A Review: Constructing Priors that Penalizes the Complexity of Gaussian Random Fields (Fuglstad, et al (2019))

Yidan (Eden) Xu, Max Schneider
Department of Statistics, University of Washington
yx2516@uw.edu, maxs15@uw.edu

Abstract

We present a discussion of [1] by Fuglstad et al., complete the proof for Thm 2.1 and replicate several results from the paper. In Section 1, we give an overview of the problem addressed in the paper. We make explicit the contributions of the paper and provide a summary of theoretical and simulation results in Section 2, 3 and 4.

1 Problem Overview

1.1 Model Based Geostatistics with Gaussian Random Fields (GRF)

Gaussian random fields (GRFs) introduce spatial second order dependence into hierarchical models (Eq. 1) that are applied extensively in spatial statistics. It is a "simple yet powerful tool", according to the authors, whose construction only involves the specification of a correlation function. Stationary GRFs are controlled only by two parameters, the (local) range ρ and marginal variance σ^2, but their estimation and inference remains a computational challenge.

$$\begin{aligned} &Y = X\boldsymbol{\beta} + u(S) + \varepsilon,\ Y_i|u(S_i) \text{ are iid} \\ &u(S) \sim N(0, \sigma^2\Sigma(\rho)), \\ &\varepsilon \sim N(0, \sigma_\varepsilon^2 I_n) \\ &Y \in \mathbb{R}^n,\ X \in \mathbb{R}^{n\times n},\ \beta \in \mathbb{R}^p,\ \Sigma \in \mathbb{R}^{n\times n},\ \sigma^2 \in \mathbb{R}^+,\ \sigma_\varepsilon^2 \in \mathbb{R}^+,\ \rho \in \mathbb{R}^+ \end{aligned} \tag{1}$$

Frequentist methods to estimate these parameters include (profile) maximum likelihood and restricted maximum likelihood. However, maximizing likelihoods for spatial models often requires gradient-based numerical optimization. This can be highly sensitive to parameter initialization, especially when the likelihood function is multimodal or flat, as often the case for hierarchical models.

1.2 Bayesian Inference and Prior Selection

Bayesian inference provides an alternative approach, by estimating the posterior distribution of the parameters, provided with some choice of priors. In most cases, posteriors are sampled using Markov Chain Monte Carlo (MCMC) or approximated using deterministic methods, such as integrated nested Laplace approximation (INLA). The summary statistics of the posterior are able to provide a rich quantification of parameter uncertainty.

However, selecting a sensible prior is non-trivial. A well-chosen prior stabilizes the inference and improves the predictive power; however, in the words of Fuglstad et al., "a poorly chosen prior may lead to catastrophe." When detailed prior knowledge on parameters is available, one may construct a subjective prior such that it represents this existing information and therefore informs the penalization of the model fit. However, as models grow more complex and expert knowledge may be unavailable, the difficulty in specifying subjective priors on the parameters increases.

One alternative is to use a non-subjective prior, such as Jeffreys' prior. The posterior estimation is then based on the information from the data only, via the likelihood, and does not incorporate any

additional a priori information. Such an approach may be suboptimal, especially in the case of GRFs with Matérn covariance function $c : [0, \infty) \to \mathbb{R}$ as in Eq. 2.

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

Here, $\sigma^2 > 0$ denotes the marginal variance, $\rho > 0$ the range and ν the smoothness. $\mathcal{K}_\nu$ denotes the modified Bessel function of the second kind with order ν and $r = |s - t|$ for spatial locations $s, t \in \mathcal{D} \subset \mathbb{R}^d$, $\mathcal{D}$ a bounded spatial domain, dimension $d \leq 3$.

In this case, the value of σ^2 and ρ are coupled, leading to a ridge in the likelihood function (Warnes and Ripley, 1987). This means that the likelihood is the same for different pairs of σ^2 and ρ that may vary dramatically. Consequently, for a bounded domain, there do not exist consistent maximum likelihood estimators for both parameters even with asymptotically infinite sample size. An uninformative prior will allow the likelihood to dominate the estimation, often leading to unstable estimates. In other words, the prior affects the behavior of the posterior of the parameters, even under in-fill asymptotics. It is essential to select the prior carefully.

1.3 Penalized Complexity Prior

Simpson et al. [2] introduced a principled way of constructing weakly informative priors for parameters in additive hierarchical models. In Simpson's framework, a base (simplest possible) model is identified for a given parameter, and a Penalized Complexity (PC) prior is set as a more informative prior (called the flexible model). Constructing such a prior often requires expert knowledge. However, a general algorithmic approach can optimize the level of information between the flexible and base model by using four key principles. These are:

1. *Occam's Razor:* a simple model is preferred;
2. *Measure of Complexity:* Kullback-Leibler Divergence (KLD) is used to measure the increased complexity from the base model;
3. *Constant Rate Penalization:* assume a constant decay rate of the PC prior as a distance based on the KLD increases and
4. *User Defined Scaling:* user defines values setting the tail probability of the prior, which can be used to select hyperparameters for the PC prior.

2 Constructing Priors for GRFs with Matérn Covariance

Based on the reparameterization of the Matérn covariance function in Eq. 3 below, Fuglstad et al. derive a PC prior for $\tau|\kappa$ and κ with fixed smoothness parameter ν in a GRF of dimension $d \leq 3$.

$$c_\nu(r; \tau, \kappa) = \tau^2 \frac{\kappa^\nu r^\nu \mathcal{K}_\nu(\kappa r)}{(4\pi)^{d/2} 2^{\nu-1} \Gamma(\nu + d/2)} \tag{3}$$

where $\kappa = \sqrt{8\nu}/\rho$ the reparameterized scale parameter, $\tau^2 = \sigma^2 \kappa^{2\nu} \frac{\Gamma(\nu+d/2)(4\pi)^{d/2}}{\Gamma(\nu)}$ the reparameterized marginal variance.

This gives an extension to the PC prior framework from [2], who only considered the case for additive hierarchical models as in Eq. 1. This is important as the Matérn covariance function is quite general (many common covariance functions are special cases of it), and is ubiquitous in spatial hierarchical modelling.

2.1 Construction Steps

To derive the PC prior, the first step is to derive KL distance d from the base model to the flexible model. (We note that d was used before to denote the dimension of the GRF; this notational issue is kept, to maintain comparison with Fuglstad, et al.) Denoting the Gaussian measures for base model and flexible model as P_0 and P over the set $\mathcal{X}$ respectively, where P is absolutely continuous w.r.t. P_0, then the KL distance is defined by $d(P \parallel P_0) = \sqrt{2KLD(P \parallel P_0)}$, where

$$KLD(P \parallel P_0) = \int_{\mathcal{X}} \log \frac{\mathrm{d}P}{\mathrm{d}P_0} \mathrm{d}P,$$

where $\frac{\mathrm{d}P}{\mathrm{d}P_0}$ is the Radon-Nikodyn derivative of P w.r.t. P_0. Notably, this construction of d corresponds to the change in complexity from the flexible model to the base model.

Next, the authors define the prior on d using the other three principles, in particular the constant decay rate

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

for a constant $r \in (0,1)$. They note that the exponential distribution is the only continuous distribution with this property if $r = \exp(-\lambda)$ is chosen; therefore, $\pi(d) = \lambda \exp(-\lambda d)$ for $d > 0$. Such a construction naturally connects the distance from flexible to base model with how the change in the flexible model's distribution. And the hyperparameter λ is chosen by user-defined upper or lower tail probability, i.e. $P(Q(d) > U) = \alpha$ or $P(Q(d) < L) = \alpha$. Here, $Q(d) = \sigma$ or ρ is found by change of variables to the original parameterization of the Matérn covariance function. This enables the user to specify prior belief on the geometry of the parameter space, by selecting the tail values (U, L) of the tail probabilities in the prior distribution.

2.2 PC prior for $\tau|\kappa$ and κ

In deriving PC priors for Matérn covariance parameters, Fuglstad et al. start by considering $\tau|\kappa$, taking the base model as $\tau = 0$ conditioned on κ. Then,

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

where $\lambda > 0$ is a hyperparameter satisfying $\mathbb{P}(\sigma > \sigma_0|\kappa) = \alpha$. We obtain that

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

The derivation to $\pi(\tau|\kappa)$ is provided in the answer to Q1 in Section **??**.

The PC prior constructed for κ is based on the infinite-dimensional GRF instead of a finite-dimensional GRF, when τ is set fixed. When the base model takes $\kappa = 0$, which corresponds to $\rho = \infty$,

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

The hyperparameter $\lambda > 0$ is again chosen by specifying $\mathbb{P}(\rho < \rho_0) = \alpha$. Then,

$$\lambda = -\left(\frac{\rho_0}{\sqrt{9\nu}}\right)^{d/2}\log(\alpha).$$

Lastly, the authors provide the joint PC prior for (τ, κ) and consequently (σ, ρ), by again transforming the parameters, taking a base model with infinite range and zero marginal variance,

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

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

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

2.3 Complete proof for Theorem 2.1 [1]

We present the proof for Theorem 2.1 in Fuglstad et al. (2019) with the four principles specified for PC prior in [2].

Theorem 1 (Theorem 2.1 [1])**.** *Let u be a GRF defined on $\mathcal{D} \subset \mathbb{R}^d$ with a Matérn covariance function with parameters (τ, κ, ν). If the GRF is observed on $s_1, s_2, \cdots, s_n \in \mathcal{D}$, then conditionally on κ, the PC prior for τ with base model $\tau = 0$ is*

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

with hyperparameter $\lambda > 0$.

Proof. Let $\boldsymbol{u}|\tau,\kappa \sim \mathcal{N}^{(n)}(0,\Sigma)$ with $\Sigma \in \mathbb{R}^n$, where $\boldsymbol{u} = (u(s_1),\cdots,u(s_n))$. $\Sigma = \tau^2 R(\kappa,\nu)$ is the Matérn covariance matrix parameterised as before, where R is a fixed correlation matrix with fixed κ and ν. Then the KLD from the flexible model $\mathcal{N}^{(n)}(0,\Sigma)$ to the base model $\mathcal{N}_0^{(n)}(0,\Sigma_0)$ is

$$KLD(\mathcal{N}^{(n)} \parallel \mathcal{N}_0^{(n)}) = \frac{1}{2}\left\{tr(\Sigma_0^{-1}\Sigma) - n - \ln\left(\frac{|\Sigma|}{|\Sigma_0|}\right)\right\}.$$

And the base model $\Sigma_0 = 0$ when $\tau_0^2 = 0$. For simplicity, assume that R has full rank, then the KLD becomes

$$KLD(\mathcal{N}^{(n)} \parallel \mathcal{N}_0^{(n)}) = \frac{n}{2}\frac{\tau^2}{\tau_0^2}(1 + \frac{\tau_0^2}{\tau^2}\ln(\frac{\tau_0^2}{\tau^2}) - \frac{\tau_0^2}{\tau^2}) \to \frac{n}{2}\frac{\tau^2}{\tau_0^2}$$

for $\tau_0^2 \ll \tau^2$. Therefore the KLD between the two distribution is $d(\tau) = \sqrt{2KLD(\mathcal{N}^{(n)} \parallel \mathcal{N}_0^{(n)})} = \sqrt{n\tau^2/\tau_0^2} = \sqrt{n/\tau_0^2}\tau$. And the prior defined for the distance $d(\tau)$ is the exponential

$$\pi(d) = \theta \exp\{-\theta d\},$$

which is derived from the assumption of constant decay rate $r = \exp\{-\theta\}$ of the flexible model from the base model (Principle 3). In other words, the prior for distance d satisfies

$$\frac{\pi_d(d+\delta)}{\pi_d(d)} = r^\delta.$$

By transformation of random variables from distance d to marginal variance τ^2, we get that

$$\pi(\tau|\kappa) = \theta e^{-\theta d(\tau)}\left|\frac{\partial d(\tau)}{\partial \tau^2}\right| = \theta\sqrt{n/\tau_0^2}e^{-\theta\sqrt{n/\tau_0^2}\tau} = \lambda e^{-\lambda\tau}$$

where $\tau > 0$, and $\lambda = \theta\sqrt{n/\tau_0^2}$. And the hyperparameter θ can be defined by user via specifying the tail value τ_0^2 defining the tail probability of the PC prior. □

2.4 Discussion on PC prior for κ

Due to the computational burden on deriving the PC prior for κ based on the finite-dimensional distributions corresponding to the observation locations, Fuglstad et al. instead propose to derive the prior based on the infinite dimensional GRF. They claim that "this is possible because the changes in κ result in finite values for the KLD for the infinite-dimensional GRF when τ is fixed." We will elucidate this sentence below.

We start with the following theorem from [3], referenced in [1], which gives the sufficient conditions for equivalence of two Gaussian measures with Matérn covariances on a bounded domain.

Theorem 2 (Theorem 2 [3])**.** *Let P_i, i=1,2 be two probability measures such that under P_i, the process $u(s)$, $s \in \mathbb{R}^d$, where $d = 1,2,3$, is stationary Gaussian with mean 0 and an isotropic Matérn covariance in $\mathbb{R}^d$ with variance σ_i^2 and scale parameter κ_i, $i = 1,2$ and the same smoothness parameter ν. For any bounded infinite set $\mathcal{D} \subset \mathbb{R}^d$, $P_1 \equiv P_2$ on the path $u(s)$, $s \in \mathcal{D}$ if and only if $\sigma_1^2\kappa_1^{2\nu} = \sigma_2^2\kappa_2^{2\nu}$ i.e. $\tau_1 = \tau_2$.*

The upshot of Theorem 2 is that it provides a direct link between the Gaussian measure of a GRF and parameters of Matérn covariance function. Therefore, if $\tau_1 = \tau_2$ with changing κ (that is, τ is fixed), the two Gaussian measures are equivalent and the KLD between the corresponding GRFs is finite. Therefore, the PC prior for κ is well-defined. On the other hand, if $\tau_1 \neq \tau_2$, i.e. when τ is changing, the two Gaussian measures are orthogonal on the path $u(s)$. This is to say that the absolute difference of the covariance for the process $u(s)$ at any indices $s,t \in \mathcal{D}$ for the Gaussian measures $P_1(\sigma_1^2,\kappa_1,\nu)$ and $P_2(\sigma_1^2,\kappa_2,\nu)$ is not bounded, i.e.

$$|\mathbb{E}_{P_1}(u(s)u(t)) - \mathbb{E}_{P_2}(u(s)u(t))| = \infty.$$

Consequently, the construction of the PC prior with infinite-dimensional GRF is possible for κ only if τ is fixed. This is because the theorem above guarantees a finite KLD, meaning this can be used as the measure of complexity (Principle 2).

Finally, note that τ can be consistently estimated as the number of observations increases; however, κ cannot be. We can show this with a proof sketch by contradiction. Note that if both are consistently

Name	Description	Expression
PriorPC	**PC Prior**, where $\rho, \sigma > 0$ and with hyperparameters $\rho_0, \alpha_\rho, \sigma_0, \alpha_\sigma$	$\pi(\rho, \sigma) = \lambda_1 \lambda_2 \rho^{-2} \exp(\lambda_1 \rho^{-1} - \lambda_2 \sigma)$
PriorJe	**Jeffreys' Rule Prior**, where $\rho, \sigma > 0$ and with no hyperparameters. $U = (\frac{\partial}{\partial \rho}\Sigma)\Sigma^{-1}$, with Σ the correlation matrix of the observations (Berger et al (2001)	$\pi(\rho, \sigma) = \sigma^{-1}(tr(U^2) - \frac{1}{n}tr(U)^2)^{1/2}$
PriorUn1	**Uniform Prior 1**: ρ follows a Uniform prior on bounded interval $[A, B]$ and $\sigma > 0$ follows Jeffreys' prior	$\pi(\rho, \sigma) \propto \sigma^{-1}$
PriorUn2	**Uniform Prior 2**: $\log(\rho)$ follows a Uniform prior on bounded interval $[A, B]$ and $\sigma > 0$ follows Jeffreys' prior	$\pi(\rho, \sigma) \propto \sigma^{-1}\rho^{-1}$

Table 1: Priors investigated in simulation study (modified version of the authors' Table S1).

estimable, then σ^2 is consistently estimable. By Theorem 2, if we take $\boldsymbol{\theta} = (\sigma^2, \kappa, \nu)$ and $\boldsymbol{\theta'} = (2^{2\nu}\sigma^2, \kappa/2, \nu)$, then $\sigma_1^2\kappa_1^{2\nu} = \sigma_2^2\kappa_2^{2\nu}$, and the two corresponding measures are equivalent: $P_{\boldsymbol{\theta}} = P_{\boldsymbol{\theta'}}$. If there exists a weakly consistent estimator σ_k^2 that converges in probability $P_{\boldsymbol{\theta}}$ to σ^2, then it also converges to $2^{2\nu}\sigma^2$. This means that σ^2 cannot be consistently estimated, contradicting the consistency assumption made above. Since a consistent estimator for σ^2 does not exist, neither does one for κ.

3 Comparison to Other Priors via Simulation

The PC prior framework described above is implemented on simulated data and compared to several alternative priors. Spatial data is simulated as follows:

1. Select 25 locations, $s_1, \ldots, s_{25}$ at random from the unit square: [0,1]^2
2. Generate realizations $\mathbf{u} = \{u(s_1), \ldots, u(s_{25})\}$ from two GRFs with exponential covariance $c(r) = \exp(\frac{-2r}{R_0})$, where the true range R_0 is set to 0.1 or 1, respectively. For both cases, the true σ^2 is set to 1. The authors choose only to vary the true range value in this simulation; interesingly, in the application (see Section 4), both range and variance parameters are varied.
3. Fit the simulated data using a GRF with exponential covariance $c(r) = \exp(\frac{-2r}{\rho})$, where the range and marginal variance are estimated from the data. Use one of four priors in this estimation, described in Table 1.

The hyperparameter selection for the PC prior requires setting probability bounds on the two parameters. That is, we select hyperparameters such that $P(\rho < \rho_0) = \alpha_\rho$ and $P(\sigma^2 > \sigma_0^2) = \alpha_\sigma$, choosing several values of ρ_0 and σ_0.

The targets of investigation in this simulation study are the coverage and lengths of the credible intervals produced by the different priors. The coverage of a credible interval is the proportion of credible intervals produced by posterior sampling that cover the true parameter value. If the posterior sampling has good frequentist properties, it should be close to the nominal value (so 0.95 for a 95% credible interval).

In this study, two kinds of credible intervals are computed:

- the standard quantile-based equal-tailed intervals,
- highest posterior density (HPD) intervals, the shortest possible intervals that contain the desired α level of the posterior distribution.

The equal-tailed intervals are preferred, and will be discussed throughout the results, because they have closer coverage probabilities to the nominal values for the PC and Jeffreys' prior, and are less sensitive to hyperparameter values for the PC prior. The authors justify using the equal-tailed intervals in light of the high skewness of the posteriors (which would motivate the HPD intervals), because both interval types yield similar results when comparing alternative priors.

The joint posterior of range and standard deviation is examined under the PC and Jeffreys' priors. For the PC prior, the hyperparameter is selected by setting $P(\rho < 0.1) = 0.05$ and $P(\sigma > 10) = 0.05$. As we demonstrate in Figure 1, the sampler will explore more of the extremes of the posterior under the Jeffreys' prior, than under the PC prior. This indicates that the PC prior "can be used to achieve credible intervals that are more reasonable", in light of, e.g., the spatial scale of the problem, or previous analysis of a subset or different set of data (this would be an empirical Bayes approach).

Next, authors examine the performance of estimation under the PC prior. The authors' Table S2 shows that coverage is good even when the σ_0 is allowed to be far greater than the true standard deviation (up to 40x greater) or ρ_0 is set to be much less than the true range (down to 1/10 of the true range). But a σ_0 that is too low (i.e., 0.625 times the true value), as well as a ρ_0 that is too high (i.e., 40 times the true value) results in credible intervals with too low coverage. Results are similar between the two true ranges ($R_0 = 0.1$ or 1), with the only difference being that for $R_0 = 1$, a ρ_0 that was 40x the true value still led to good coverage.

The authors then compare coverage and the length of the credible intervals across the 4 alternative priors. The PC prior has much shorter credible intervals than the Jeffreys' prior, though they both have reasonable coverages across the true ranges and for the two parameters. For PriorUn1, both these performance metrics vary based on the upper limit set (meaning that a poorly set upper limit leads to poor coverage and excessively wide intervals). For PriorUn2, only the interval lengths depend on the upper limit set and the coverage is close to nominal throughout. In summary, since both coverage probability and short credible intervals are of import, the PC prior outperforms the three alternatives across a range of hyperparameter values and for all experiments performed.

3.1 Answer to Q3

We reproduce the results in Fuglstad et al's Figure 2 and its discussion in the text (see our Figure 1). The plot compares draws from the joint posterior distribution for ρ and σ, under two different priors. We first simulate observations from a GRF as described in Section 3, using true ρ = 1 and true $\sigma = 1$. We then use PriorJe (black circles in Figure 1) and PriorPC (red circles) to estimate the joint posterior of the parameters. For PriorPC, we use a hyperparameter such that $P(\rho < 0.1)$ = 0.05 and $P(\sigma > 10)$ = 0.05.

The authors claim to take 1,000,000 draws from the joint posterior distribution, which leads to very extreme posterior values under PriorJe but not under PriorPC. In fact, the largest values under PriorJe are several orders of magnitude higher than those under PriorPC. However, running the MCMC sampler to draw from this complicated joint posterior (under either prior specification) requires several numerical steps (for example, in taking a Cholesky decomposition of the covariance matrix). This meant that producing many posterior draws resulted in numerical errors in R, as the matrix system was treated as singular. It took hundreds of attempts before a run of 10,000 draws from the posterior could be accomplished without a numerical error. We experimented with a larger posterior sample but this failed to run, even with thousands of attempts.

Even with a smaller number of draws, we get the same pattern when comparing PriorJe and PriorPC, and can conclude, as the authors do, that PriorPC yields posterior estimates that are more reasonable. For example, a posterior range of 100 or more (produced by PriorJe but not PriorPC) is quite unbelievable when the true range is 1. This indicates that even weakly informative priors yield better (more stable) results than non-informative priors like the Jeffreys' rule. This effect will be even stronger when the likelihood has more features (e.g., flatness, multimodality) that make it difficult to work with numerically.

4 Application to Nonstationary Precipitation Data

The PC prior framework established in Section 2 can also be used to fit nonstationary models to data. This is useful when predicting a process whose second-order structure varies over time, its spatial domain or both (e.g., many environmental processes under climate change). An example is implemented by the authors on total precipitation levels for a one-year period, across 233 measurement stations in southern Norway. The guiding question is whether a nonstationary model, which can more flexibly explain detailed structure in the data, provides better precipitation *predictions* than a stationary model. Guided by previous research [4], the authors specify a linear geostatistical model with intercept and elevation fixed effect, a spatial random effect and an iid normally distributed nugget

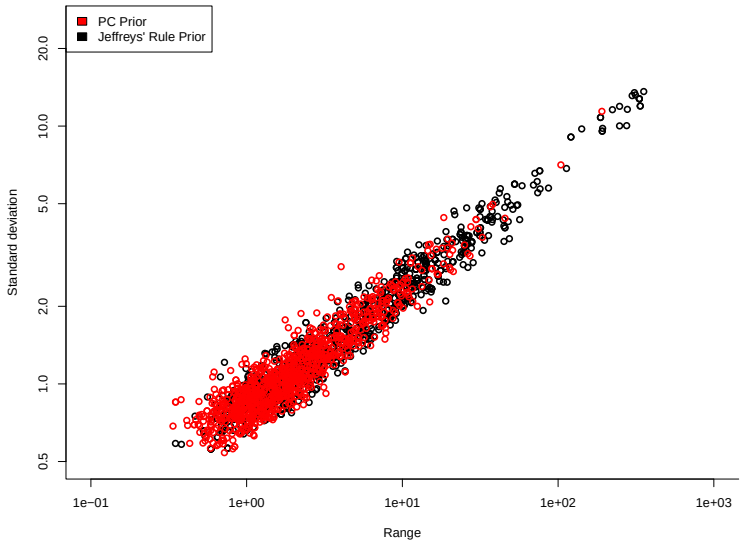

Figure 1: Replication of Figure 2 from [1].

effect, that is

$$y_i = \beta_0 + x_1\beta_1 + u(s_i) + \varepsilon_i,$$

where y_i are precipitation values, x_i are elevation values at station i, $u(s_i)$ is a GRF parametrized by local range ρ and marginal variance σ^2 and the nugget $\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$. This model assumes that the precipitation process is second-order stationary, with a constant mean and a covariance structure specified entirely by two parameters: ρ and σ^2. This **stationary model** describes the first-order structure of the precipitation process.

PC priors are constructed on the parameters; hyperparameters for the PC priors are set such that ρ should not have high probability to be under 10 km and σ^2 (both for the spatial field and nugget precision) should not exceed 3 km. These are justified "based on the spatial scale [the authors] are working on."

Building a **nonstationary model** requires specifying a first-order structure, which is borrowed from the stationary model, and the second-order structure, or how the local range and marginal variance vary over space. These are specified by functions $R(\cdot)$ and $S(\cdot)$, respectively. For both parameters, the authors use a linear function of elevation and the magnitude of gradient of the elevation (without explaining why, outside of an apparent visible similarity between the latter covariate and the pattern of precipitation). Functions $R(\cdot)$ and $S(\cdot)$ are written as a sum of a set of basis functions scaled by some coefficients, which are grouped into vectors $\boldsymbol{\theta_1}$ and $\boldsymbol{\theta_2}$, respectively. Using a g-prior approach [5], these are modelled by a Normal distribution with precision controlled by precision parameters (τ_1 and τ_2, respectively) multiplied by a Gramian matrix of the basis function set. Each of these two parameters now gets an exponential PC prior placed on it, and this flexible model is shrunk towards a base model of zero variance (no second-order structure). In general, the hyperparameter for the exponential distribution can be selected by expert knowledge or, as the authors do, by shrinking down to the base model.

The base model of zero variance is exactly the stationary model. So, the hyperparameters can be set such that they yield good inferences when fitting data from a stationary process. To do this, the authors fit a stationary model to the dataset and then many processes are simulated from the dataset and fit by the non-stationary model under different hyperparameters. Credible intervals for each parameter under these different nonstationary models are compared with the parameter estimate from the stationary model. The hyperparameter is set to be the one that provides coverage close to the nominal coverage to the stationary estimates. In the Norwegian precipitation example, the resulting best-covering hyperparameters are $\lambda_1 = \lambda_2 = 20$.

This nonstationary model is fitted to the precipiation data using Markov Chain Monte Carlo simulations from the posterior. It is compared to the stationary model on two standard scores of predictive performance: the log score and the Continuous Ranked Probability Score (CRPS). In both, a leave-one-out cross validation scheme is used to see how well the model built on the remaining stations predicts the one left out. The log-score is higher (better) for the nonstationary model and the CRPS is lower (better) as well. It is also found that CRPS improves when the second-order structure is modelled only by $R(\cdot)$ in the non-stationary model. That is, not including the marginal variance in the second-order structure leads to better predictions than a model that has both. This is left undiscussed in the article, but may be due to the greater importance of local range in describing the spatial pattern of precipitation values, which drop quickly when moving away from the coastline.

References

[1] Geir-Arne Fuglstad, Daniel Simpson, Finn Lindgren, and Håvard Rue. Constructing priors that penalize the complexity of gaussian random fields. *Journal of the American Statistical Association*, 114(525):445–452, 2019.

[2] Daniel Simpson, Håvard Rue, Andrea Riebler, Thiago G Martins, Sigrunn H Sørbye, et al. Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical science*, 32(1):1–28, 2017.

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

[4] Rikke Ingebrigtsen, Finn Lindgren, Ingelin Steinsland, and Sara Martino. Estimation of a non-stationary model for annual precipitation in southern norway using replicates of the spatial field. *Spatial Statistics*, 14:338–364, 2015.

[5] Arnold Zellner. On assessing prior distributions and bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques*, 1986.