High-dimensional Bayesian Geostatistics

Sudipto Banerjee

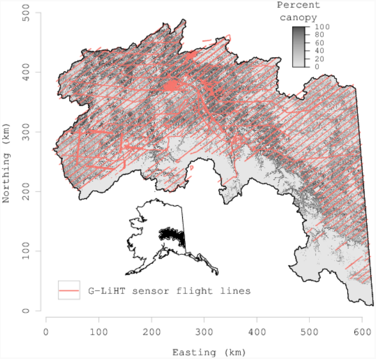
August 15, 2017

Department of Biostatistics, Fielding School of Public Health, University of California, Los Angeles

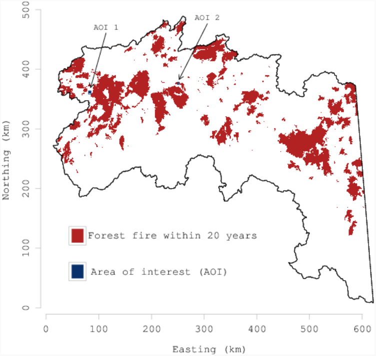
Based upon projects involving:

- Abhirup Datta (Johns Hopkins University)
- Lu Zhang (UCLA)
- Andrew O. Finley (Michigan State University)
- Alan. E. Gelfand (Duke University)

Case Study: Alaska Tanana Valley Forest Height Dataset



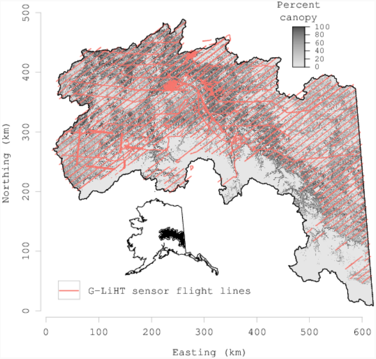
Forest height and tree cover



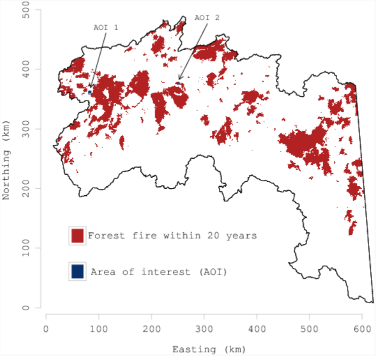
Forest fire history

- Forest height (red lines) data from LiDAR at 5×10^6 locations
- Knowledge of forest height is important for biomass assessment, carbon management etc

Case Study: Alaska Tanana Valley Forest Height Dataset



Forest height and tree cover



Forest fire history

- Goal: High-resolution domainwide prediction maps of forest height
- Covariates: Domainwide tree cover (grey) and forest fire history (red patches) in the last 20 years

Analyzing the data

Models used:

- Non-spatial regression: $y_{FH} = \beta_0 + \beta_{tree}x_{tree} + \beta_{fire}x_{fire} + \epsilon$

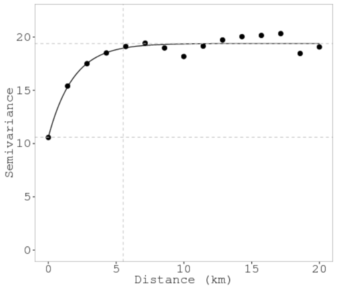


Figure: Variogram of the residuals from non-spatial regression indicates strong spatial pattern

Geostatistical models

- $y_{FH}(\ell) = \beta_0 + \beta_{tree} x_{tree}(\ell) + \beta_{fire} x_{fire}(\ell) + w(\ell) + \epsilon(\ell)$
- $w(\ell) \sim GP(0, C(\cdot, \cdot \,|\, \sigma^2, \phi))$
- $y_{FH} \sim N(X\beta, K_\theta)$ where K_θ is the spatial covariance matrix:

$$K_\theta = C_{(\sigma,\phi)} + \tau^2 I \ , \quad \text{where} \ \ \theta = \{\sigma, \phi, \tau\}$$

where $C_{(\sigma^2,\phi)}$ is the GP covariance matrix derived from $C(\cdot, \cdot \,|\, \sigma^2, \phi)$.

Likelihood from (full rank) GP models

- $\mathscr{L} = \{\ell_1, \ell_2, \ldots, \ell_n\}$ 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:

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

- Customary: $K_\theta = C_{(\sigma,\phi)} + D_\tau$, where $\theta = \{\sigma, \phi, \tau\}$
- Bayesian inference: Priors on $\{\beta, \theta\}$
- Challenges: Storage and $\texttt{chol}(K_\theta) = LDL^\top$.

Computational Details

- Compute the quadratic form and determinant (for any given $\{\beta, \theta\}$):

$$\begin{array}{ll} \text{Solve for } u\text{:} & K_\theta u = y - X\beta \;\; \text{(expensive)} \;\; ; \\ \text{Quadratic form:} & (y - X\beta)^\top u \; ; \\ \text{Determinant:} & \det(K_\theta) \;\; \text{(expensive)} \;\; . \end{array}$$

- Compute the quadratic form and determinant (for any given $\{\beta, \theta\}$):

$$\begin{array}{ll} \text{Cholesky:} & \texttt{chol}(K_\theta) = LDL^\top \;\; \text{(expensive)} \;\; ; \\ \text{Solve for } v\text{:} & v = \texttt{trsolve}(L, y - X\beta) \; ; \\ \text{Quadratic form:} & v^\top D^{-1} v = \sum_{i=1}^n v_i^2 / d_{ii} \; ; \\ \text{Determinant:} & \log\det(K_\theta) = \sum_{i=1}^n \log d_{ii} \; . \end{array}$$

- Log-likelihood (up to a constant):

$$-\frac{1}{2}\sum_{i=1}^n \log d_{ii} - \frac{1}{2}\sum_{i=1}^n v_i^2 / d_{ii}$$

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 for Prediction

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

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

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

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

- 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)
- 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)
- 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)

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 $\tilde{w}(\ell)$

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

  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;
  - $\{\ell_1^*, \ell_2^*, \dots, \ell_r^*\}$ 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.)

- $\tilde{w} = (\tilde{w}(\ell_1), \tilde{w}(\ell_2), \ldots, \tilde{w}(\ell_n))^\top$ 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:

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

  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 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:

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

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 $\mathscr{L}^* = \{\ell_1^*, \ell_2^*, \ldots, \ell_r^*\}$,

  and then taking the conditional expectation:

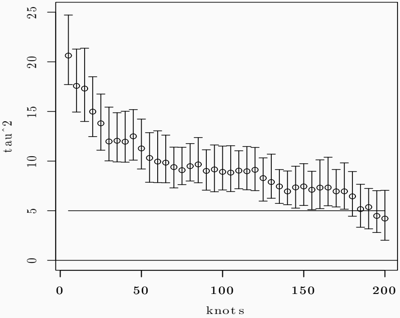
$$\tilde{w}(\ell) = \mathrm{E}[w(\ell) \mid w^*] = b_\theta^\top(\ell) z \; .$$

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

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

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,

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

Bias-adjusted or modified predictive processes

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

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

- Remedy:

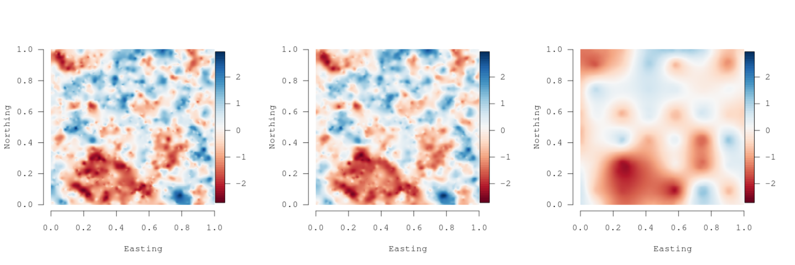
$$\tilde{w}_\epsilon(\ell) = \tilde{w}(\ell) + \tilde{\epsilon}(\ell)\ ,$$

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

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

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

Oversmoothing in low rank models



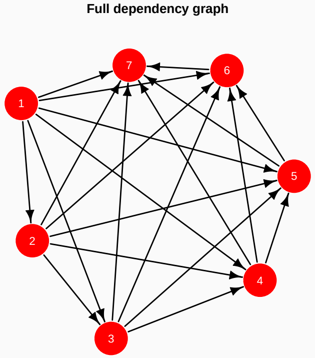
True w

Full GP

PPGP 64 knots

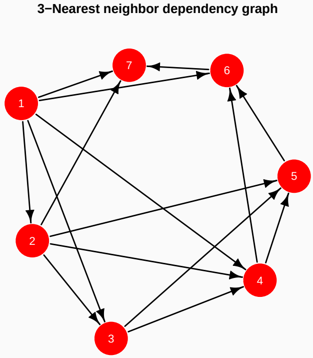
Figure: Comparing full GP vs low-rank GP with 2500 locations. Figure (1c) exhibits oversmoothing by a low-rank process (predictive process with 64 knots)

Introducing sparsity through conditional independence



$$p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2)p(w_4 \mid w_1, w_2, w_3)$$
$$\times\, p(w_5 \mid w_1, w_2, w_3, w_4)p(w_6 \mid w_1, w_2, \ldots, w_5)p(w_7 \mid w_1, w_2, \ldots, w_6)\ .$$

Simple method of introducing sparsity (e.g. graphical models)



$$p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2)p(w_4 \mid w_1, w_2, w_3)$$

$$p(w_5 \mid \cancel{w_1}, w_2, w_3, w_4)p(w_6 \mid w_1, \cancel{w_2}, \cancel{w_3}, w_4, w_5)p(w_7 \mid w_1, w_2, \cancel{w_3}, \cancel{w_4}, \cancel{w_5}, w_6)$$

Gaussian graphical models: linearity

- Write a joint density $p(w) = p(w_1, w_2, \ldots, w_n)$ as:

$$p(w_1)p(w_2 \mid w_1)p(w_3 \mid w_1, w_2) \cdots p(w_n \mid w_1, w_2, \ldots, w_{n-1})$$

- Example: For Gaussian distribution $N(w \mid 0, K_\theta)$, we have a linear model

$$\begin{aligned} w_1 &= 0 + \eta_1; \\ w_2 &= a_{21} w_1 + \eta_2; \\ w_3 &= a_{31} w_1 + a_{32} w_2 + \eta_3; \\ w_i &= a_{i1} w_1 + a_{i2} w_2 + \cdots + a_{i,i-1} w_{i-1} + \eta_i; \quad i = 4, \ldots, n\ . \end{aligned}$$

- More compactly: $w = Aw + \eta\ ; \qquad \eta \sim N(0, D)$.

Simple method of introducing sparsity (e.g. graphical models)

- Assume $w \sim N(0, K_\theta)$. Introduce sparsity by modeling $\texttt{chol}(K_\theta)$

$$K_\theta = (I - A)^{-1} D (I - A)^{-\top} \; ; \quad D = \mathrm{diag}(\mathrm{var}\{w_i \mid w_{\{j<i\}}\})$$

- If L is from $\texttt{chol}(K_\theta) = LDL^\top$, then $L^{-1} = I - A$.

- a_{ij}'s obtained from $n-1$ linear systems by comparing coefficients of w_j's in

$$\sum_{j<i} a_{ij} w_j = \mathrm{E}[w_i \mid w_{\{j<i\}}] \quad i = 2, \ldots, n$$

- Example:

```
for(i in 1:(n-1)) {
  a[i+1,1:i] = solve(K[1:i,1:i], K[1:i,i+1])
  d[i+1,i+1] = K[i+1,i+1] - dot(K[i+1,1:i],a[i+1,1:i])
}
```

- Let $a_{ij} = 0$ for all but m nearest neighbors of node i implies solving

$$\sum_{j \in N[i]} a_{ij} w_j = \mathrm{E}[w_i \,|\, w_{\{j \in N[i]\}}] \quad i = 2, \ldots, n \,,$$

  where $N[i] = \{j < i : j \sim i\}$ are indices for neighbors of i.

- Example:

```
for(i in 1:(n-1) {
  Pa = N[i+1] # neighbors of i+1
  a[i+1,Pa] = solve(K[Pa,Pa], K[i+1, Pa])
  d[i+1,i+1] = K[i+1,i+1] - dot(K[i+1, Pa],a[i+1,Pa])
}
```

- We need to solve $n - 1$ linear systems of size at most $m \times m$. Trivially parallelizable!
- **Storage and flops linear in n.**

Sparse likelihood approximations (Vecchia, 1988)

- Let $\mathscr{R} = \{\ell_1, \ell_2, \ldots, \ell_r\}$

- With $w(\ell) \sim GP(0, K_\theta(\cdot))$, write the joint density $p(w_{\mathscr{R}})$ as:

$$\begin{aligned} N(w_{\mathscr{R}} \,|\, 0, K_\theta) &= \prod_{i=1}^{r} p(w(\ell_i) \,|\, w_{H(\ell_i)}) \\ &\approx \prod_{i=1}^{r} p(w(\ell_i) \,|\, w_{N(\ell_i)}) = N(w_{\mathscr{R}} \,|\, 0, \tilde{K}_\theta)\ . \end{aligned}$$

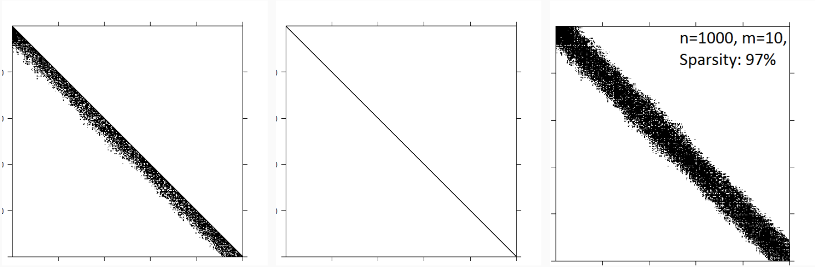
  where $N(\ell_i) \subseteq H(\ell_i)$.

- Shrinkage: Choose $N(\ell)$ as the set of "m nearest-neighbors" among $H(\ell_i)$. Theory: "Screening" effect (Stein, 2002).

- $\tilde{K}_\theta^{-1}$ depends on K_θ, but is *sparser* with at most nm^2 non-zero entries

Sparse precision matrices (e.g., graphical Gaussian models)

$$N(w \mid 0, K_\theta) \approx N(w \mid 0, \tilde{K}_\theta) \; ; \tilde{K}_\theta^{-1} = (I - A)^\top D^{-1} (I - A)$$



$I - A$ D^{-1} $\tilde{K}_\theta^{-1}$

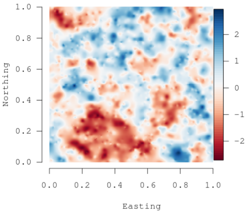
- $\det(\tilde{K}_\theta^{-1}) = \prod_{i=1}^{n} D_{ii}^{-1}$, $\tilde{K}_\theta^{-1}$ is sparse with $O(nm^2)$ entries

Extension to a GP (Datta et al., *JASA*, 2016)

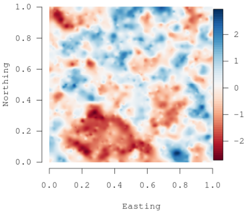
- Fix "reference" set $\mathscr{R} = \{\ell_1, \ell_2, \ldots, \ell_r\}$ (e.g. observed points)
- $N(\ell)$ is the set of *m*-nearest neighbors of ℓ in $\mathscr{R}$
- This completes the consistent extension to a process $w(\ell) \sim GP$:

$$p(w_{\mathscr{R}}, w(\ell) \,|\, \theta) = N(w_{\mathscr{R}} \,|\, 0, \tilde{K}_\theta) \times p(w(\ell) \,|\, \{w(\ell_i) : \ell_i \in N(\ell)\}, \theta)\ .$$

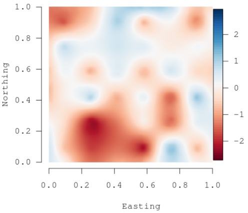
- For any $\ell, \ell' \notin \mathscr{R}$, conditional indep: $w(\ell) \perp w(\ell') \,|\, w_{\mathscr{R}}$
- Finite-dimensional realizations of $w(\ell)$ (given $\mathscr{R}$) will enjoy sparse precision matrices
- Call this NNGP. In hierarchical models, substitute NNGP for GP and achieve MASSIVE scalability.



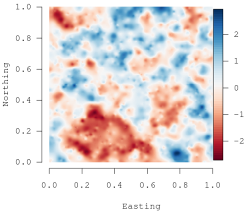
True w



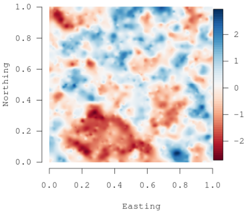
Full GP



PPGP 64 knots



NNGP, $m = 10$



NNGP, $m = 20$

NNGP models

- Collapsed NNGP:
    - $y_{FH}(\ell) = \beta_0 + \beta_{tree} x_{tree}(\ell) + \beta_{fire} x_{fire}(\ell) + w(\ell) + \epsilon(\ell)$
    - $w(\ell) \sim NNGP(0, C(\cdot, \cdot \,|\, \sigma^2, \phi))$
    - $y_{FH} \sim N(X\beta, \tilde{C} + \tau^2 I)$ where $\tilde{C}$ is the NNGP covariance matrix derived from C
- Response NNGP:
    - $y_{FH}(\ell) \sim NNGP(\beta_0 + \beta_{tree} x_{tree}(\ell) + \beta_{fire} x_{fire}(\ell), \Sigma(\cdot, \cdot \,|\, \sigma^2, \phi, \tau^2))$
    - $y_{FH} \sim N(X\beta, \tilde{\Sigma})$ where $\tilde{\Sigma}$ is the NNGP covariance matrix derived from $\Sigma = C + \tau^2 I$

NNGP models

	Non-spatial regression	Collapsed NNGP	Response NNGP
CRPS	2.3	0.86	0.86
RMSPE	4.2	1.73	1.72
CP	93%	94%	94%
CIW	16.3	6.6	6.6

Table: Model comparison metrics for the Tanana valley dataset

- NNGP models perform significantly better than the non-spatial model
- MCMC run time for the NNGP models:
    - Collapsed model: 319 hours
    - Response model: 38 hours
- For massive spatial data, full Bayesian output for even NNGP models require substantial time

Another look at the response model

- Original full GP model: $y(\ell) \overset{ind}{\sim} N(x^\top(\ell)\beta + w(\ell), \tau^2)$
- $w(\ell) \sim GP$ with a stationary covariance function $C(\cdot, \cdot \,|\, \sigma^2, \phi)$
- $Cov(w) = \sigma^2 R(\phi)$
- Full GP model: $y \sim N(X\beta, \Sigma)$ where $\Sigma = \sigma^2 M$
- $M = R(\phi) + \alpha I$
- $\alpha = \tau^2/\sigma^2$ is the ratio of the noise to signal variance
- Response NNGP model: $y \sim N(X\beta, \tilde{\Sigma})$
- $\tilde{\Sigma} = \sigma^2 \tilde{M}$ where $\tilde{M}$ is the NNGP approximation for M

Conjugate NNGP

- $y \sim N(X\beta, \sigma^2 \tilde{M})$
- If ϕ and α are known, M, and hence $\tilde{M}$, are known matrices
- The model becomes a standard Bayesian linear model
- Assume a *Normal Inverse Gamma (NIG)* prior for $\{\beta, \sigma^2\}$
- $\{\beta, \sigma^2\} \sim NIG(\mu_\beta, V_\beta, a_\sigma, b_\sigma)$, i.e.,

$$\beta \,|\, \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta) \;\text{ and }\; \sigma^2 \sim IG(a_\sigma, b_\sigma)\ .$$

Conjugate NNGP

- $y \sim N(X\beta, \sigma^2 \tilde{M})$, $\tilde{M}$ is known

Joint likelihood:

$$N(y \,|\, X\beta, \sigma^2 \tilde{M}) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma)$$

Conjugate NNGP

- $y \sim N(X\beta, \sigma^2 \tilde{M})$, $\tilde{M}$ is known

Joint likelihood:

$$N(y \,|\, X\beta, \sigma^2 \tilde{M}) \times N(\beta \,|\, \mu_\beta, \sigma^2 V_\beta) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma)$$

- Conjugate posterior distribution $\{\beta, \sigma^2\} \,|\, y \sim NIG(\mu^*_\beta, V^*_\beta, a^*_\sigma, b^*_\sigma)$
- Expressions for μ^*_β, V^*_β, a^*_σ and b^*_σ can be calculated in $O(n)$ time

Conjugate NNGP

- $\{\beta, \sigma^2\} \,|\, y \sim NIG(\mu_\beta^*, V_\beta^*, a_\sigma^*, b_\sigma^*)$
- Marginal posterior: $\beta \,|\, y \sim MVt_{2a_\sigma^*}(\mu_\beta^*, \frac{b_\sigma^*}{a_\sigma^*} V_\beta^*)$
- $MVt_k(m, V)$ is the *multivariate t* distribution with degrees of k, mean m and scale matrix V
- $E(\beta \,|\, y) = \mu_\beta^*$, $Var(\beta \,|\, y) = \frac{b_\sigma^*}{a_\sigma^* - 1} V_\beta^*$
- Marginal posterior: $\sigma^2 \,|\, y \sim IG(a_\sigma^*, b_\sigma^*)$
- $E(\sigma^2 \,|\, y) = \frac{b_\sigma^*}{a_\sigma^* - 1}$, $Var(\sigma^2 \,|\, y) = \frac{b_\sigma^{*2}}{(a_\sigma^* - 1)^2 (a_\sigma^* - 2)}$
- Exact posterior distributions of β and σ^2 are available

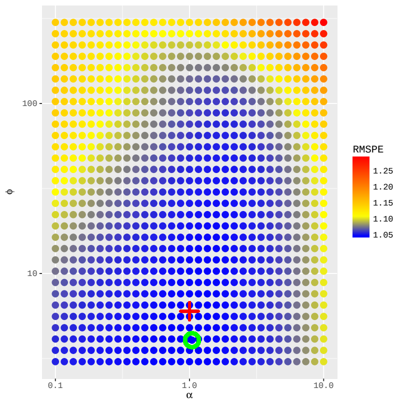
Predictive distributions

- $y(\ell) \,|\, y \sim t_{2a_\sigma^*}(m(\ell), \frac{b_\sigma^*}{a_\sigma^*} v(\ell))$
- $E(y(\ell) \,|\, y) = m(\ell)$, $Var(y(\ell) \,|\, y) = \frac{b_\sigma^*}{a_\sigma^* - 1} v(\ell)$
- $m(\ell)$ and $v(\ell)$ can be computed using $O(m)$ flops
- Exact posterior predictive distributions of $y(\ell) \,|\, y$ for any ℓ
- No MCMC required for parameter estimation or prediction

Choosing α and ϕ

- ϕ and α are chosen using K-fold cross validation over a grid of possible values
- Unlike MCMC, cross-validation can be completely parallelized
- Resolution of the grid for ϕ and α can be decided based on computing resources available
- In practice, a reasonably coarse grid often suffices

Choosing α and ϕ



RMSPE

Figure: Simulation experiment: True value (+) of (α, ϕ) and estimated value (○) using 5-fold cross validation

Scalability

- Computation and storage requirements are $O(n)$
- One evaluation time similar to the response NNGP model
- Unlike response NNGP, does not involve any serial MCMC iterations
- For K fold cross validation and G combinations of ϕ and α, total number of evaluations is KG
- Embarassingly parallel: Each of the KG evaluations can proceed in parallel

Alaska Tanana Valley dataset

	Conjugate NNGP	Collapsed NNGP	Response NNGP
β_0	2.51	2.41 (2.35, 2.47)	2.37 (2.31,2.42)
β_{TC}	0.02	0.02 (0.02, 0.02)	0.02 (0.02, 0.02)
β_{Fire}	0.35	0.39 (0.34, 0.43)	0.43 (0.39, 0.48)
σ^2	23.21	18.67 (18.50, 18.81)	17.29 (17.13, 17.41)
τ^2	1.21	1.56 (1.55, 1.56)	1.55 (1.54, 1.55)
ϕ	3.83	3.73 (3.70, 3.77)	4.15 (4.13, 4.19)
CRPS	0.84	0.86	0.86
RMSPE	1.71	1.73	1.72
time (hrs.)	0.002	319	38

Table: Parameter estimates and model comparison metrics for the Tanana valley dataset

- Conjugate model produces estimates and model comparison numbers very similar to the MCMC based NNGP models
- For 5×10^6 locations, conjugate model takes 7 seconds

Summary

- MCMC free exact Bayesian approach by fixing some covariance parameters
- Conjugate posterior distributions of the parameters and posterior predictive distributions available in closed form
- Embarassingly parallel cross validation to identify best choices for fixed parameters
- Runs in seconds for massive spatial dataset with millions of locations
- Available in the spNNGP package in R

Concluding remarks

- Model-based solution for spatial “BIG DATA”
- Algorithms: Gibbs, RWM, HMC, VB, INLA. HMC-NUTS is especially promising on `STAN`.
- Compare with scalable multi-resolution frameworks (Katzfuss, 2016)
- Enhance scalability using META-KRIGING approaches (e.g., Rajarshi Guhaniyogi, 2017)
- Challenges: Nonstationary models; High-dimensional outcomes; High-dimensional domains; Smoother process approximations.