A Model for Spatial Data

We consider modeling spatially indexed data $X(s)$ observed at n locations $s_1, \ldots, s_n$. We propose the model

$$X(s) = \mu(s) + v(s) + \varepsilon(s)$$

where μ corresponds to the large scale variability, and is usually associated with some covariates, like location. $v(s)$ corresponds to small scale variability, and is modeled using a Gaussian process. $\varepsilon(s)$ corresponds to measurement error, and is usually modeled with white noise.

A Model for Spatial Data

We consider modeling spatially indexed data $X(s)$ observed at n locations $s_1, \ldots, s_n$. We propose the model

$$X(s) = \mu(s) + v(s) + \varepsilon(s)$$

where μ corresponds to the large scale variability, and is usually associated with some covariates, like location. $v(s)$ corresponds to small scale variability, and is modeled using a Gaussian process. $\varepsilon(s)$ corresponds to measurement error, and is usually modeled with white noise.

Cressie proposes the addition of yet another term to the above model

$$X(s) = \mu(s) + v(s) + \omega(s) + \varepsilon(s)$$

where $\omega(s)$ captures micro-scale variability. This additional white noise corresponds to the Nugget Effect.

Hierarchical Formulation

The previous model can be expressed hierarchically in vector notation as

$$\begin{aligned} X &= \mu + \varepsilon, \qquad & \varepsilon \sim N(0, \tau^2 I) \\ \mu &= D\beta + \upsilon, \qquad & \upsilon \sim N(0, \sigma^2 R(\psi)) \ . \end{aligned}$$

One way to read this hierarchical model is that the first equation corresponds to the observation equation. The second equation corresponds to the “process” equation.

The former results in $X \sim N(D\beta, \tau^2 I + \sigma^2 R(\psi))$.

Hierarchical Formulation

The previous model can be expressed hierarchically in vector notation as

$$\begin{aligned} X &= \mu + \varepsilon, \qquad & \varepsilon &\sim N(0, \tau^2 I) \\ \mu &= D\beta + \upsilon, \qquad & \upsilon &\sim N(0, \sigma^2 R(\psi)) \ . \end{aligned}$$

One way to read this hierarchical model is that the first equation corresponds to the observation equation. The second equation corresponds to the "process" equation.

The former results in $X \sim N(D\beta, \tau^2 I + \sigma^2 R(\psi))$.

The nugget would add yet another variance parameters to the model

$$X \sim N(D\beta, \kappa^2 I + \tau^2 I + \sigma^2 R(\psi))$$

Estimating τ^2 and κ^2

Spatial data usually have no replicates. This implies that the information related to the observational error is usually very scarce. Thus τ^2 is usually either considered known, or a strong informative prior is provided.

ESTIMATING τ^2 AND κ^2

Spatial data usually have no replicates. This implies that the information related to the observational error is usually very scarce. Thus τ^2 is usually either considered known, or a strong informative prior is provided.

The estimation of the microscale variability requires observations that are very close together. As in the case of the observational error, these are usually scarce. So estimating the nugget precisely can be difficult. Moreover, the observational error and the nugget are likely to be confounded.

Note that the presence of the nugget makes the resulting stochastic process non-differentiable.

Bayesian Inference

We consider a vector $X = (X(s_1), \ldots, X(s_m))'$ normally distributed with mean $D\beta$ and covariance matrix $\sigma^2 R(\psi) + \tau^2 I$, where $R(\psi)_{ij} = C(s_i, s_j)$. Thus

$$L(\beta, \psi, \sigma^2, \tau^2) \propto |\sigma^2 R(\psi) + \tau^2 I|^{-1/2}$$

$$\exp\left\{-\frac{1}{2}(X - D\beta)'(\sigma^2 R(\psi) + \tau^2 I)^{-1}(X - D\beta)\right\}$$

Bayesian Inference

We consider a vector $X = (X(s_1), \ldots, X(s_m))'$ normally distributed with mean $D\beta$ and covariance matrix $\sigma^2 R(\psi) + \tau^2 I$, where $R(\psi)_{ij} = C(s_i, s_j)$. Thus

$$L(\beta, \psi, \sigma^2, \tau^2) \propto |\sigma^2 R(\psi) + \tau^2 I|^{-1/2}$$

$$\exp\left\{-\frac{1}{2}(X - D\beta)'(\sigma^2 R(\psi) + \tau^2 I)^{-1}(X - D\beta)\right\}$$

Given a prior distribution $p(\beta, \psi, \sigma^2, \tau^2)$ we can obtain the posterior distribution for all parameters in the model. This is usually done using Monte Carlo methods, as there is no possibility of obtaining conjugate priors for all parameters.

Bayesian Inference

For convenience, parametrize as $\gamma^2 = \tau^2/\sigma^2$, so that $X \sim N(D\beta, \tau^2(I + 1/\gamma^2 R(\psi)))$. Then τ^2 can be factorized in the likelihood.

$$L(\beta, \psi, \tau^2, \gamma^2) \propto |1/\gamma^2 R(\psi) + I|^{-1/2} (\tau^2)^{-m/2}$$

$$\exp\left\{ -\frac{1}{2\tau^2} (X - D\beta)' \left(\frac{1}{\gamma^2} R(\psi) + I \right)^{-1} (X - D\beta) \right\}$$

In this formulation τ^2 and β play the role of the variance and the regression parameters in a linear model with general error covariance.

Posterior Distribution

For simplicity, assume that there is no nugget and no observational error. Let $p(\beta, \psi, \sigma^2) \propto p(\psi) IG(\sigma^2 | a, b)$, where σ^2 is the scale parameter. The posterior distribution is

$$\pi(\beta, \psi, \sigma^2 | X) = \pi(\beta | \sigma^2, \psi, X) \pi(\sigma^2 | \psi, X) \pi(\psi | X)$$

where

Posterior Distribution

For simplicity, assume that there is no nugget and no observational error. Let $p(\beta, \psi, \sigma^2) \propto p(\psi) IG(\sigma^2|a, b)$, where σ^2 is the scale parameter. The posterior distribution is

$$\pi(\beta, \psi, \sigma^2|X) = \pi(\beta|\sigma^2, \psi, X)\pi(\sigma^2|\psi, X)\pi(\psi|X)$$

where

$$p(\beta|\psi, \sigma^2, X) \propto \exp\left\{-\frac{1}{2\sigma^2}\left((\beta - \hat{\beta})' D' R(\psi)^{-1} D(\beta - \hat{\beta})\right)\right\}$$

Posterior Distribution

For simplicity, assume that there is no nugget and no observational error. Let $p(\beta, \psi, \sigma^2) \propto p(\psi) IG(\sigma^2 | a, b)$, where σ^2 is the scale parameter. The posterior distribution is

$$\pi(\beta, \psi, \sigma^2 | X) = \pi(\beta | \sigma^2, \psi, X) \pi(\sigma^2 | \psi, X) \pi(\psi | X)$$

where

$$p(\beta | \psi, \sigma^2, X) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left((\beta - \hat{\beta})' D' R(\psi)^{-1} D(\beta - \hat{\beta}) \right) \right\}$$

$$p(\sigma^2 | \psi, X) \propto \left(\frac{1}{\sigma^2} \right)^{(m-k)/2+a+1} \exp\left\{ -\frac{1}{2\sigma^2} \left(S(\psi)^2 + 2b \right) \right\}$$

Posterior Distribution

For simplicity, assume that there is no nugget and no observational error. Let $p(\beta, \psi, \sigma^2) \propto p(\psi) IG(\sigma^2 | a, b)$, where σ^2 is the scale parameter. The posterior distribution is

$$\pi(\beta, \psi, \sigma^2 | X) = \pi(\beta | \sigma^2, \psi, X) \pi(\sigma^2 | \psi, X) \pi(\psi | X)$$

where

$$p(\beta | \psi, \sigma^2, X) \propto \exp\left\{ -\frac{1}{2\sigma^2} \left((\beta - \hat{\beta})' D' R(\psi)^{-1} D (\beta - \hat{\beta}) \right) \right\}$$

$$p(\sigma^2 | \psi, X) \propto \left(\frac{1}{\sigma^2} \right)^{(m-k)/2+a+1} \exp\left\{ -\frac{1}{2\sigma^2} \left(S(\psi)^2 + 2b \right) \right\}$$

$$p(\psi | X) \propto |R(\psi)|^{-1/2} |D' R(\psi)^{-1} D|^{-1/2} \left(S(\psi)^2 + 2b \right)^{-\frac{m-k}{2} - a} p(\psi)$$

MCMC Block Sampling

The former suggest a simple blocking strategy to sample from the posterior distribution of model parameters. Choose a jumping distribution for g_1 for ψ and define the block jumping distribution as

$$g(\beta, \sigma^2, \psi) = \pi(\beta|\sigma^2, \psi, X)\pi(\sigma^2|\psi, X)g_1(\psi)$$

MCMC Block Sampling

The former suggest a simple blocking strategy to sample from the posterior distribution of model parameters. Choose a jumping distribution for g_1 for ψ and define the block jumping distribution as

$$g(\beta, \sigma^2, \psi) = \pi(\beta|\sigma^2, \psi, X)\pi(\sigma^2|\psi, X)g_1(\psi)$$

Then, if ψ_c is the current sample and ψ_p the proposed one, the acceptance probability is

$$\frac{\pi(\psi_p|X)g(\psi_c)}{\pi(\psi_c|X)g(\psi_p)}$$

MCMC Block Sampling

By performing block sampling we achieve:

- Savings in computing times.
- Better mixing of the Markov chain.

MCMC Block Sampling

By performing block sampling we achieve:

- Savings in computing times.
- Better mixing of the Markov chain.

When ψ_p is rejected, there is no need to sample new β and σ^2. β is usually of small dimension, but obtaining a sample of β involves computations of $R(\psi)$, which can be a very large matrix.

MCMC Block Sampling

By performing block sampling we achieve:

- Savings in computing times.
- Better mixing of the Markov chain.

When ψ_p is rejected, there is no need to sample new β and σ^2. β is usually of small dimension, but obtaining a sample of β involves computations of $R(\psi)$, which can be a very large matrix.

Methods based on slice sampling have recently been proposed. These can be very efficient but they can have two drawbacks: they may be strongly dependent on the prior and they may require many evaluations of the target distribution. But, given the increasing popularity of GPU computations, this may have to be revisited.

Inference for Gaussian Processes

Selecting a prior for ψ can be tricky. For the range parameter with smoothness parameter fixed, consider priors of the type $\pi(\beta, \psi, \sigma^2) \propto \pi(\psi)/(\sigma^2)^a$. Then $\pi(\psi) = 1/\psi$ implies that the posterior is improper for any choice of a. Choosing $\pi(\psi) \propto 1$ gives an improper posterior unless a has a large enough value (see Berger et al. 2001).

Inference for Gaussian Processes

Selecting a prior for ψ can be tricky. For the range parameter with smoothness parameter fixed, consider priors of the type $\pi(\beta, \psi, \sigma^2) \propto \pi(\psi)/(\sigma^2)^a$. Then $\pi(\psi) = 1/\psi$ implies that the posterior is improper for any choice of a. Choosing $\pi(\psi) \propto 1$ gives an improper posterior unless a has a large enough value (see Berger et al. 2001).

The former implies that the prior on ψ has substantial influence on the posterior. There is an infinite amount of posterior mass above any choice of upper limit for ϕ. So, using a uniform distribution over a given interval would not work conceptually.

Reference Prior

Discussions in the literature point at the fact that the information on the range provided by the likelihood is weak. Integrating or fixing the other parameters is required. The reference prior proposed by Berger, De Oliveira and Sansó (BDS) is obtained from the integrated likelihood.

Reference Prior

The BDS prior, has the form $\pi(\beta, \sigma^2, \psi) \propto 1/\sigma^2 \pi^R(\psi)$, where

$$\pi^R(\psi) \propto \left\{ tr[H_\psi^2] - \frac{1}{(n-p)} (tr[H_\psi])^2 \right\}^{1/2},$$

$$H_\psi = \left(\frac{\partial}{\partial \psi} R_\psi \right) R_\psi^{-1} P_\psi^R, \quad P_\psi^R = I - D(D' R_\psi^{-1} D)^{-1} D' R_\psi^{-1}$$

Reference Prior

The resulting marginal posterior for ψ is

$$\pi(\psi|X) \propto |R(\psi)|^{-1/2}|D'R(\psi)^{-1}D|^{-1/2}(S^2(\psi))^{-(m-k)/2}\pi^R(\psi)$$

This posterior is proper, moreover, the marginal $\pi^R(\psi)$ is an integrable function. This fact can be used for model comparison. That is, consider $m_1^R(X)$ and $m_1^R(X)$ two marginals for the sample X obtained using the reference prior corresponding to two different values of a second component of ψ. Then the Bayes factor $m_1^R(X)/m_2^R(X)$ is well calibrated. Thus, a strategy for the estimation of, say, the smoothness parameters of the Matèrn family is to maximize the marginal distribution $m^R(X)$.

Reference Prior

The resulting marginal posterior for ψ is

$$\pi(\psi|X) \propto |R(\psi)|^{-1/2}|D'R(\psi)^{-1}D|^{-1/2}(S^2(\psi))^{-(m-k)/2}\pi^R(\psi)$$

This posterior is proper, moreover, the marginal $\pi^R(\psi)$ is an integrable function. This fact can be used for model comparison. That is, consider $m_1^R(X)$ and $m_1^R(X)$ two marginals for the sample X obtained using the reference prior corresponding to two different values of a second component of ψ. Then the Bayes factor $m_1^R(X)/m_2^R(X)$ is well calibrated. Thus, a strategy for the estimation of, say, the smoothness parameters of the Matèrn family is to maximize the marginal distribution $m^R(X)$.

Posterior intervals based on the BDS prior are likely to provide better coverage than other non-informative priors as well as methods based on maximum likelihood.

Priors for Other Parameters

We can elicit a prior for the ratio of nugget variance to total variance, $(1 + 1/\gamma^2)^{-1}$. We have found that the prior for γ^2 has a strong influence. The prior for τ^2 does not.

Priors for Other Parameters

We can elicit a prior for the ratio of nugget variance to total variance, $(1+1/\gamma^2)^{-1}$. We have found that the prior for γ^2 has a strong influence. The prior for τ^2 does not.

Conjecture: If $\pi(\gamma^2)$ is proper, $\pi(\beta, \gamma^2, \tau^2, \psi) \propto \pi(\gamma^2)\pi^R(\psi)/\tau^2$ produces a proper posterior.

Priors for Other Parameters

We can elicit a prior for the ratio of nugget variance to total variance, $(1 + 1/\gamma^2)^{-1}$. We have found that the prior for γ^2 has a strong influence. The prior for τ^2 does not.

Conjecture: If $\pi(\gamma^2)$ is proper, $\pi(\beta, \gamma^2, \tau^2, \psi) \propto \pi(\gamma^2)\pi^R(\psi)/\tau^2$ produces a proper posterior.

Eliciting a prior for the smoothness parameter in the Matern family is usually easier as the data usually provide substantial information about it. A possible strategy is to use point masses on, for example, .5; 1; 1.5; 2.5; 3.5, whose correlation functions have simple functional forms.

Calculating π^R

Unfortunately the formula for π^R is numerically unstable for large values of the range parameter. A formula that depends on higher order terms is preferred.

Let

$$B_\psi = \frac{1}{\nu(\psi)}(R_\psi - \mathbb{1}\mathbb{1}'), \quad P_\psi = I - D(D'B_\psi^{-1}D)^{-1}D'B_\psi^{-1},$$

$$Q_\psi = \frac{1}{\nu'(\psi)}\left(\frac{\partial}{\partial\psi}R_\psi\right) - B_\psi$$

then

$$\pi^R(\psi) \propto \left|\frac{\nu'(\psi)}{\nu(\psi)}\right| \left\{\mathrm{tr}([Q_\psi B_\psi^{-1} P_\psi]^2) - \frac{1}{m-k}[\mathrm{tr}(Q_\psi B_\psi^{-1} P_\psi]^2\right\}^{1/2}$$

CALCULATING π^R

For the Matèrn family, let ψ_1 denote the range and ψ_2 denote the smoothness parameter, then we have that

- $0 < \psi_2 < 1$, then $\nu(\psi_1) = \psi_1^{-2\psi_2}$
- $\psi_2 = 1$, then $\nu(\psi_1) = \frac{\log \psi_1}{\psi_1^2}$
- $\psi_2 > 1$, then $\nu(\psi_1) = \psi_1^{-2}$

For the spherical family, $\nu(\psi) = \psi^{-1}$. For the power exponential $\nu(\psi_1) = \psi_1^{-\psi_2}$, for the rational quadratic $\nu(\psi) = \psi^{-2}$.

Predictive Inference

Consider Z as the vector that corresponds to a new set of locations. Assume that the joint distribution of Z and X is

$$\begin{pmatrix} Z \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} D_Z \\ D_X \end{pmatrix}\beta, \sigma^2 \begin{pmatrix} R_Z(\psi) & R_{ZX}(\psi) \\ R_{XZ}(\psi) & R_X(\psi) \end{pmatrix}\right)$$

Usually Z corresponds to a grid of locations used to interpolate the random field.

Predictive Inference

We can use the posterior predictive distribution of Z given X to make inference on Z. In particular, a point estimator is given by

$$\int\int\int [D_Z\beta + R_{ZX}(\psi)R_X^{-1}(\psi)(X - D_X\beta)]\pi(\beta, \sigma^2, \psi|X)d\beta d\sigma^2 d\psi$$

which is the predictive expectation of Z given X. Notice that $E(X|X) = X$, and so the spatial estimation is actually interpolating the observations.