Constructing Priors that Penalize the Complexity of Gaussian Random Fields

Geir-Arne Fuglstad[1], Daniel Simpson[2], Finn Lindgren[3], and Håvard Rue[4]

[1]Department of Mathematical Sciences, NTNU, Norway
[2]Department of Statistical Sciences, University of Toronto, Canada
[3]School of Mathematics, University of Edinburgh, United Kingdom
[4]CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia

December 6, 2017

Abstract

Priors are important for achieving proper posteriors with physically meaningful covariance structures for Gaussian random fields (GRFs) since the likelihood typically only provides limited information about the covariance structure under in-fill asymptotics. We extend the recent Penalised Complexity prior framework and develop a principled joint prior for the range and the marginal variance of one-dimensional, two-dimensional and three-dimensional Matérn GRFs with fixed smoothness. The prior is weakly informative and penalises complexity by shrinking the range towards infinity and the marginal variance towards zero. We propose guidelines for selecting the hyperparameters, and a simulation study shows that the new prior provides a principled alternative to reference priors that can leverage prior knowledge to achieve shorter credible intervals while maintaining good coverage.

We extend the prior to a non-stationary GRF parametrized through local ranges and marginal standard deviations, and introduce a scheme for selecting the hyperparameters based on the coverage of the parameters when fitting simulated stationary

data. The approach is applied to a dataset of annual precipitation in southern Norway and the scheme for selecting the hyperparameters leads to concervative estimates of non-stationarity and improved predictive performance over the stationary model.

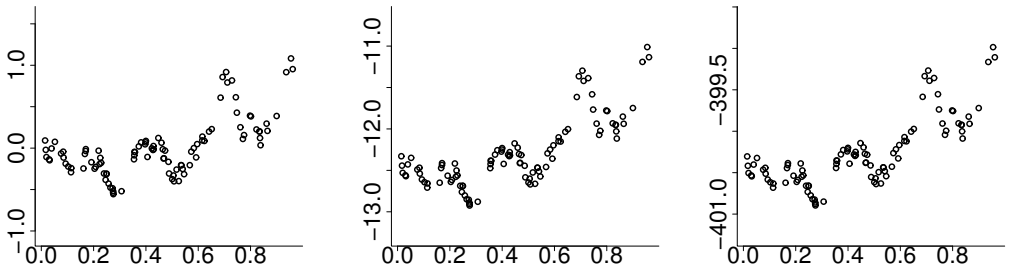
Keywords: Bayesian, Penalised Complexity, Priors, Spatial models, Range, Non-stationary

1 Introduction

Gaussian random fields (GRFs) provide a simple and powerful tool for introducing spatial or temporal dependence in Bayesian hierarchical models and are fundamental building blocks in spatial statistics and non-parametric modelling, but even for stationary GRFs controlled only by range and marginal variance, the choice of prior distribution 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 next to catastrophic. The main focus in this paper is one-dimensional, two-dimensional and three-dimensional GRFs with Matérn covariance functions with fixed smoothness, but we also discuss how to extend the prior to non-stationary covariance structures.

The Matérn covariance function leads a ridge in the likelihood for the range and the marginal variance (Warnes and Ripley, 1987), and there is no consistent estimator under in-fill asymptotics for these parameters when the base space of the GRF is of dimension three or lower (Stein, 1999; Zhang, 2004). For these GRFs only a limited amount of information can be learned about the parameters from a bounded domain and the prior affects the behaviour of the posterior of the parameters even under in-fill asymptotics. For example, for a one-dimensional GRF with an exponential covariance function observed on the interval $[0, 1]$, it is only the ratio of the range and the marginal variance that can be estimated consistently, and not the range or the marginal variance separately (Ying, 1991).

This ratio also determines the asymptotic properties of predictions under in-fill asymptotics with the exponential covariance function (Stein, 1999), but predictive distributions are not the only target for inference. Figure 1 shows that moves along the ridge in the likelihood when using the exponential covariance function, changes the level of the simulated observations, but that the pattern of the values around the level remains stable. These



(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.

choices of parameters lead to similar predictive distributions conditional on the observed data, but simulating unconditionally from GRFs with these parameters lead to highly different realizations. In a real application where the values in Figure 1a were observed, the practitioner will likely know that the ranges and marginal variances that generate Figures 1b and 1c are not physically meaningful even if the spreads of values are consistent with the observed pattern. Therefore, we believe the practitioner should be provided with a principled prior that allows him/her to include expert knowledge, in an interpretable way, about the range of parameters that are physically meaningful.

But 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. Their work has been extended by several authors (Paulo, 2005; Kazianka and Pilz, 2012; Kazianka, 2013) and, critically, Oliveira (2007) allowed for Gaussian observation noise. 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.

However, reference priors aim to be objective and are built on the fundamental principle of being the least informative priors, in an information-theoretic sense, for Bayesian inference (Berger et al., 2009), and GRFs are often embedded in Bayesian hierarchical models that are too complex for deriving the reference priors. Therefore, we propose a different construction that leads to a weakly informative prior that can leverage prior knowledge and is appropriate for hierarchical models where models components are combined linearly in the latent part of the model. In this setting, the model construction tends to be modular and priors should be constructed separately for each model component. The GRFs are used to achieve the desired second-order structure while the first-order structure of the model is handled by separate model components, and we must construct a joint prior for the range and marginal variance of a zero-mean Matérn GRF.

This setting is similar to structured additive regression models where Klein and Kneib (2016) has shown that the Penalised Complexity (PC) prior framework developed by Simpson et al. (2017) behaves well when used for the components of the models. This motivates the desire to use the PC prior framework to construct a joint prior for the range and the marginal variance of a Matérn GRF, but there are three questions that must be answered. Is the PC prior framework suitable for infinite-dimensional model components? How can we deal with the fact that the KLD between Matérn GRFs in general is infinite? And how can we construct a multivariate PC prior that properly accounts for the intrinsic link between range and marginal variance due to the ridge in the likelihood?

In this paper we extend Simpson et al. (2017) by answering these questions, and we show that the principles of the PC prior framework can be applied to construct a prior

for Matérn GRFs that is independent of the observation process. This is technically more demanding than the direct approach, which would be to construct the PC prior based on the finite-dimensional observation process, but the rewards are a prior that can be applied for any spatial design and any observation process, and is computationally inexpensive and has a much simpler form than the reference priors for GRFs in published literature. The resulting prior is weakly informative and shrinks towards a *base model* with infinite range and zero marginal variance through hyperparameters that indicate how strongly the user wishes to shrink towards the base model.

The stationary Matérn GRF can be extended to a non-stationary GRF by adding extra flexibility in the covariance function, but since the covariance structure of a GRF is only observed indirectly, the estimated covariance structure can be highly sensitive to the type of flexibility allowed and the prior used on the flexibility. We show that the PC prior developed for the stationary Matérn GRF can be extended further to a prior for a non-stationary GRF where the non-stationarity is controlled by covariates. The prior is motivated by g-priors and shrinkage properties, and we consider one scheme for selecting the hyperparameters that reduces the risk of overfitting the non-stationary GRF.

The joint PC prior for the range and the marginal variance of a Matérn GRF with a fixed smoothness parameter is derived in Section 2. Then in Section 3 a small simulation study is performed to evaluate the frequentist properties of the credible intervals and the behaviour of the joint posterior, and we demonstrate that the prior is applicable also for logistic spatial regression where the observation process is highly non-Gaussian. In Section 4 we discuss how to extend the PC prior to a conservative prior for a non-stationarity model for annual precipitation in southern Norway. The paper ends with discussion and conclusions in Section 5. The Supplementary Material contains proofs of the theorems, computer code, technical details and further discussion of many of the topics addressed,

and there are multiple references to it throughout the paper.

2 Penalised Complexity prior

2.1 Framework

Including a GRF in a model may lead to overfitting by, for example, estimating spurious spatial trends or spurious temporal trends. Simpson et al. (2017) suggest to handle the issue of overfitting by viewing model components, such as GRFs, as flexible extensions of simpler, less flexible *base models* and then developing priors that shrink the components towards their base models. For example, they view a random effect with non-zero variance as an extension of a random effect with zero variance, and construct a prior that shrinks the variance of the random effect towards zero.

The first step of their approach is to derive a distance from the base model to its flexible extension using the Kullback-Leibler divergence (KLD). The purpose of the distance is to provide a better parametrization of the model component where the size of the change in the parameter corresponds to the size of the change in the difference between the model component and its base model. In the setting of this paper, this can be done by describing the base model for the GRF by the Gaussian measure P_0 and the flexible model by the Gaussian measure P, and then defining the distance by $\text{dist}(P||P_0) = \sqrt{2\text{KL}(P||P_0)}$, where $\text{KL}(P||P_0)$ is the KLD from P_0 to P and is defined as follows.

Definition 2.1 (Kullback-Leibler divergence)**.** Let P_0 and P be measures over the set χ, where P is absolutely continuous with respect to P_0, then the Kullback-Leibler divergence from P_0 to P is defined as

$$\text{KL}(P||P_0) = \int_\chi \log \frac{\text{d}P}{\text{d}P_0} \text{d}P,$$

where $\mathrm{d}P/\mathrm{d}P_0$ is the Radon-Nikodym derivative of P with respect to P_0.

The KLD is used by Simpson et al. (2017) and has the benefits that it has an information-theoretical interpretation as the information lost when using the base model P_0 to approximate P and that it is an asymmetric distance from the "preferred" base model to the flexible extension. The square root is used in the definition of the distance to bring the distance to the correct scale (Simpson et al., 2017).

The second step of the prior construction is to define the prior on the derived distance using three principles: Occam's razor, constant-rate penalisation and user-defined scaling. Occam's razor means that the prior penalises more and more strongly the further one is from the base model and can be achieved by using constant-rate penalisation, where the prior on the distance, t, satisfies

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

for a constant decay-rate $0 < r < 1$. The only continuous distribution with this property is the exponential distribution $\pi(t) = \lambda \exp(-\lambda t)$, for $t > 0$, where the relative change in the prior when the distance increases by δ does not depend on the current distance t. The justification for using a simple prior on distance is that the parametrization corresponds directly to the size of the changes in the distribution of the model component.

The prior has a hyperparameter λ that must be set by the user and the principle of user-defined scaling is used to provide an interpretable way to set its value. The distance itself is typically not directly interpretable by the user and must be transformed to an interpretable size $Q(t)$. The prior information can then be included through, for example, tail probabilities $P(Q(d) > U) = \alpha$ or $P(Q(d) < L) = \alpha$, where U or L is an upper or lower limit, respectively, and α is the upper or lower tail probability of the prior distribution. Through this construction the PC prior combines the geometry of the parameter space

with prior belief about an interpretable size.

2.2 Derivation

The Matérn covariance function has been studied extensively (Stein, 1999), and it is isotropic and can be defined as a function of the distance between locations.

Definition 2.2 (Matérn covariance function)**.** A Matérn covariance function $c : [0, \infty) \to \mathbb{R}$ can be parametrized through a marginal standard deviation σ, a range parameter ρ, and a smoothness parameter ν, and is given by

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

where K_ν is the modified Bessel function of the second kind, order ν.

The choice of $\sqrt{8\nu}$ in the definition follows Lindgren et al. (2011) and makes ρ the distance at which the correlation is approximately 0.1. This parametrization of the Matérn covariance function has convenient interpretations for the parameters, but the parametrization is not convenient for deriving a PC prior. Therefore, we introduce an alternative parametrization.

Definition 2.3 (Alternative parametrization of the Matérn covariance function)**.** Assume that the base space is $\mathbb{R}^d$ and introduce

$$\kappa = \sqrt{8\nu}/\rho \quad \text{and} \quad \tau = \sigma \kappa^\nu \sqrt{\frac{\Gamma(\nu + d/2)(4\pi)^{d/2}}{\Gamma(\nu)}}. \tag{1}$$

This parametrization has the benefit that it describes what can, τ, and what cannot, κ, be consistently estimated under in-fill asymptotics when the dimension of the base space $d \leq 3$. When κ is changed, but τ is fixed, the resulting Gaussian measures are

equivalent and the KLD between the GRFs is finite, but if τ is changed, the resulting Gaussian measures are singular and the KLD between the GRFs is infinite (Zhang, 2004). By assumption ν is fixed, and the joint prior is derived in two steps: first $\pi(\tau|\kappa)$ and then $\pi(\kappa)$. The parameter τ can be consistently estimated under in-fill asymptotics, so the derivation of the PC prior for $\tau|\kappa$ must be based on a finite-dimensional observation (but will not depend on the spatial design).

Theorem 2.1 (PC prior for $\tau|\kappa$)**.** *Let u be a GRF defined on $\mathcal{D} \subset \mathbb{R}^d$ with a Matérn covariance function with parameters τ, κ and ν. If the GRF is observed on $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n \in \mathcal{D}$, then conditionally on κ the PC prior for τ with base model $\tau = 0$ is*

$$\pi(\tau|\kappa) = \lambda \exp(-\lambda\tau), \quad \tau > 0,$$

where $\lambda > 0$ is a hyperparameter.

Proof. See Section S1.1 in the online supplementary material. □

Since the prior shrinks towards zero variance conditionally on κ, we suggest to select the hyperparameter λ by limiting the upper tail probability α that the marginal standard deviation of the GRF will exceed σ_0. That is by selecting σ_0 and α such that $\mathrm{P}(\sigma > \sigma_0|\kappa) = \alpha$, where σ is the marginal standard deviation corresponding to τ and κ. Alternatively, one can set the hyperparameter by selecting the tail probability that the GRF at an arbitrary location exceeds a chosen value, but this does not lead to a simple analytic expression.

Theorem 2.2. *The PC prior for $\tau|\kappa$ satisfies $\mathrm{P}(\sigma > \sigma_0|\kappa) = \alpha$ if*

$$\lambda(\kappa) = -\kappa^{-\nu}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}}}\frac{\log(\alpha)}{\sigma_0}.$$

Proof. See Section S1.2 in the online supplementary material. □

The PC prior for κ can also be based on the finite-dimensional distribution corresponding to the observation locations, but this would lead to a computationally expensive prior because calculating KLDs between Gaussian distributions with dense covariance matrices has a cubic complexity in the number of observation locations. We seek to overcome this challenge by constructing the PC prior for κ using the infinite-dimensional GRF instead of the finite-dimensional observations. This is possible because changes in κ result in finite values for the KLD for the infinite-dimensional GRF when τ is fixed. In the proofs it is assumed that the GRF itself exists on an arbitrarily large ambient domain. In the next section we discuss how the prior could be derived under the assumption that the GRF only exists on the area from which the observations were made.

Theorem 2.3 (PC prior for κ)**.** *Let u be a GRF defined on $\mathbb{R}^d$, where $d \leq 3$, with a Matérn covariance function with parameters τ, κ and ν. The PC prior for κ with base model $\kappa = 0$ is*

$$\pi(\kappa) = \frac{d}{2}\lambda\kappa^{d/2-1}\exp\left(-\lambda\kappa^{d/2}\right), \quad \kappa > 0,$$

where $\lambda > 0$ is a hyperparameter.

Proof. See Section S1.3 in the online supplementary material. □

The prior in Theorem 2.3 is a Weibull distribution with shape parameter $d/2$ and scale parameter $\lambda^{-d/2}$, and is a heavy-tailed distribution. Since the prior shrinks the range towards infinity ($\kappa = 0$), we suggest to set the hyperparameter by controlling the tail probability that the range is below a certain limit.

Theorem 2.4. *The prior for κ satisfies $\mathrm{P}(\rho < \rho_0) = \alpha$ if*

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

Proof. See Section S1.4 in the online supplementary material. □

Combining the priors for $\tau|\kappa$ and κ provides the main results of this paper, which are the joint PC prior for (κ, τ) and the joint PC prior for (ρ, σ).

Theorem 2.5 (PC prior for the Matérn $(\kappa,\ \tau)$)**.** *Let u be a GRF defined on $\mathbb{R}^d$, where $d \leq 3$, with a Matérn covariance function with parameters τ, κ and ν. The joint PC prior based on the base models $\tau = 0$ and $\kappa = 0$ is*

$$\pi(\kappa, \tau) = \frac{d}{2}\lambda_1\lambda_2(\kappa)\kappa^{d/2-1}\exp(-\lambda_1\kappa^{d/2} - \lambda_2(\kappa)\tau), \quad \kappa > 0, \tau > 0,$$

where $\mathrm{P}(\rho < \rho_0) = \alpha_1$ and $\mathrm{P}(\sigma > \sigma_0|\kappa) = \alpha_2$ are achieved by

$$\lambda_1 = -\left(\frac{\rho_0}{\sqrt{8\nu}}\right)^{d/2}\log(\alpha_1) \quad \textit{and} \quad \lambda_2(\kappa) = -\kappa^{-\nu}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu + d/2)(4\pi)^{d/2}}}\frac{\log(\alpha_2)}{\sigma_0}.$$

Proof. See Section S1.5 in the online supplementary material. □

Theorem 2.6 (PC prior for the Matérn $(\rho,\ \sigma)$)**.** *Let u be a GRF defined on $\mathbb{R}^d$, where $d \leq 3$, with a Matérn covariance function with parameters σ, ρ and ν. Then the joint PC prior corresponding to a base model with infinite range and zero variance is*

$$\pi(\sigma, \rho) = \frac{d}{2}\tilde{\lambda}_1\tilde{\lambda}_2\rho^{-d/2-1}\exp(-\tilde{\lambda}_1\rho^{-d/2} - \tilde{\lambda}_2\sigma), \quad \sigma > 0, \rho > 0,$$

where $\mathrm{P}(\rho < \rho_0) = \alpha_1$ and $\mathrm{P}(\sigma > \sigma_0) = \alpha_2$ are achieved by

$$\tilde{\lambda}_1 = -\log(\alpha_1)\rho_0^{d/2} \quad \textit{and} \quad \tilde{\lambda}_2 = -\frac{\log(\alpha_2)}{\sigma_0}.$$

Proof. See Section S1.6 in the online supplementary material. □

The prior is easy and fast to compute regardless of the number of observations and $d = 2$ provides the two-dimensional spatial case that is used in Sections 3 and 4.

2.3 Restrictions and extensions

The results derived in the previous section do not hold when $d > 4$ since in this case both the range and the marginal variance are consistently estimable under in-fill asymptotics and it is not possible to make moves in the parameter space for which the KLD is finite. It is unknown whether the results hold for $d = 4$ since it is an open question whether the parameters can be consistently estimated for that case (Anderes, 2010). This means that the assumption on the dimension, $d \leq 3$, used to derive the joint prior is important and cannot be removed.

Most of the technical difficulties in the previous section is caused by the desire to work with continuously indexed GRFs instead of discretely indexed observation processes. The benefit is that the prior is not dependent on the spatial design, which is a good property because the GRF also exists on other locations than on those it was observed. In particular, the prior does not need to be changed if data is made available at new observation locations and the prior is meaningful when predictions are made at a higher resolution than the observed data or for a larger observation area. In the former case there is more difference between small ranges than a prior based on the observed locations would indicate and in the latter case there is a larger difference between large ranges than a prior based on the observed locations would indicate.

Similarly, if the GRF were assumed to exists only on the area on which the observations were made, the upper tail behaviour of the prior for the range would be wrong if the posterior is used to make predictions on a larger domain. A longer discussion is provided in Section S2 of the Supplementary Material, but the short story is: when the range changes, the properties of the GRF change even if those changes cannot be detected on the arbitrary observation locations or observation domain, and the construction of the prior should account for these changes.

In most applications the covariance function is chosen from the Matérn family of covariance functions and a prior applicable only for this family is of great interest. However, the approach in the paper could be extended to other isotropic families of covariance functions that are defined through a marginal variance and a spatial scale. If the spatial scale is consistently estimable, the techniques in the paper are not applicable. If the spatial scale is not consistently estimable, the main challenge is to know which combination of the parameters that is consistently estimable. When this information is known, one can let κ be the spatial scale and let τ be the consistently estimable parameter, and one can likely use a similar proof as in this paper. However, it is, in general, not known which parameters are consistently estimable for different families of covariance functions and it is outside the scope of this paper to go investigate further.

3 Simulation study

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. A prior intended for use as a default prior should lead to good frequentist properties such as frequentist coverage of the equal-tailed $100(1-\alpha)\%$ Bayesian credible intervals that is close to the nominal $100(1-\alpha)\%$. In this paper, the study is replicated with one key difference: no covariates are included. This choice is made because the PC prior is derived for a zero-mean GRF, and if a mean were desired, it would be handled by extending the hierarchical model with another latent component that had its own, separate prior. Without covariates the reference prior approach results in the Jeffreys' rule prior as there are no nuisance parameters to integrate out when constructing the spatial reference prior. Furthermore, we compute the $100(1-\alpha)\%$ highest posterior density (HPD) intervals (Chen and Shao,

1999) to investigate whether skewness of the posteriors result in substantially different conclusions for HPD credible intervals compared to quantile-based credible intervals.

We start by selecting 25 locations, $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{25}$, at random in $[0,1]^2$ and generate realizations, $\boldsymbol{u} = (u(\boldsymbol{s}_1), u(\boldsymbol{s_2}), \ldots, u(\boldsymbol{s}_{25}))$, using a GRF with an exponential covariance function $c(r) = \exp(-2r/R_0)$ for true ranges $R_0 = 0.1$ and $R_0 = 1$. The data is then fitted using a GRF with the exponential covariance function $c(r) = \sigma^2 \exp(-2r/\rho)$, where the unknown parameters are marginal variance σ^2 and range ρ. Four priors are considered: the PC prior (PriorPC), the Jeffreys' rule prior (PriorJe), and the Jeffreys prior for variance combined with a bounded uniform prior on range (PriorUn1) and a bounded uniform prior on the logarithm of range (PriorUn2). The most important and interesting results are presented in this section, while the full details of the simulation study are provided in Section S4 of the Supplementary Material.

We begin with a general discussion on the differences in results observed between quantile-based credible intervals and HPD credible intervals, and then proceed with discussion about specific results. In general, the marginal posteriors are highly skew and the HPD credible intervals are substantially shorter than the equal-tailed credible intervals, but comparisions of average lengths remain consistent between the two approaches because the relative differences are similar. Further, the coverage was further away from the nominal level for the HPD credible intervals than the quantile-based credible intervals for PriorJe and PriorPC, and the coverage of the credible intervals was more sensitive to the hyperparameters of PriorPC for HPD credible intervals than for quantile-based credible intervals. The coverage of the HPD credible intervals was closer to the nominal level than the quantile-based credible intervals for PriorUn1 and PriorUn2, but since our main focus are PriorPC and PriorJe we use the equal-tailed 95% credible intervals in what follows.

First, one observation with true range equal to 1.0 is selected and the model is fitted

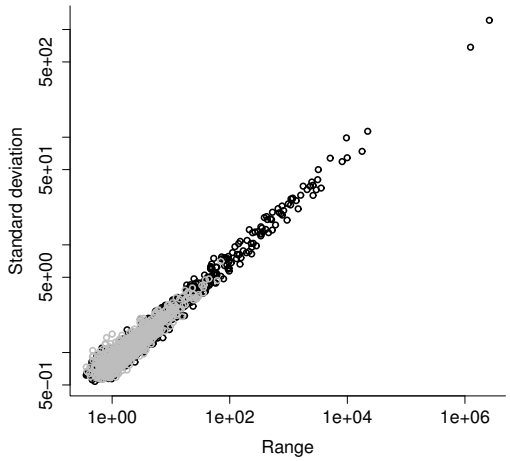


Figure 2: Samples from the joint posterior of range and marginal standard deviation. The grey circles are samples using the PC-prior and the black circles are samples using the Jeffreys' rule prior.

with PriorJe, and with PriorPC with hyperparameters selected such that $\mathrm{P}(\rho < 0.1) = 0.05$ and $\mathrm{P}(\sigma > 10) = 0.05$. The latter corresponds to a probability of 0.025 that the value of the GRF at an arbitrary location will exceed 10. The resulting samples from the posterior are shown in Figure 2 and the figure shows that when PriorJe is used, the MCMC sampler explores areas far out in the tail, whereas when PriorPC is used, the prior restricts the movement away from the upper tail. This means that when prior knowledge is available, PriorPC can be used to achieve credible intervals that are more reasonable.

Second, we study the sensitivity of the coverage and the lengths of the credible intervals to the choice of the hyperparameters in PriorPC and look for general guidelines for selecting the hyperparameters. We choose to set the hyperparameters in PriorPC through $\mathrm{P}(\rho < \rho_0) = 0.05$ and $\mathrm{P}(\sigma > \sigma_0) = 0.05$. The results show that choosing σ_0 lower than the true

standard deviation or ρ_0 higher than the true range results in too low coverage for both the marginal variance and the range. Selecting σ_0 to be 2.5, 10 or 40 times the true standard deviation and ρ_0 to be 1/10 or 1/2.5 times the true range results in good coverage both for the marginal variance and the range for both values of the true range. Selecting ρ_0 to be 1/40 times the true range degrades the coverage for the range when the true range is 0.1, but leads to good coverage when the true range is 1.0, while the coverage for the marginal variance is good for both values of the true range. Thus the study indicates that good coverage properties are achieved when σ_0 is selected between 2.5 to 40 times the true standard deviation and ρ_0 is set to between 1/10 and 1/2.5 times the true range. Further, shorter credible intervals are achieved for smaller values of σ_0 and smaller values of ρ_0, and for the values tested the best balance between good coverage and shortest lengths of the credible intervals is achieved for σ_0 equal to 2.5 times the true standard deviation and ρ_0 equal to 1/10 times the true range.

Third, we compare the properties when using PriorPC, PriorJe, PriorUn1 and PriorUn2. PriorJe results in 98.3% coverage with average length of the credible intervals of 0.78 for range and 96.7% coverage and average length of the credible intervals of 2.6 for marginal variance for true range $R_0 = 0.1$, and 95.6% coverage with average length of the credible intervals of 376 for range and 95.6% coverage with average length of the credible intervals of 295 for variance for $R_0 = 1$. The lengths of the credible intervals are shorter when using PriorPC with the hyperparameters suggested in the previous paragraph than when using PriorJe. The average lengths of the credible intervals are around 1.4 and 3.1 for marginal variance for true ranges 0.1 and 1.0, respectively. Note that the use of HPD intervals significantly reduces the average length of the credible intervals for range and marginal variance for PriorJe to 95 and 75, respectively, but they are still long and there are no hyperparameters that can be used to reduce them. For PriorUn1 the coverage and average

lengths of the credible intervals are sensitive to the upper limit on range, and for PriorUn2 the coverage is good and has little sensitivity to the lower and upper limit on range, while the average lengths of the credible intervals are sensitive to the upper limit.

Fourth, we investigate whether the behaviour found for PriorPC changes when the observation process is changed to a less informative observation process. For each realization with true range $R_0 = 0.1$ probabilities are calculated through a probit link, $\text{probit}(p_i) = u(\boldsymbol{s}_i)$, and binomial data is simulated using $y_i|p_i \sim \text{Binomial}(20, p_i)$. The data is then fitted using the true logistic spatial regression model and the coverage and average lengths of the credible intervals are estimated for marginal variance and range. The results show that the properties found using direct observations of the spatial field also holds for the spatial logistic regression, and the only significant difference is that the average lengths of the credible intervals are larger.

Overall, the simulation study shows that with respect to computation time and ease of use versus coverage and lengths of the credible 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. Furthermore, choosing the optimal values for σ_0 and ρ_0 or missing the optimal values by less than one order provides good coverage and lengths of the credible intervals.

4 Example: Extending to non-stationarity

Neither stationary nor non-stationary GRFs provide true representations of reality, but the extra flexibility in the covariance structure of a non-stationary GRF may provide a better fit to the data than a stationary GRF. Therefore, we consider how to extend the prior for

the stationary model to a prior for a non-stationary model, with the goal of improving predictions, using a dataset of annual precipitation. The details are technical and can be found in Section S7 of the Supplementary Material, but this section provides a condensed version.

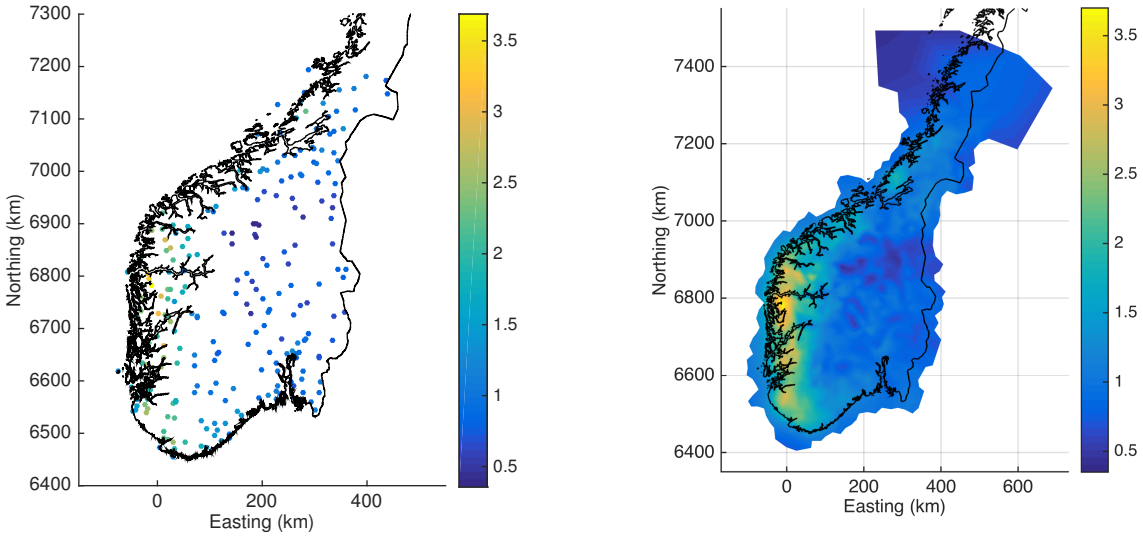
We use a dataset consisting of total annual precipitation for the one year period September 1, 2008, to August 31, 2009, for the 233 measurement stations in southern Norway shown in Figure 3. The dataset has previously been used by Ingebrigtsen et al. (2014, 2015) to study the use of elevation as a covariate in the covariance structure and associated priors. They used an intercept and a linear effect of the elevations of the stations in the first-order structure and used the elevation as a covariate in the second-order structure. We will follow their choice of covariates in the first-order structure, but use two covariates in the second-order structure: elevation and the magnitude of the gradient of the elevation.

We use the simple geostatistical model

$$y_i = \beta_0 + x_i\beta_1 + u(\boldsymbol{s}_i) + \epsilon_i, \qquad i = 1, 2, \ldots, 233, \tag{2}$$

where for station i, y_i is the observation made at location $\boldsymbol{s}_i$, x_i is the elevation of the station, (β_0, β_1) are the coefficients of the fixed effects, $u(\cdot)$ is the spatial effect, and ϵ_i is the nugget effect. The nuggets are i.i.d. $\epsilon_i \sim \mathcal{N}(0, \sigma_{\mathrm{N}}^2)$, and the spatial effect is constructed with the SPDE approach of Lindgren et al. (2011) and the stationary version has two parameters: spatial range ρ and marginal variance of the spatial field σ^2. The non-stationary version is constructed as shown in the Supplementary Material and uses the two covariates shown in Figure 4 in the second-order structure. The spatial field is orthogonalized against the intercept and the two covariates in the second-order structure to avoid confounding between the first-order structure and the second-order structure.

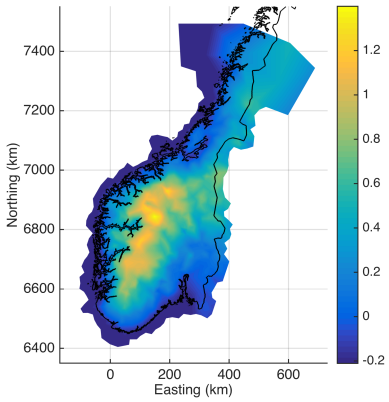
The stationary model uses the PC prior developed in this paper for the spatial field and the PC prior for precision parameter from Simpson et al. (2017) for the precision



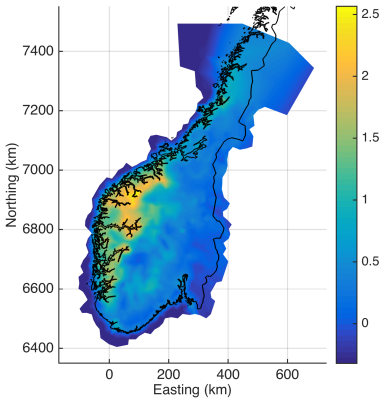
(a) Observation

(b) Prediction with non-stationary model

Figure 3: Total precipitation for the one year period September 1, 2008, to August 31, 2009, for 233 measurement stations in southern Norway measured in meters in a) and predictions from the non-stationary model in b). Coordinate system is UTM33.



(a) Elevation (km)



(b) Magnitude of gradient (100m/km)

Figure 4: The covariates (a) elevation and (b) magnitude of the gradient used for the covariance structure.

of the nugget effect. The hyperparameters are selected to satisfy $\mathrm{P}(\rho < 10) = 0.05$, $\mathrm{P}(\sigma > 3) = 0.05$ and $\mathrm{P}(\sigma_{\mathrm{n}} > 3) = 0.05$, and the model is fitted to the data with INLA (Rue et al., 2009). With this prior we consider a standard deviation greater than 3 large for both the GRF and the nugget effect, and a range less than 10 km unlikely based on the spatial scale that we are working on. The MAP estimates are $\hat{\sigma}_{\mathrm{N}} = 0.13$, $\hat{\rho} = 219$ and $\hat{\sigma} = 0.72$, and will be used in our scheme for setting the hyperparameters in the prior for the non-stationarity.

The non-stationarity is described by a function $R(\cdot)$ that describes how the local range varies and a function $S(\cdot)$ that describes how the marginal variance varies. The two covariates in the second-order linear structure enters linearly in $\log(R(\cdot))$ and $\log(S(\cdot))$, and the

coefficients, $\boldsymbol{\theta}_1$, of the two linear covariates in $\log(R(\cdot))$ are given the prior

$$\boldsymbol{\theta}_1|\tau_1 \sim \mathcal{N}(\mathbf{0}, S_1/\sqrt{\tau_1})$$
$$\tau_1 \sim \frac{\lambda_1}{2}\tau_1^{-3/2}\mathrm{e}^{-\lambda_1/\sqrt{\tau_1}}$$

and the coefficients, $\boldsymbol{\theta}_2$, of the two linear covariates in $\log(S(\cdot))$ are given a similar prior, but with hyperparameter λ_2. Further details are found in the Supplementary Material.

The hyperparameters λ_1 and λ_2 are selected based on the frequentist coverages of the non-stationarity parameters when fitting the non-stationary model to stationary data. Specifically, we use the MAP estimates of the stationary model to simulate 100 datasets from the stationary model with $\beta_0 = \beta_1 = 0$, set values for the hyperparameters λ_1 and λ_2, fit a non-stationary model with $\beta_0 = \beta_1 = 0$ to each of datasets, and calculate the frequentist coverage of the the 95% credible intervals of the non-stationarity parameters. It is overly expensive to run the model 100 times and we use a cheaper approximation in INLA that is conservative. We tried several values for the hyperparameters λ_1 and λ_2 and found that $\lambda_1 = \lambda_2 = 20$ provides coverage that is close to the nominal 95% for $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$.

The non-stationary model was then fitted using an MCMC sampler and the resulting posterior means of the range and the standard deviation are shown in the Supplementary Material, and they are not included here since the focus is on improving predictions. The figures show that the non-stationary model moves away from the stationary model even under the conservative prior.

The leave-one-out log-score is estimated from the samples of the MCMC sampler, and we find the score 0.13 for the stationary model and 0.22 for the non-stationary model. The leave-one-out estimates for the continuous rank probability score (CRPS) are 0.092 for the stationary model and 0.083 for the non-stationary model. Experimentation with the strictness of the prior showed that further improvements were possible by making the

prior weaker, but that making the prior too weak leads to worse scores. The prior and the procedure for selecting the hyperparameters appears to introduce a reasonable level of conservativeness for this dataset.

If we run the model with the same hyperparameters and remove the non-stationarity in the local range, the CRPS is 0.086, and if we remove the non-stationarity in the marginal standard deviations, the CRPS is 0.081. This shows that the covariates in the local range appear to be contributing more to the improved predictions than the covariates in the standard deviation, and that using all four covariates has degraded the performance slightly compared to using only a non-stationary local range. This demonstrates that guaranteeing improvements when including more covariates in the second-order structure is difficult. So a procedure for constructing conservative priors are critically important for non-stationary models, but the prediction scores of the models must be compared to ensure that the non-stationary model improves the predictions.

5 Discussion

The main challenge for constructing multivariate PC priors based on a measure of distance from a base model is that a joint prior for the parameters cannot be uniquely determined from a prior for the distance. Simpson et al. (2017) present a general approach where conditional on the value of the distance, D, the probability density is uniformly distributed on the set of parameters that specify models that are a distance D from the base model, but the approach is not parametrization-invariant and it is not clear for which parametrization of range and marginal variance that this approach would be appropriate. However, in this paper we have shown that the key for properly extending the PC prior framework to a joint prior for range and marginal variance in Matérn GRFs with fixed smoothness is to

use knowledge about the parameter space to split the construction of the multivariate PC prior into a sequential construction of univariate PC priors. This demonstrates that the principles of the PC prior framework are applicable for model components with complex parameter spaces that contain intrinsically linked parameters, but that the simple idea of distance from a base model must be combined with careful consideration of the parameter space.

The construction of the joint prior based on the infinite-dimensional distribution of the GRF instead of the finite-dimensional distribution of an observation from the GRF is technically more challenging that the finite-dimensional examples in Simpson et al. (2017). But the calculation of the KLD can be handled using the spectrum of the GRF and the fact that the KLD is infinite for general changes in the parameters can be overcome by careful reparametrization and a sequential construction of the prior. The benefits gained from the extra difficulty are that the PC prior for Matérn GRFs with fixed smoothness and the extension to the non-stationary GRF are computationally inexpensive since they have simple forms, are appropriate for hierarchical models since they work with any observation process, and can be applied for sequential analysis of data since they do not depend on the design of the experiment.

Setting the hyperparameters for the stationary part of the model can be done based on statements about what constitutes a large standard deviation or a large deviation from zero for the spatial field, and what constitutes a small range. This allows the users to choose to limit the preference for intrinsic models and thus provide more sensible posterior inference for the problem at hand. In the simulation study we observe good coverage of the equal-tailed 95% credible intervals when the prior satisfies $\mathrm{P}(\sigma > \sigma_0) = 0.05$ and $\mathrm{P}(\rho < \rho_0) = 0.05$, where σ_0 is between 2.5 to 40 times the true marginal standard deviation and ρ_0 is between 1/10 and 1/2.5 of the true range. The lengths of the credible intervals

depend on the values chosen for σ_0 and ρ_0, but are shorter than for the reference prior and consistent with the information put into the prior. The recommendations are based on the quantile-based credible intervals because the coverage of the 95% HPD credible intervals is further away from the nominal level and more sensitive to hyperparameters than the equal-tailed 95% credible intervals when the PC prior is used.

It is difficult to elicit expert knowledge about the hyperparameters for a non-stationary GRF since the second-order structure is not observed directly, and we discuss an alternative way to set the hyperparameters based on the frequentist coverage of the credible intervals. Using the new prior and the associated scheme for selecting the hyperparameters, we find a better fit for the non-stationary GRF than with the stationary GRF when applied to the dataset of annual precipitation in southern Norway measured both with leave-one-out CRPS and log scores.

The paper shows that the PC prior framework provides a useful tool for deriving a principled joint prior for the range and the marginal variance of a Matérn GRF with fixed smoothness, and that the ideas of the framework are useful for constructing priors that limit flexibility also for non-stationary GRFs where exact derivation is not possible.

6 Acknowledgements

Fuglstad was supported by project number 240873/F20 from the Research Council of Norway.

References

Anderes, E. (2010). On the consistent separation of scale and variance for gaussian random fields. *Ann. Statist.*, 38(2):870–893.

Berger, J. O., Bernardo, J. M., and Sun, D. (2009). The formal definition of reference priors. *Ann. Statist.*, 37(2):905–938.

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.

Chen, M.-H. and Shao, Q.-M. (1999). Monte carlo estimation of bayesian credible and hpd intervals. *Journal of Computational and Graphical Statistics*, 8(1):69–92.

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

Ingebrigtsen, R., Lindgren, F., Steinsland, I., and Martino, S. (2015). Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics*, 14, Part C:338–364.

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.

Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Anal.*, 11(4):1071–1106.

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.

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.

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.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statist. Sci.*, 32(1):1–28.

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.

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