High-dimensional Bayesian Geostatistics (on your laptop!)

Sudipto Banerjee

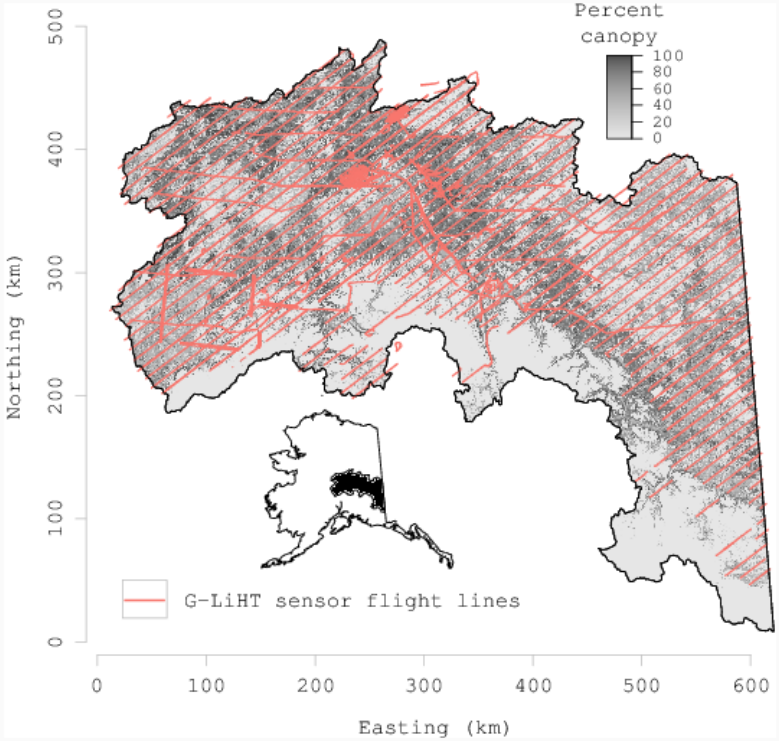
CIRM Nov 26-30, 2018

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

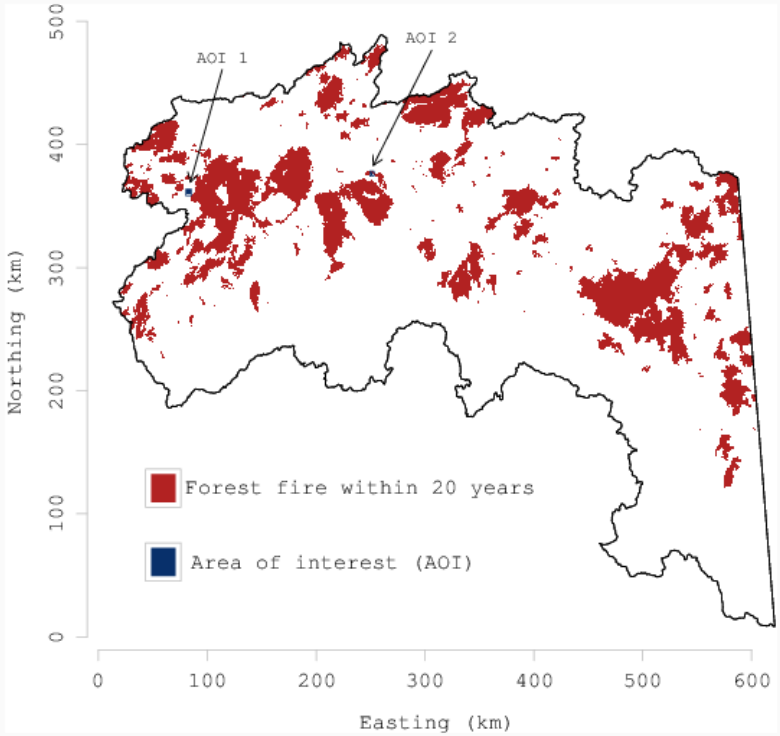
Based upon projects involving:

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

Case Study: Alaska Tanana Valley Forest Height Dataset



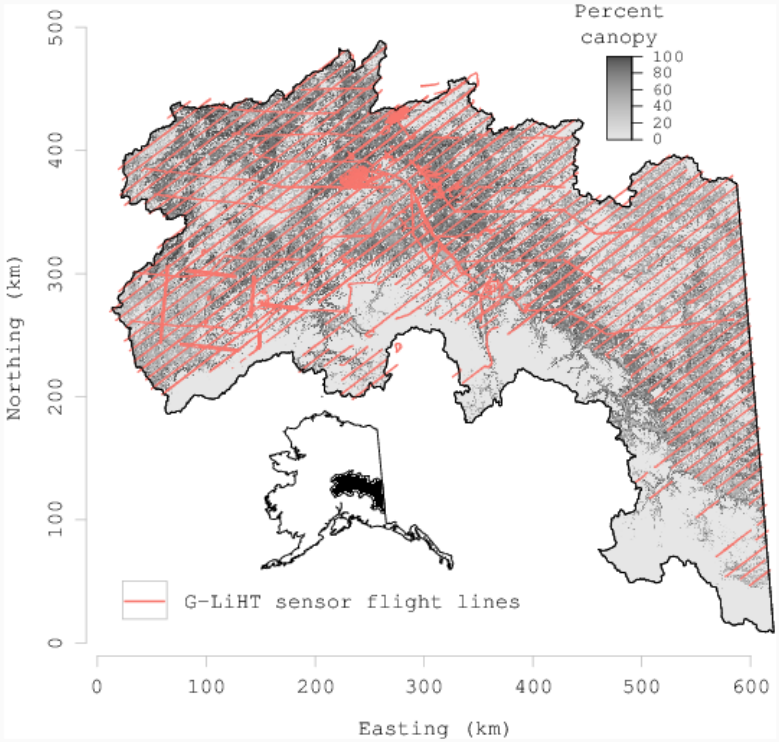
Forest height and tree cover



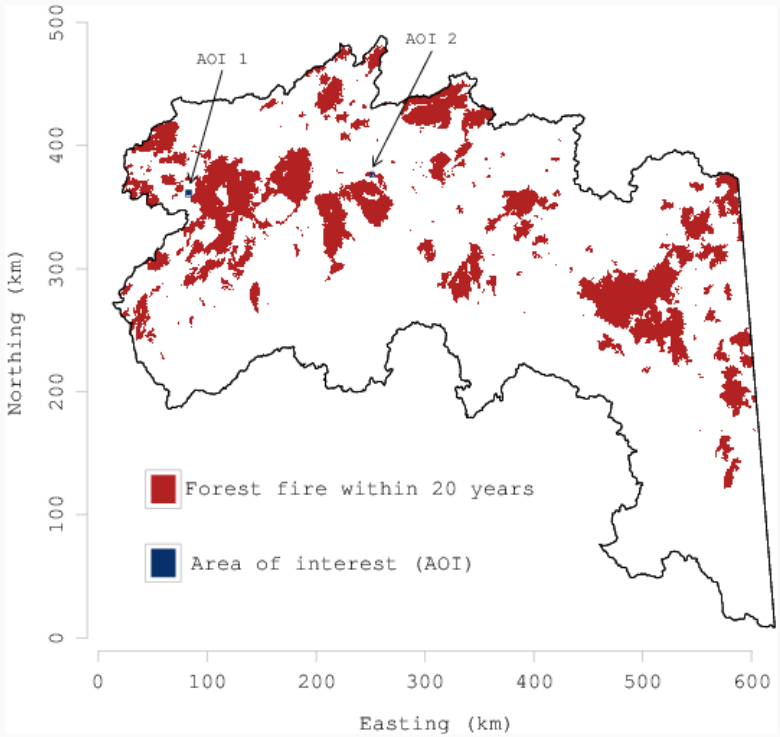
Forest fire history

- Forest height (red lines) data from LiDAR at 10×10^6 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

Bayesian regression for BIG DATA

- Conjugate Bayesian hierarchical linear model:

$$y_i \,|\, \beta, \sigma^2 \overset{ind}{\sim} N(x_i^\top \beta, \sigma^2) \;,\; i = 1, 2, \ldots, n \;;$$
$$\beta \,|\, \sigma^2 \sim N(\mu_\beta, \sigma^2 V_\beta) \;; \quad \sigma^2 \sim IG(a, b) \;.$$

- Exact Bayesian inference:

$$\sigma^2 \,|\, y \sim IG(a^*, b^*) \quad \beta \,|\, \sigma^2, y \sim N(Mm, \sigma^2 M) \;, \quad \text{where}$$
$$m = V_\beta^{-1} \mu_\beta + X^\top y \;, \quad M^{-1} = V_\beta^{-1} + X^\top X \;,$$
$$a^* = a + n/2 \;, \quad b^* = \mu_\beta^\top V_\beta^{-1} \mu_\beta + y^\top y - m^\top M^{-1} m \;.$$

- What if the data cannot be stored/loaded into available workspace?
- HADOOP: Map-Reduce framework (Divide & Conquer) with cloud computing.

Bayesian regression on HADOOP

- Partition data as $\{y_k, X_k\}$, $k = 1, 2, \ldots, K$, where each y_k is $n_k \times 1$, X_k is $n_k \times p$ and $N = \sum_{k=1}^{K} n_k$.

- For each subset compute:

$$m_k = V_\beta^{-1} + X_k^\top y_k \text{ and } M_k^{-1} = V_\beta^{-1} + X_k^\top X_k \ .$$

- Then

$$m = \sum_{k=1}^{K} (m_k - (1 - 1/K)\mu_\beta) \text{ and } M^{-1} = \sum_{k=1}^{K} (M_k^{-1} - (1 - 1/K)V_\beta^{-1}) \ .$$

- Depends (crucially) on independence across subsets; not suitable for spatial random fields.

- Meta-Kriging (GB, *Technometrics*, 2018+): find convex combination of subset-posteriors closest to the full posterior.

Geostatistical models (Cressie, 1993; Stein, 1999)

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

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

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

High-dimensional outcomes: Jointly modeling LiDAR and survey data with factor models (TDFB *Statistica Sinica*, 2018+)

- $y(\ell)$: Field observed (survey) data (multivariate outcomes);
- $z(\ell)$: LiDAR signals (high-dimensional) vectors at each location ℓ.

$$\text{Stage 1: } z(\ell) = X_z(\ell)^\top \beta_z + \Lambda_z w(\ell) + \epsilon_z(\ell)$$
$$\text{Stage 2: } y(\ell) = X_y(\ell)^\top \beta_y + \Lambda_y w(\ell) + \Gamma v(\ell) + \epsilon_y(\ell)\ .$$

- $w(\cdot)$ and $v(\cdot)$ are spatial processes;
- Λ_z and Λ_y are loadings extracting "principal process components";
- $\epsilon_z(\cdot)$ and $\epsilon_y(\cdot)$ are additional (perhaps white-noise) processes to capture unstructured or micro-structured variation.

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:

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

- Bayesian inference: Priors on $\{\beta, \theta\}$
- Computing: (i) $\texttt{chol}(K_\theta) = LDL^\top$, (ii) $v = \texttt{trsolve}(L, y - X\beta)$,

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

- 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)

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

- Bayesian "kriging" computes (simulates) posterior predictive density:

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

Bayesian low rank models (Wikle, *HSS*, 2010)

- Hierarchical Bayesian regression models are *naturally* low-rank:

$$\begin{aligned} y \,|\, \beta, z, \theta, \tau &\sim N(X\beta + B_\theta z, D_\tau) \; ; \\ z \,|\, \theta &\sim N(0, V_{z,\theta}) \; ; \\ \beta \,|\, \mu_\beta, V_\beta &\sim N(0, V_\beta) \; ; \\ \theta, \tau &\sim p(\theta, \tau) = p(\theta) \times p(\tau) \; . \end{aligned}$$

- Posterior distribution:

$$p(\theta) \times p(\tau) \times N(\beta \,|\, \mu_\beta, V_\beta) \times N(z \,|\, 0, V_{z,\theta}) \times N(y \,|\, X\beta + B_\theta z, D_\tau) \; .$$

- $B_\theta z$? Start with a *parent process* $w(\ell)$ and construct $\tilde{w}(\ell)$

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

- Example: $\tilde{w}(\ell) = \mathrm{E}[w(\ell) \,|\, w^*] = \sum_{j=1}^{r} b_\theta(\ell, \ell_j^*) w(\ell_j^*)$

Implementing low-rank Bayesian models

- The ubiquituous Sherman-Woodbury-Morrison formulas (discovered and rediscovered through the ages!)
- Computing var(y) in two different ways yields (Lindley & Smith, *JRSS-B*, 1972)

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

- A companion formula for the determinant:

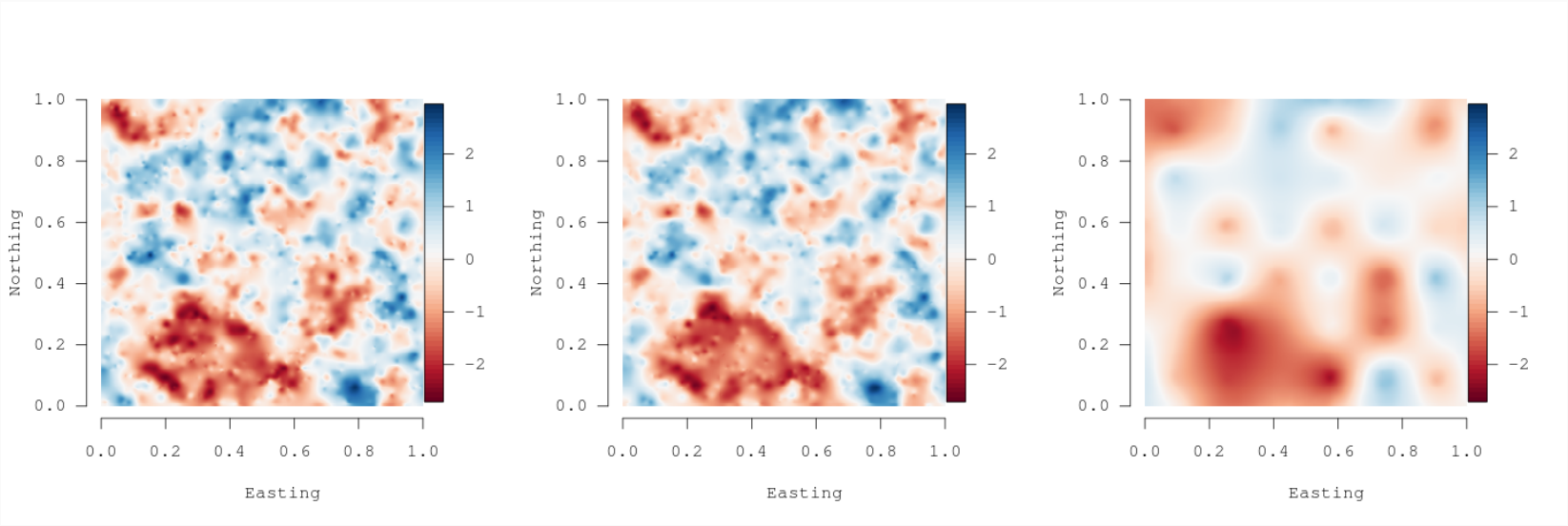
$$\det(D_\tau + B_\theta V_z B_\theta^\top) = \det(V_z) \det(D_\tau) \det(V_z^{-1} + B_\theta^\top D_\tau^{-1} B_\theta) \; .$$

- For BIG DATA computations avoid directly computing the above formulas; use optimized functions (Banerjee, *Bayesian Anal., 2017*):

$$L = \texttt{chol}(V) \;\; \text{and} \;\; W = \texttt{trsolve}(T, B) \; .$$

- Complexity: $O(nr^2 + r^3) \approx O(nr^2)$.

Oversmoothing in low rank models (Banerjee, 2017; *BA*)



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 (with $r = 64$)

- Can be explained: $P_{[B_1:B_2]} = P_{B_1} + P_{[(I-P_{B_1})B_2]}$
- Fixes and improvements: MRA (e.g., Katzfuss, *JASA*, 2016).
- Sparse approximations or sparsity-inducing processes.

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:

$$N(w_{\mathscr{R}} \mid 0, K_\theta) = \prod_{i=1}^{r} p(w(\ell_i) \mid w_{H(\ell_i)}) \approx \prod_{i=1}^{r} p(w(\ell_i) \mid w_{N(\ell_i)}) \, .$$

  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).

Sparse likelihood approximations (Datta et al, 2016)

- 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:

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

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

- $\tilde{K}_\theta^{-1}$ is *sparser* with at most nm^2 non-zero entries
- $\tilde{K}_\theta$ is a Nearest-Neighbor (NN) approximation for K_θ.

Gaussian graphical models: linearity

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

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

- For Gaussian distribution $N(w \mid 0, K_\theta)$, we have a sequence of linear models

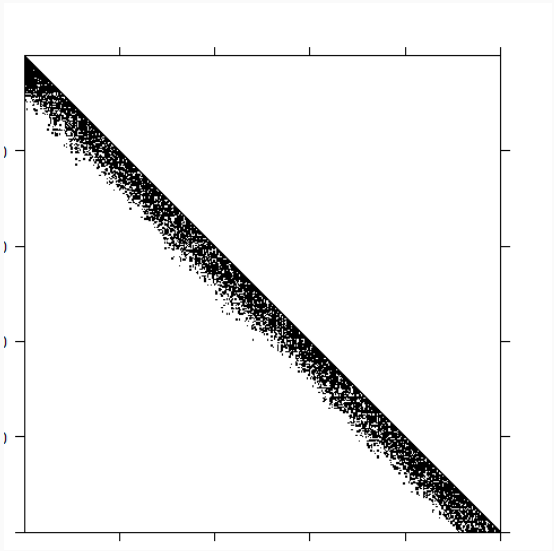
$$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 & \ldots & 0 & 0 \\ a_{21} & 0 & 0 & \ldots & 0 & 0 \\ a_{31} & a_{32} & 0 & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \ldots & a_{n,n-1} & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ w_3 \\ \vdots \\ w_n \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \\ \vdots \\ \eta_n \end{bmatrix}$$

$$\Longrightarrow w = Aw + \eta; \quad \eta \sim N(0, D) \; .$$

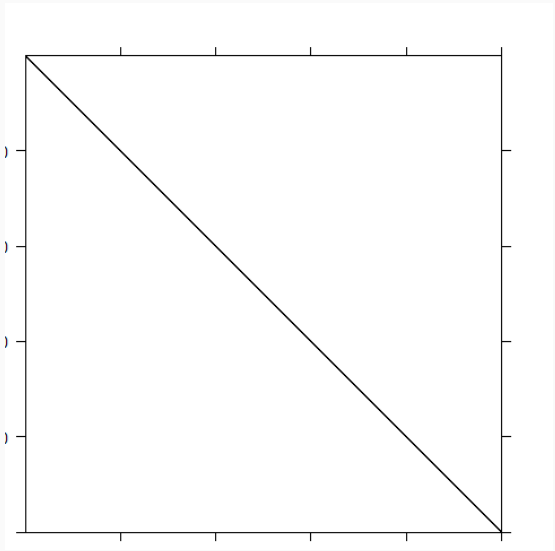
- $\eta = (I - A)w = Lw$ (L is the decorrelation transformation).
- $a_{ij} = 0$ introduces conditional independence and sparsity.

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

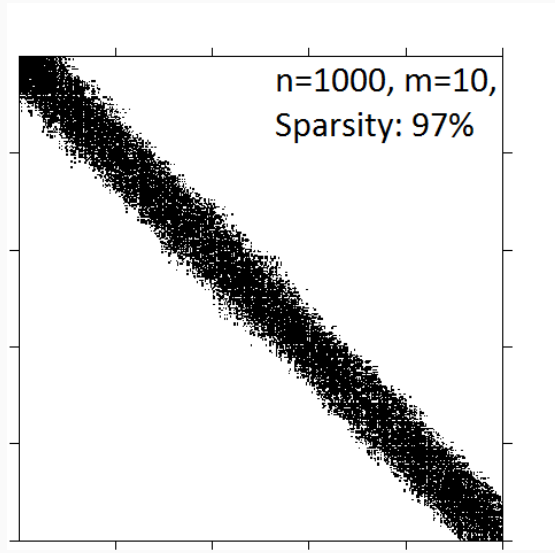
$$N(w \mid 0, K_\theta) \approx N(w \mid 0, \tilde{K}_\theta) \; ; \tilde{K}_\theta^{-1} = (I - A)^\top D^{-1} (I - A)$$



$I - A$



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

- Computing A and D

```
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!
- Quadratic form:

```
qf(u,v,A,D) = u[1]*v[1] / D[1,1]
for(i in 2:n) {
    qf(u,v,A,D) = qf(u,v,A,D) + (u[i] - dot(A[i,N(i)], u[N(i)]))
        *(v[i] - dot(A[i,N(i)], v[N(i)]))/D[i,i]
    }
```

- Determinant: $\det(\tilde{K}_\theta) = \prod_{i=1}^{n} \texttt{d[i,i]}$

Nearest-Neighbor GP (Datta et al., 2016)

- Fix "reference" set $\mathscr{R} = \{\ell_1, \ell_2, \ldots, \ell_r\}$ (e.g. observed points)

$$N(\ell_i) = \begin{cases} \text{empty set for } i = 1 \\ \{\ell_1, \ell_2, \ldots, \ell_{i-1}\} \text{ for } 2 \leq i \leq m \\ m \text{ nearest neighbors of } \ell_i \text{ among } \{\ell_1, \ell_2, \ldots, \ell_{i-1}\} \text{ for } i > m \end{cases}$$

- $N(\ell_i)$ is the set of at most m nearest neighbors of ℓ_i among $\{\ell_1, \ell_2, \ldots, \ell_{i-1}\}$.

- $N(\ell)$ is the set of m-nearest neighbors of ℓ in $\mathscr{R}$

Nearest-Neighbor GP (Datta et al., 2016)

- 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.

NNGP as BLUPs: Pourahmadi *Biometrika*, 1999; Stein et al. *JRSS B*, 2004; Guinness *Technometrics*, 2018

- Let $a_{ij} = 0$ for all but m nearest neighbors of node i. Then,

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

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

- w_i is the "projection" onto a subset of $\{w_1, w_2, \ldots, w_{i-1}\}$

- So any "optimality" of the resulting distribution will depend upon the "order" of $\{w_1, w_2, \ldots, w_{i-1}\}$.

- If $N^{(1)}[i] \subseteq N^{(2)}[i] \subseteq \cdots \subseteq N^{(m)}[i]$ for all $i = 1, 2, \ldots, n$

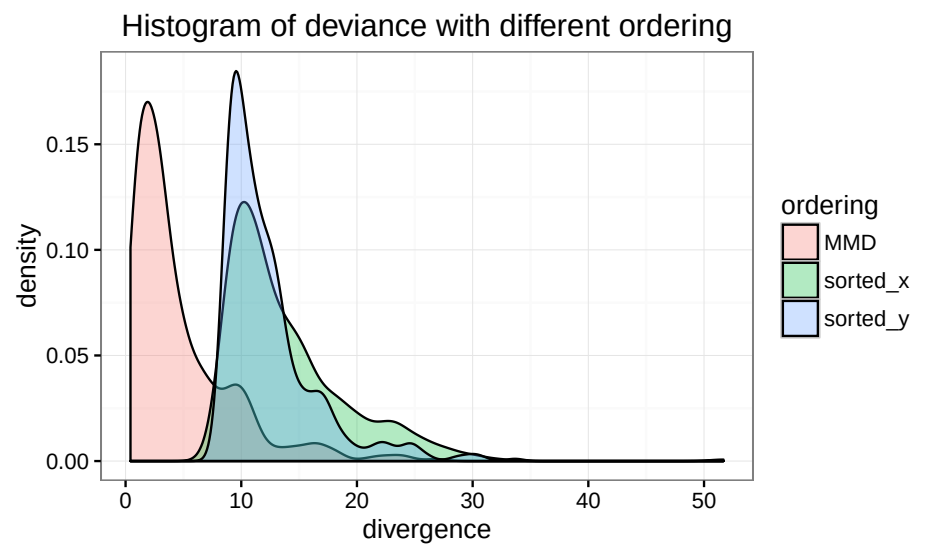
$$KL(N(0, K_\theta) || N(0, \tilde{K}_\theta^{(1)})) \geq \cdots \geq KL(N(0, K_\theta) || N(0, \tilde{K}_\theta^{(m)}))$$

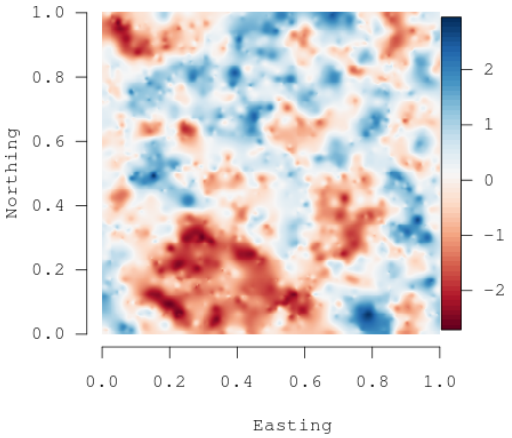
Effect of topological ordering (Guinness, *Technometrics*, 2018)

- Cholesky decompositions are not invariant to ordering

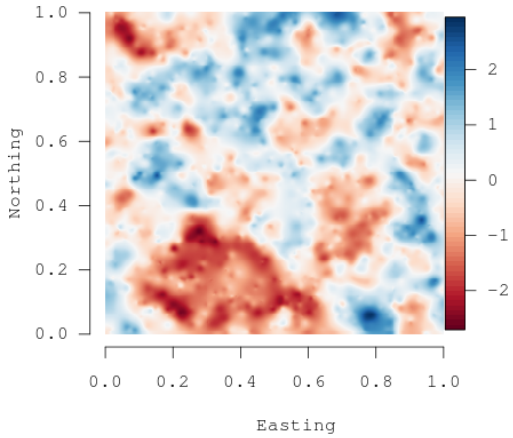
- Robust estimates and RMSPE:

Parameter	True values	Sorted by $x + y$	MMD
Sill	1.41	0.74 (0.54, 1.15)	0.76 (0.52, 1.23)
Range	0.20	0.15 (0.11, 0.23)	0.15 (0.10, 0.25)
RMSPE		0.1565937	0.1565853

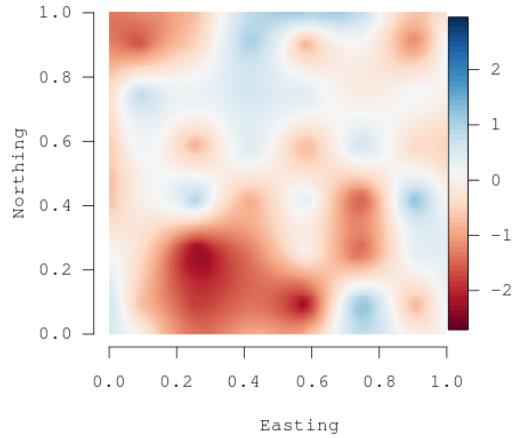




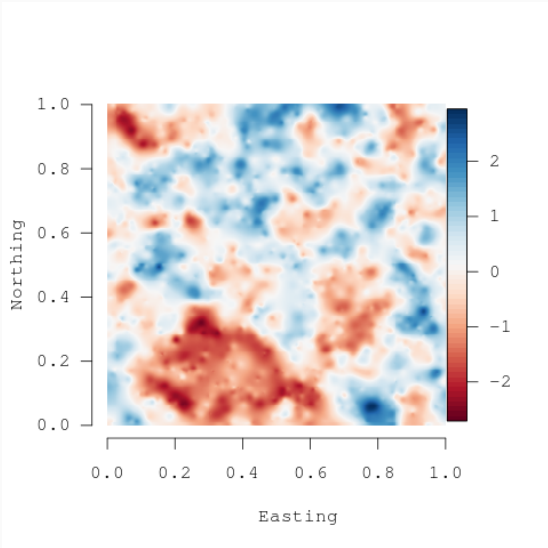
True w



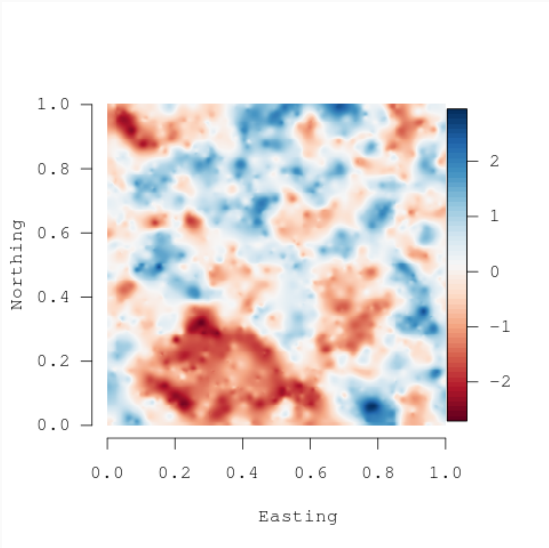
Full GP



PPGP 64 knots



NNGP, $m = 10$



NNGP, $m = 20$

NNGP models

- Collapsed (latent) NNGP:
  - $y_{FH}(\ell) = \beta_0 + \beta_{tree} x_{tree}(\ell) + \beta_{fire} x_{fire}(\ell) + w(\ell) + \epsilon(\ell)$
  - $w(\ell) \sim NNGP(0, C(\cdot, \cdot \mid \sigma^2, \phi))$
  - $y_{FH} \sim N(X\beta, \tilde{C} + \tau^2 I)$ where $\tilde{C}$ is the NNGP covariance matrix derived from C
- Response NNGP:
  - $y_{FH}(\ell) \sim NNGP(\beta_0 + \beta_{tree} x_{tree}(\ell) + \beta_{fire} x_{fire}(\ell), \Sigma(\cdot, \cdot \mid \sigma^2, \phi, \tau^2))$
  - $y_{FH} \sim N(X\beta, \tilde{\Sigma})$ where $\tilde{\Sigma}$ is the NNGP covariance matrix derived from $\Sigma = C + \tau^2 I$
- Generalized-Vecchia (Katzfuss and Guinness, 2017).

Conjugate Response NNGP (FDCMAB, *JCGS*, 2018+)

- Full GP covariance matrix: $K_\theta = \sigma^2 M$, where $M = R(\phi) + \delta^2 I$
- If ϕ and δ^2 are known, so are M and its NNGP approximation $\tilde{M}$
- 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.,

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

- The model becomes a conjugate Bayesian linear model:

$$\begin{aligned} p(\beta, \sigma^2 \,|\, y) &\propto NIG(\beta, \sigma^2 \,|\, \mu_\beta, V_\beta, a_\sigma, b_\sigma) \times N(y \,|\, X\beta, \sigma^2 \tilde{M}) \\ &= NIG(\beta, \sigma^2 \,|\, \mu_\beta^*, V_\beta^*, a_\sigma^*, b_\sigma^*) \end{aligned}$$

- Exact posterior predictive distribution: $y(\ell) \,|\, y \sim t_{2a_\sigma^*}(m(\ell), \frac{b_\sigma^*}{a_\sigma^*} v(\ell))$
- Fully closed-form Bayesian inference; all computations in $O(n)$ time.

Response VS Latent Model

- Compare KL-divergence of response and latent NNGP models from the full GP model

$$D_{KL}(P||Q) = \int \log \frac{dP}{dQ} dP \ .$$

- Theoretically: both NNGP models are "admissible"...
- ...but, in practice, the latent NNGP model tends to (not always!) outperform the response NNGP:

$$\begin{aligned}(C + \tau^2 I)^{-1} - (\tilde{C} + \tau^2 I)^{-1} &= C^{-1} - C^{-1}M^{-1}C^{-1} - \tilde{C}^{-1} + \tilde{C}^{-1}M^{*-1}\tilde{C}^{-1} \\ &= \underbrace{E - EM^{-1}\tilde{C}^{-1} - \tilde{C}^{-1}M^{-1}E - \tilde{C}^{-1}(M^{-1} - M^{*-1})\tilde{C}^{-1}}_{B} - \underbrace{EM^{-1}E}_{\mathcal{O}(E^2)}\end{aligned}$$

- E is the error from Vecchia (or response NNGP) approximation of full GP
- The leading matrix B tends to shrink the order of the approximation further...

$$||B||_F \leq ||E||_F$$

Conjugate latent NNGP models

$$[\text{data} \,|\, \text{process}] \times [\text{process} \,|\, \text{parameters}] \times [\text{parameters}] \,.$$

$$y(\ell_i) \overset{ind}{\sim} N(x(\ell_i)^\top \beta + w(\ell_i), \sigma^2\delta^2) \,, i = 1, 2, \ldots, n$$
$$w = \{w(\ell_i)\} \sim N(0, \sigma^2 \tilde{M}); \quad \{\beta, \sigma^2\} \sim NIG(\mu_\beta, V_\beta, a_\sigma, b_\sigma)$$

Hierarchical linear model:

$$\underbrace{\begin{bmatrix} \frac{1}{\delta} y \\ L_\beta^{-1} \mu_\beta \\ 0 \end{bmatrix}}_{y_*} = \underbrace{\begin{bmatrix} \frac{1}{\delta} X & \frac{1}{\delta} I_n \\ L_\beta^{-1} & O \\ O & D^{-\frac{1}{2}}(I - A) \end{bmatrix}}_{X_*} \underbrace{\begin{bmatrix} \beta \\ w \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}}_{\eta}$$

The posterior distribution of γ and σ^2 is

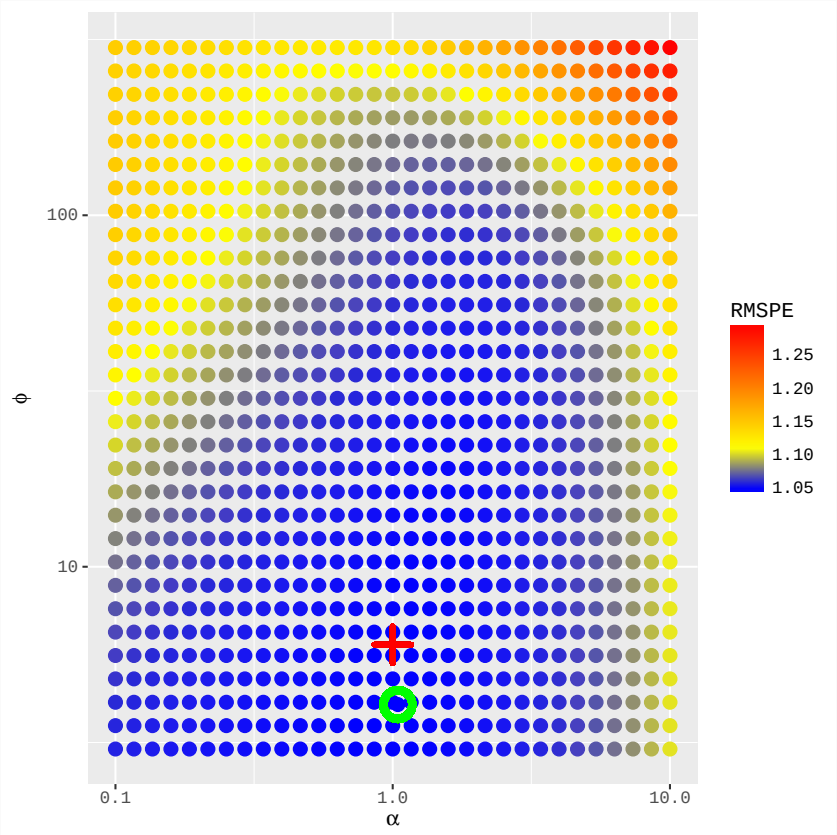
$$p(\gamma, \sigma^2 \,|\, y) \propto IG(\sigma^2 \,|\, a_*, b_*) \times N(\gamma \,|\, \hat{\gamma}, \sigma^2 (X_*^\top X_*)^{-1})$$

Storage and computational complexity $O(n(m+1)^2)$.

Choosing δ^2 and ϕ

- ϕ and δ^2 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 δ^2 can be decided based on computing resources available
- In practice, a reasonably coarse grid often suffices

Choosing δ^2 and ϕ



RMSPE

Figure: Simulation experiment: True value ($+$) of (δ^2, ϕ) and estimated value ($\circ$) using 5-fold cross validation

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

Comparison of computing times for different NNGP algorithms

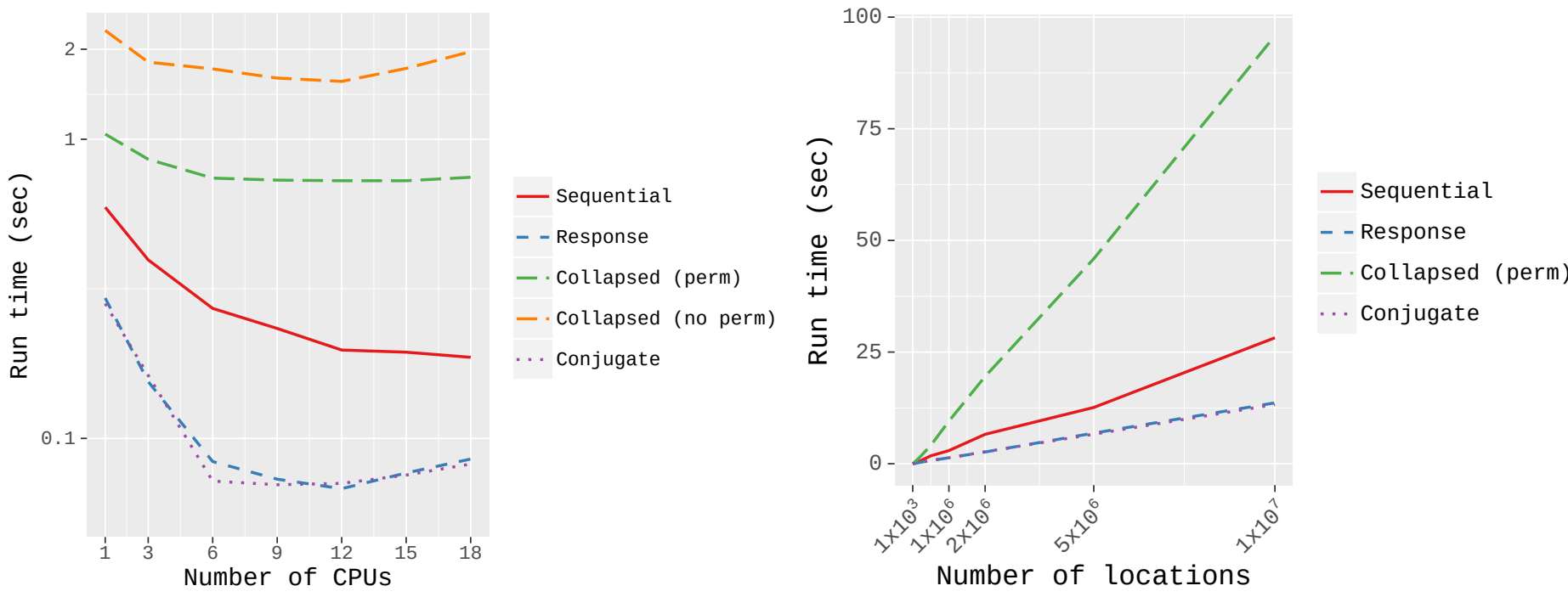


Figure: (a) Run time required for one sampler iteration using $n{=}5 \times 10^4$ by number of CPUs (y-axis is on the log scale). (b) Run time required for one sampler iteration by number of locations.

Concluding remarks

- Model-based solution for spatial "BIG DATA"
- Available in the spNNGP package in `R`
- Algorithms: Gibbs, RWM, HMC, VB or INLA; HMC is especially promising on `STAN`.
- Multivariate Geostatistics: Conjugate NNGP models using Matrix-variate Normal-IW family.
- Challenges: Nonstationary models; High-dimensional outcomes; High-dimensional domains; Smoother process approximations.

Hamiltonian Monte Carlo Implementation of NNGP models

NNGP using Hamiltonian Monte Carlo

http://mc-stan.org/users/documentation/case-studies/nngp.html

- The Metropolis-Hastings algorithm: Sample from any *target* probability density, e.g., posterior density $p(\theta \mid y) \propto p(\theta) \times f(y \mid \theta)$
- Start with a initial value for $\theta = \theta^{(0)}$. Repeat for $j = 1, 2, \ldots, M$:
  1. Propose $\theta^* \sim Q(\cdot \mid \theta^{(j-1)})$. For example, $Q(\cdot \mid \theta^{(j-1)}) = N(\cdot \mid \theta^{(j-1)}, \nu)$.
  2. Compute

$$A(\theta^* \mid \theta^{(j-1)}) = \min\left(1, \frac{p(\theta^* \mid y) Q(\theta^{(j-1)} \mid \theta^*)}{p(\theta^{(j-1)} \mid y) Q(\theta^* \mid \theta^{(j-1)})}\right)$$

  3. *Accept* $\theta^{(j)} = \theta^*$ with probability $A(\theta^* \mid \theta^{(j-1)})$.
- MH works because it leaves the target invariant (satisfies detailed balance):

$$p(\theta \mid y) T(\theta' \mid \theta) = p(\theta' \mid y) T(\theta \mid \theta')$$

- Hamiltonian Monte Carlo: Use (discretized) Hamiltonian dynamics using *symplectic integrators* to propose in MH.

Selected papers

- Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, 12, 583–614.
- Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian predictive process models for large spatial datasets. *Journal of the Royal Statistical Society Series B*, 70, 825–848.
- Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). Hierarchical Nearest-Neighbor Gaussian Process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111, 800–812.
- Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A. S., and Schaap, M. (2016b). Non-separable dynamic Nearest-Neighbor Gaussian Process models for large spatio-temporal data with an application to particulate matter analysis. *Annals of Applied Statistics*, 10, 1286–1316.
- Finley, A. O., Datta, A., Cook, B. C., Morton, D. C., Andersen, H. E., and Banerjee, S. (in press). Applying Nearest Neighbor Gaussian Processes to massive spatial data sets: Forest canopy height prediction across Tanana Valley Alaska. *Journal of Computational and Graphical Statistics arXiv:1702.00434*.
- Guinness, J. (2016). Permutation methods for sharpening Gaussian Process approximations. *arXiv:1609.05372*.
- Heaton, M., Datta, A., Finley, A., Furrer, R., Guhaniyogi, R., Gerber, F., Hammerling, D., Katzfuss, M., Lindgren, F., Nychka, D., and Zammit-Mangion, A. (2017). Methods for analyzing large spatial data: A review and comparison. *arXiv:1710.05013*.
- Katzfuss, M. (2017). A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112, 201–214.
- Katzfuss, M. and Guinness, J. (2017). A general framework for vecchia approximations of gaussian processes. *arXiv preprint arXiv:1708.06302*.
- Nychka, D., Bandyopadhyay, S., Hammerling, D., Lindgren, F., and Sain, S. (2015). A multiresolution gaussian process model for the analysis of large spatial datasets. *Journal of Computational and Graphical Statistics*, 24, 579–599.
- Nychka, D., Wikle, C., and Royle, J. A. (2002). Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2, 315–331.
- Zhang, L., Datta, A. and Banerjee, S. (2018). Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments. *arXiv:1802.00495*.

Thank You!