Low-Rank and Predictive Process Models

Abhirup Datta, Andrew O. Finley and Sudipto Banerjee

University of California, Los Angeles, USA

Multivariate Gaussian likelihoods for geostatistical models

- $\mathcal{L} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ are locations where data is observed
- $y(\ell_i)$ is outcome at the i-th location, $y = (y(\ell_1), y(\ell_2), \ldots, y(\ell_n))^\top$
- Model: $y \sim N(X\beta, K_\theta)$
- Estimating process parameters from the likelihood:

$$-\frac{1}{2}\log\det(K_\theta) - \frac{1}{2}(y - X\beta)^\top K_\theta^{-1}(y - X\beta)$$

- K_θ is usually dense with no exploitable structure
- Bayesian inference: Priors on $\{\beta, \theta\}$
- Challenges: Storage and $\texttt{chol}(K_\theta) = LDL^\top$.

Prediction and interpolation

- Conditional predictive density

$$p(y(\ell_0)\,|\,y,\theta,\beta) = N\left(y(\ell_0)\,\middle|\mu(\ell_0),\sigma^2(\ell_0)\right)\ .$$

- “Kriging” (spatial prediction/interpolation)

$$\begin{aligned}\mu(\ell_0) = \mathrm{E}[y(\ell_0)\,|\,y,\theta] &= x^\top(\ell_0)\beta + k_\theta^\top(\ell_0)K_\theta^{-1}(y - X\beta)\ ,\\ \sigma^2(\ell_0) = \mathrm{var}[y(\ell_0)\,|\,y,\theta] &= K_\theta(\ell_0,\ell_0) - k_\theta^\top(\ell_0)K_\theta^{-1}k_\theta(\ell_0)\ .\end{aligned}$$

- Bayesian “kriging” computes (simulates) posterior predictive density:

$$p(y(\ell_0)\,|\,y) = \int p(y(\ell_0)\,|\,y,\theta,\beta)p(\beta,\theta\,|\,y)\mathrm{d}\beta\mathrm{d}\theta$$

Computational Details

- Compute the mean and variance (for any given $\{\beta, \theta\}$ and ℓ_0):

$$\begin{array}{ll} \text{Solve for } u: & K_\theta u = k_\theta(\ell_0)\ ; \\ \text{Predictive mean:} & x^{\mathrm{T}}(\ell_0)\beta + u^\top(y - X\beta)\ ; \\ \text{Predictive variance:} & K_\theta(\ell_0, \ell_0) - u^\top k_\theta(\ell_0)\ . \end{array}$$

- Compute the mean and variance (for any given $\{\beta, \theta\}$ and ℓ_0):

$$\begin{array}{ll} \text{Cholesky:} & \texttt{chol}(K_\theta) = LDL^\top\ ; \\ \text{Solve for } v: & v = \texttt{trsolve}(L, k_\theta(\ell_0))\ ; \\ \text{Solve for } u: & u = \texttt{trsolve}(L^\top, D^{-1}v)\ ; \\ \text{Predictive mean:} & x^{\mathrm{T}}(\ell_0)\beta + u^\top(y - X\beta)\ ; \\ \text{Predictive variance:} & K_\theta(\ell_0, \ell_0) - u^\top k_\theta(\ell_0)\ . \end{array}$$

- Primary bottleneck is chol$(\cdot)$

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)

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* $w(\ell)$ and construct $\tilde{w}(\ell)$

$$w(\ell) \approx \tilde{w}(\ell) = \sum_{j=1}^{r} b_\theta(\ell, \ell_j^*) z(\ell_j^*) = b_\theta^{\mathrm{T}}(\ell) z,$$

  where

  - $z(\ell)$ is *any* well-defined process (could be same as $w(\ell)$);
  - $b_\theta(\ell, \ell')$ is a family of basis functions indexed by parameters θ;
  - $\{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$ are the knots;
  - $b_\theta(\ell)$ and z are $r \times 1$ vectors with components $b_\theta(\ell, \ell_j^*)$ and $z(\ell_j^*)$, respectively.

Bayesian low rank models (contd.)

- $\tilde{w} = (\tilde{w}(\ell_1), \tilde{w}(\ell_2), \ldots, \tilde{w}(\ell_n))^{\mathrm{T}}$ is represented as $\tilde{w} = B_\theta z$
- B_θ is $n \times r$ with (i, j)-th element $b_\theta(\ell_i, \ell_j^*)$
- Irrespective of how big n is, we now have to work with the r (instead of n) $z(\ell_j^*)$'s and the $n \times r$ matrix B_θ.
- Since $r << n$, the consequential dimension reduction is evident.
- $\tilde{w}$ is a valid stochastic process in r-dimensions space with covariance:

  $$\mathrm{cov}(\tilde{w}(\ell), \tilde{w}(\ell')) = b_\theta^{\mathrm{T}}(\ell) V_z b_\theta(\ell') \ ,$$

  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.

The Sherman-Woodbury-Morrison formulas

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

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

  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

  $$(D_\tau + B_\theta V_z B_\theta^{\mathrm{T}})^{-1} = D_\tau^{-1} - D_\tau^{-1} B_\theta (V_z^{-1} + B_\theta^{\mathrm{T}} D_\tau^{-1} B_\theta)^{-1} B_\theta^{\mathrm{T}} D_\tau^{-1}\ .$$
- A companion formula for the determinant:

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

Practical implementation for Bayesian low rank models

- In practical implementation, better to avoid SWM formulas.

$$\underbrace{\begin{bmatrix} D_\tau^{-1/2} y \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} D_\tau^{-1/2} B_\theta \\ V_z^{-1/2} \end{bmatrix}}_{B_*} z + \underbrace{\begin{bmatrix} e_1 \\ e_2 \end{bmatrix}}_{e_*} .$$

- $e_* \sim N(0, I_{n+r})$.
- $V_z^{1/2}$ and $D_\tau^{1/2}$ are matrix square roots of of V_z and D_τ, respectively.
- If D_τ is diagonal (as is common), then $D_\tau^{1/2}$ is simply the square root of the diagonal elements of D_τ.
- $V_z^{1/2} = \texttt{chol}(V_z)$ is the triangular (upper or lower) Cholesky factor of the $r \times r$ matrix V_z.
- Use `backsolve` to efficiently obtain $V_z^{-1/2} z$

Practical implementation for Bayesian low rank models (contd.)

- The marginal density of $p(y_* \,|\, \theta, \tau)$ after integrating out z now corresponds to the normal linear model

  $$y_* = B_* \hat{z} + e_* \ ,$$

  where $\hat{z}$ is the ordinary least-square estimate of z.

- Use `lm` function to compute $\hat{z}$ applying the QR decomposition to B_*.

- Thus, we estimate the Bayesian linear model

  $$p(\theta, \tau) \times N(y_* \,|\, B_* \hat{z}, I_{n+r})$$

- MCMC will generate posterior samples for $\{\theta, \tau\}$.

- *Recover* the posterior samples for z from those of $\{\theta, \tau\}$:

  $$p(z \,|\, y) = \int N(z \,|\, \hat{z}, M) \times p(\theta, \tau \,|\, y) \mathrm{d}\theta \mathrm{d}\tau$$

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

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

- A particular low-rank model emerges by taking
  - $z(\ell) = w(\ell)$
  - $z = (w(\ell_1^*), w(\ell_2^*), \ldots, w(\ell_r^*))^\top$ as the realizations of the parent process $w(\ell)$ over the set of knots $\mathcal{L}^* = \{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$,

  and then taking the conditional expectation:

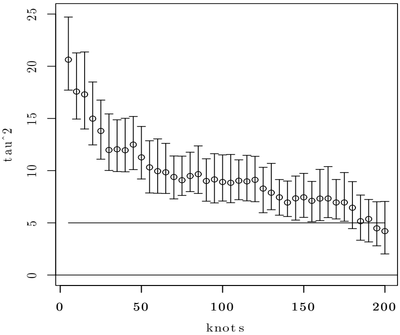
$$\tilde{w}(\ell) = \mathrm{E}[w(\ell) \,|\, w^*] = b_\theta^\top(\ell) z \; .$$

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

$$b_\theta^\top(\ell) = \mathrm{cov}\{w(\ell), w^*\}\mathrm{var}^{-1}\{w^*\} = K_\theta(\ell, \mathcal{L}^*)K_\theta^{-1}(\mathcal{L}^*, \mathcal{L}^*) \; .$$

Biases in low-rank models

- In low-rank processes, $w(\ell) = \tilde{w}(\ell) + \eta(\ell)$. What is lost in $\eta(\ell)$?



- For the predictive process,

$$\mathrm{var}\{w(\ell)\} = \mathrm{var}\{\mathrm{E}[w(\ell)\,|\,w^*]\}+\mathrm{E}\{\mathrm{var}[w(\ell)\,|\,w^*]\} \geq \mathrm{var}\{\mathrm{E}[w(\ell)\,|\,w^*]\}\ .$$

Bias-adjusted or modified predictive processes

- $\eta(\ell)$ is a Gaussian process with covariance structure

$$\begin{aligned}\text{Cov}\{\eta(\ell), \eta(\ell')\} &= K_{\eta,\theta}(\ell, \ell') \\ &= K_{\theta}(\ell, \ell') - K_{\theta}(\ell, \mathcal{L}^*) K_{\theta}^{-1}(\mathcal{L}^*, \mathcal{L}^*) K_{\theta}(\mathcal{L}^*, \ell') \; .\end{aligned}$$

- Remedy:

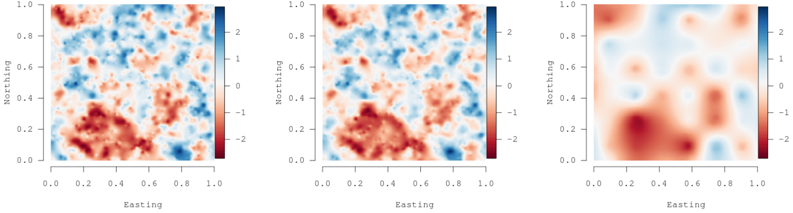
$$\tilde{w}_{\epsilon}(\ell) = \tilde{w}(\ell) + \tilde{\epsilon}(\ell) \; ,$$

  where $\tilde{\epsilon}(\ell) \overset{ind}{\sim} N(0, \delta^2(\ell))$ and

$$\delta^2(\ell) = \text{var}\{\eta(\ell)\} = K_{\theta}(\ell, \ell) - K_{\theta}(\ell, \mathcal{L}^*) K_{\theta}^{-1}(\mathcal{L}^*, \mathcal{L}^*) K_{\theta}(\mathcal{L}^*, \ell) \; .$$

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

Oversmoothing in low rank models



(a) True w

(b) Full GP

(c) PPGP 64 knots

Figure: Comparing full GP vs low-rank GP with 2500 locations. Figure (1(c)) exhibits oversmoothing by a low-rank process (predictive process with 64 knots)