Mathematical Details for NPVecchia package

Brian Kidd[*] Matthias Katzfuss[*†]

1 Methodology

1.1 Notation

All of these will be introduced in the text, but for a quick reference they are listed here.

n : the number of (spatial) locations with observations

N : the number of observations per location

m : the number of neighbors for calculation

α_i : the shape parameter for the IG prior in the i'th regression

β_i : the scale parameter for the IG prior in the i'th regression

a_i, b_i : posterior IG parameters

$\mathbf{\Gamma}_i$: the prior variance on the coefficients in the i'th regression

$\mathbf{G}_i$: posterior variance

$\hat{\mathbf{U}}$: Cholesky of the precision matrix

1.2 A spatial model and the screening effect

Assume we have $N \geq 1$ observations of a continuous spatial process at n locations (in low dimensional space). We model the detrended (i.e., centered) data as

$$\mathbf{z}^{(\ell)}|\mathbf{\Sigma} \stackrel{iid}{\sim} \mathcal{N}_n(\mathbf{0}, \mathbf{\Sigma}), \qquad \ell = 1, \ldots, N, \tag{1}$$

where $\mathbf{z}^{(\ell)} = (z_1^{(\ell)}, \ldots, z_n^{(\ell)})'$, and $z_i^{(\ell)}$ is observed at spatial location $\mathbf{s}_i$. We denote by $\mathbf{z}$ all observations $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(N)}$ stacked into a long vector. We assume that the locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and hence the corresponding variables $z_i^{(\ell)}$ in $\mathbf{z}^{(\ell)}$, are ordered according to a maximin ordering (Guinness, 2018; Schäfer et al., 2017).

Our goal is to make inference on the spatial covariance matrix $\mathbf{\Sigma}$ based on the data $\mathbf{z}$, in the case where n is large (in the hundreds or even hundreds of thousands) and N is relatively small. Typically, a parametric, and often isotropic, covariance function is assumed

[*]Department of Statistics, Texas A&M University

[†]Corresponding author: `katzfuss@gmail.com`

to determine $\mathbf{\Sigma}$ such that it only is a function of a very small number of parameters, which can then be estimated relatively easily. Here, we avoid explicit assumptions of stationarity and isotropy.

Instead, we assume that a spatial screening effect holds, such that

$$p(z_i^{(\ell)}|\mathbf{z}_{1:i-1}^{(\ell)}, \mathbf{\Sigma}) = p(z_i^{(\ell)}|\mathbf{z}_{g_m(i)}^{(\ell)}, \mathbf{\Sigma}), \tag{2}$$

where $g_m(i) \subset (1, \ldots, i-1)$ is an index vector consisting of the indices of the $\min(m, i-1)$ nearest neighbors to $\mathbf{s}_i$ among those ordered previously; that is, $\mathbf{s}_{(g_m(i))_j}$ is the jth nearest neighbor of $\mathbf{s}_i$. The equation (2) always holds trivially holds for $m = n-1$, but for many covariances, it even holds (at least approximately) for $m \ll n$ due to the so-called screening effect. Assume for now that m is fixed and known.

Consider the modified Cholesky decomposition of the inverse of $\mathbf{\Sigma}$ (i.e., the precision matrix):

$$\mathbf{\Sigma}^{-1} = \mathbf{U}\mathbf{D}^{-1}\mathbf{U}', \tag{3}$$

where $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ is a diagonal matrix with positive entries $d_i > 0$, and $\mathbf{U}$ is an upper triangular matrix with unit diagonal (i.e., $\mathbf{U}_{ii} = 1$). The screening effect in (2) implies that $\mathbf{U}$ is sparse, with at most m nonzero off-diagonal elements per column (e.g., Katzfuss and Guinness, 2017, Prop. 3.1). We define $\mathbf{u}_i = \mathbf{U}_{g_i,i}$ as the nonzero off-diagonal entries in the ith column.

1.3 Inference conditional on hyperparameters $\boldsymbol{\theta}$

From (3), we see that we can estimate $\mathbf{\Sigma}$ by inferring $d_1, \ldots, d_n$ and $\mathbf{u}_1, \ldots, \mathbf{u}_n$. To do so, note that our data model (1) can be written as a series of linear regression models (Huang et al., 2006):

$$p(\mathbf{z}|\mathbf{\Sigma}) = \prod_{i=1}^{n} p(\mathbf{y}_i|\mathbf{y}_{1:i-1}, \mathbf{\Sigma}) = \prod_{i=1}^{n} \mathcal{N}_N(\mathbf{y}_i|\mathbf{X}_i\mathbf{u}_i, d_i\mathbf{I}_N), \tag{4}$$

where $\mathbf{y}_i = (z_i^{(1)}, \ldots, z_i^{(N)})'$, and $\mathbf{X}_i$ is an $N \times m$ matrix with ℓth row $-\mathbf{z}_{g_i}^{(\ell)\prime}$. Note the negative sign for the entries of $\mathbf{X}_i$. Further details on why this and (3) are pushed to Section 2.

For the regression models in (4), we assume the standard, conjugate priors to form a series of Bayesian regression models:

$$\mathbf{u}_i|d_i, \boldsymbol{\theta} \overset{ind.}{\sim} \mathcal{N}(\mathbf{0}, d_i\mathbf{\Gamma}_i), \qquad d_i|\boldsymbol{\theta} \overset{ind.}{\sim} \mathcal{IG}(\alpha_i, \beta_i),$$

where $\boldsymbol{\theta}$ is a vector of hyperparameters determining m, $\mathbf{\Gamma}_i$, α_i, and β_i, which will be discussed further in Section 1.4 below.

Due to conjugacy, the posterior distribution (conditional on $\boldsymbol{\theta}$) is available in closed form:

$$\begin{aligned} p(\mathbf{u}_1, \ldots, \mathbf{u}_n, d_1, \ldots, d_n|\mathbf{z}, \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(\mathbf{u}_i, d_i|\mathbf{z}, \boldsymbol{\theta}) = \prod_{i=1}^{n} p(\mathbf{u}_i|d_i, \mathbf{z}, \boldsymbol{\theta})\, p(d_i|\mathbf{z}, \boldsymbol{\theta}) \\ &= \prod_{i=1}^{n} \mathcal{N}(\mathbf{u}_i|\hat{\mathbf{u}}_i, d_i\mathbf{G}_i)\, \mathcal{IG}(d_i|a_i, b_i), \end{aligned} \tag{5}$$

where $\hat{\mathbf{u}}_i = \mathbf{G}_i\mathbf{X}_i'\mathbf{y}_i$, $\mathbf{G}_i = (\mathbf{X}_i'\mathbf{X}_i + \boldsymbol{\Gamma}_i^{-1})^{-1}$, $a_i = \alpha_i + N/2$, and $b_i = \beta_i + (\mathbf{y}_i'(\mathbf{I}_N + \mathbf{X}_i\boldsymbol{\Gamma}_i\mathbf{X}_i')^{-1}\mathbf{y}_i)/2 = \beta_i + (\mathbf{y}_i'\mathbf{y}_i - \hat{\mathbf{u}}_i'\mathbf{G}_i^{-1}\hat{\mathbf{u}}_i')/2$.

Using (5), we can easily obtain samples or posterior summaries of the entries of $\mathbf{U}$ and $\mathbf{D}$ conditional on $\boldsymbol{\theta}$. However, in many applications, primary interest will be in computing posterior summaries of $\boldsymbol{\Sigma}$ and other quantities. If n is not too large ($n < 10^4$, say), we can simply compute $\boldsymbol{\Sigma}^{-1}$ and hence $\boldsymbol{\Sigma}$ from $\mathbf{U}$ and $\mathbf{D}$. For large n, it is usually not possible to even hold the entire dense matrix $\boldsymbol{\Sigma}$ in memory, but we can quickly compute useful summaries of it based on $\mathbf{U}$ and $\mathbf{D}$.

In many applications, including climate-model emulation, it is of interest to sample new spatial fields from the model. We can sample $\mathbf{z}^\star \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ by sampling $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and then setting $\mathbf{z}^\star = (\mathbf{U}\mathbf{D}^{-1/2})^{-1}\mathbf{w}$. If $\mathbf{U}$ and $\mathbf{D}$ are sampled from their posterior distribution given $\mathbf{z}$, then we have obtained a sample from the posterior predictive distribution $p(\mathbf{z}^\star|\mathbf{z})$.

1.4 Inference on the hyperparameters $\boldsymbol{\theta}$

Previously, we have assumed the hyperparameters $\boldsymbol{\theta}$ determining m, $\boldsymbol{\Gamma}_i$, α_i, and β_i to be fixed. We now discuss the inference of these hyperparameters.

First, assuming a hyperprior $p(\boldsymbol{\theta})$ has been specified, the goal is to obtain the posterior distribution $p(\boldsymbol{\theta}|\mathbf{z}) \propto p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta})$. While this distribution cannot be obtained analytically, we can sample from the posterior using the Metropolis-Hastings algorithm using the closed form of the marginal or integrated likelihood,

$$p(\mathbf{z}|\boldsymbol{\theta}) \propto \prod_{i=1}^{n} \sqrt{|\mathbf{G}_i|/|\boldsymbol{\Gamma}_i|} \times \beta_i^{\alpha_i}/b_i^{a_i} \times \Gamma(a_i)/\Gamma(\alpha_i),$$

where the (non-bold) Γ denotes the gamma function. Given the posterior distributions of $\mathbf{U}, \mathbf{D}$, these evaluations are cheap computationally. Another alternative is to optimize these hyperparameters with this likelihood.

We now parameterize the prior distributions from Section 1.3 in terms of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)'$, such that the resulting model shrinks toward an isotropic Matérn-type covariance. The parameter θ_1 will play the role of a marginal variance, while θ_2 and θ_3 are related to the range and smoothness. For the package, we concatenate the prior parameters α_i, β_i into n-dimensional vectors a, b, and the prior variance parameter $\boldsymbol{\Gamma}_i$ (which is diagonal) into a matrix $\mathbf{G}$, with each row of $\mathbf{G} =$ temp, of dimension n by m as follows:

$$\begin{aligned}
a &= 6 \\
b &= 5e^{\theta_1}\left[1 - \exp\left(-\frac{e^{\theta_2}}{\sqrt{0:(n-1)}}\right)\right] \\
\text{temp} &= \exp\left(-e^{\theta_3} * (1:m)\right) \\
\text{each row of } \mathbf{G} &= \frac{\text{temp}}{b_i/(a_i-1)}
\end{aligned}$$

For the method, we also provide a guideline for choosing m. Our solution is to tie m to the decay of the elements of $\mathbf{U}$. To allow the data to choose m within the MCMC algorithm or optimization, we deterministically link the number of neighbors to θ_3 (for our experiments we use $\exp(\theta_3 * j) < 0.001$, where j denotes the neighbor number). This coincides to the amount of variation expected to be learnable from the data. By allowing m to change within the MCMC, an incorrect m will negatively influence the integrated likelihood so the data can reject it.

2 Why (3) and (4) hold

This section is based on Section 2.2.4 of (Pourahmadi, 2011).

First consider an autoregressive model, then move all elements to the same side.

$$\mathbf{y}_i = \sum_{j \in g_i} \phi_{ij} z_j + \epsilon_i$$
$$\mathbf{y}_i - \sum_{j \in g_i} \phi_{ij} z_j = \epsilon_i$$

Now, it can be written in matrix form as $\epsilon = T\mathbf{X}$, where

$$T = \begin{pmatrix} 1 & & & & \\ -\phi_{21} & 1 & & & \\ -\phi_{31} & -\phi_{32} & 1 & & \\ \dots & & & \dots & \\ -\phi_{n1} & -\phi_{n2} & \cdots & -\phi_{nn-1} & 1 \end{pmatrix}$$

However, for notational simplicity, we absolve the negative sign into the coefficient matrix $\mathbf{X}$ Now, to see that it is indeed the valid covariance function:

$$cov(\epsilon) = D^2 = cov(TY) = T\Sigma T'$$
$$\Sigma = T^{-1} D^2 T'^{-1}$$
$$\Sigma^{-1} = T' D^{-2} T$$

References

Guinness, J. (2018). Permutation methods for sharpening Gaussian process approximations. *Technometrics*, 60(4):415–429.

Huang, J., Liu, N., Pourahmadi, M., and Liu, L. (2006). Covariance matrix selection and estimation via penalized normal likelihood. *Biometrika*, 93(1):85–98.

Katzfuss, M. and Guinness, J. (2017). A general framework for Vecchia approximations of Gaussian processes. *arXiv:1708.06302*.

Pourahmadi, M. (2011). Covariance estimation: The glm and regularization perspectives. *Statistical Science*, pages 369–387.

Schäfer, F., Sullivan, T. J., and Owhadi, H. (2017). Compression, inversion, and approximate PCA of dense kernel matrices at near-linear computational complexity. *arXiv:1706.02205*.