Review of Constructing Priors that Penalize the Complexity of Gaussian Random Fields (2019)

Peter Liu and Houjie Wang

March 26, 2021

Motivation

- Gaussian random field (GRFs):

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

Motivation

- Frequestist approaches exist, yet might not be optimal.

Motivation

- Frequestist approaches exist, yet might not be optimal.
- Bayesian approaches are preferred in practice!

Motivation

- Frequestist approaches exist, yet might not be optimal.
- Bayesian approaches are preferred in practice!
- However, choosing a proper prior is highly non-trivial!

Motivation

- Frequestist approaches exist, yet might not be optimal.
- Bayesian approaches are preferred in practice!
- However, choosing a proper prior is highly non-trivial!
- Issue with GRFs: model overfitting.

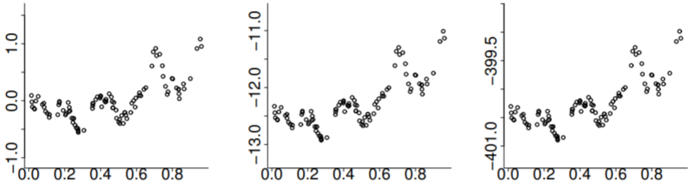
Motivation

- In the case of Matern covariance function, only the ratio of the parameters can be estimated consistently via in-fill asymptotics.

Motivation

- In the case of Matern covariance function, only the ratio of the parameters can be estimated consistently via in-fill asymptotics.
- Choosing an uninformative prior will allow the likelihood to dominate the estimation.

Motivation



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

Figure 1: Simulations with the exponential covariance function $c(d) = \sigma^2 e^{-d/\rho}$ 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.

Motivation

A desirable prior should be ...

Motivation

A desirable prior should be ...

- Penalize complexity.

Motivation

A desirable prior should be ...

- Penalize complexity.
- Robust and precise.

Motivation

A desirable prior should be ...

- Penalize complexity.
- Robust and precise.
- **User friendly!**
  - Interpretable,

Motivation

A desirable prior should be ...

- Penalize complexity.
- Robust and precise.
- **User friendly!**
  - Interpretable,
  - modularized,

Motivation

A desirable prior should be ...

- Penalize complexity.
- Robust and precise.
- **User friendly!**
  - Interpretable,
  - modularized,
  - efficiency.

Methodology

- **Base model** is the simplest possible model;

Methodology

- **Base model** is the simplest possible model;
- **Flexible model** (GRFs) is set to be more informative than the base model;

Methodology

- **Base model** is the simplest possible model;
- **Flexible model** (GRFs) is set to be more informative than the base model;
- **Goal**: Construct prior that shrink the components towards their base model.

Methodology - Step one

- Define a proper distance metric between the models.

Methodology - Step one

- Define a proper distance metric between the models.
- Construct via properly scaled KL divergence:

$$\text{dist}(P||P_0) = \sqrt{2\text{KL}(P||P_0)}.$$

Methodology - Step two

Three general principles are introduced as the keystone of prior construction:

Methodology - Step two

Three general principles are introduced as the keystone of prior construction:

- *Occam's Razor:* The simpler, the better;

Methodology - Step two

Three general principles are introduced as the keystone of prior construction:

- *Occam's Razor:* The simpler, the better;
- *Constant-rate penalization:* Constant decay rate depends only on distance;

Methodology - Step two

Three general principles are introduced as the keystone of prior construction:

- *Occam's Razor:* The simpler, the better;
- *Constant-rate penalization:* Constant decay rate depends only on distance;
- *User Defined Scaling:* A interpretable way to set hyperparameter λ.

Methodology - Step two

- *User Defined Scaling:* Set by the practitioners!

Methodology - Step two

- *User Defined Scaling:* Set by the practitioners!
- Transform into interpretable formulations.

Methodology - Step two

- *User Defined Scaling:* Set by the practitioners!
- Transform into interpretable formulations.
- Through tail probabilities

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

where U, L are the upper/lower tail limits.

Main result

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

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

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

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

Numerical Experiment

Table 1: The four different priors used in the simulation study. 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	$\pi_2(\rho,\sigma) = \sigma^{-1}\left(\mathrm{tr}(U^2) - \frac{1}{n}\mathrm{tr}(U)^2\right)^{1/2}$	$\rho, \sigma > 0$ 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

Apply PC Prior to Leman

- Create 20 pairs of training (70%) and testing data by randomly sampling leman.
- For each pair, we fit a universal kriging model with linear trend and exponential covariance via MLE and obtain range and marginal variance estimates.
- For each pair, we apply PC prior and compute MAP estimates of range and marginal variance based on the posterior samples obtained by adaptive MCMC.
- We also perform spatial predictions at testing locations with parameter estimates from MLE and PC prior and compute the MSE.

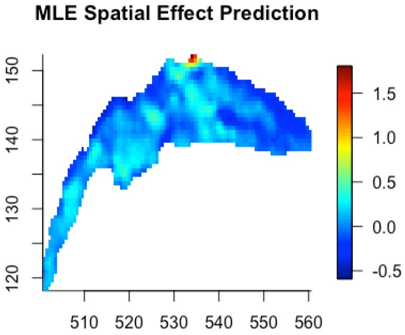
Apply PC Prior to Leman

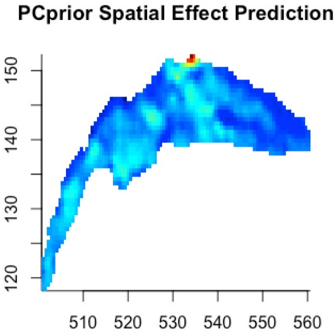
	MLE	PCprior
var	0.0690	0.0685
range	2.5175	2.4599

Table: Mean estimates of 20 pairs

	MLE	PCprior
MSE	2.9321	2.9342

Table: Mean MSE of 20 pairs





PC Prior with Different Covariance Structures

- We generate observations based on the 25 locations on $[0, 1]^2$ used in paper with exponential, Matern ($\nu = 2.5$) and spherical covariance functions with range($\rho = 1$) and variance ($\sigma^2 = 1$).
- We compute the posterior medians, MAPs, credible interval and HPD with respect to ρ and σ based on the samples obtained by adaptive MCMC.
- We repeat such experiment by 1000 times and compute the coverage probabilities and mean width of credible intervals.

PC Prior with Different Covariance Structures

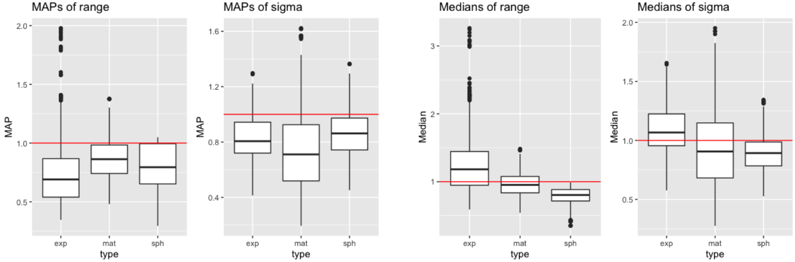


Figure: MAPs and Medians of range and standard deviation

	exp	mat	sph
CI ρ	0.988	0.972	0.945
CI σ	0.996	0.972	0.900
HPD ρ	1.000	0.965	0.934
HPD σ	1.000	0.965	0.866

Table: Coverage Probabilities

	exp	mat	sph
CI ρ	5.308	0.709	0.539
CI σ	1.588	1.643	0.583
HPD ρ	4.064	0.686	0.513
HPD σ	1.419	1.476	0.565

Table: Interval Lengths

PC Prior with Different Covariance Structures

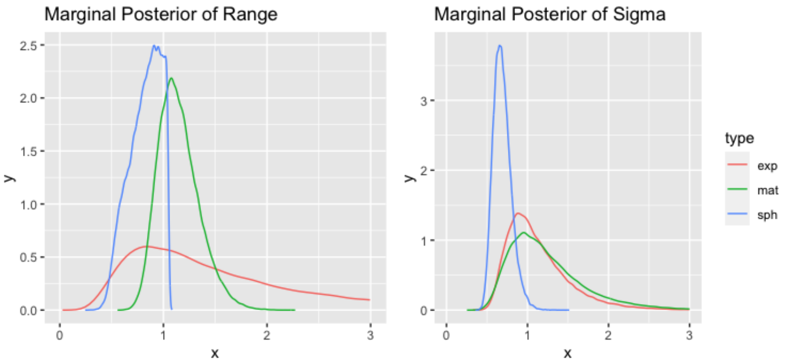


Figure: Posterior in One Experiment

Appendix: KLD

The Kullback-Leibler divergence (KLD) is defined as

$$\mathrm{KL}(P||P_0) = \int_\Omega \log \frac{dP}{dP_0} dP,$$

where $P \ll P_0$. For continuous Gaussian processes $\mathcal{N}_1$, $\mathcal{N}_0$ with density p_1 and p_0 respectively, the KLD is rewritten as

$$\mathrm{KL}\left(\mathcal{N}_1||\mathcal{N}_0\right) = \int_\Omega \log \frac{p_1(x)}{p_0(x)} p_1(x) dx.$$