Interpretable Priors for Hyperparameters for Gaussian Random Fields

Geir-Arne Fuglstad[1], Daniel Simpson[2], Finn Lindgren[3], and Håvard Rue[1]

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Statistics, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, United Kingdom
[3]Department of Mathematical Sciences, University of Bath, United Kingdom

Abstract

Gaussian random fields (GRFs) are important building blocks in hierarchical models for spatial data, but there is no practically useful, principled approach for selecting the prior on their hyperparameters. The prior is typically chosen in an *ad-hoc* manner, which lacks theoretical justification, despite the fact that we know that the hyperparameters are not consistently estimable from a single realization and that there is sensitivity to the choice of the prior.

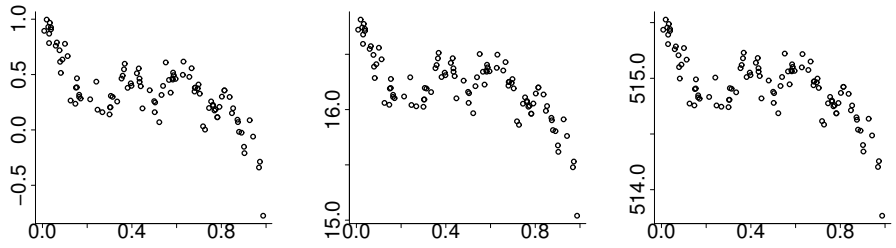
We first use the recent Penalised Complexity prior framework to construct a practically useful, tunable, weakly informative joint prior on the range and the marginal variance for Matérn GRFs with fixed smoothness. We then discuss how to extend this prior to a prior for a non-stationary GRF with covariates in the covariance structure.

Keywords: Bayesian, Gaussian random fields, Spatial models, Priors, Range, Variance, Penalised Complexity, Non-stationary

1 Introduction

Gaussian random fields (GRFs) are fundamental building blocks in spatial statistics and non-parametric modelling. They provide a simple and powerful tool for modelling data with spatial or temporal dependence, but the Gaussian assumption is in many cases too stringent and they are embedded within a hierarchical structure as one of multiple components that controls the behaviour of the observations. In this context, the behaviour of the GRF is usually controlled through a few parameters such as range, marginal variance and smoothness, but, even though GRFs are a standard modelling tool, the choice of prior distribution for the parameters remains a challenge. The prior is difficult to choose: a well-chosen prior will stabilise the inference and improve the predictive performance, whereas a poorly chosen prior can be catastrophic. Due to the infinite-dimensional nature of GRFs, it is difficult to construct a good prior and in most applications the prior is chosen in an *ad-hoc* fashion. In this paper we focus on Matérn GRFs with fixed smoothness, but the methods we develop are more widely applicable.

The lack of practically useful, theoretically founded priors is troubling since there is a ridge in the likelihood along which the value of the likelihood decreases slowly (Warnes and Ripley, 1987), and since the range and the marginal variance for the Matérn family of covariance functions cannot be estimated consistently under in-fill asymptotics (Zhang, 2004). The behaviour of the



(a) $\rho = \sigma^2 = 1$ (b) $\rho = \sigma^2 = 1000$ (c) $\rho = \sigma^2 = 1000000$

Figure 1: Simulations with the exponential covariance function $c(d) = \sigma^2 \mathrm{e}^{-d/\rho}$ for different values of $\rho = \sigma^2$ using the same underlying realization of independent standard Gaussian random variables. The patterns of the values are almost the same, but the levels differ.

prior on the ridge will strongly affect the behaviour of the posterior on the ridge and no matter how many points are observed in a bounded observation window, there is a limit to the amount of information that can be learned about these parameters. For example, if a one-dimensional GRF with an exponential covariance function is observed on $[0, 1]$, it is only the ratio of the range and the marginal variance that can be estimated consistently, and not the range and the marginal variance separately (Ying, 1991). When we move along the ridge towards large values of the range and the marginal variance, we change the distribution of the level of the points, but the distribution of the spread around the level changes only slightly. Figure 1 shows how the level moves, but the pattern of the points around the level remains stable for increasing values of the range and the marginal variance.

In some sense, the lack of identifiability is alleviated by the fact that there is a connection between what we can learn from observations and what can affect the predictive distributions, and in this example it is the ratio of the range and the marginal variance that is the important quantity for the asymptotic properties of the predictions (Stein, 1999). But even though there is a place for intrinsic models in spatial statistics, a practitioner who observes the values in Figure 1a is unlikely to believe that the ranges and marginal variances that can generate Figures 1b and 1c are correct even if the spread is consistent with the observed pattern. In this problem the likelihood by itself is not informative enough to properly control the sizes of the credible intervals, therefore, the practitioner should be provided with a prior that allows control, in an interpretable way, of how far the posterior is allowed to move along the ridge.

Despite this, to our knowledge, the only principled approach to prior selection for GRFs was introduced by Berger et al. (2001), who derived reference priors for a GRF partially observed with no noise. These priors fundamentally depend on the design of the experiment, which makes them inappropriate as "blind" default priors or when data is being analysed in a sequential fashion. This work has been extended by several authors (Paulo, 2005; Kazianka and Pilz, 2012; Kazianka, 2013) – critically Oliveira (2007) allowed for Gaussian observation noise – however, these papers have the same design dependence as the original work. Furthermore, the priors are not applicable as default priors for hierarchical models because the assumption of Gaussian observation noise is insufficient in many situations and there are currently no extensions of spatial reference priors to other observation processes. In the more restricted case of a GRF with a Gaussian covariance function van der Vaart and van Zanten (2009) showed that the inference asymptotically behaves well with an inverse gamma distribution on range, but they provide no

guidance on which hyperparameters should be selected for the prior.

In practice, the range is commonly given a uniform distribution on a bounded interval, a log-uniform distribution on a bounded interval or an inverse gamma distribution with infinite variance and the mean placed at an appropriate location. These priors have little theoretical foundation and are *ad-hoc* choices based on the idea that they will allow reasonable ranges, but Berger et al. (2001) noted that the posterior inference can be sensitive to the choice of cut-off for the uniform prior, and it is necessary with careful sensitivity analysis. The bounded intervals are necessary because improper prior distributions cannot be applied without great care as they tend to lead to improper posterior distributions. From a Bayesian modelling perspective this is an unsatisfactory situation because the prior is supposed to encode the user's uncertainty about the parameters and not be an *ad-hoc* choice made out of convenience without theoretical justification.

We apply the recent Penalised Complexity (PC) prior framework developed by Simpson et al. (2014) to construct a new, principled joint prior for the range and the marginal variance of a Matérn GRF. The PC prior framework ignores the observation process entirely and focuses instead on the geometry of the parameter space induced by the infinite-dimensional GRF. This is more technically demanding than considering only the finite-dimensional observation, like for the reference priors, but we are able to use the resulting prior for any spatial design and any observation process. The second key difference between the reference priors and the PC prior approach is that while the former is "non-informative" in a technical sense, PC priors are weakly informative and, therefore, require specific information from the user. In particular, PC priors need a point in the parameter space, considered a *base model*, and hyperparameters indicating how strongly the user wishes to shrink towards the base model. Simpson et al. (2014) showed that the resulting inference was quite robust against the specification of the hyperparameters.

The reference prior makes the posterior decay slowly along the ridge since predictions are not heavily influenced by the near intrinsicness in the level, but the PC-prior is weakly informative and allows the user to force the posterior to decay quicker along the ridge. When we incorporate a prior belief that the marginal standard deviation is below a specific value, we cannot move much past this value on the ridge without violating the prior belief. In this way it is possible to obtain more realistic parameter estimates and smaller and more meaningful credible intervals than with the reference priors. The reference priors are fully based on the likelihood and have no options for controlling how far the spatial model is allowed to move towards near-intrinsic models with large ranges and large variances even if they do not make sense for the application at hand.

The drawbacks and insufficiencies of noninformative priors for spatial models have already been commented by other authors and the arguments are well summarized by Palacios and Steel (2006) who wrote:

> Thus we need to think carefully about our priors and try to use as much information as we have available in eliciting reasonable prior distributions. In this particular context [Bayesian geostatistical models], we feel that this strategy is preferable to relying on automatic noninformative priors like the reference prior (if such priors are at all available; ...).

The prior for stationary GRFs provides a strong foundation for the development of priors for non-stationary GRFs. The covariance structure of a GRF is only observed indirectly through the values of the process and for locations without observations there is no information about the covariances. Therefore, the estimated covariance structure can be highly model-dependent and it would be useful and important to have an interpretable prior that provides understanding about the *a priori* assumptions that we put into the non-stationary model. We extend the prior

for the stationary GRF to a prior for a non-stationary GRF with covariates in the covariance structure. The prior is motivated by the PC prior framework, but has *ad-hoc* components.

We start by deriving the new, joint PC prior for the range and the marginal variance for a Matérn GRF with fixed smoothness parameter in Section 2. Then in Section 3 the frequentist coverage is studied through a simulation study and compared with the coverage when using the Jeffreys' rule prior and *ad-hoc* uniform and log-uniform priors. In Section 4 we study the behaviour of the joint posterior under the PC-prior and the Jeffreys' rule prior and discuss the difference in behaviour. Then the frequentist properties of spatial logistic regression are studied in Section 5 to demonstrate the applicability of the PC prior for a non-Gaussian observation process. In Section 6 we discuss how to extend the prior for the stationary model to a prior for a non-stationary GRF. The paper ends with discussion and concluding remarks in Section 7.

2 Penalised complexity prior

2.1 Background

The principle idea of the PC-prior framework is to think of a model component as a flexible extension of the *base model*, which is chosen to be the simplest or least flexible state of the model component. For example, a random effect is an extension of a random effect with zero variance, i.e. no random effect. After selecting the base model, one derives a distance measure from the base model to the models described by other parameter values. This distance from the base model describes how much more flexible each model is than the base model and provides a measure of complexity for the model component, and the prior is set directly on the distance from the base model instead of on the parameters of the model. This provides a useful tool for setting priors on parameters for which it is hard to have intuition. For example, correlation parameters close the border values -1, 0 and 1, or the range in spatial models.

To put this idea into practice, it is necessary to decide which measure of complexity to use and which prior to put on the resulting distance. We measure the extra complexity of each model compared to the base model through the Kullback-Leibler divergence (KLD). The KLD of the probability density f from the probability density g is defined by

$$D_{\mathrm{KL}}(f||g) = \int_{\chi} f(\boldsymbol{x}) \log\left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x},$$

and expresses the information lost when g is used to approximate f. The asymmetry of the KLD fits well with the choice of the base model as the favoured model, and we turn the KLD into a uni-directional distance from the base model g to the model f through $d(f||g) = \sqrt{2\mathrm{KLD}(f||g)}$.

The remaining key point is which distribution to put on the derived distance and Simpson et al. (2014) provide three principles for selecting the prior on the distance: Occam's razor, constant rate penalisation and user-defined scaling. Occam's razor is achieved by constructing a prior that penalises deviations from the base model and favours the base model until the data provides evidence against it. This suggests that the prior density should have its peak at distance 0 and less and less density for higher distances. The constant rate penalisation is achieved by making the prior on the distance, d, satisfy the relationship

$$\frac{\pi(d+\delta)}{\pi(d)} = r^{\delta}, \qquad d, \delta \geq 0,$$

for a constant decay-rate $0 < r < 1$. This means that the relative change in the prior when the distance increases by δ does not depend on the current distance d, and leads to the exponential

distribution

$$\pi(d) = \lambda \exp(-\lambda d).$$

After deciding on this distribution we apply the final principle of user-defined scaling to determine the hyperparameter λ. We transform the distance back to an interpretable size $Q(d)$ and include prior information through

$$\mathrm{P}(Q(d) > U) = \alpha \quad \text{or} \quad \mathrm{P}(Q(d) < L) = \alpha,$$

where U or L is an upper or lower limit, respectively, and α is the probability in the upper or lower tail of the prior distribution. By selecting U or L, and α the user combines prior belief with a prior derived from the geometry of the parameter space.

We want to extend the approach outlined above to Gaussian Matérn fields with fixed smoothness. These GRFs have the covariance function

$$C(d) = \sigma^2 \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{8\nu}d}{\rho}\right)^{\nu} K_\nu \left(\frac{\sqrt{8\nu}d}{\rho}\right),$$

where ρ is the spatial range, σ^2 is the marginal variance and ν is a fixed smoothness. The first thing we need to decide is what the base model should be; what type of model do we want to shrink towards? It seems clear that we want to shrink towards zero standard deviation, or no effect, but the GRF is controlled by two parameters and we need to describe how the range behaves as the marginal variance goes to zero. We choose to shrink the range simultaneously towards infinity with the goal of achieving shrinkage towards a constant random field with variance zero.

The GRFs possess a technical difficulty not present for finite dimensional distributions. Imagine that the spatial field is observed at all points in a bounded observation window, for example, $[0, 1]^2$, then the KLD between the distributions specified by two choices of parameters (ρ_0, σ_0^2) and (ρ_1, σ_1^2) is in general infinite. This means that we must be careful in our prior construction and understand which changes corresponds to infinite KLD. To facilitate the construction of the prior, we select a parametrization of the spatial field that more accurately describes what can be and what cannot be estimated from a bounded observation window. We introduce the parameters $\kappa = \sqrt{8\nu}/\rho$ and

$$\tau = \frac{\Gamma(\nu)}{(4\pi)^{d/2}\Gamma(\nu + d/2)\sigma^2\kappa^{2\nu}}.$$

These parameters arise from a slight re-parametrization of the SPDE in Lindgren et al. (2011),

$$(\kappa^2 - \Delta)^{\alpha/2}(\sqrt{\tau}u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{1}$$

where $\Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}$ is the Laplacian and $\mathcal{W}$ is standard Gaussian white noise. If κ and τ are chosen as above, this SPDE specifies a Matérn GRF with range ρ, marginal variance σ^2 and smoothness $\nu = \alpha - d/2$. In this parametrization τ can be consistently estimated from a bounded observation window, whereas κ cannot. If κ is kept fixed and the value of τ changes, the KLD is infinite, but if the value of τ is kept fixed and the value of κ changes, the KLD is finite. This parametrization allows us to use a two-step procedure where we first set a prior on κ based on the distribution of the GRF through the KLD and then set a prior on τ given the value of κ through a consideration of finite-dimensional observations.

2.2 Joint prior on range and marginal variance

We first construct a joint prior for κ and τ through the densities $\pi(\kappa)$ and $\pi(\tau|\kappa)$, and then transform the resulting joint prior to a joint prior on ρ and σ^2. The prior on κ is constructed using the limit of a re-scaled form of the KLD, but the details are technical and the derivation is, therefore, given in Appendix A. The result is a description of how much more complex the models with $\kappa > 0$ are compared to the intrinsic model $\kappa = 0$ for a fixed τ. If we ignore the multiplicative constants, the resulting distance is

$$d(\kappa) = \kappa^{d/2}. \tag{2}$$

The distance in Equation (2) expresses how far the distribution of the GRF is from the intrinsic GRF as a function of κ and solves the highly non-trivial problem of describing how much a Matérn GRF varies as a function of range. Since $\rho = \sqrt{8\nu}/\kappa$, the distance in Equation (2) implies that the range can be made arbitrarily large without making large changes in the distribution if the marginal variance increases in such a way that τ is kept constant. This increase in marginal variance will be controlled by the prior for $\tau|\kappa$, which will disallow unreasonably large marginal variances and thus near-intrinsic models.

An exponential prior on the distance from the base model gives

$$\begin{aligned}\pi(\kappa) &= \lambda_1 \exp(-\lambda_1 d(\kappa)) \left| \frac{\mathrm{d}}{\mathrm{d}\kappa} d(\kappa) \right| \\ &= \frac{\lambda_1 d}{2} \kappa^{d/2-1} \exp(-\lambda_1 \kappa^{d/2}), \quad \kappa > 0,\end{aligned} \tag{3}$$

where λ_1 is determined by controlling the *a priori* probability that the range is below a specific limit,

$$\mathrm{P}\left(\frac{\sqrt{8\nu}}{\kappa} < \rho_0 \right) = \alpha_1,$$

i.e.

$$\lambda_1 = -\log(\alpha_1) \left(\frac{\rho_0}{\sqrt{8\nu}} \right)^{d/2}.$$

The calibration of the prior requires the selection of two values: the lower range, ρ_0, and the probability in the lower tail, α_1.

The prior on τ cannot be derived from the distribution of the process on a bounded observation window since this parameter is completely determined by the values of the process on the observation window. It would be meaningless to put a prior on τ if we observed all values in the observation window, and we must instead choose a situation in which a prior is necessary. We make the assumption that we are interested in observing finite-dimensional quantities from the spatial field.

With κ fixed the joint distribution of a finite number of observations is a multivariate Gaussian distribution of the form

$$\pi(\boldsymbol{u}) \propto \exp\left(-\frac{\tau}{2} \boldsymbol{u}^{\mathrm{T}} \Sigma^{-1} \boldsymbol{u} \right),$$

where Σ is a fixed matrix. In this distribution τ acts as a precision parameter and we can use the prior constructed by (Simpson et al., 2014),

$$\pi(\tau) = \frac{\lambda_2}{2} \tau^{-3/2} \exp(-\lambda_2 \tau^{-1/2}), \quad \tau > 0, \tag{4}$$

where λ_2 is determined by controlling the *a priori* probability that the marginal variance exceeds a specific level,

$$\mathrm{P}\left(\frac{C(\nu)}{\tau\kappa^{2\nu}} > \sigma_0^2 \middle| \kappa\right) = \alpha_2, \tag{5}$$

where

$$C(\nu) = \frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}}$$

is the constant needed to make the left-hand side of the inequality equal to the marginal variance of the GRF.

Since the calibration criterion in Equation (5) is conditional on the value of κ, it introduces dependence between κ and τ in the joint prior. We write Equation (5) as

$$\mathrm{P}\left(\tau < \frac{C(\nu)}{\kappa^{2\nu}\sigma_0^2} \middle| \kappa\right) = \alpha_2$$

and find

$$\exp\left(-\lambda_2 \left(\frac{C(\nu)}{\kappa^{2\nu}\sigma_0^2}\right)^{-1/2}\right) = \alpha_2,$$

$$\lambda_2 = \frac{\lambda_3}{\kappa^\nu},$$

where λ_3 absorbs the other constants in λ_2. We insert this into Equation (4) and find the conditional distribution

$$\pi(\tau|\kappa) = \frac{\lambda_3 \tau^{-3/2}}{\kappa^\nu} \exp(-\lambda_3 \kappa^{-\nu} \tau^{-1/2}). \tag{6}$$

This implies that the dependence between κ and τ is affected by the value of the smoothness ν.

The joint prior on κ and τ is found by combining Equation (3) and Equation (6), and is given by

$$\begin{aligned}\pi(\kappa, \tau) &= \pi(\kappa)\pi(\tau|\kappa) \\ &= \frac{\lambda_1 \lambda_3 d}{2} \tau^{-3/2} \kappa^{d/2-1-\nu} \exp(-\lambda_1 \kappa^{d/2} - \lambda_3 \kappa^{-\nu} \tau^{-1/2}).\end{aligned}$$

There is a one-to-one correspondence between κ and τ, and ρ and σ^2,

$$\begin{bmatrix}\rho \\ \sigma^2\end{bmatrix} = \begin{bmatrix}\frac{\sqrt{8\nu}}{\kappa} \\ \frac{C(\nu)}{\kappa^{2\nu}\tau}\end{bmatrix},$$

which can be exploited to transform the joint prior for κ and τ to the joint prior for ρ and σ^2,

$$\pi(\rho, \sigma^2) = \left[\frac{d\lambda_4}{2} \rho^{-1-d/2} \exp\left(-\lambda_4 \rho^{-d/2}\right)\right] \left[\frac{\lambda_5}{2} \sigma^{-1} \exp\left(-\lambda_5 \sigma\right)\right], \tag{7}$$

where λ_4 and λ_5 are selected according to the *a priori* statements

$$\mathrm{P}(\rho < \rho_0) = \alpha_4 \quad \text{and} \quad \mathrm{P}(\sigma^2 > \sigma_0^2) = \alpha_5,$$

which give

$$\lambda_4 = -\rho_0^{d/2} \log(\alpha_4) \quad \text{and} \quad \lambda_5 = -\frac{\log(\alpha_5)}{\sigma_0}.$$

Table 1: The four different priors used in the study of frequentist coverage. The Jeffreys' rule prior uses the spatial design of the problem through $U = (\frac{\partial}{\partial\rho}\Sigma)\Sigma^{-1}$, where Σ is the correlation matrix of the observations (See Berger et al. (2001)).

Prior	**Expression**	**Parameters**
PriorPC	$\pi_1(\rho,\sigma) = \lambda_1\lambda_2\rho^{-2}\exp\left(-\lambda_1\rho^{-1} - \lambda_2\sigma\right)$	$\rho,\sigma > 0$ Hyperparameters: α_ρ, ρ_0, α_σ, σ_0
PriorJe	$\pi_2(\rho,\sigma) = \sigma^{-1}\left(\text{tr}(U^2) - \frac{1}{n}\text{tr}(U)^2\right)^{1/2}$	$\rho,\sigma > 0$ Hyperparameters: None
PriorUn1	$\pi_3(\rho,\sigma) \propto \sigma^{-1}$	$\rho \in [A,B]$, $\sigma > 0$ Hyperparameters: A, B
PriorUn2	$\pi_4(\rho,\sigma) \propto \sigma^{-1}\cdot\rho^{-1}$	$\rho \in [A,B]$, $\sigma > 0$ Hyperparameters: A, B

3 Frequentist coverage

The series of papers on reference priors for GRFs starting with Berger et al. (2001) evaluated the priors by studying frequentist properties of the resulting Bayesian inference. If a prior is intended as a default prior, it should lead to good frequentist properties such as a frequentist coverage of the equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals that is close to the nominal $100(1-\alpha)\%$. We replicate their simulation study with one key difference: we do not include covariates and measurement noise. The reference priors are not proper distributions and the goal of the series of papers was to derive them for different situations such as a spatial field combined with covariates, and a spatial field combined with covariates and Gaussian measurement noise. However, in this paper we construct a prior for the GRF component itself and we are *not* constructing a prior for the GRF together with covariates or together with covariates and Gaussian measurement noise. This is possible because the PC-prior is a proper distribution and can be applied to a spatial field together with covariates and arbitrary observation processes without worrying about the properness of the posterior.

The study uses an isotropic GRF, u, with an exponential covariance function $c(d) = \exp(-2d/\rho_0)$ observed at the locations shown in Figure 2. The observation locations were randomly selected within the domain $[0,1]^2$ and are distributed in an irregular pattern. The study is performed for two values of the nominal range: a short range, $\rho_0 = 0.1$, and a long range, $\rho_0 = 1$. We generate multiple realizations and for each realization we assume that the field is observed directly and fit the model

$$y_i = u(\boldsymbol{s}_i), \qquad i = 1, 2, \ldots, 25,$$

where u is a GRF with an exponential covariance function with parameters ρ and σ^2. We apply four different priors: the PC-prior (PriorPC), the Jeffreys' rule prior (PriorJe), a uniform prior on range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn1) and a uniform prior on the log-range on a bounded interval combined with the Jeffreys' prior for variance (PriorUn2). The full expressions for the priors are given in Table 1.

For each choice of prior and hyperparameters we generate 1000 observation vectors $\boldsymbol{y} =$

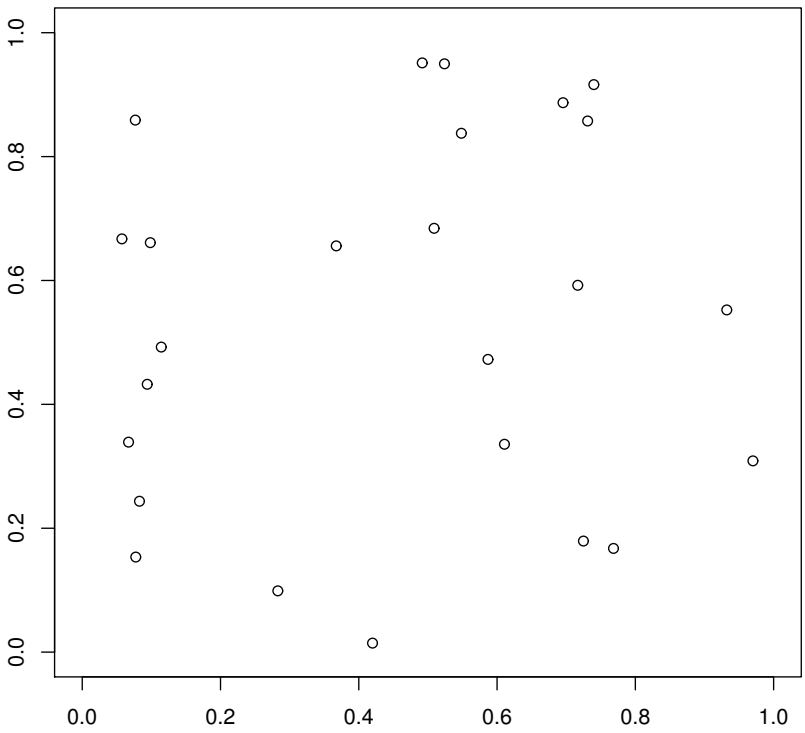


Figure 2: Spatial design for the simulation study of frequentist coverage.

Table 2: Frequentist coverage of 95% credible intervals for range and marginal variance when the range $\rho_0 = 0.1$ using PriorPC, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.768 [0.25]	0.749 [0.24]	0.760 [0.20]	0.693 [0.17]
0.1	0.965 [0.35]	0.976 [0.29]	0.961 [0.27]	0.937 [0.21]
0.4	0.990 [0.45]	0.989 [0.41]	0.993 [0.33]	0.987 [0.25]
1.6	0.717 [0.98]	0.692 [0.82]	0.756 [0.54]	0.807 [0.34]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.941 [1.5]	0.952 [1.4]	0.957 [1.3]	0.918 [0.97]
0.1	0.953 [1.6]	0.966 [1.5]	0.944 [1.4]	0.927 [0.98]
0.4	0.953 [2.0]	0.952 [1.8]	0.960 [1.5]	0.943 [1.1]
1.6	0.904 [3.9]	0.906 [3.2]	0.939 [2.2]	0.972 [1.3]

$(y_1, y_2, \ldots, y_{25})$ and estimate the equal-tailed 95% credible interval for each observation by running an MCMC-chain. The number of times the true value is contained within the estimated credible interval is divided by 1000 and given as the estimate of the frequentist coverage. We tried MCMC chains of length 25000 with 5000 iterations burn-in and MCMC chains of length 125000 with 25000 burn-in. The results for PriorPC, PriorUn1 and Prior2 were stable, but the chains for PriorJe showed in some cases notoriously high autocorrelation and unstable results and we re-ran with MCMC chains of length 1500000 with 300000 iterations as burn-in.

PriorJe has no hyperparameters, but PriorPC, PriorUn1 and PriorUn2 each has hyperparameters that need to be set before using the prior. For PriorUn1 and PriorUn2 it is hard to give guidelines about which values should be selected since the main purpose of limiting the prior distributions to a bounded interval is to avoid an improper posterior and the choice tends to be *ad-hoc*. For PriorPC, on the other hand, there is a calibration criterion to help choosing the hyperparameters, which helps give an idea about which prior assumptions the chosen hyperparameters are expressing.

For PriorPC we make a decision about the scales of the range and the marginal variance. The prior is set through four hyperparameters that describe our prior beliefs about the spatial field. We use

$$\mathrm{P}(\rho < \rho_0) = 0.05$$

for $\rho_0 = 0.025\rho_{\mathrm{T}}$, $\rho_0 = 0.1\rho_{\mathrm{T}}$, $\rho_0 = 0.4\rho_{\mathrm{T}}$ and $\rho_0 = 1.6\rho_{\mathrm{T}}$, where ρ_{T} is the true range. This covers a prior where ρ_0 is much smaller than the true range, two priors where ρ_0 is smaller than the true range, but not far away, and one prior where ρ_0 is higher than the true range. For the marginal variance we use

$$\mathrm{P}(\sigma^2 > \sigma_0^2) = 0.05,$$

for $\sigma_0 = 0.625$, $\sigma_0 = 2.5$, $\sigma_0 = 10$ and $\sigma_0 = 40$. We follow the same logic as for range and cover too small and too large σ_0 and two reasonable values. For PriorUn1 and PriorUn2, we set the lower and upper limits for the nominal range according to the values $A = 0.05$, $A = 0.005$ and $A = 0.0005$, and $B = 2$, $B = 20$ and $B = 200$. Some of the values are intentionally extreme to see the effect of misspecification.

Table 3: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 0.1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.901 [0.95]	0.901 [8.6]	0.847 [122]
$5 \cdot 10^{-3}$	0.935 [0.92]	0.918 [7.7]	0.887 [110]
$5 \cdot 10^{-4}$	0.948 [0.93]	0.929 [7.9]	0.893 [110]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.952 [3.5]	0.941 [29]	0.895 [460]
$5 \cdot 10^{-3}$	0.945 [3.3]	0.937 [27]	0.907 [410]
$5 \cdot 10^{-4}$	0.953 [3.3]	0.925 [27]	0.921 [412]

Table 4: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 0.1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.986 [0.47]	0.979 [0.84]	0.988 [1.1]
$5 \cdot 10^{-3}$	0.976 [0.44]	0.950 [0.81]	0.966 [1.0]
$5 \cdot 10^{-4}$	0.932 [0.40]	0.945 [0.70]	0.944 [1.3]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.949 [2.0]	0.962 [2.9]	0.965 [3.6]
$5 \cdot 10^{-3}$	0.968 [1.8]	0.960 [2.6]	0.959 [3.2]
$5 \cdot 10^{-4}$	0.948 [1.7]	0.960 [2.4]	0.949 [3.7]

Table 5: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorPC, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.950 [12]	0.945 [7.1]	0.906 [3.2]	0.821 [1.4]
0.1	0.977 [15]	0.966 [8.2]	0.962 [3.6]	0.866 [1.5]
0.4	0.965 [26]	0.981 [13]	0.992 [5.1]	0.988 [1.8]
1.6	0.159 [74]	0.349 [31]	0.700 [11]	0.954 [3.3]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.944 [11]	0.956 [6.2]	0.933 [2.8]	0.797 [1.1]
0.1	0.957 [13]	0.966 [7.2]	0.954 [3.1]	0.865 [1.2]
0.4	0.943 [23]	0.957 [11]	0.987 [4.4]	0.972 [1.5]
1.6	0.441 [68]	0.534 [29]	0.797 [9.1]	0.984 [2.5]

Table 6: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorUn1, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.995 [1.5]	0.831 [18]	0.593 [188]
$5 \cdot 10^{-3}$	0.996 [1.5]	0.818 [18]	0.539 [188]
$5 \cdot 10^{-4}$	0.994 [1.5]	0.844 [18]	0.537 [188]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.979 [2.0]	0.857 [20]	0.614 [208]
$5 \cdot 10^{-3}$	0.979 [2.0]	0.821 [20]	0.585 [205]
$5 \cdot 10^{-4}$	0.969 [2.0]	0.828 [20]	0.561 [206]

Table 7: Frequentist coverage of 95% credible intervals for range and marginal variance when the true range $\rho_0 = 1$ using PriorUn2, where the average lengths of the credible intervals are shown in brackets.

(a) Range

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.980 [1.5]	0.959 [12]	0.933 [69]
$5 \cdot 10^{-3}$	0.974 [1.5]	0.954 [12]	0.954 [67]
$5 \cdot 10^{-4}$	0.964 [1.5]	0.953 [13]	0.956 [68]

(b) Marginal variance

$A \backslash B$	2	20	200
$5 \cdot 10^{-2}$	0.955 [1.8]	0.952 [12]	0.945 [61]
$5 \cdot 10^{-3}$	0.962 [1.8]	0.943 [12]	0.941 [60]
$5 \cdot 10^{-4}$	0.939 [1.8]	0.946 [12]	0.953 [60]

The results for PriorPC, PriorUn1 and PriorUn2 are shown in Tables 2 and 5, Tables 3 and 6, and Tables 4 and 7, respectively. The results for PriorJe was 97.0% coverage with average length of the credible intervals of 0.86 for range and 96.0% coverage and average length of the credible intervals of 2.7 for marginal variance for $\rho_0 = 0.1$, and 95.4% coverage with average length of the credible intervals of 445 for range and 94.4% coverage with average length of the credible intervals of 355 for variance for $\rho_0 = 1$. It is clear from the tables that for PriorPC, PriorUn1 and PriorUn2 the coverage and the length of the credible intervals are dependent on the choice of hyperparameters. This is not surprising since there are few observation and there is a ridge in the likelihood where the behaviour is strongly dependent on the the prior. The length of the credible intervals are, in general, more well-behaved for $\rho_0 = 0.1$ than for $\rho_0 = 1$ because there is more information available about range when the range is short compared to the domain size.

For PriorUn1 the coverage and the length of the credible intervals is strongly dependent on the upper limit in the prior. The prior has the undesirable property of including stronger and stronger prior belief in high ranges when the upper limit is increased. One might argue that the upper limit would never be selected as extreme as in the example, but it verifies the observation of Berger et al. (2001) that the inference is sensitive to the hyperparameters for this prior. For PriorUn2 the coverage is good in both the short range and long range situation, but the lengths of the credible intervals are sensitive to the upper limit of the prior. The new PriorPC exhibit sensitivity in the coverage and the lengths of the credible intervals, but for this prior it is caused by explicitly including information that conflicts with the true value, whereas for PriorUn1 and PriorUn2 it is not immediately clear what information is included through the different choices of hyperparameters.

The coverage of PriorJe is good, but the credible intervals seem excessively long and the prior is more computationally expensive than the other priors. PriorJe is only computationally feasible for low amounts of points since there is a cubic increase in complexity as a function of the number of observations. The average length of the credible intervals for $\rho_0 = 1$ for marginal variance is 355, which imply unreasonably high standard deviations. The high standard deviations do not seem consistent with an observation with values contained between -3 and 3. We study the credible intervals for PriorPC and PriorJe closer for a specific realization in the next section to gain intuition about why this happen.

With respect to computation time and easy of use versus coverage and length of credible

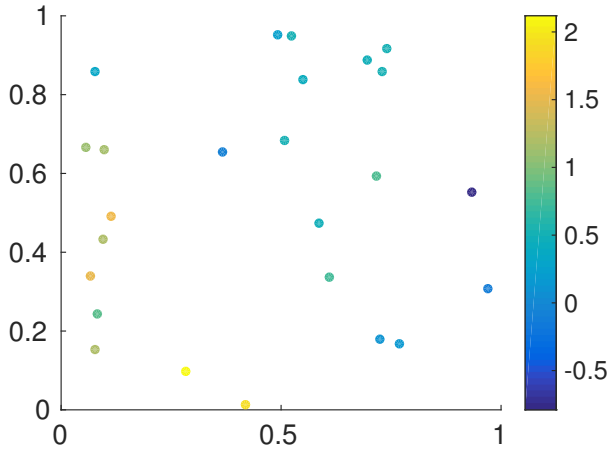


Figure 3: One realization of a GRF with the covariance function $c(d) = \exp(-2d)$ at 25 selected locations.

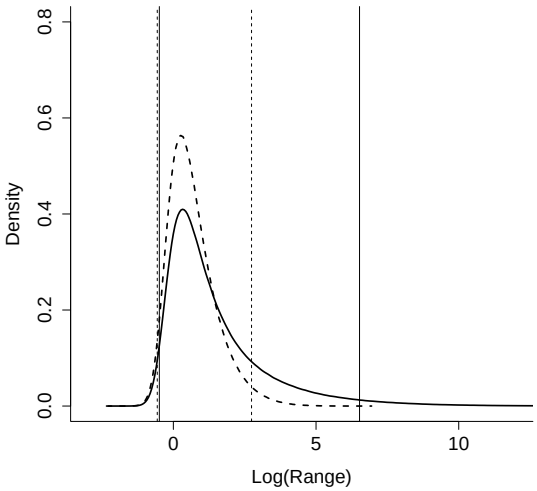
intervals PriorUn2 and PriorPC appear to be the best choices. If coverage is the only concern, PriorUn2 performs the best, but if one also wants to control the length of the credible intervals by disallowing unreasonably high variances, PriorPC offers the most interpretable alternative. Based on one of the realizations in this simulation study we are unlikely to believe that the spatial field could have a standard deviation greater than 4, and by encoding this information in PriorPC we can limit the upper limits of the credible intervals both for range and marginal variance.

4 Behaviour of the joint posterior

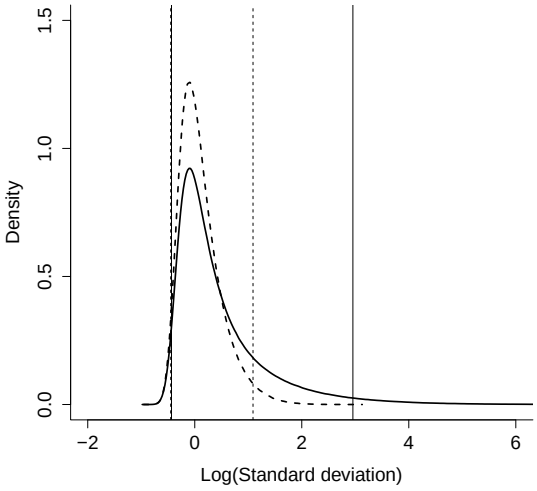
The sensitivity of the length of the credible intervals to the prior and the extreme length of the credible intervals seen for the Jeffreys' rule prior are not entirely surprising due to the ridge in the likelihood, but they are troubling. In the previous section we only looked at properties of the marginal credible intervals, but these do not tell the entire story because there is strong dependence between range and marginal variance in the joint posterior distribution. We study this dependence by studying the posterior distribution for the realization shown in Figure 3. The true range used to simulate the realization is 1. We draw samples from the joint posterior using the PC-prior with parameters $\alpha_\rho = 0.05$, $\rho_0 = 0.1$, $\alpha_\sigma = 0.05$ and $\sigma_0 = 10$, and we draw samples from the joint posterior using the Jeffreys' rule prior.

Figure 4 shows that the upper tails of the posteriors when the Jeffreys' rule prior is used are heavier than the upper tails of the posteriors when the PC-prior is used. The lower endpoints of the credible intervals are similar for both priors, but there is a large difference in the upper limits because the likelihood decays slowly along the ridge and the behaviour of the prior on the ridge is important for the behaviour of the posterior. The marginal posterior distributions do not show the full story about the inference on range and marginal variance because the two parameters are strongly dependent in the posterior distribution. The PC-prior for range has a heavy upper tail for range and the upper tail of the posterior of range is controlled through the prior on marginal variances. The large difference in the marginal posterior for the nominal range in Figure 4a can be explained by the behaviour of the joint posterior.

Figure 5 shows the strong posterior dependence between nominal range and standard de-



(a) Posterior for the logarithm of range



(b) Posterior for the logarithm of marginal standard deviation

Figure 4: Marginal posteriors of the logarithm of range and the logarithm of marginal standard deviation. The dashed lines shows the posterior and the credible intervals when the PC-prior is used and the solid line shows the posterior and the credible intervals when the Jeffreys' rule prior is used.

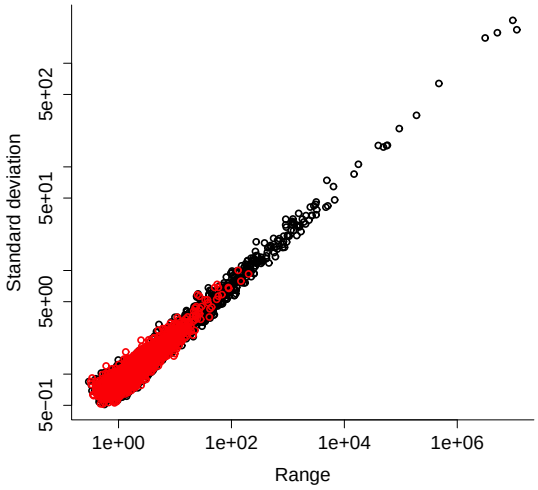


Figure 5: Samples from the joint posterior of range and marginal standard deviation. The red circles are samples using the PC-prior and the black circles are samples using the Jeffreys' rule prior.

viation in the tail of the distribution. The extreme tail of the Jeffreys' rule prior corresponds to movement far along the ridge in the likelihood. Stein (1999) showed that the ratio of range and marginal variance is the important quantity for asymptotic predictions with the exponential covariance function, which means that long tails are not a major concern for predictions, but for interpretability of range and marginal variance this heavy tail presents a problem. The values of all the observations in Figure 3 lie in the range -1 to 3 and it is unlikely that the true standard deviation should be on the order of 20. After conditioning on data the effect of using a near intrinsic GRF with simultaneously large values for range and marginal variance is almost the same as a GRF with meaningful values for range and marginal variance. Intrinsic models have a place in statistics, but the results show that the Jeffreys' rule prior has the, potentially, undesirable behaviour of favouring intrinsic GRFs with large marginal standard deviations and ranges. The PC prior offers a way to introduce prior belief about the marginal standard deviations, and thus a way to avoid the intrinsic GRFs and keep the standard deviation at reasonable (according to prior belief) values.

5 Example: Spatial logistic regression

What makes the PC prior more practically useful than the reference prior, beyond the computational benefits and interpretability, is that the prior is applicable in any hierarchical model and does not have to be re-derived each time a component is removed or added, or the observation process is changed. We consider a simple spatial logistic regression example to demonstrate the applicability of the PC prior beyond direct observations or Gaussian measurement noise.

We select the 25 locations in Figure 2 and generate realizations from the model

$$y_i|p_i \sim \text{Binomial}(20, p_i), \quad i = 1, 2, \ldots, 25,$$

where

$$\text{probit}(p_i) = u(\boldsymbol{s}_i),$$

Table 8: Frequentist coverage of the 95% credible intervals for range and marginal variance when the true range is 0.1 and true marginal variance is 1, where the average length of the credible intervals are given in brackets, for the spatial logistic regression example.

(a) Range

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.804 [0.29]	0.790 [0.24]	0.774 [0.22]	0.726 [0.19]
0.1	0.974 [0.41]	0.986 [0.37]	0.974 [0.33]	0.956 [0.24]
0.4	0.996 [0.61]	0.982 [0.57]	0.996 [0.43]	0.992 [0.30]
1.6	0.648 [1.4]	0.604 [1.2]	0.722 [0.67]	0.762 [0.44]

(b) Marginal variance

$\rho_0 \backslash \sigma_0$	40	10	2.5	0.625
0.025	0.942 [2.0]	0.946 [1.9]	0.948 [1.7]	0.912 [1.2]
0.1	0.920 [2.3]	0.942 [2.0]	0.964 [1.8]	0.922 [1.2]
0.4	0.952 [2.7]	0.962 [2.4]	0.968 [1.9]	0.928 [1.2]
1.6	0.904 [5.3]	0.936 [4.1]	0.966 [2.7]	0.982 [1.5]

where u is a GRF with the exponential covariance function with parameters $\rho = 0.1$ and $\sigma = 1$. For each realization the parameters ρ and σ^2 are assumed unknown and must be estimated. The posterior of the parameters is estimated with an MCMC chain and the equal-tailed 95% credible intervals are estimated from the samples of the MCMC-chain after burn-in. We repeat the procedure above 500 times and report the number of times the true value is contained in the credible interval and the average length of the credible interval.

The experiment is repeated for 64 different settings of the prior: the hyperparameter ρ_0 varies over $\rho_0 = 0.0025, 0.01, 0.04, 0.16$ and the hyperparameter σ_0 varies over $\sigma_0 = 40, 10, 2.5, 0.625$. This covers a broad range of values from too small to too large. The values in Table 8 are similar to the values in Table 2 except that the credible intervals are slightly longer. The longer credible intervals are reasonable since the binomial likelihood gives less information about the spatial field than direct observation of the spatial field. The coverage for marginal variance is good even for grossly miscalibrated priors, but the coverage for range is sensitive to bad calibration for range and the coverage is somewhat higher than nominal for the well-calibrated priors. This is a feature also seen in the directly observed case in Section 3.

6 Priors on non-stationarity

The development of practically useful, interpretable priors for stationary GRFs is important and useful, but the need for such priors is even stronger for non-stationary GRFs. The covariance structure estimated with a non-stationary GRF can be strongly dependent on the *a priori* assumptions on the non-stationarity. It can be difficult to understand the implications of the *a priori* assumptions that we put into non-stationary models because it is difficult to understand how the distribution of a GRF varies as a function of the parameters. The two main challenges are to construct a prior, which accounts for the highly non-trivial geometry of the parameter space, and to calibrate the prior in an interpretable way. Therefore, the PC prior framework is an appealing starting point with properties that fit well for developing such a prior.

There exists different models for non-stationary data and they incorporate non-stationarity in

different ways. For example, in the deformation method (Sampson and Guttorp, 1992; Schmidt and O'Hagan, 2003; Damian et al., 2001, 2003) a stationary GRF is made non-stationary through a spatial deformation, in the process convolution method (Haas, 1990b,a; Paciorek and Schervish, 2006) a spatially varying kernel function is convolved with Gaussian white noise, and in the stochastic partial differential equation (SPDE) approach (Bolin and Lindgren, 2011; Fuglstad et al., 2015a,b) a non-stationary GRF is specified through an SPDE with spatially varying coefficients. Each of these types of models has had extensions to covariates in the covariance structure (Schmidt et al., 2011; Neto et al., 2014; Ingebrigtsen et al., 2014a,b).

Ideally, we would derive a prior that could deal with any type of non-stationarity and be applicable for any model for non-stationarity, but, in practice, this is not feasible. For the purpose of this discussion the starting point is the sub-class of the SPDE models (Lindgren et al., 2011) consisting of the model discussed in Ingebrigtsen et al. (2014a). This model uses covariates, and thus needs fewer parameters and is less computationally expensive than a model with a more flexible covariance structure. The model is an extension of the stationary SPDE in Equation (1) with a slightly different parametrization and coefficients that vary spatially,

$$[\kappa(\boldsymbol{s})^2 - \Delta](\tau(\boldsymbol{s})u(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \qquad \mathcal{D} \subset \mathbb{R}^d, \tag{8}$$

with Neumann boundary conditions. We have fixed $\alpha = 2$ to get a practically feasible model, but with a spatially varying range it is unlikely to pose a large practical limitation to fix the smoothness $\nu = 1$.

We make the assumption that the priors on the correlation structure and the marginal variances can be set independently in an analogous way to the stationary GRF. This means we must solve two challenges: covariates must be included separately in the correlation structure and in the marginal variances, and practically useful, interpretable priors must be developed for the covariates in the correlation structure and for the covariates in the marginal variances. Thus setting priors on non-stationarity is not only a question about which prior to set after the parametrization is decided, but a question of how to parametrize the non-stationarity *and* how to set priors on the parameters in the parametrization.

6.1 Parametrizing the non-stationarity

Ingebrigtsen et al. (2014a) expands $\log(\kappa(\cdot))$ and $\log(\tau(\cdot))$ in Equation (1) into low-dimensional bases, but experience numerical problems and prior sensitivity to the priors for the weights in the basis expansions. Ingebrigtsen et al. (2014b) attempt to solve this by setting the hyperparameters of the priors based on the properties of the spatially varying local ranges and marginal variances. The procedure improves the calibration step of the prior specification compared to Ingebrigtsen et al. (2014a), but does not solve the inherent problem that $\kappa(\cdot)$ affects both the correlation structure and the marginal variances of the spatial field. We aim to improve their procedure by first improving the parametrization of the non-stationarity, and then setting and calibrating the prior using the improved parametrization.

The model used by Ingebrigtsen et al. (2014a) introduces spatial variation in the covariance structure by varying the coefficients of the SPDE, but there exists another way to introduce non-stationarity. Instead of varying the coefficients of the SPDE, one can vary the geometry of the space in a similar way as the deformation method. If E is the Euclidean space $\mathbb{R}^2$, the simple SPDE

$$(1 - \Delta_E)u(\boldsymbol{s}) = \sqrt{4\pi}\mathcal{W}_E(\boldsymbol{s}), \quad \boldsymbol{s} \in E, \tag{9}$$

generates a stationary Matérn GRF with range $\rho = \sqrt{8}$, marginal variance $\sigma^2 = 1$ and smoothness $\nu = 1$. Instead of introducing spatially varying coefficients, we introduce spatially varying

distances in the space on which the SPDE is defined. We take a two-dimensional manifold $E = \mathbb{R}^2$ and give the space geometric structure according to the metric tensor $\mathbf{g}(\boldsymbol{s}) = R(\boldsymbol{s})^{-2}\mathbf{I}_2$, where $R(\cdot)$ is a strictly positive scalar function. This means that distances are locally scaled by a factor $R(\boldsymbol{s})^{-1}$, or more specifically,

$$\mathrm{d}\sigma^2 = \begin{bmatrix}\mathrm{d}s_1 & \mathrm{d}s_2\end{bmatrix} \mathbf{g}(\boldsymbol{s}) \begin{bmatrix}\mathrm{d}s_1 \\ \mathrm{d}s_2\end{bmatrix} = R(\boldsymbol{s})^{-2}(\mathrm{d}s_1^2 + \mathrm{d}s_2^2), \tag{10}$$

where $\mathrm{d}\sigma$ is the line element, and s_1 and s_2 are the two coordinates of $E = \mathbb{R}^2$.

The line element in two-dimensional Euclidean space $\sqrt{\mathrm{d}s_1^2 + \mathrm{d}s_2^2}$ is everywhere scaled according to the function $R(\cdot)$, and Equation (10) describes the non-stationary through a spatially varying geometry, which results in a curved two-dimensional manifold that must be embedded dimension higher than 2 to exist in Euclidean space. The SPDE is not stationary on this space and does not lead to constant marginal variance because the curvature of the space is non-constant unless $R(\cdot)$ is a constant function, but there will be less interaction between $R(\cdot)$ and the marginal variance than $\kappa(\cdot)$ and the marginal variance. And for a slowly varying $R(\cdot)$ the variation in marginal variances is small.

The above construction gives geometric intuition about what type of non-stationarity the equation can generate, but it is not directly useful for implementation. We can relate the Laplace-Beltrami operator in E to the usual Laplacian in $\mathbb{R}^2$ through

$$\Delta_E = \frac{1}{\sqrt{\det(g)}} \nabla_{\mathbb{R}^2} \cdot (\sqrt{\det(g)} g^{-1} \nabla_{\mathbb{R}^2}) = R(\boldsymbol{s})^2 \Delta_{\mathbb{R}^2},$$

and the Gaussian standard white noise in E to the Gaussian standard white noise in $\mathbb{R}^2$ through

$$\mathcal{W}_E(\boldsymbol{s}) = \det(g)^{1/4} \mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}) = R(s)^{-1} \mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}).$$

Thus the equivalent SPDE in $\mathbb{R}^2$ can be written as

$$R(\boldsymbol{s})^{-2} \left[1 - R(\boldsymbol{s})^2 \Delta_{\mathbb{R}^2}\right] u(\boldsymbol{s}) = R(s)^{-1} \sqrt{4\pi} \mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2$$

where the first factor is needed because the volume elements of the spaces differ, $\mathrm{d}V_E = \sqrt{\det(g)}\mathrm{d}V_{\mathrm{R}^2}$. We use the SPDE

$$(R(s)^{-2} - \Delta_{\mathbb{R}^2}) u(\boldsymbol{s}) = \sqrt{4\pi} R(s)^{-1} \mathcal{W}_{\mathbb{R}^2}, \quad \boldsymbol{s} \in \mathbb{R}^2, \tag{11}$$

in Euclidean space, but can interpret the SPDE through the implied metric tensor. The SPDE is similar to setting $\kappa(\cdot) = R(\cdot)^{-1}$ in Equation (8), but has an extra factor on the right-hand side of the equation to reduce the variability of the marginal variances.

For example, the space $[0, 9] \times [0, 3]$ with the Euclidean distance metric can be visualized as a rectangle, which exists in $\mathbb{R}^2$, or as a half cylinder with radius $3/\pi$ and height 9, which exists in $\mathbb{R}^3$, but if the space is given the spatially varying metric tensor according to the local range function

$$R(s_1, s_2) = \begin{cases} 1 & 0 \le s_1 < 3, 0 \le s_2 \le \pi, \\ (s_1 - 2) & 3 \le s_1 < 6, 0 \le s_2 \le \pi, \\ 4 & 6 \le s_1 \le 9, 0 \le s_2 \le \pi, \end{cases} \tag{12}$$

the space cannot be embedded in $\mathbb{R}^2$. With this metric tensor, the space is no longer flat, but it can be embedded in $\mathbb{R}^3$ as the deformed cylinder shown in Figure 6. Thus, solving Equation (11) with the spatially varying coefficient is the same as solving Equation (9) on the deformed space.

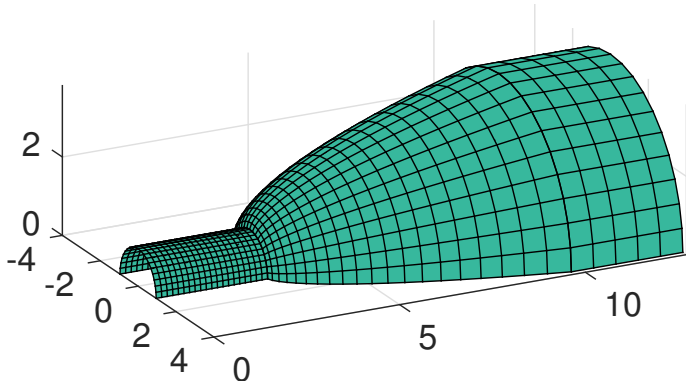


Figure 6: Half cylinder deformed according to the spatially varying metric tensor. The lines formed a regular grid on the half cylinder before deformation.

This means that unlike the deformation method, a spatially varying $R(\cdot)$ does not correspond to a deformation of $\mathbb{R}^2$ to $\mathbb{R}^2$, but rather from $\mathbb{R}^2$ to a higher-dimensional space.

However, the SPDE does not completely describe a deformation since solving the "stationary" SPDE on a curved space leads to changes in the marginal variances, but if $R(\cdot)$ does not vary too much, the marginal variances are close to 1. We can, therefore, introduce a separate function $S(\cdot)$ that controls the marginal variances of the process and limit the SPDE to a region of interest, $\mathcal{D}$, with Neumann boundary conditions,

$$(R(s)^{-2} - \Delta_{\mathbb{R}^2}) \left(\frac{u(\boldsymbol{s})}{\sqrt{S(\boldsymbol{s})}} \right) = \sqrt{4\pi} R(s)^{-1} \mathcal{W}_{\mathbb{R}^2}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^2.$$

This introduces boundary effects as was discussed in the paper by Lindgren et al. (2011), but we will not discuss the effects of the boundary in this paper.

This SPDE is different than the SPDE in Equation (8) beyond a re-parametrization, and allows for greater separation of the parameters that affect correlation structure and the parameters that affect marginal variances than the SPDE in Equation (8). This demonstrates that even though two SPDEs are similar and able to capture similar types of behaviour, one can be more useful for setting priors. The SPDE derived based on the metric tensor allows for separate priors for correlation structure and marginal variances through expansions of $\log(R(\cdot))$ and $\log(S(\cdot))$ into bases.

6.2 Setting priors on the non-stationarity

A stationary GRF described through a range ρ and a marginal variance σ^2 will constitute the base model when we work with non-stationarity. We want to shrink the non-stationary GRF towards the stationary GRF that has the PC prior developed in Section 2 for the range and the marginal variance. Denote the parameters that describe the departure from the base model by $\boldsymbol{\theta}$, where $\boldsymbol{\theta} = \mathbf{0}$ corresponds to the stationary GRF. Following the idea of the PC prior framework, we want to give the "distance" from stationarity a prior conditional on the current stationary model $\pi(\boldsymbol{\theta}|\rho, \sigma^2)$. The construction will be based on the ideas of the PC prior framework, but will not be based on a distance calculated from a formal measure of complexity, and the prior will be an *ad-hoc* prior that is motivated by theoretical principles.

We parametrize the local distance factor, $R(\cdot)$, and the approximate marginal variances, $S(\cdot)$, through

$$\begin{aligned} \log(R(\boldsymbol{s})) &= \log\left(\frac{\rho}{\sqrt{8}}\right) + \sum_{i=1}^{n_1} \theta_{1,i} f_{1,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \\ \log(S(\boldsymbol{s})) &= \log(\sigma^2) + \sum_{i=1}^{n_2} \theta_{2,i} f_{2,i}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathcal{D}, \end{aligned} \tag{13}$$

where $\{f_{1,i}\}$ is a set of basis functions for the local range centred such that $\langle f_{1,i}, 1\rangle_{\mathcal{D}} = 0$, for $i = 1, 2, \ldots, n_1$, and $\{f_{2,i}\}$ is a set of basis functions for the marginal variances centred such that $\langle f_{2,i}, 1\rangle_{\mathcal{D}} = 0$ for $i = 1, 2, \ldots, n_2$. We collect the parameters in vectors $\boldsymbol{\theta}_1 = (\theta_{1,1}, \ldots, \theta_{1,n_1})$ and $\boldsymbol{\theta}_2 = (\theta_{2,1}, \ldots, \theta_{2,n_2})$ such that $\boldsymbol{\theta}_1$ controls the local ranges and $\boldsymbol{\theta}_2$ controls the marginal variances.

A simple way to account for different scales and dependencies among the basis functions is to give the non-stationary effect in the correlation structure and the non-stationary effect in the marginal variances independent g-priors (Zellner, 1986) with $g = 1$,

$$\boldsymbol{\theta}_1 \sim \mathcal{N}(\mathbf{0}, \tau_1^{-1}\mathbf{S}_1^{-1}) \quad \text{and} \quad \boldsymbol{\theta}_2 \sim \mathcal{N}(\mathbf{0}, \tau_2^{-1}\mathbf{S}_2^{-1})$$

where S_1 is the Gramian,

$$S_{1,i,j} = \langle f_{1,i}, f_{1,j}\rangle_{\mathcal{D}}, \quad \text{for } i, j = 1, 2, \ldots, n_1,$$

and S_2 is the Gramian,

$$S_{2,i,j} = \langle f_{2,i}, f_{2,j}\rangle_{\mathcal{D}}, \quad \text{for } i, j = 1, 2, \ldots, n_2.$$

In this set-up the Gramians account for the structures of the basis functions and the strengths of the effects are reduced to two precisions parameters τ_1 and τ_2. We choose to give the precision parameters the PC prior for precision parameters for Gaussian distributions developed by Simpson et al. (2014), which is designed to shrink towards the base model of zero effect. Because of our *a priori* ansatz of *a priori* independence between the correlation structure and the marginal variances, we set independent priors

$$\pi(\tau_1) = \frac{\lambda_1}{2}\tau_1^{-3/2}\exp\left(-\lambda\tau_1^{-1/2}\right) \text{ and } \pi(\tau_2) = \frac{\lambda_2}{2}\tau_2^{-3/2}\exp\left(-\lambda_2\tau_2^{-1/2}\right).$$

In this way we have implicitly described the effect in the correlation structure and the effect in the marginal variances through an *ad-hoc* distance and shrunk the effects towards the base model, which is stationarity.

We calibrate the priors based on the *a priori* relative variations they allow for the local range and for the marginal variance through the *a priori* statements,

$$\text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{R(\boldsymbol{s})}{\rho/\sqrt{8}}\right)\right| > C_1\right) = \alpha_1,$$

$$\text{Prob}\left(\max_{\boldsymbol{s}\in\mathcal{D}}\left|\log\left(\frac{S(\boldsymbol{s})}{\sigma^2}\right)\right| > C_2\right) = \alpha_2.$$

These statements are only based on the relative differences in the local range and the marginal variance from a stationary model, and we can see from Equation (13) that the relative differences

do not depend on the parameters of the stationary model. This means that the prior on the non-stationary GRF separates as

$$\begin{aligned}\pi(\rho, \sigma^2, \boldsymbol{\theta}) &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}|\rho, \sigma^2) \\ &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \\ &= \pi(\rho)\pi(\sigma^2)\pi(\boldsymbol{\theta}_1)\pi(\boldsymbol{\theta}_2),\end{aligned}$$

where the first equality uses the conditional prior constructed, the second equality uses that the calibration does not introduce dependence with the parameters for the stationary part of the GRF, and the last equality uses that the priors on the spatially varying part of the local range and the marginal variance are independent.

The above conditions control the probability that the relative differences from the stationary model in the local range and the marginal variance exceed pre-specified levels, and allow the user to control the priors based on beliefs about the variability expected in the local range and the marginal variance. The calibration is slightly different than the one used by Ingebrigtsen et al. (2014b). The approach derived is fundamentally *ad-hoc*, but makes several theoretical improvements over the approach in Ingebrigtsen et al. (2014b) due to the new form of the SPDE that reduces the interaction between the correlation structure and the marginal variances in the prior specification, and due to the use of priors on the effects that follow the PC prior framework principle of shrinking towards the base model.

7 Discussion

In this paper we have presented an answer to an important, open question in Bayesian spatial statistics that previously had no satisfactory answer: which prior should we put on the range and the marginal variance for a Matérn GRF? The range and the marginal variance have seemingly clear interpretations and ideally one would hope that it were possible to infer them from a single observation of the spatial field, but in reality there are no consistent estimators of the range and the marginal variance under in-fill asymptotics and the posterior distributions do not contract even for a complete observation of the process in a bounded observation window. There is a ridge in the likelihood where the posterior distribution of the parameters always will be affected by the prior on the parameters.

For Matérn GRFs objectivity, through noninformative priors such as the reference priors, is not necessarily the correct answer because it can lead to posterior inference that is not sensible. For in-sample predictions, the near-intrinsic models do not negatively affect predictions and it is possible to use the reference priors if they are applicable for the model. However, if the purpose is to infer the range and the marginal variance, to do out-of-sample predictions, or to generate new scenarios from the model using the posterior distribution of the parameters as the prior distribution, the objective approach can lead to meaningless answers that are not compatible with subjective beliefs about the model. The three GRFs in Figure 1 are similar if they are used to predict the unobserved values within the interval $[0, 1]$, but they are highly different if we want to understand the process that generated the data or use this process to generate new data using the parameter posterior.

In practice, we are likely to have subjective knowledge making high marginal variances unreasonable even when the value of the range is high. With the PC-prior developed in this paper this knowledge can be combined with the geometry of the parameter space through two statements of prior belief about the spatial field. In this way it is possible to not just encode the information about geometry contained in the likelihood, but also prior belief about the range and the

marginal variance. For example, it is possible to encode prior belief that the marginal standard deviations are unlikely to exceed a specific upper limit. This information disallows the near-intrinsic models far along the ridge of the likelihood and lead to shorter and more meaningful credible intervals than an objective prior such as the reference prior.

The PC prior is weakly informative and like all subjective priors there is a danger that putting prior mass in the wrong place can negatively affect the inference. The calibration of the PC-prior will by design affect the posterior distribution, but since the prior is not just an *ad-hoc* choice, the hyperparameters of the prior have a clearly defined meaning through information that could potentially be elicited from experts. The study of frequentist properties showed that there were negative consequences of being one order wrong in the prior specification, but the examples only had 25 observations and in realistical settings there will be more information available than this, and with more observations, the sensitivity to the prior specification is likely to be less severe. Further, it is when the true range is long compared to the domain size that the likelihood is insufficient for getting physically meaningful estimates and that it is most important to limit the behaviour of the posterior through an interpretable prior.

The greatest benefits of the PC-prior over the reference priors are that there is no dependence on the sampling design, it works with any observation process and can be used in hierarchical models, it is easy to implement, and it is computationally cheap. The benefits over the *ad-hoc* priors is that it has theoretical justification, and that the hyperparameters are interpretable and connected to prior belief about the scale that the parameters are on. This makes the PC prior useful in practice as opposed to the impractical reference priors, while at the same time having the theoretically justification that is lacked by the practical, but *ad-hoc* priors.

The PC prior is extended to a prior for a non-stationary GRF based on an SPDE model using ideas from the PC prior framework, but the extension has *ad-hoc* elements and is specialized to a GRF that can de-couple the parameters controlling the correlation structure and the parameters controlling the marginal variances. It would be desirable to derive a distance from stationarity to non-stationarity by using the KLD in a similar way as for the stationary GRF and not be restricted to a specific non-stationary GRF, but this is difficult because certain properties of the covariance structure of GRFs are identifiable under in-fill asymptotics while others are not.

For SPDE models, a seemingly desirable way to be independent of the parametrization would be to restrict oneself to the approximation of the model used for computations. In computations, one uses a finite-dimensional approximation of the GRF derived through a finite element approximation on a triangulation of the domain. The multivariate Gaussian distribution of the values at the nodes in the triangulation completely describes the distribution of the approximation of the spatial field and we can compute the distances between non-stationary spatial fields by computing the KLDs between multivariate Gaussian distributions.

However, if we do this, a change in $\kappa(\cdot)$ can be handled, but a change in $\tau(\cdot)$ makes the KLD diverge to infinity as the mesh is refined. Thus, we can use the KLD to construct a prior for the parameters in $\kappa(\cdot)$, but there is still work left in understanding how the KLD behaves as a function of parameters in $\tau(\cdot)$ when the mesh is refined. If one can understand more about the identifiability of $\tau(\cdot)$ as the mesh is refined, one can use this knowledge to re-scale the KLD in a meaningful way, and separate out the constants and the asymptotic behaviour as a function of $\tau(\cdot)$, but this remains an unsolved challenge.

For stationary GRFs, the paper makes significant progress by finding sensible priors for Matérn GRFs through a practically useful, weakly informative joint prior on range and marginal variance. The important remaining question is shared with the other priors derived with the PC prior framework, namely, how easy is it for a user to set the hyperparameters? The hyperparameters have a clear connection to understandable quantities, but the users are still required to gain intuition about setting priors based on the probability of exceeding or being below a chosen

value. This is a question we expect to gain experience with when the new prior is implemented within the INLA package (Rue et al., 2009). For non-stationary GRFs the paper makes progress by providing a motivated, but *ad-hoc* construction of a prior. However, a construction fully based on the PC prior framework remains future work.

A The theoretical details for the derivation of the prior for κ

If we consider the distribution of a Matérn GRF, parametrized through $\kappa = \sqrt{8\nu}/\rho$ and

$$\tau = \frac{\Gamma(\nu)}{(4\pi)^{d/2}\Gamma(\nu + d/2)\sigma^2\kappa^{2\nu}},$$

where ρ is the range, σ^2 is the marginal variance, and ν is the smoothness, on a bounded observation window, the KLD between the distributions for two choices of parameters $\kappa, \kappa_0 > 0$ is always finite and it is possible to use the KLD to describe how different the distributions are. The distributions are not absolutely continuous with respect to a Lebesgue measure and we need to describe the KLD in terms of measures. The KLD of the probability measure Q from the probability measure P is defined by

$$D_{\mathrm{KL}}(P||Q) = \int_{\chi} \log\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) \mathrm{d}P, \tag{14}$$

where $\mathrm{d}P/\mathrm{d}Q$ is the Radon-Nikodym derivative of P with respect to Q, and expresses the information lost when Q is used to approximate P.

We base the constructions in this section on spectral densities and not directly on covariance functions. Fix τ and ν and consider the Matérn GRF u_κ for different values of κ. Lindgren et al. (2011) showed that this GRF can be expressed as a solution of the stochastic partial differential equation (SPDE)

$$(\kappa^2 - \Delta)^{\alpha/2}(\sqrt{\tau}u_\kappa(\boldsymbol{s})) = \mathcal{W}(\boldsymbol{s}), \quad \boldsymbol{s} \in \mathbb{R}^d, \tag{15}$$

where $\alpha = \nu + d/2$ and $\mathcal{W}$ is standard Gaussian white noise. The SPDE can be used to show that the spectral density of u_κ is given by

$$f_\kappa(\boldsymbol{w}) = \left(\frac{1}{2\pi}\right)^d \frac{1}{\tau(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}. \tag{16}$$

The continuous spectrum in Equation (16) is difficult to use directly and we make an intermediate step through a periodic approximation of the GRF. Restrict SPDE (15) to the domain $\mathcal{D} = [-L/2, L/2]^d$ and apply periodic boundary conditions. This leads to an approximation $\tilde{u}_\kappa$ of u_κ that can be written as

$$\tilde{u}_\kappa(\boldsymbol{s}) = \sum_{\boldsymbol{k}\in\mathbb{Z}^d} z_{\boldsymbol{k}} \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L, \boldsymbol{s}\rangle},$$

where $\{z_{\boldsymbol{k}}\}$ are independent Gaussian random variables with variances given by

$$\begin{aligned}
\lambda_{\boldsymbol{k}}(\kappa) &= \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} \frac{\mathrm{Var}[\langle \mathcal{W}, \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L, \boldsymbol{s}\rangle}\rangle_{\mathcal{D}}]}{\langle \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L, \boldsymbol{s}\rangle}, \mathrm{e}^{i\langle 2\pi\boldsymbol{k}/L, \boldsymbol{s}\rangle}\rangle_{\mathcal{D}}^2} \\
&= \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} \frac{L^d}{L^{2d}} \\
&= \frac{1}{L^d} \frac{1}{\tau(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}.
\end{aligned} \tag{17}$$

Using this approximation we calculate the KLD between $\tilde{u}_\kappa$ and $\tilde{u}_{\kappa_0}$ (based on Bogachev (1998, Thm. 6.4.6)),

$$\begin{aligned} 2\mathrm{KLD}(\kappa||\kappa_0) &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left[\frac{\lambda_{\boldsymbol{k}}(\kappa_0)}{\lambda_{\boldsymbol{k}}(\kappa)} - 1 - \log\frac{\lambda_{\boldsymbol{k}}(\kappa_0)}{\lambda_{\boldsymbol{k}}(\kappa)}\right] \\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left[\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} - 1 - \log\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}\right], \end{aligned} \tag{18}$$

which is a simple expression involving only the spectral densities of the processes. If we add scaling with step-size, it becomes a Riemann sum, and we can write

$$\begin{aligned} &2\left(\frac{2\pi}{L}\right)^d \mathrm{KLD}(\kappa||\kappa_0) \\ &= \sum_{\boldsymbol{k}\in\mathbb{Z}^d} \left(\frac{2\pi}{L}\right)^d \left[\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha} - 1 - \log\frac{(\kappa_0^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}{(\kappa^2 + ||2\pi\boldsymbol{k}/L||^2)^\alpha}\right] \\ &= \int_{\mathbb{R}^d} \left[\frac{f_\kappa(\boldsymbol{w})}{f_{\kappa_0}(\boldsymbol{w})} - 1 - \log\frac{f_\kappa(\boldsymbol{w})}{f_{\kappa_0}(\boldsymbol{w})}\right] \mathrm{d}\boldsymbol{w} + E(L,\kappa_0), \end{aligned}$$

where $E(L,\kappa_0)$ is the error in the Riemann sum.

Since we want the base model $\kappa_0 = 0$, which corresponds to infinite range, we need to be careful about how the error $E(L,\kappa_0)$ behaves as $\kappa_0 \to 0$. If L is fixed, the zero frequency gives an infinite term in the summand. Thus the rate at which L tends to infinity must be related to the rate at which κ_0 tends to zero. If the summand for $\boldsymbol{k} = 0$ tends to zero, the Riemann sum converges and $E(L,\kappa_0) \to 0$. The zero-frequency term

$$\left(\frac{2\pi}{L}\right)^d \left[\left(\frac{\kappa_0^2}{\kappa^2}\right)^\alpha - 1 - \alpha\log\frac{\kappa_0^2}{\kappa^2}\right],$$

converges to zero if $L = o(\kappa_0^{-1})$. We apply this relationship between L and κ_0 and introduce the scaled KLD

$$\begin{aligned} \tilde{\mathrm{KLD}}(\kappa||0) &= \lim_{\kappa_0\to 0} \left(\frac{2\pi}{L}\right)^d \mathrm{KLD}(\kappa||\kappa_0) \\ &= \frac{1}{2}\int_{\mathbb{R}^d} \left[\frac{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}{(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha} - 1 - \log\frac{(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}{(\kappa^2 + \boldsymbol{w}^{\mathrm{T}}\boldsymbol{w})^\alpha}\right] \mathrm{d}\boldsymbol{w}. \end{aligned}$$

We perform the change variables $\boldsymbol{w} = \kappa\boldsymbol{y}$ and find

$$\begin{aligned} \tilde{\mathrm{KLD}}(\kappa||0) &= \frac{1}{2}\int_{\mathbb{R}^d} \left[\frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}{(1 + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha} - 1 - \log\frac{(\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}{(1 + \boldsymbol{y}^{\mathrm{T}}\boldsymbol{y})^\alpha}\right] \kappa^d \mathrm{d}\boldsymbol{y} \\ &= \kappa^d \tilde{\mathrm{KLD}}(1||0) \\ &\propto \kappa^d, \end{aligned} \tag{19}$$

if $\tilde{\mathrm{KLD}}(1||0)$ exists.

However, $\tilde{\mathrm{KLD}}(1||0)$ does not exist for all dimensions d. Perform a change of coordinates to n-dimensional spherical coordinates to find

$$\tilde{\mathrm{KLD}}(1||0) = C_d \int_0^\infty \left[\left(\frac{r^2}{1+r^2}\right)^\alpha - 1 - \log\left(\frac{r^2}{1+r^2}\right)^\alpha\right] r^{d-1}\mathrm{d}r, \tag{20}$$

where C_d is a constant that varies with dimension. There are two issues: the behaviour for small r and the behaviour for large r. For $d = 1$,

$$\tilde{\mathrm{KLD}}(1||0) \leq -C_1\alpha \int_0^\infty \log \frac{r^2}{1+r^2} \mathrm{d}r = \pi\alpha C_1,$$

and we can conclude that the behaviour around 0 is not a problem for any $d \geq 1$. The behaviour for large r can be studied through an expansion in $(1+r^2)^{-1}$. The integrand in Equation (20) behaves as

$$\frac{\alpha^2}{2}\frac{1}{(1+r^2)^2} + \mathcal{O}\left(\frac{1}{(1+r^2)^3}\right).$$

This means that we can find an $0 < r_0 < \infty$ such that

$$\begin{aligned}\int_0^\infty \left[\left(\frac{r^2}{1+r^2}\right)^\alpha - 1 - \log\left(\frac{r^2}{1+r^2}\right)^\alpha\right] r^{d-1}\mathrm{d}r \\ \leq \text{Const} + \int_{r_0}^\infty \left[\frac{\alpha^2}{2}\frac{1}{(1+r^2)^2} + \frac{C}{(1+r^2)^3}\right] \mathrm{d}r,\end{aligned}$$

where $C \geq 0$ is a constant. For $d \leq 3$ both terms on the right hand side are finite and based on this and the boundedness for $d = 1$, we can conclude that $\tilde{\mathrm{KLD}}(1||0)$ is finite for $d \leq 3$.

B Calculation of the Kullback-Leibler divergence for a one-dimensional GRF with exponential covariance function

B.1 Goal

Let u_κ be a stationary GRF with the exponential covariance function,

$$c(d) = \frac{1}{2\kappa}\mathrm{e}^{-\kappa d}, \tag{21}$$

where $\kappa > 0$. This way of writing the exponential covariance function differs from the traditional parametrization using the range and the marginal variance, and is chosen because the KLD between the distributions described by different values $\kappa > 0$ is finite. The parametrization describes how to move in the parameter space while keeping the KLD finite. The goal of this appendix is to calculate the KLD between the distributions of u_κ and u_{κ_0} on the interval $[0, L]$

B.2 Discretization

The direct computations for the interval $[0, L]$ are difficult. So we first consider the KLD for the distributions of u_κ and u_{κ_0} at the observation points $t_i = i\Delta t$, for $i = 0, 1, \ldots, N$, where $\Delta t = L/N$. The spatial field u_κ can be described as a stationary solution of the stochastic differential equation

$$\mathrm{d}u_\kappa(t) = -\kappa u_\kappa(t)\mathrm{d}t + \mathrm{d}W(t),$$

where W is a standard Wiener processes, and written explicitly as

$$u_\kappa(t) = \int_{-\infty}^t \mathrm{e}^{-\kappa(t-s)}\,\mathrm{d}W(s).$$

This expression shows that

$$u_\kappa(t_{i+1})|u_\kappa(t_i) \sim \mathcal{N}(\mathrm{e}^{-\kappa\Delta t}u_\kappa(t_i), \sigma_\kappa^2),$$

where

$$\sigma_\kappa^2 = \mathrm{Var}[u_\kappa(t+\Delta t)|u_\kappa(t)] = \int_t^{t+\Delta t} \mathrm{e}^{-2\kappa(t+\Delta t-s)}\,\mathrm{d}s = \frac{1-\mathrm{e}^{-2\kappa\Delta t}}{2\kappa}.$$

This is an AR(1) process with initial condition $u_\kappa(t_0) \sim \mathcal{N}(0,(2\kappa)^{-1})$, which means that $\boldsymbol{u}_\kappa = (u_\kappa(t_0),\ldots,u_\kappa(t_N))$ has a multivariate Gaussian distribution with mean $\mathbf{0}$ and precision matrix

$$\mathbf{Q}_\kappa = \frac{1}{\sigma_\kappa^2}\begin{bmatrix} 1 & -\mathrm{e}^{-\kappa\Delta t} & & & \\ -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} & & \\ & \ddots & \ddots & \ddots & \\ & & -\mathrm{e}^{-\kappa\Delta t} & 1+\mathrm{e}^{-2\kappa\Delta t} & -\mathrm{e}^{-\kappa\Delta t} \\ & & & -\mathrm{e}^{-\kappa\Delta t} & 1 \end{bmatrix}. \tag{22}$$

B.3 Kullback-Leibler divergence

The vectors $\boldsymbol{u}_{\kappa_0}$ and $\boldsymbol{u}_\kappa$ have multivariate Gaussian distributions and the KLD from the distribution described by κ_0 to the distribution described by κ is

$$\mathrm{KLD}(\kappa||\kappa_0) = \frac{1}{2}\left[\mathrm{tr}(\mathbf{Q}_{\kappa_0}\mathbf{Q}_\kappa^{-1}) - (N+1) - \log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_\kappa|}\right)\right].$$

We are interested in taking the limit $\Delta t \to 0$ to find the value corresponding to the KLD from u_{κ_0} to u_κ. This is done in two steps: first we consider the trace and the $N+1$ term, and then we consider the log-determinant term.

B.3.1 Step 1

Let $f_\kappa = 1/\sigma_\kappa^2$, then the trace term can be written as

$$\begin{aligned}
&\mathrm{tr}(\mathbf{Q}_{\kappa_0}\Sigma_\kappa) \\
&\quad = f_{\kappa_0}\left[2c_\kappa(0) + \sum_{i=1}^{N-1}(1+\mathrm{e}^{-2\kappa_0\Delta t})c_\kappa(0) - 2\sum_{i=1}^{N}\mathrm{e}^{-\kappa_0\Delta t}c_\kappa(\Delta t)\right] \\
&\quad = f_{\kappa_0}\left[2c_\kappa(0) + (N-1)(1+\mathrm{e}^{-2\kappa_0\Delta t})c_\kappa(0) - 2N\mathrm{e}^{-\kappa_0\Delta t}c_\kappa(\Delta t)\right].
\end{aligned}$$

We extract the first summand and parts of the last summand, and combine with 2 from the $N+1$ term, to find the limit

$$\begin{aligned}
2f_{\kappa_0}[c_\kappa(0) - \mathrm{e}^{-\kappa_0\Delta t}c\kappa(\Delta t)] - 2 &= 2f_{\kappa_0}\frac{1-\mathrm{e}^{-(\kappa+\kappa_0)\Delta t}}{2\kappa} - 2 \\
&= \frac{\kappa+\kappa_0}{\kappa}\frac{f_{\kappa_0}/\Delta t}{f_{\kappa+\kappa_0}/\Delta t} - 2 \\
&\to \frac{\kappa_0-\kappa}{\kappa}.
\end{aligned}$$

For the remaining summands and the remaining $N-1$ from the $N+1$ term, we can simplify the expression as

$$\begin{aligned}
&S_3(\Delta t)\\
&\quad= (N-1)f_{\kappa_0}\left[\left(1+\mathrm{e}^{-2\kappa_0\Delta t}\right)c_\kappa(0) - 2\mathrm{e}^{-\kappa_0\Delta t}c_\kappa(\Delta t)\right] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\left[\left(1+\mathrm{e}^{-2\kappa_0\Delta t}\right)\frac{1}{2\kappa} - 2\frac{\mathrm{e}^{-(\kappa_0+\kappa)\Delta t}}{2\kappa}\right] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\frac{1}{2\kappa}\Bigg[1 + (1 - 2\kappa_0\Delta t + \frac{4\kappa_0^2(\Delta t)^2}{2})\\
&\qquad\quad - 2(1-(\kappa_0+\kappa)\Delta t + \frac{(\kappa_0+\kappa)^2(\Delta t)^2}{2}) + o((\Delta t)^2)\Bigg] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\frac{1}{2\kappa}\Big[(-2\kappa_0 + 2(\kappa_0+\kappa))\Delta t\\
&\qquad\quad + (2\kappa_0^2 - (\kappa_0+\kappa)^2)(\Delta t)^2 + o((\Delta t)^2)\Big] - (N-1)\\
&\quad= (N-1)f_{\kappa_0}\left[\Delta t + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa}(\Delta t)^2 + o((\Delta t)^2)\right] - (N-1)\\
&\quad= \left(\frac{L}{\Delta t} - 1\right)\left(\frac{1}{\Delta t} + \kappa_0 + o(1)\right)\Big[\Delta t\\
&\qquad\quad + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa}(\Delta t)^2 + o((\Delta t)^2)\Big] - \left(\frac{L}{\Delta t} - 1\right),
\end{aligned}$$

and see that the products involving $o(1)$ tend to zero

$$\begin{aligned}
S_3(\Delta t) &= L\left[\frac{1}{\Delta t} + \frac{2\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa} - \frac{1}{\Delta t}\right] + L\kappa_0 - [1+o(1)] + 1\\
&= L\frac{4\kappa_0^2 - (\kappa_0+\kappa)^2}{2\kappa} + L\kappa_0 + o(1)\\
&= L\left(\kappa_0 + \frac{\kappa_0^2}{2\kappa} - \kappa_0 - \frac{\kappa}{2}\right) + o(1).
\end{aligned}$$

Thus the limit is

$$\mathrm{tr}(\mathbf{Q}_{\kappa_0}\Sigma_\kappa) - (N+1) \to \frac{\kappa_0}{\kappa} - 1 + L\left(\frac{\kappa_0^2}{2\kappa} - \frac{\kappa}{2}\right).$$

B.3.2 Step 2

The determinant of the matrix in Equation (22) can be found by summing rows upwards, and we see that

$$|\mathrm{Q}| = \sigma^{-2(N+1)}(1-\mathrm{e}^{-2\kappa\Delta t}) = 2\kappa\sigma^{-2N}.$$

Note that in the limit $\kappa \to 0$, $f \to \Delta t$ so the determinant behaves asymptotically as κ. This means that

$$\begin{aligned}
\log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_\kappa|}\right) &= \log\left(\frac{2\kappa_0 f_{\kappa_0}^N}{2\kappa f_\kappa^N}\right)\\
&= \log\left(\frac{\kappa_0}{\kappa}\right) + N\log\left(\frac{f_{\kappa_0}}{f_\kappa}\right)
\end{aligned}$$

and we need to find the limit of the second part,

$$\begin{aligned}
&N \log\left(\frac{f_{\kappa_0}}{f_\kappa}\right) \\
&= \frac{L}{\Delta t}\left[\log\frac{1}{f_\kappa} - \log\frac{1}{f_{\kappa_0}}\right] \\
&= \frac{L}{\Delta t}\left[\log\left(\frac{1}{2\kappa}\left(1-\mathrm{e}^{-2\kappa\Delta t}\right)\right) - \log\left(\frac{1}{2\kappa_0}\left(1-\mathrm{e}^{-2\kappa_0\Delta t}\right)\right)\right] \\
&= \frac{L}{\Delta t}\left[\log\left(\Delta t - \kappa(\Delta t)^2 + o((\Delta t)^2)\right) - \log\left(\Delta t - \kappa_0(\Delta t)^2 + o((\Delta t)^2)\right)\right] \\
&= \frac{L}{\Delta t}\left[\log\left(1 - \kappa\Delta t + o(\Delta t)\right) - \log\left(1 - \kappa_0\Delta t + o(\Delta t)\right)\right] \\
&= \frac{L}{\Delta t}\left[-\kappa\Delta t + \kappa_0\Delta t + o(\Delta t)\right]
\end{aligned}$$

Thus the limit is

$$\log\left(\frac{|\mathbf{Q}_{\kappa_0}|}{|\mathbf{Q}_\kappa|}\right) \to \log\left(\frac{\kappa_0}{\kappa}\right) + L(\kappa_0 - \kappa)$$

B.4 Full KLD

The combination of the limits from the two steps gives the full KLD,

$$\begin{aligned}
\mathrm{KLD}(\kappa||\kappa_0) &= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 + L\left(\frac{\kappa_0^2}{2\kappa} - \frac{\kappa}{2}\right) - \log\left(\frac{\kappa_0}{\kappa}\right) - L(\kappa_0 - \kappa)\right] \\
&= \frac{1}{2}\left[\frac{\kappa_0}{\kappa} - 1 - \log\left(\frac{\kappa_0}{\kappa}\right) + L\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right)\right]. \qquad (23)
\end{aligned}$$

B.5 Comparison with the integral expression

The integral in Appendix A gives the expression

$$\frac{1}{2}\int_{-\infty}^{\infty}\left(\left(\frac{\kappa_0^2 + w^2}{\kappa^2 + w^2}\right) - 1 - \log\left(\frac{\kappa_0^2 + w^2}{\kappa + w^2}\right)\right)\mathrm{d}w = \pi\left(\frac{\kappa_0^2}{2\kappa} - \kappa_0 + \frac{\kappa}{2}\right).$$

If we divide by 2π, this is the same expression as the one that is multiplied with L in Equation (23). This is what we would expect because the integral is derived under the assumption that $L \gg 1/\kappa_0$ and "absorbs" the constant $2\pi/L$.

References

Berger, J. O., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374.

Bogachev, V. I. (1998). *Gaussian measures*. Number 62. American Mathematical Soc.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics*, 5(1):523–550.

Damian, D., Sampson, P. D., and Guttorp, P. (2001). Bayesian estimation of semi-parametric non-stationary spatial covariance structures. *Environmetrics*, 12(2):161–178.

Damian, D., Sampson, P. D., and Guttorp, P. (2003). Variance modeling for nonstationary spatial processes with temporal replications. *Journal of Geophysical Research: Atmospheres*, 108(D24).

Fuglstad, G.-A., Lindgren, F., Simpson, D., and Rue, H. (2015a). Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25:115–133.

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2015b). Does non-stationary spatial data always require non-stationary random fields? *arXiv preprint arXiv:1409.0743*.

Haas, T. C. (1990a). Kriging and automated variogram modeling within a moving window. *Atmospheric Environment. Part A. General Topics*, 24(7):1759 – 1769.

Haas, T. C. (1990b). Lognormal and moving window methods of estimating acid deposition. *Journal of the American Statistical Association*, 85(412):950–963.

Ingebrigtsen, R., Lindgren, F., and Steinsland, I. (2014a). Spatial models with explanatory variables in the dependence structure. *Spatial Statistics*, 8:20–38.

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2014b). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *arXiv preprint arXiv:1412.2798*.

Kazianka, H. (2013). Objective Bayesian analysis of geometrically anisotropic spatial data. *Journal of Agricultural, Biological, and Environmental Statistics*, 18(4):514–537.

Kazianka, H. and Pilz, J. (2012). Objective Bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, 40(2):304–327.

Lindgren, F., Rue, H., and Lindström, J. (2011). An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4):423–498.

Neto, J. H. V., Schmidt, A. M., and Guttorp, P. (2014). Accounting for spatially varying directional effects in spatial covariance structures. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(1):103–122.

Oliveira, V. d. (2007). Objective Bayesian analysis of spatial data with measurement error. *The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 35(2):pp. 283–301.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Palacios, M. B. and Steel, M. F. J. (2006). Non-gaussian Bayesian geostatistical modeling. *Journal of the American Statistical Association*, 101(474):604–618.

Paulo, R. (2005). Default priors for Gaussian processes. *The Annals of Statistics*, 33(2):556–582.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Sampson, P. D. and Guttorp, P. (1992). Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119.

Schmidt, A. M., Guttorp, P., and O'Hagan, A. (2011). Considering covariates in the covariance structure of spatial processes. *Environmetrics*, 22(4):487–500.

Schmidt, A. M. and O'Hagan, A. (2003). Bayesian inference for non-stationary spatial covariance structure via spatial deformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(3):743–758.

Simpson, D. P., Martins, T. G., Riebler, A., Rue, H., Fuglstad, G.-A., and Sørbye, S. H. (2014). Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:1403.4630*.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

van der Vaart, A. W. and van Zanten, J. H. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse Gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675.

Warnes, J. and Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74(3):640–642.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280 – 296.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

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