

Scalable methods for large spatial data: Low rank predictive processes

Abhi Datta

Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland

abhidatta.com

[@dattascience](https://twitter.com/dattascience)

Multivariate Gaussian likelihoods for geostatistical models

- $\mathcal{S} = \{s_1, s_2, \dots, s_n\}$ are locations where data is observed
- $y(s_i)$ is outcome at the i -th location,
 $y = (y(s_1), y(s_2), \dots, y(s_n))^{\top}$
- Model: $y \sim N(X\beta, C_{\theta})$
- Estimating process parameters from the likelihood:
$$-\frac{1}{2} \log \det(C_{\theta}) - \frac{1}{2} (y - X\beta)^{\top} C_{\theta}^{-1} (y - X\beta)$$
- C_{θ} is usually **dense** with no exploitable structure
- Bayesian inference: Priors on $\{\beta, \theta\}$
- Need to calculate **$\det(C_{\theta})$** and **quadratic forms of C_{θ}^{-1}**

- Conditional predictive density

$$p(y(s_0) | y, \theta, \beta) = N \left(y(s_0) \mid \mu(s_0), \sigma^2(s_0) \right) .$$

- Kriging

$$\begin{aligned} \mu(s_0) &= E[y(s_0) | y, \theta] = x^\top(s_0)\beta + c_\theta^\top(s_0)C_\theta^{-1}(y - X\beta) , \\ \sigma^2(s_0) &= \text{var}[y(s_0) | y, \theta] = C_\theta(s_0, s_0) - c_\theta^\top(s_0)C_\theta^{-1}c_\theta(s_0) . \end{aligned}$$

- Again need to evaluate **quadratic forms of C_θ^{-1}**

Computational Details

- **Cholesky decomposition**: Any symmetric matrix A can be factorized as $A = LDL^\top$ where L is **lower triangular** and D is **diagonal**
- Both $\det(C_\theta)$ and quadratic forms of C_θ^{-1} are best obtained via Cholesky decomposition of C_θ

Cholesky:	$\text{chol}(C_\theta) = LDL^\top ;$
Determinant:	$\det(C_\theta) = \prod_{i=1}^n d_{ii} ;$
Quadratic forms $a' C_\theta^{-1} b$	$v = \text{trsolve}(L, a) ;$ $C_\theta^{-1} a = u = \text{trsolve}(L^\top, D^{-1} v) ;$ $a' C_\theta^{-1} b = u^\top b ;$

- Primary **bottleneck** is $\text{chol}(\cdot)$ requiring $O(n^2)$ **storage** and $O(n^3)$ **memory**
- **Not feasible** for large n

Burgeoning literature on spatial big data

- Low-rank models (Wahba, 1990; Higdon, 2002; Kamman & Wand, 2003; Paciorek, 2007; Rasmussen & Williams, 2006; Stein 2007, 2008; Cressie & Johannesson, 2008; Banerjee et al., 2008; 2010; Gramacy & Lee 2008; Sang et al., 2011, 2012; Lemos et al., 2011; Guhaniyogi et al., 2011, 2013; Salazar et al., 2013; Katzfuss, 2016)
- Spectral approximations and composite likelihoods: (Fuentes 2007; Paciorek, 2007; Eidsvik et al. 2016)
- Multi-resolution approaches (Nychka, 2002; Johannesson et al., 2007; Matsuo et al., 2010; Tzeng & Huang, 2015; Katzfuss, 2016)
- Sparsity: (Solve $Ax = b$ by (i) sparse A , or (ii) sparse A^{-1})
 1. Covariance tapering (Furrer et al. 2006; Du et al. 2009; Kaufman et al., 2009; Shaby and Ruppert, 2013)
 2. GMRFs to GPs: INLA (Rue et al. 2009; Lindgren et al., 2011)
 3. LAGP (Gramacy et al. 2014; Gramacy and Apley, 2015)
 4. Nearest-neighbor models (Vecchia 1988; Stein et al. 2004; Stroud et al 2014; Datta et al., 2016)

Heaton, M.J., Datta, A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R.B., Hammerling, D., Katzfuss, M. and Lindgren, F., 2018. *A case study competition among methods for analyzing large spatial data*. Journal of Agricultural, Biological and Environmental Statistics, pp.1-28.

Bayesian low rank models

- A *low rank* or *reduced rank* process approximates a *parent* process over a smaller set of points (*knots*).
- Start with a *parent process* $w(s)$ and construct $\tilde{w}(s)$

$$w(s) \approx \tilde{w}(s) = \sum_{j=1}^r b_{\theta}(s, s_j^*) z(s_j^*) = b_{\theta}^{\top}(s) z,$$

where

- $z(s)$ is *any* well-defined process (could be same as $w(s)$);
- $b_{\theta}(s, s')$ is a family of basis functions indexed by parameters θ ;
- $\{s_1^*, s_2^*, \dots, s_r^*\}$ are the knots;
- $b_{\theta}(s)$ and z are $r \times 1$ vectors with components $b_{\theta}(s, s_j^*)$ and $z(s_j^*)$, respectively.

Bayesian low rank models (contd.)

- $\tilde{w} = (\tilde{w}(s_1), \tilde{w}(s_2), \dots, \tilde{w}(s_n))^T$ is represented as $\tilde{w} = B_\theta z$
- B_θ is $n \times r$ with (i, j) -th element $b_\theta(s_i, s_j^*)$
- Irrespective of how big n is, we now have to work with the r (instead of n) $z(s_j^*)$'s and the $n \times r$ matrix B_θ .
- Since $r \ll n$, the consequential dimension reduction is evident.
- \tilde{w} is a valid stochastic process in r -dimensions space with covariance:

$$\text{cov}(\tilde{w}(s), \tilde{w}(s')) = b_\theta^\top(s) V_z b_\theta(s'),$$

where V_z is the variance-covariance matrix (also depends upon parameter θ) for z .

- When $n > r$, the joint distribution of \tilde{w} is singular (hence the name **fixed-rank**).

The Sherman-Woodbury-Morrison formulas

- Low-rank dimension reduction is similar to Bayesian linear regression
- Consider a simple hierarchical model (with $\beta = 0$):

$$N(z \mid 0, V_z) \times N(y \mid B_\theta z, D_\tau) ,$$

where y is $n \times 1$, z is $r \times 1$, D_τ and V_z are positive definite matrices of sizes $n \times n$ and $r \times r$, respectively, and B_θ is $n \times r$.

- The low rank specification is $B_\theta z$ and the prior on z .
- D_τ (usually diagonal) has the residual variance components.
- Computing $\text{var}(y)$ in two different ways yields

$$(D_\tau + B_\theta V_z B_\theta^\top)^{-1} = D_\tau^{-1} - D_\tau^{-1} B_\theta (V_z^{-1} + B_\theta^\top D_\tau^{-1} B_\theta)^{-1} B_\theta^\top D_\tau^{-1} .$$

- A companion formula for the determinant:

$$\det(D_\tau + B_\theta V_z B_\theta^\top) = \det(V_z) \det(D_\tau) \det(V_z^{-1} + B_\theta^\top D_\tau^{-1} B_\theta) .$$

Predictive process models (Banerjee et al., *JRSS-B*, 2008)

- A particular low-rank model emerges by taking
 - $z(s) = w(s)$
 - $z = (w(s_1^*), w(s_2^*), \dots, w(s_r^*))^\top$ as the realizations of the parent process $w(s)$ over the set of knots $\mathcal{S}^* = \{s_1^*, s_2^*, \dots, s_r^*\}$,

and then taking the conditional expectation:

$$\tilde{w}(s) = E[w(s) \mid w^*] = b_\theta^\top(s)z .$$

Predictive process models (Banerjee et al., *JRSS-B*, 2008)

- A particular low-rank model emerges by taking
 - $z(s) = w(s)$
 - $z = (w(s_1^*), w(s_2^*), \dots, w(s_r^*))^\top$ as the realizations of the parent process $w(s)$ over the set of knots $\mathcal{S}^* = \{s_1^*, s_2^*, \dots, s_r^*\}$,

and then taking the conditional expectation:

$$\tilde{w}(s) = E[w(s) \mid w^*] = b_\theta^\top(s)z .$$

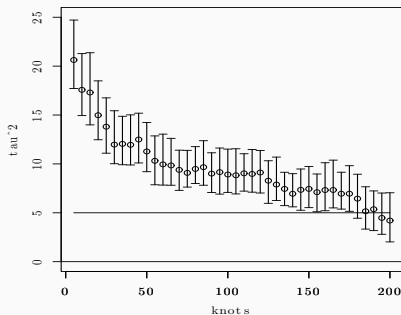
- The basis functions are *automatically* derived from the spatial covariance structure of the parent process $w(s)$:

$$b_\theta^\top(s) = \text{cov}\{w(s), w^*\} \text{var}^{-1}\{w^*\} = c_\theta(s, \mathcal{S}^*) C_\theta^{-1}(\mathcal{S}^*, \mathcal{S}^*) .$$

Biases in low-rank models

- For the predictive process,

$$\begin{aligned}\text{var}\{w(s)\} &= \text{var}\{E[w(s) \mid w^*]\} + E\{\text{var}[w(s) \mid w^*]\} \\ &\geq \text{var}\{E[w(s) \mid w^*]\} = \text{var}(\tilde{w}(s)) .\end{aligned}$$



- Leads to **overestimation** of the nugget

Bias-adjusted or modified predictive processes

- Note that $w(s) = \tilde{w}(s) + \eta(s)$. In low-rank processes, the residual process $\eta(s)$ is ignored.
- $\eta(s)$ is a Gaussian process with covariance structure

$$\begin{aligned}\text{Cov}\{\eta(s), \eta(s')\} &= K_{\eta, \theta}(s, s') \\ &= C_{\theta}(s, s') - c_{\theta}(s, \mathcal{S}^*) C_{\theta}^{-1}(\mathcal{S}^*, \mathcal{S}^*) c_{\theta}(\mathcal{S}^*, s') .\end{aligned}$$

- Remedy: **low-rank + sparse approximations**

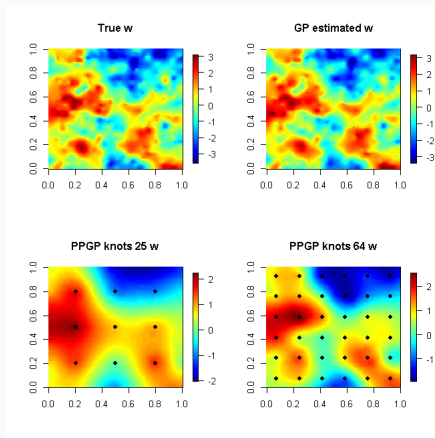
$$\tilde{w}_{\epsilon}(s) = \tilde{w}(s) + \tilde{\epsilon}(s) ,$$

where $\tilde{\epsilon}(s) \stackrel{\text{ind}}{\sim} N(0, \delta^2(s))$ and

$$\delta^2(s) = \text{var}\{\eta(s)\} = C_{\theta}(s, s) - c_{\theta}(s, \mathcal{S}^*) C_{\theta}^{-1}(\mathcal{S}^*, \mathcal{S}^*) c_{\theta}(\mathcal{S}^*, s) .$$

- Other improvements suggested by Sang et al. (2011, 2012) and Katzfuss (2017).

Oversmoothing in low rank models



Low rank models **oversmooths** unless we use more knots which becomes computationally expensive