results in Figs. 6 and 7 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 Sect. 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 $\boldsymbol{H}$ 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 Fig. 5. Again, we let $\delta_0 = 0.01, 0.1, 1, 10, 100$ and vary ϕ_0 such that $3/\phi_0 \in (0, \sqrt{2}]$.

The results for the CP and the NCP are given in Figs. 8 and 9 respectively. We can see that changing ψ has very little effect on the convergence rates of either parameterisation as expected since d_{ij}^* is free of ψ. However, because we apply the rotation matrix first the subsequent stretch effectively acts in a different direction, that direction depending on ψ, and the resulting values for d_{ij}^* may be different. Take the example we used here. Let $\boldsymbol{s} = (s_1, s_2)^{\mathrm{T}}$ and $\boldsymbol{s}_i - \boldsymbol{s}_j = (s_{i1} - s_{j1}, s_{i2} - s_{j2})^{\mathrm{T}} = (l_1, l_2)^{\mathrm{T}}$. For $\psi = \pi/4$, $d_{ij}^* = \sqrt{0.5[(l_1 - l_2)^2 + \alpha^2(l_1 + l_2)^2]}$, whereas if $\psi = \pi/2$ then $d_{ij}^* = \sqrt{l_1^2 + \alpha^2 l_2^2}$. The different values of d_{ij}^* may lead to different rates of convergence. Further investigation is needed to determine whether similar results hold for patterned sampling locations.

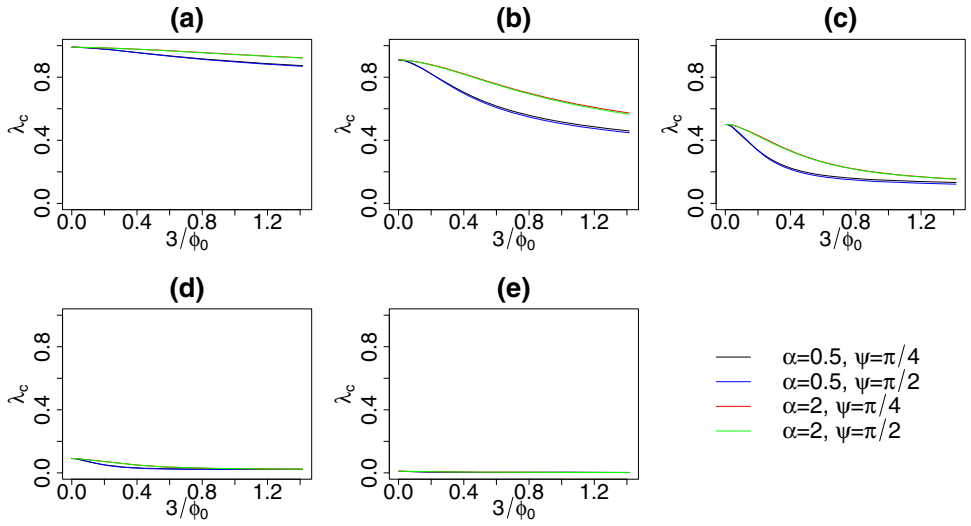

Fig. 8 Convergence rates for the CP of model (6) with an anisotropic exponential correlation function for different values of α and ψ. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

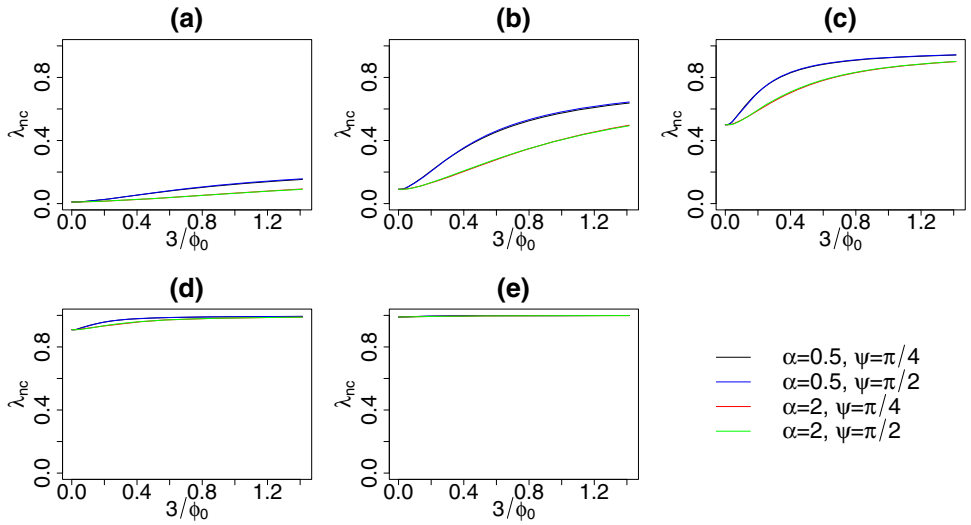

Fig. 9 Convergence rates for the NCP of model (6) with an anisotropic exponential correlation function for different values of α and ψ. **a** $\delta_0 = 0.01$, **b** $\delta_0 = 0.1$, **c** $\delta_0 = 1$, **d** $\delta_0 = 10$, **e** $\delta_0 = 100$

3.5 Effect of introducing 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. One strategy to deal with this is 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 (6) with an exponential correlation function for $\boldsymbol{R}_0$ and compare the convergence rates found in Sect. 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 $\chi = -\phi_0/\log(0.05)$. Therefore,

$$(\boldsymbol{R}_{Tap})_{ij} = \begin{cases} \exp(-\phi_0 d_{ij})\left(1 - \dfrac{3d_{ij}\chi}{2} + \dfrac{d_{ij}^3\chi^3}{2}\right) & \text{if } d_{ij} < d_0,\ \phi_0 > 0,\ \chi > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $d_0 = -\log(0.05)/\phi_0$ is the effective range.

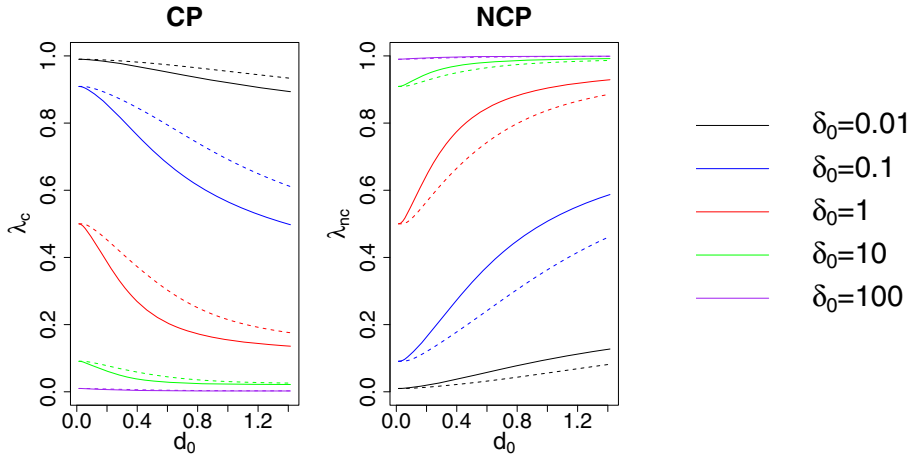

Fig. 10 Convergence rates with tapered covariance matrices for the CP and the NCP at different levels of δ_0

We let $\delta_0 = 0.01, 0.1, 1, 10$ and 100 and vary d_0 between 0 and $\sqrt{2}$. The convergence rates for the CP and the NCP are given in Fig. 10. 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 Fig. 1. 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

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

Therefore, $\rho_{Tap} \leq \rho$, with equality attained only when $\rho = 0$. We know from Eqs. (16) and (17) that for equi-correlated random effects if ρ decreases, λ_c is increased and λ_{nc} is decreased. In other words, when $n = 2$, tapering can only worsen the performance of CP and improve the performance of the NCP.

3.6 Effect of covariates

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

$$Y(\boldsymbol{s}_i) = \{\theta_1 + \beta_1(\boldsymbol{s}_i)\}x_1(\boldsymbol{s}_i) + \epsilon(\boldsymbol{s}_i) \quad (i = 1, \ldots, n), \quad (24)$$

which may be found by letting $k = 1, \ldots, p-1$, and $p = 2$ in model (1). Recalling that $\tilde{\boldsymbol{\beta}}_1 = \{\tilde{\beta}_1(\boldsymbol{s}_1), \ldots, \tilde{\beta}_1(\boldsymbol{s}_n)\}^{\mathrm{T}}$, where $\tilde{\beta}_1(\boldsymbol{s}_i) = \beta_1(\boldsymbol{s}_i) + \theta_1$, and $\boldsymbol{x}_1 = \{x_1(\boldsymbol{s}_1), \ldots, x_1(\boldsymbol{s}_n)\}^{\mathrm{T}}$ and $\boldsymbol{D}_1 = \mathrm{diag}(\boldsymbol{x}_1)$, we can write model (24) in the following form

$$\begin{aligned} \boldsymbol{Y} \mid \tilde{\boldsymbol{\beta}}_1 &\sim N\left(\boldsymbol{D}_1\tilde{\boldsymbol{\beta}}_1, \sigma_\epsilon^2\boldsymbol{I}\right) \\ \tilde{\boldsymbol{\beta}}_1 \mid \theta_1 &\sim N\left(\theta_1\boldsymbol{1}, \sigma_1^2\boldsymbol{R}_1\right) \\ \theta_1 &\sim N\left(m_1, \sigma_1^2 v_1\right). \end{aligned} \quad (25)$$

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

First suppose that random effects are independent. This can be considered the limiting case for weakening spatial correlation. For the sake of notational clarity, under the assumption of spatial independence we write $x(\boldsymbol{s}_i) = x_i$ $(i = 1, \ldots, n)$. The convergence rate for the CP is

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

If we let $1/v_1 = 0$, thus implying and improper prior for θ_1, we can write λ_c as

$$\lambda_c = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{1 + \left(\sigma_1^2/\sigma_\epsilon^2\right)x_i^2}. \quad (26)$$

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 (24) 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

$$u_i = \frac{x_i - \bar{\boldsymbol{x}}}{sd_x}, \quad i = 1, \ldots, n, \quad (27)$$

and $\bar{\boldsymbol{x}}$ and sd_x are the sample mean and sample standard deviation of $\boldsymbol{x}$ respectively. Substituting Eq. (27) into Eq. (26) we have

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

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 the NCP, the convergence rate is

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

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

$$\lambda_{nc} = \frac{1}{\sum_{i=1}^{n} x_i^2} \sum_{i=1}^{n} \frac{x_i^4}{\left(\sigma_\epsilon^2/\sigma_1^2\right) + x_i^2}. \tag{28}$$

For fixed $\boldsymbol{x}$, if $\sigma_\epsilon^2/\sigma_1^2$ goes to zero then λ_{nc} goes to one. In contrast, as the data variance dominates that of the random effects, the convergence rate falls. Note that in general $\lambda_c + \lambda_{nc} \neq 1$.

To see the effect of the scale of $\boldsymbol{x}$ upon λ_{nc} we substitute Eq. (27) into equation (28). Then we have

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

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

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

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.

Now 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 $(\boldsymbol{R}_1)_{ij} = \exp(-\phi_1 d_{ij})$ and so the effective range $d_1 = -\log(0.05)/\phi_1$. 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 (25) 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 Fig. 11. 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. Therefore, the variance ratio δ_1 and the decay parameter ϕ_1 have same influence on the convergence rates of the CP and NCP as δ_0 and ϕ_0.

4 Practical examples with unknown covariance parameters

In this section, 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 centring 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 Sect. 3, the CP and the NCP are compared in terms of the exact convergence rates of the associated Gibbs samplers. 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. The full conditional distributions needed to construct the Gibbs samplers are given in Appendix 2.

We employ two diagnostic statistics to compare parameterisations. The first statistic we use is based on the multivariate potential scale reduction factor (MPSRF) (Brooks and Gelman 1998). We define the $\text{MPSRF}_M(1.1)$ to be the number of iterations required for the MPSRF to fall below 1.1. To compute the $\text{MPSRF}_M(1.1)$ 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 MPSRF 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 the model parameters (Robert and Casella 2004). 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. 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.

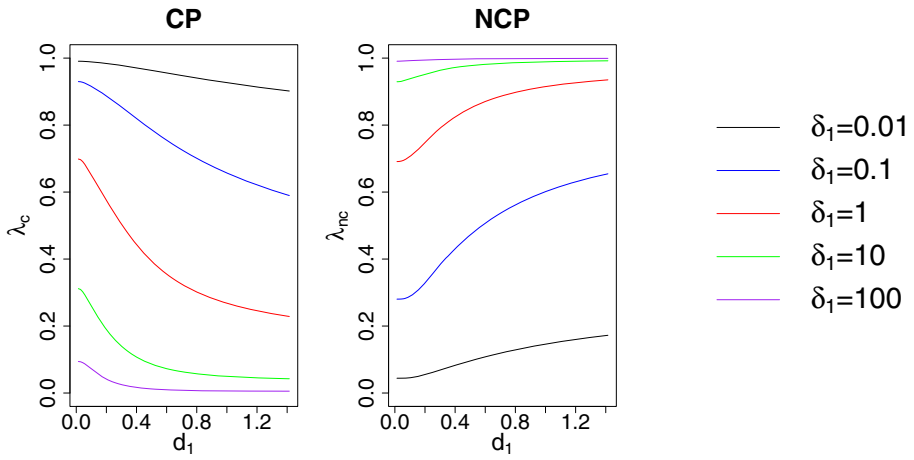

Fig. 11 A comparison of convergence rates for the CP and the NCP at different levels of δ_1

4.1 A simulation study

We simulate data from model (6) for $n = 40$ randomly chosen locations across the unit square assuming an exponential correlation function for the spatial process. 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 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 datasets in total.

We fix the decay parameter at its known value and sample from the marginal posterior distributions of θ_0, σ_0^2 and σ_ϵ^2. We let hyperparameters $m_0 = 0$ and $v_0 = 10^4$. 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. (2007, 2010), Gelfand et al. (2003).

Figure 12 shows boxplots of the $\text{MPSRF}_M(1.1)$ (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. 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 Fig. 13. 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 exact convergence rates when the variance parameters were assumed to be known in Sect. 3.2.

Figure 14 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 15 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 for σ_ϵ^2 under the CP and is similarly explained.

4.2 Real-data example

In this section, we compare the sampling efficiency of the CP and the NCP when they are fitted to a real dataset. We have annual PM10 concentrations, in micrograms per cubic metre (μg/m^3), taken from 70 monitoring sites in Greater London, UK. We use data from 50 sites for model fitting, leaving out data from 20 sites for model validation, see Fig.

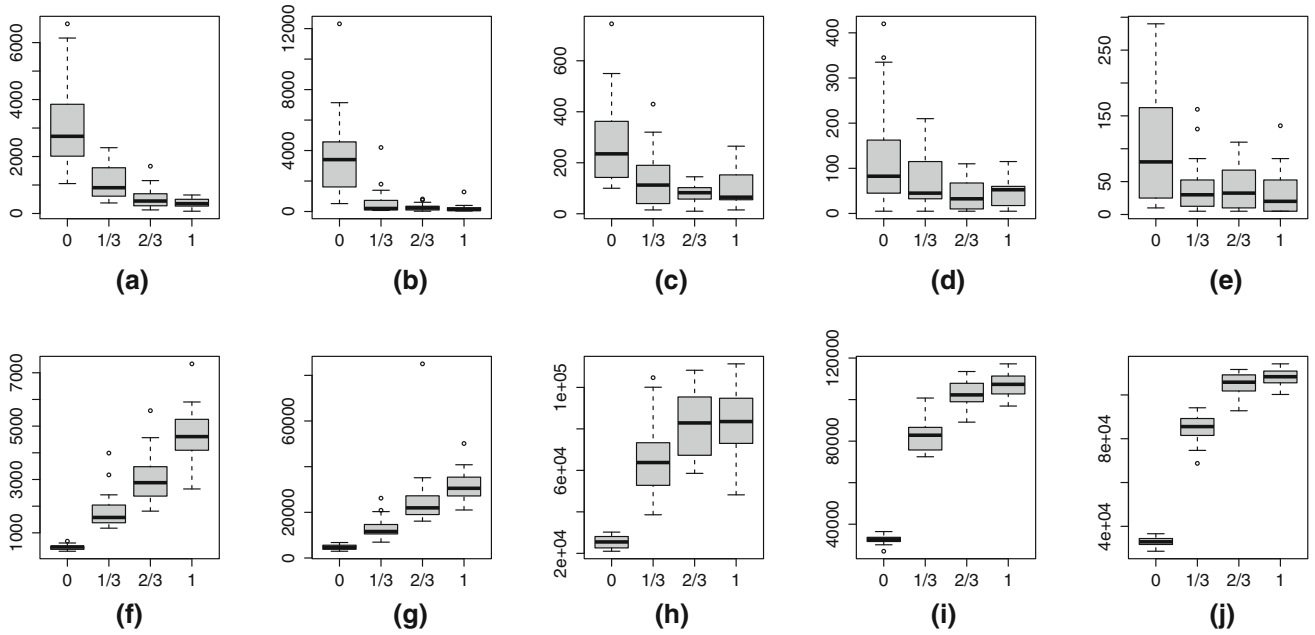

Fig. 12 MPSRF$_M(1.1)$, panels (**a**–**e**), and the ESS of θ_0, plots (**f**–**j**), for the CP

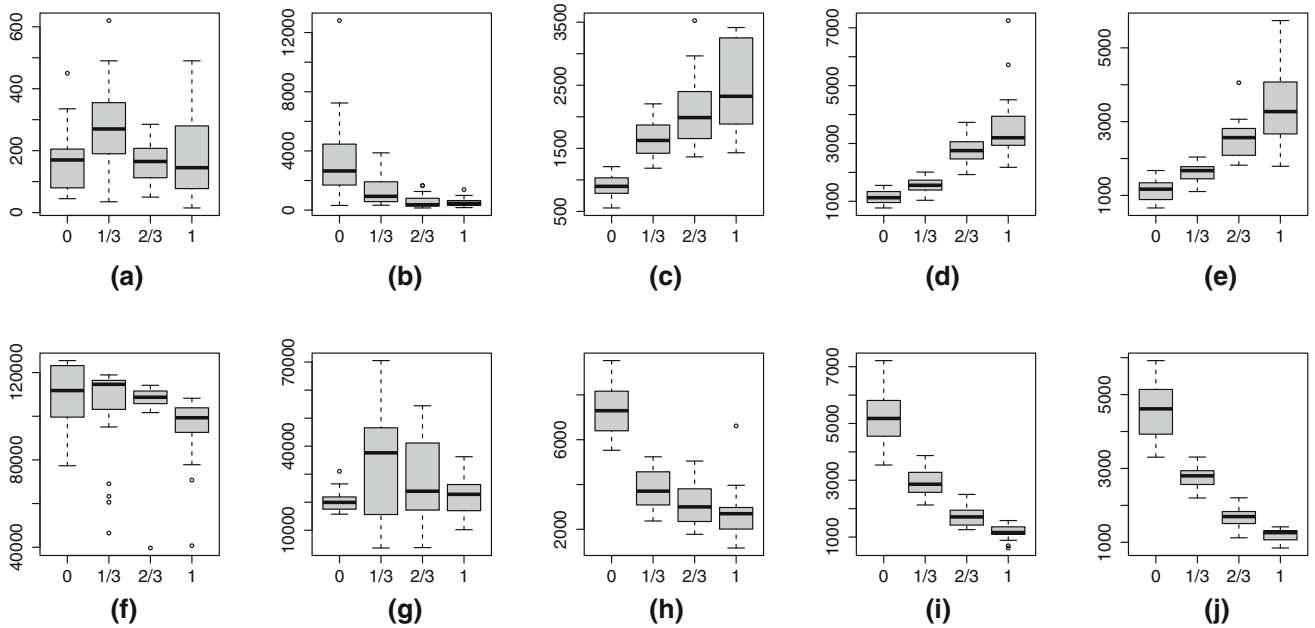

Fig. 13 MPSRF$_M(1.1)$, panels (**a**–**e**), and the ESS of θ_0, plots (**f**–**j**), for the NCP

16. The mean and standard deviation for the 50 data sites is 26.47 and 5.00 μg/m^3 respectively.

In addition to the observed data, we have 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 $\boldsymbol{s}$. Therefore, we use a downscaler model as employed by Berrocal et al. (2010).

To stabilise the variance, we model the data on the square root scale. However, in order not to underestimate their variability, predictions are obtained on the original scale.

We fit model (1) with $p = 2$ and so we have two processes, an intercept and a slope process. Each process has a corresponding global variance and decay parameters and so $\boldsymbol{\theta} = (\theta_0, \theta_1)^{\mathrm{T}}$, $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)^{\mathrm{T}}$ and $\boldsymbol{\phi} = (\phi_0, \phi_1)^{\mathrm{T}}$. In addition, we have the data variance, σ_ϵ^2. For the prior distribution of $\boldsymbol{\theta}$, we let $\boldsymbol{m} = (0, 0)^{\mathrm{T}}$ and $v_0 = v_1 = 10^4$. We

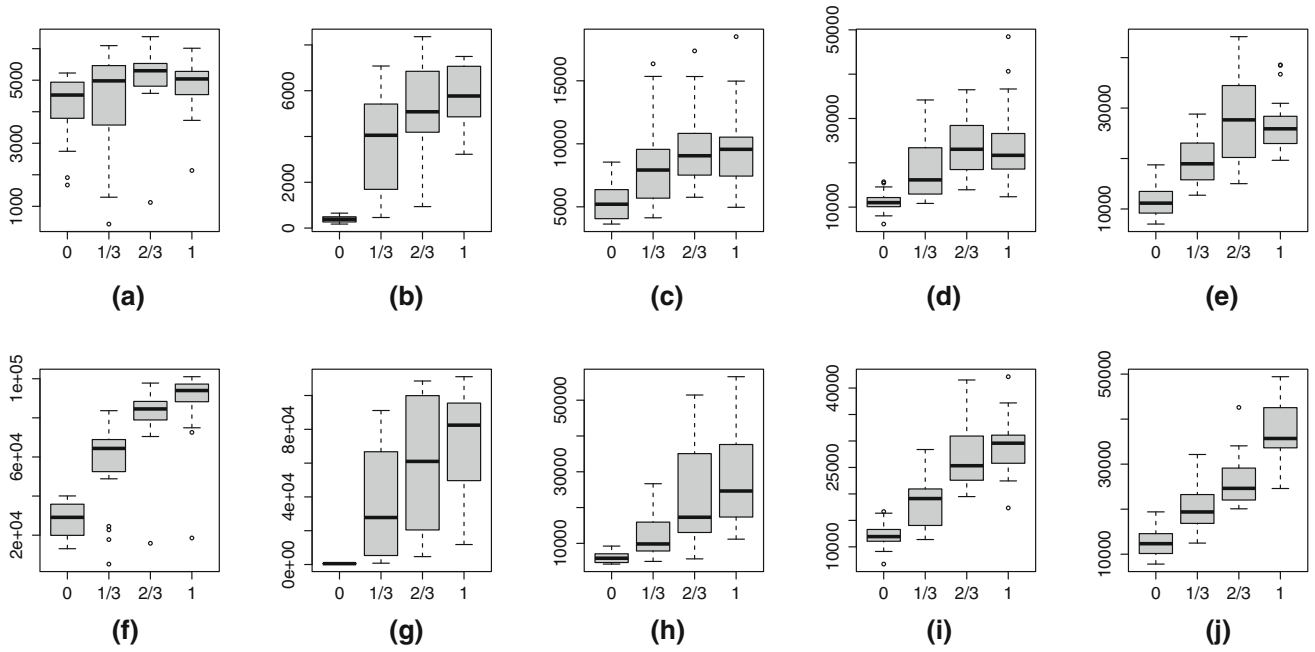

Fig. 14 ESS of σ_0^2, panels (**a**–**e**), and σ_ϵ^2, panels (**f**–**j**) for the CP

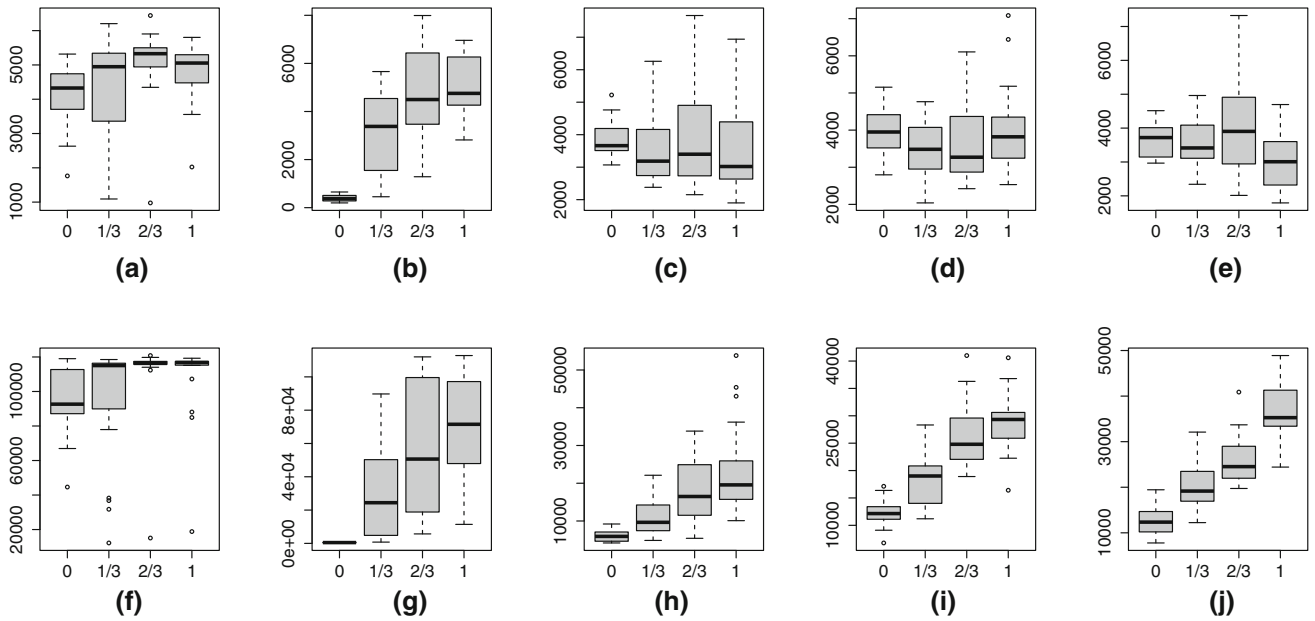

Fig. 15 ESS of σ_0^2, panels (**a**–**e**), and σ_ϵ^2, panels (**f**–**j**) for the NCP

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.

We begin by estimating the decay parameters using an empirical Bayes method by performing a grid search over a small number of values, then choosing those values that minimise some calibration criterion. This is a common approach adopted by many authors, e.g., Sahu et al. (2011), Berrocal et al. (2010), Sahu et al. (2007) since Zhang (2004) showed that it is not possible to consistently estimate these in a model with Matérn covariance function in the presence of other unknown parameters. In our Bayesian inference setting, this will imply weak identifiability in the posterior distribution when non-informative prior distributions are assumed.

The greatest distance between any two of the 70 monitoring stations is 96.2 kilometres (km), and so we select values

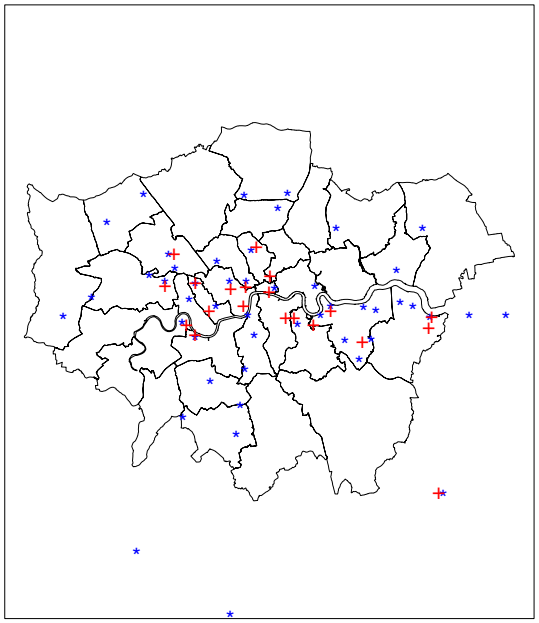

Fig. 16 Sampling locations for PM10 concentration data. *Blue stars* indicate locations used for model fitting and *red crosses* indicate locations used for model validation

of ϕ_0 and ϕ_1 corresponding to effective ranges of 5, 10, 25, 50 and 100 km. Predictions are made at the 20 validations sites and we compute the values of three measures of prediction error; the mean absolute prediction error (MAPE), the root mean square prediction error (RMSPE) and the continuous ranked probability score (CRPS), see for example Gneiting et al. (2007).

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. The values of the MAPE, RMSPE and CRPS are given in Table 1. Recall that d_0 and d_1 denote the effective ranges implied by ϕ_0 and ϕ_1 respectively. By all three measures, the prediction error is minimised when $d_0 = 5$ and $d_1 = 50$, and so our estimates for the spatial decay parameters are

$$\hat{\phi}_0 = -\log(0.05)/5 \approx 0.6, \quad \hat{\phi}_1 = -\log(0.05)/50 \approx 0.06.$$

We now compare the sampling efficiencies of the CP and NCP for the London PM10 data when the values of the decay parameters are fixed at the above optimal values. For each parameterisation, 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 of $\boldsymbol{\theta} = (\theta_0, \theta_1)'$, $\boldsymbol{\sigma^2} = (\sigma_0^2, \sigma_1^2)'$ and σ_ϵ^2 are computed and given in Table 2. We can see that the CP requires far fewer iterations for the MPSRF to drop below 1.1 than the NCP: 275 versus 1985. The ESS of the mean parameters is greater for the CP than the NCP, especially for θ_1 reflecting the stronger spatial correlation for the slope process and the estimate for δ_1. For the variance parameters σ_0^2 and σ_1^2, the ESS for the CP is nearly double that of the NCP. The NCP achieves better mixing in the σ_ϵ^2 coordinate than does the CP.

Table 1 Prediction error for different combinations of d_0 and d_1

d_0	d_1	MAPE	RMSPE	CRPS
5	5	5.494	6.224	3.720
	10	5.476	6.200	3.704
	25	5.416	6.164	3.701
	50	**5.375**	**6.126**	**3.695**
	100	5.418	6.160	3.735
10	5	5.534	6.281	3.740
	10	5.497	6.256	3.728
	25	5.480	6.230	3.733
	50	5.436	6.207	3.736
	100	5.452	6.193	3.753
25	5	5.618	6.534	3.878
	10	5.585	6.492	3.862
	25	5.492	6.351	3.801
	50	5.485	6.330	3.809
	100	5.476	6.314	3.828
50	5	5.620	6.741	4.003
	10	5.615	6.711	4.002
	25	5.549	6.572	3.944
	50	5.505	6.500	3.924
	100	5.499	6.470	3.938
100	5	5.644	6.910	4.129
	10	5.586	6.820	4.093
	25	5.541	6.658	4.035
	50	5.491	6.595	4.015
	100	5.482	6.581	4.046

For making inference, we generate a single long chain for 50,000 iterations and discard the first 10,000. Parameter estimates and their 95% credible intervals are given in Table 3. An estimate for the global intercept θ_0 of 5.148 reflects that the data are modelled on the square root scale. The global regression parameter θ_1 for the AQUM output is not significant given the inclusion in the model of the regression process. Of particular interest to us are the estimates of the variance parameters. Weak spatial correlation in the intercept process means that the estimate for σ_0^2 is the smallest of the three variance parameters. More of the spatial variability is explained by the intercept process, and so the estimate for σ_1^2 is the greatest of the three variance parameters. 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 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
CP	275	22,161	63,137	31,876	16,689	24,171
NCP	1985	15,596	2958	15,407	8275	31,806

Table 3 Parameter estimates and their 95% credible intervals (CI)

Parameter	Estimate	95% CI
θ_0	5.148	(4.942, 5.352)
θ_1	0.093	(−0.396, 0572)
σ_0^2	0.172	(0.093, 0.299)
σ_1^2	0.224	(0.101, 0.469)
σ_ϵ^2	0.177	(0.096, 0.307)
δ_0	1.070	(0.412, 2.268)
δ_1	1.376	(0.487, 3.256)

5 Conclusion

We have compared the efficiencies of the CP and the NCP of spatial models. We find that in addition to the ratio of the variance parameters, the correlation structure between the random effects plays a key role in determining the rate of convergence.

For known variance and correlation parameters, the exact rate of convergence has been examined. We have shown that for spatial models with an exponential correlation function, increasing the variance of the random effects relative to that of the data, 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 favoured. 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 favoured and the NCP hindered. Both 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, and in the presence of moderate spatial correlation, both parameterisations will fail to converge when the sample size is large enough.

When the variance parameters are unknown, the sampling efficiencies of the parameterisations are compared via the $\text{MPSRF}_M(1.1)$ and the ESS of the unknown model parameters. We have seen that the relationships between the sampling efficiencies 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. In contrast to this, the NCP performs best when the data are less informative and the correlation is weak.

Appendix 1: Computing the exact convergences for the CP and NCP

For Gibbs samplers with Gaussian target distributions with known precision matrices, we have analytic results for the exact convergence rate (Roberts and Sahu 1997). We let $\boldsymbol{\xi}$ denote the set of all mean parameters in the model: $\boldsymbol{\xi} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}})^{\mathrm{T}}$. Suppose that $\boldsymbol{\xi} \mid \boldsymbol{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and let $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$ be the posterior precision matrix (PPM). Further suppose that $\boldsymbol{\xi}$ is partitioned into l blocks for updating within the Gibbs sampler. To compute the convergence rate first, partition $\boldsymbol{Q}$ according to the l blocks where the ijth block is denoted by $\boldsymbol{Q}_{ij}, i, j = 1, \ldots l$.

Let $\boldsymbol{A} = \boldsymbol{I} - \text{diag}(\boldsymbol{Q}_{11}^{-1}, \ldots, \boldsymbol{Q}_{ll}^{-1})\boldsymbol{Q}$ and $\boldsymbol{F} = (\boldsymbol{I} - \boldsymbol{L})^{-1}\boldsymbol{U}$, where $\boldsymbol{L}$ is the block lower triangular matrix of $\boldsymbol{A}$, and $\boldsymbol{U} = \boldsymbol{A} - \boldsymbol{L}$. Roberts and Sahu (1997) show that the Markov chain induced by the Gibbs sampler with components block updated according to the above blocking scheme, 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:

Theorem 5.1 (Roberts and Sahu 1997, Theorem 1) A Markov chain with transition density

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

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

Corollary 5.2 *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

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

We use Theorem 5.1 to compare the convergence rates of Gibbs samplers associated with the CP and the NCP. First, we must compute the PPM for each parameterisation. To identify the PPM for the CP write

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

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

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

By corollary 5.2, if we update all random effects as one block and all global effects as another, then the convergence rate for the CP is the maximum modulus eigenvalue of

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

Similarly for the NCP, we find the PPM by writing

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

and hence, we have

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

By Corollary 5.2, the convergence rate of the Gibbs sampler for the NCP is the maximum modulus eigenvalue of

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

Appendix 2: Full conditional distributions

(a) Posterior distributions for the CP

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

$$\begin{aligned}
\pi(\boldsymbol{\xi}|\boldsymbol{y}) &\propto \pi(\boldsymbol{Y}|\tilde{\boldsymbol{\beta}}, \sigma_\epsilon^2)\pi(\tilde{\boldsymbol{\beta}}|\boldsymbol{\theta}, \boldsymbol{\sigma^2}, \boldsymbol{\phi})\pi(\boldsymbol{\theta}|\boldsymbol{\sigma^2})\pi(\boldsymbol{\sigma^2})\pi(\sigma_\epsilon^2)\pi(\boldsymbol{\phi}) \\
&\propto \prod_{k=0}^{p-1}\left(\sigma_k^2\right)^{-(n/2+1/2+a_k+1)}|\boldsymbol{R}_k|^{-1/2}\left(\sigma_\epsilon^2\right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp\left\{ -\frac{1}{2\sigma_\epsilon^2}\left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1}\boldsymbol{D}_k\tilde{\boldsymbol{\beta}}_k \right)^{\mathrm{T}} \left(\boldsymbol{Y} - \sum_{k=0}^{p-1}\boldsymbol{D}_k\tilde{\boldsymbol{\beta}}_k \right) + 2b_\epsilon \right] \right\} \\
&\quad \exp\left\{ -\frac{1}{2}\sum_{k=0}^{p-1}\left(\tilde{\boldsymbol{\beta}}_k - \theta_k\boldsymbol{1}\right)^{\mathrm{T}}\boldsymbol{\Sigma}_k^{-1}\left(\tilde{\boldsymbol{\beta}}_k - \theta_k\boldsymbol{1}\right) \right\} \\
&\quad \exp\left\{ -\frac{1}{2}\sum_{k=0}^{p-1}\frac{1}{\sigma_k^2}\left(\frac{(\theta_k - m_k)^2}{v_k} + 2b_k \right) \right\}\prod_{k=0}^{p-1}\pi(\phi_k), \qquad (29)
\end{aligned}$$

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

We use Gibbs sampling to sample from $\pi(\boldsymbol{\xi}|\boldsymbol{y})$ for the CP, given in (29). 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 that 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

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

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

$$\boldsymbol{\Sigma}_k^* = \left(\frac{1}{\sigma_\epsilon^2} \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k + \boldsymbol{\Sigma}_k^{-1} \right)^{-1},$$
$$\boldsymbol{m}_k^* = \boldsymbol{\Sigma}_k^* \left[\frac{1}{\sigma_\epsilon^2} \boldsymbol{D}_k \left(\boldsymbol{y} - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{D}_j \tilde{\boldsymbol{\beta}}_j \right) + \boldsymbol{\Sigma}_k^{-1} \theta_k \mathbf{1} \right].$$

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

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

where

$$v_k^* = \left(\mathbf{1}^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \mathbf{1} + \frac{1}{\sigma_k^2 v_k} \right)^{-1},$$
$$m_k^* = v_k^* \left(\mathbf{1}^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \tilde{\boldsymbol{\beta}}_k + \frac{m_k}{\sigma_k^2 v_k} \right).$$

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

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

- The full conditional distribution for data variance σ_ϵ^2 for the CP is

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

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 $\boldsymbol{\xi} = (\boldsymbol{\beta}^{\mathrm{T}}, \boldsymbol{\theta}^{\mathrm{T}}, \boldsymbol{\sigma^2}^{\mathrm{T}}, \sigma_\epsilon^2, \boldsymbol{\phi}^{\mathrm{T}})^{\mathrm{T}}$, and

$$\begin{aligned}
\pi(\boldsymbol{\xi} | \boldsymbol{y}) &\propto \pi(\boldsymbol{Y} | \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma_\epsilon^2) \\
&\quad \pi(\boldsymbol{\beta} | \boldsymbol{\sigma^2}, \boldsymbol{\phi}) \pi(\boldsymbol{\theta} | \boldsymbol{\sigma^2}) \pi(\boldsymbol{\sigma^2}) \pi(\sigma_\epsilon^2) \pi(\boldsymbol{\phi}) \\
&\propto \prod_{k=0}^{p-1} (\sigma_k^2)^{-(n/2+1/2+a_k+1)} |\boldsymbol{R}_k|^{-1/2} \left(\sigma_\epsilon^2 \right)^{-(n/2+a_\epsilon+1)} \\
&\quad \exp \left\{ -\frac{1}{2\sigma_\epsilon^2} \left[\left(\boldsymbol{Y} - \sum_{k=0}^{p-1} (\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k) \right)^{\mathrm{T}} \right. \right. \\
&\quad \left. \left. \left(\boldsymbol{Y} - \sum_{k=0}^{p-1} (\boldsymbol{D}_k \boldsymbol{\beta}_k + \boldsymbol{x}_k \theta_k) \right) + 2b_\epsilon \right] \right\} \\
&\quad \exp \left\{ -\frac{1}{2} \sum_{k=0}^{p-1} \boldsymbol{\beta}_k^{\mathrm{T}} \boldsymbol{\Sigma}_k^{-1} \boldsymbol{\beta}_k \right\} \\
&\quad \exp \left\{ -\frac{1}{2} \sum_{k=0}^{p-1} \frac{1}{\sigma_k^2} \left(\frac{(\theta_k - m_k)^2}{v_k} + 2b_k \right) \right\} \\
&\quad \prod_{k=0}^{p-1} \pi(\phi_k),
\end{aligned} \tag{30}$$

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

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

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

where

$$\boldsymbol{\Sigma}_k^* = \left(\frac{1}{\sigma_\epsilon^2} \boldsymbol{D}_k^{\mathrm{T}} \boldsymbol{D}_k + \boldsymbol{\Sigma}_k^{-1} \right)^{-1},$$
$$\boldsymbol{m}_k^* = \boldsymbol{\Sigma}_k^* \left[\frac{1}{\sigma_\epsilon^2} \boldsymbol{x}_k^{\mathrm{T}} \left(\boldsymbol{y} - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{D}_j \boldsymbol{\beta}_j - \sum_{j=0}^{p-1} \boldsymbol{x}_j \theta_j \right) \right].$$

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

$$\theta_k | \boldsymbol{\beta}, \boldsymbol{\theta}_{-k}, \boldsymbol{\sigma^2}, \sigma_\epsilon^2, \boldsymbol{\phi}, \boldsymbol{y} \sim N \left(m_k^*, v_k^* \right),$$

where

$$v_k^* = \left(\frac{1}{\sigma_\epsilon^2} \boldsymbol{x}_k^{\mathrm{T}} \boldsymbol{x}_k + \frac{1}{\sigma_k^2 v_k} \right)^{-1},$$

$$m_k^* = v_k^* \left[\frac{1}{\sigma_\epsilon^2} \boldsymbol{x}_k^{\mathrm{T}} \left(\boldsymbol{y} - \sum_{j=0}^{p-1} \boldsymbol{D}_j \boldsymbol{\beta}_j - \sum_{\substack{j=0 \\ j \neq k}}^{p-1} \boldsymbol{x}_j \theta_j \right) + \frac{m_k}{\sigma_k^2 v_k} \right].$$

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

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

- The full conditional distribution for the data variance σ_ϵ^2 for the NCP is

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

References

Abramowitz, M., Stegun, I.A.: Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Courier Dover Publications, New York (1972)

Apputhurai, P., Stephenson, A.G.: Spatiotemporal hierarchical modelling of extreme precipitation in Western Australia using anisotropic Gaussian random fields. Environ. Ecol. Stat. **20**(4), 667–677 (2013)

Banerjee, S., Gelfand, A.E., Finley, A.O., Sang, H.: Gaussian predictive process models for large spatial data sets. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**(4), 825–848 (2008)

Banerjee, S., Finley, A.O., Waldmann, P., Ericsson, T.: Hierarchical spatial process models for multiple traits in large genetic trials. J. Am. Stat. Assoc. **105**(490), 506–521 (2010)

Banerjee, S., Carlin, B.P., Gelfand, A.E.: Hierarchical Modeling and Analysis for Spatial Data, 2nd edn. CRC Press, Boca Raton (2015)

Bass, M.R., Sahu, S.K.: Dynamically updated spatially varying parameterizations of hierarchical Bayesian models for spatially correlated data. Submitted (2016)

Berrocal, V.J., Gelfand, A.E., Holland, D.M.: A spatio-temporal downscaler for output from numerical models. J. Agric. Biol. Environ. Stat. **15**(2), 176–197 (2010)

Brooks, S.P., Gelman, A.: General methods for monitoring convergence of iterative simulations. J. Comput. Graph. Stat. **7**(4), 434–455 (1998)

Cressie, N., Johannesson, G.: Fixed rank kriging for very large spatial data sets. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **70**(1), 209–226 (2008)

Ecker, M.D., Gelfand, A.E.: Bayesian modeling and inference for geometrically anisotropic spatial data. Math. Geol. **31**(1), 67–83 (1999)

Filippone, M., Zhong, M., Girolami, M.: A comparative evaluation of stochastic-based inference methods for gaussian process models. Mach. Learn. **93**(1), 93–114 (2013)

Furrer, R., Genton, M.G., Nychka, D.: Covariance tapering for interpolation of large spatial datasets. J. Comput. Graph. Stat. **15**(3), 502–523 (2006)

Gelfand, A.E., Smith, A.F.: Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85**(410), 398–409 (1990)

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

Gelfand, A.E., Sahu, S.K., Carlin, B.P.: Efficient parameterizations for generalized linear mixed models, (with discussion). In: Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M. (eds.) Bayesian Statistics 5, pp. 165–180. Oxford University Press, Oxford (1996)

Gelfand, A.E., Kim, H.-J., Sirmans, C., Banerjee, S.: Spatial modeling with spatially varying coefficient processes. J. Am. Stat. Assoc. **98**(462), 387–396 (2003)

Gelman, A., Rubin, D.B.: Inference from iterative simulation using multiple sequences. Stat. Sci. **7**(4), 457–472 (1992)

Gneiting, T., Balabdaoui, F., Raftery, A.E.: Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **69**(2), 243–268 (2007)

Handcock, M.S., Stein, M.L.: A Bayesian analysis of kriging. Technometrics **35**(4), 403–410 (1993)

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

Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (2012)

Huerta, G., Sansó, B., Stroud, J.R.: A spatiotemporal model for Mexico City ozone levels. J. R. Stat. Soc. Ser. C (Appl. Stat.) **53**(2), 231–248 (2004)

Imai, K., van Dyk, D.: A bayesian analysis of the multinomial probit model using marginal data augmentation. J. Econom. **124**, 311–334 (2005)

Kaufman, C.G., Schervish, M.J., Nychka, D.W.: Covariance tapering for likelihood-based estimation in large spatial data sets. J. Am. Stat. Assoc. **103**(484), 1545–1555 (2008)

Liu, J.S., Wu, Y.N.: Parameter expansion for data augmentation. J. Am. Stat. Assoc. **94**, 1264–1274 (1999)

Matérn, B.: Spatial Variation, 2nd edn. Springer, Berlin (1986)

Meyn, S.P., Tweedie, R.L.: Markov Chains and Stochastic Stability. Springer, London (1993)

Papaspiliopoulos, O., Roberts, G.: Stability of the Gibbs sampler for Bayesian hierarchical models. Ann. Stat. **36**(1), 95–117 (2008)

Papaspiliopoulos, O., Roberts, G.O., Sköld, M.: Non-centered parameterisations for hierarchical models and data augmentation (with discussion). 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, pp. 307–326. Oxford University Press, USA (2003)

Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT press, Cambridge, MA (2006)

Robert, C.O., Casella, G.: Monte Carlo Statistical Methods, 2nd edn. Springer, New York (2004)

Roberts, G.O., Sahu, S.K.: Updating schemes, correlation structure, blocking and parameterization for the Gibbs sampler. J. R. Stat. Soc. Ser. B (Stat. Methodol.) **59**(2), 291–317 (1997)

Sahu, S.K., Roberts, G.O.: On convergence of the em algorithm and the gibbs sampler. Stat. Comput. **9**, 55–64 (1999)

Sahu, S.K., Gelfand, A.E., Holland, D.M.: High resolution space-time ozone modeling for assessing trends. J. Am. Stat. Assoc. **102**(480), 1221–1234 (2007)

Sahu, S.K., Gelfand, A.E., Holland, D.M.: Fusing point and areal level space-time data with application to wet deposition. J. R. Stat. Soc. Ser. C (Appl. Stat.) **59**(1), 77–103 (2010)

Sahu, S.K., Yip, S., Holland, D.M.: A fast Bayesian method for updating and forecasting hourly ozone levels. Environ. Ecol. Stat. **18**(1), 185–207 (2011)

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

Schabenberger, O., Gotway, C.A.: Statistical Methods for Spatial Data Analysis. CRC Press, Boca Raton (2004)

Stein, M.L.: Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York (1999)

van Dyk, D., Meng, X.-L.: The art of data augmentation (with discussion). J. Comput. Graph. Stat. **10**, 1–50 (2001)

Yu, Y., Meng, X.-L.: To center or not to center: That is not the question: an Ancillarity-Sufficiency Interweaving Strategy (ASIS) for boosting MCMC efficiency. J. Comput. Graph. Stat. **20**(3), 531–570 (2011)

Zhang, H.: Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. J. Am. Stat. Assoc. **99**(465), 250–261 (2004)

Zimmerman, D.L.: Another look at anisotropy in geostatistics. Math. Geol. **25**(4), 453–470 (1993)