Approximating Likelihoods for Large Environmental Datasets

Ying Sun

Environmental Statistics Group

es.kaust.edu.sa

King Abdullah University of Science and Technology

Joint work with Michael Stein (University of Chicago) and Huang Huang (Ph.D. student from KAUST)

Large Datasets Problem

- Large observational and computer-generated datasets:
  - Often have spatial and temporal aspects.
  - Nearly global coverage.
  - High resolutions.
- Satellite measurements:
  - MODIS: data at about 60 million locations daily since 1999.
  - TOMS/OMI: measures daily total column ozone since 1978 with nearly global coverage.
- Computer model outputs: even larger datasets.
  - NCAR climate models: CCSM, CESM, NRCM.
  - EPA air quality models: CMAQ.

Outline

- How to fit a Gaussian process model to large and irregularly spaced sets of observations?
  - Existing methods
  - New estimating equations
  - Composite likelihoods
- Computational and statistical efficiency

Gaussian Processes

EnvStat
@ KAUST

- Gaussian process models can be used to
  - describe the spatial variability in the process.
  - predict unobserved values; provide prediction uncertainties.
  - serve as a building block for more complex models.
- Gaussian process Z on a domain $\mathcal{D} \subset \mathbb{R}^d$ is fully specified by
  - $\mu(x) = E\{Z(x)\}$, and
  - $K(x, y) = \text{cov}\{Z(x), Z(y)\}$, for all $x, y \in \mathcal{D}$.
- This talk focuses on:
  - Estimation of K when specified up to $\boldsymbol{\theta} \in \mathbb{R}^p$.
  - Assume $\mu = 0$ for simplicity.

Maximum Likelihood Estimation

Suppose data $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$ is observed from a stationary and isotropic Gaussian process $Z \sim GP(0, K(h; \boldsymbol{\theta}))$ at n locations.

- Goal: estimate $\boldsymbol{\theta} \in \mathbb{R}^p$ by likelihood methods. Up to an additive constant, the loglikelihood is

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T \Sigma_{n\times n}^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log\left|\Sigma_{n\times n}(\boldsymbol{\theta})\right|.$$

- The standard way:
  - Cholesky decomposition of $\Sigma_{n\times n}$.
  - Generally requires $O(n^3)$ computations and $O(n^2)$ memory.
- The covariance matrix Σ_{nxn} is
  - large: $n \times n$ for n locations.
  - unstructured: irregular spaced locations.
  - dense: non-negligible correlations.

Exact MLE Computation

- Loglikelihood:

$$\ell(\boldsymbol{\theta}) = -\frac{1}{2}\mathbf{Z}^T\Sigma_{n\times n}^{-1}(\boldsymbol{\theta})\mathbf{Z} - \frac{1}{2}\log\left|\Sigma_{n\times n}(\boldsymbol{\theta})\right|.$$

- Score equations:

$$\mathbf{Z}^T\Sigma^{-1}\Sigma_i\Sigma^{-1}\mathbf{Z} - \mathrm{tr}(\Sigma^{-1}\Sigma_i) = 0, \quad i = 1, \ldots, p,$$

  where $\Sigma_i = \partial\Sigma(\boldsymbol{\theta})/\partial\theta_i$.
- Computing $\Sigma^{-1}\mathbf{Z}$: best done by solving systems $\Sigma\mathbf{x} = \mathbf{Z}$.
- Loglikelihood:
  - Main computation is due to calculating $\log|\Sigma|$.
- Score equations:
  - Need n solves to compute $\mathrm{tr}(\Sigma^{-1}\Sigma_i)$.
  - May not be any easier than computing $\log|\Sigma|$.

Iterative Solution of Linear Equations

EnvStat
@ KAUST

- Solving $\Sigma \mathbf{x} = \mathbf{Z}$: conjugate gradient method.
- Matrix-free:
  - Never have to store an $n \times n$ matrix.
  - Computation is becoming cheap much faster than memory.
- Main computation: matrix-vector multiplication.
  - Requires $O(n^2)$ for dense and unstructured matrices.
  - This is fast, if Σ has some exploitable structures.
- Let m be the number of iterations:

$$O(n^2 \times m) \quad vs. \quad O(n^3).$$

- Preconditioning is important to reduce m, i.e., to solve

$$W\Sigma \mathbf{x} = W\mathbf{Z}.$$

Large n

- Options for large n:
  - Use models that reduce computations and/or storage.
  - Use approximate methods.
  - Both.
- Models that might allow for exact computations:
  - Compactly supported covariance functions.
  - Reduced rank covariance functions.
  - Markov models.
- Approximation methods:
  - Approximating likelihoods.
  - Approximating score equations.

Existing Methods

- Popular methods:
  - Covariance tapering:
    - Furrer et al. (2006), Kaufman et al. (2008), Stein (2013).
  - Low rank approximations:
    - Banerjee et al. (2008), Cressie and Johannesson (2008).
  - Composite likelihoods:
    - Vecchia (1988), Stein et al. (2004), Sun and Stein (2015).
  - Low rank+tapering:
    - Sang and Huang (2011).
  - Multi-resolution models:
    - Nychka et al. (2015).
  - Markov models:
    - Rue et al. (2009), Lindgren et al. (2011).
- Key ideas: use sparsity and/or low rank representation.

Assessing Approximation Quality

- Approximation methods:
  - Computational efficiency.
  - Statistical efficiency.
- How to assess the quality of an approximation: depends in part on the use for the approximate model.
  - Prediction.
  - Estimation.
  - Model approximation.
- Possible measures:
  - Godambe information.
  - Kullback-Leibler (KL) divergence.

Score Equation Approximations (Sun and Stein, JCGS, 2015)

- Score equations:

$$\mathbf{Z}^T\Sigma^{-1}\Sigma_i\Sigma^{-1}\mathbf{Z} - \mathrm{tr}(\Sigma^{-1}\Sigma_i) = 0, \quad i = 1, \ldots, p,$$

- Since $E(\mathbf{Z}^T A_i \mathbf{Z}) = \mathrm{tr}(A_i\Sigma)$, a general way is:

$$\mathbf{Z}^T A_i\mathbf{Z} - \mathrm{tr}(A_i\Sigma) = 0, \quad i = 1, \ldots, p.$$

where A_i is chosen to

  - approximate $\Sigma^{-1}\Sigma_i\Sigma^{-1}$.
  - avoid multiple solves in Σ.
  - retain as much statistical efficiency as possible.

Unbiased Estimating Equations

- Let V be a sparse approximation of Σ^{-1}.
- Estimating equations (E_1):

$$\mathbf{Z}^T V \Sigma_i \Sigma^{-1} \mathbf{Z} - \mathrm{tr}(V \Sigma_i) = 0, \quad i = 1, \ldots, p.$$

  - Quadratic term: compute $\Sigma^{-1}\mathbf{Z}$, one solve.
  - Trace term: diagonal elements of $V\Sigma_i$ with a sparse V.
  - Avoids n solves, $\mathrm{tr}(\Sigma^{-1}\Sigma_i)$, in score equations.
- Estimating equations (E_2):

$$\mathbf{Z}^T V \Sigma_i V \mathbf{Z} - \mathrm{tr}(V \Sigma_i V \Sigma) = 0, \quad i = 1, \ldots, p.$$

  - No solves are involved.
  - Trace term: generally requires more calculations.
- Linear combination (E_3): $2E_1 - E_2$

$$\mathbf{Z}^T \{2V\Sigma_i\Sigma^{-1} - V\Sigma_i V\}\mathbf{Z} - \mathrm{tr}\{2V\Sigma_i - V\Sigma_i V\Sigma\} = 0.$$

Statistical Properties of Estimating Equations

- Compare to score equations:

$$\mathrm{var}\{\mathbf{Z}^T(A_i - \Sigma^{-1}\Sigma_i\Sigma^{-1})\mathbf{Z}\}.$$

- Define V_k as the first k terms of the Taylor expansion

$$\Sigma^{-1} = \xi\{I - (I - \xi\Sigma)\}^{-1} = \xi\sum_{j=0}^{\infty}(I - \xi\Sigma)^j,$$

we have shown that

  - E_1 is better than E_2.
  - E_3 might be better than E_1 and E_2,

Comparisons

- Covariance tapering:
  - Kaufman et al. (2008) and Furrer et al. (2006).
  - Assume Σ is sparse.
  - Σ^{-1} is not generally sparse.
  - Statistical properties of covariance tapers (Stein, 2012).
  - Our methods: approximate Σ^{-1} by a sparse matrix V.
- Markov random field models:
  - Assume Σ^{-1} is actually sparse.
  - Our methods: do not necessarily replace Σ^{-1} by V everywhere.

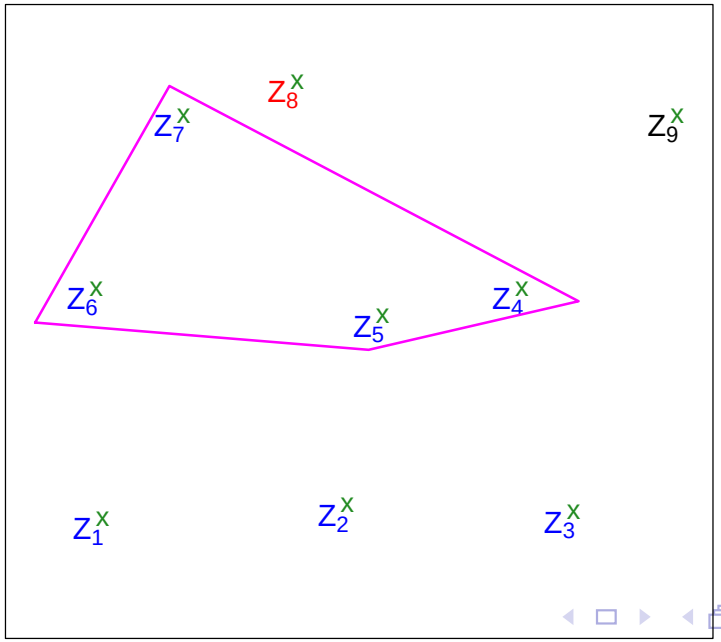
Approximating Σ^{-1}

EnvStat
@ KAUST

- Find V: sparse approximation of Σ^{-1}.
- Sparse inverse Cholesky:
  - Vecchia (1988): likelihood approximation.
  - Kolotilina and Yeremin (1993): preconditioning.
  - Sun and Stein (2015): score equation approximation.
- Links to composite likelihood methods:
  - Joint density: product of conditional densities.
  - Condition on only subset of the "past" observations.
  - Sparseness of V depends on the length of conditioning sets.
  - Obtain Σ^{-1}: use complete conditioning sets.

Nearest Neighbors

- Example: $f(\mathbf{Z}) = f(Z_1)f(Z_2|Z_1)\ldots f(Z_9|Z_1,\ldots,Z_8)$.
- Conditioning sets of Z_8:
  - Complete set: $\{Z_1,\ldots,Z_7\}$.
  - Subset: $\{Z_4, Z_5, Z_6, Z_7\}$.



Evaluating Statistical Efficiency

- $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{0}$: a set of unbiased estimating equations.
- $\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})$: $p \times p$ matrix whose jth column is $\partial \boldsymbol{\psi}(\boldsymbol{\theta}) / \partial \theta_j$.
- Godambe information matrix:

$$\mathbf{G}(\boldsymbol{\theta}) = \left(\mathrm{E}\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})\right)^T \left(\mathrm{E}\boldsymbol{\psi}(\boldsymbol{\theta})\boldsymbol{\psi}(\boldsymbol{\theta})^T\right)^{-1} \left(\mathrm{E}\dot{\boldsymbol{\psi}}(\boldsymbol{\theta})\right).$$

- $\mathbf{G}(\boldsymbol{\theta})$ is the Fisher information matrix, if $\boldsymbol{\psi}(\boldsymbol{\theta}) = \mathbf{0}$ is the exact score equation.
- $g^{ii}(\boldsymbol{\theta})$: the diagonal elements of $\mathbf{G}^{-1}(\boldsymbol{\theta})$, $i = 1, \ldots, p$.
- Evaluate statistical efficiency: $\sqrt{g^{ii}(\boldsymbol{\theta})}$.
- $\psi_i = \mathbf{Z}^T A_i \mathbf{Z} - \mathrm{tr}(A_i \Sigma)$: straightforward to compute $\mathbf{G}(\boldsymbol{\theta})$.

Conditioning Sets

- Order observations such that $\mathbf{Z} = (Z_1, \ldots, Z_n)^T$.
- Any joint density can be written as a product of conditional densities, then the loglikelihood is

$$\ell(\boldsymbol{\theta}|\mathbf{Z}) = \log f(Z_1; \boldsymbol{\theta}) + \sum_{j=2}^{n} \log f(Z_j|\mathbf{Z}_{j-1}; \boldsymbol{\theta}),$$

  where $\mathbf{Z}_{j-1} = (Z_1, \ldots, Z_{j-1})^T$.
- Is it enough to choose r nearest neighbors of Z_j?
- How to choose the subset or how to choose the conditioning set in general?

Neighbor Selection

Approximating $f(Z_j|\mathbf{Z}_{j-1};\boldsymbol{\theta})$:

- Key idea: for each $j > r$, $r \ll n$, choose the conditioning set of rank r, such that only a $r \times r$ matrix needs to be inverted.
- Let $A_{j,r}$ be a $(j-1) \times r$ matrix with full column rank r for neighbor selection.
- Each column of $A_{j,r}$ gives a combination of $\mathbf{Z}_{j-1}$ in the conditioning set of Z_j.

Examples of A

We consider 4 cases of the conditioning set:

1. Independent (**IND**): $A_{j,r}$ is a $\mathbf{0}$ matrix. No neighbors are selected.
2. r nearest neighbors (**NN**): each column of $A_{j,r}$ has one element of 1, and r selected neighbors in total.
3. r sets of nearest neighbors (**SUM**): each column of $A_{j,r}$ has $m > 1$ elements of 1, and mr selected neighbors in total.
4. r_1 nearest neighbors and r_2 sets (**NN+SUM**): $r = r_1 + r_2$, and $r_1 + mr_2$ selected neighbors in total.

Selected Neighbors

From case 2 to case 3,

- The number of selected neighbors increases from r to mr while only a matrix of rank r needs to be inverted.
- The sum of each m neighbors is considered.
- Can we find a better linear combination for the mr selected neighbors?

Hierarchical Low Rank Approximation

- **HLR** approximation: for the jth hierarchy and $j > r$, let $A_{j,mr}$ indicate the mr selected nearest neighbors.
- Approximate $A_{j,mr}^T \Sigma_{j-1} A_{j,mr}$ by a low rank representation

$$P_j L_j P_j^T + \epsilon_j^2 I_{mr},$$

where L_j is p.d. of $r \times r$, P_j is $mr \times r$ consisting of r basis functions, and ϵ_j^2 is the nugget.

- Use Sherman-Morrison-Woodbury formula for the inversion, and still only need to invert a $r \times r$ matrix.

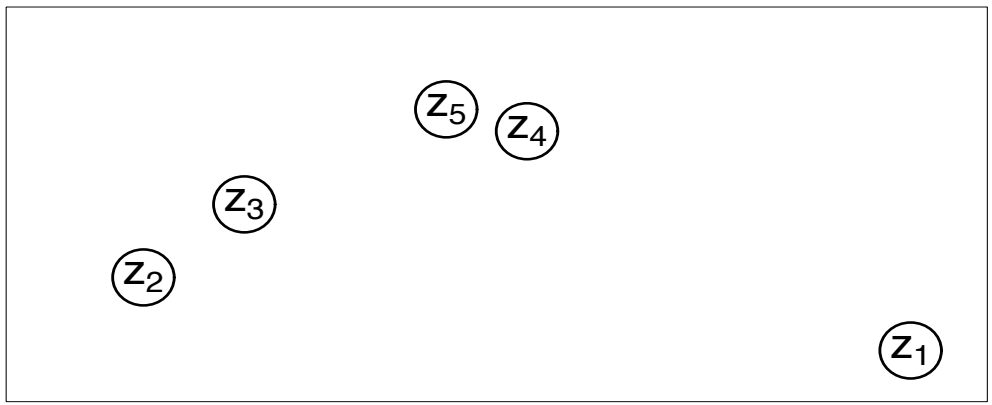
$$(P_j L_j P_j^T + \epsilon_j^2 I_{mr})^{-1} = \epsilon_j^{-2} I_{mr} - \epsilon_j^{-4} P_j (L_j^{-1} + \epsilon_j^{-2} P_j^T P_j)^{-1} P_j^T$$

Illustration

Suppose $\mathbf{Z} = (Z_1, \ldots, Z_5)^T$ and $r = 2$. For $f(Z_5 \mid Z_4, \ldots, Z_1)$,

- **NN**: $f(Z_5 \mid Z_4, Z_3)$
- **SUM**: $f(Z_5 \mid Z_4 + Z_3, Z_2 + Z_1)$
- **NN+SUM**: $f(Z_5 \mid Z_4, Z_3 + Z_2)$
- **HLR**:

$$f(Z_5 \mid a_{14}Z_4 + a_{13}Z_3 + a_{12}Z_2 + a_{11}Z_1, a_{24}Z_4 + a_{23}Z_3 + a_{22}Z_2 + a_{21}Z_1)$$



Theoretical Comparisons

We have shown that the approximation to Σ_{jj} induced by V_{jj}^{H} is better than that induced by V_{jj}^{N} in terms of the Frobenius norm.

Theorem

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{mr} > 0$ be the eigenvalues of $A_{j,mr}^T \Sigma_{jj} A_{j,mr}$. If $\epsilon_j^2 < (\lambda_r + \lambda_{mr})/2$, we have,

$$\|A_{j,mr} V_{jj}^{\mathrm{H}} A_{j,mr}^T - \Sigma_{jj}\|_F \leq \|A_{jr} V_{jj}^{\mathrm{N}} A_{jr}^T - \Sigma_{jj}\|_F,$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Numerical Studies

Consider a mean-zero GP observed at irregularly spaced $n = 900$ locations in $[0, 1] \times [0, 1]$:

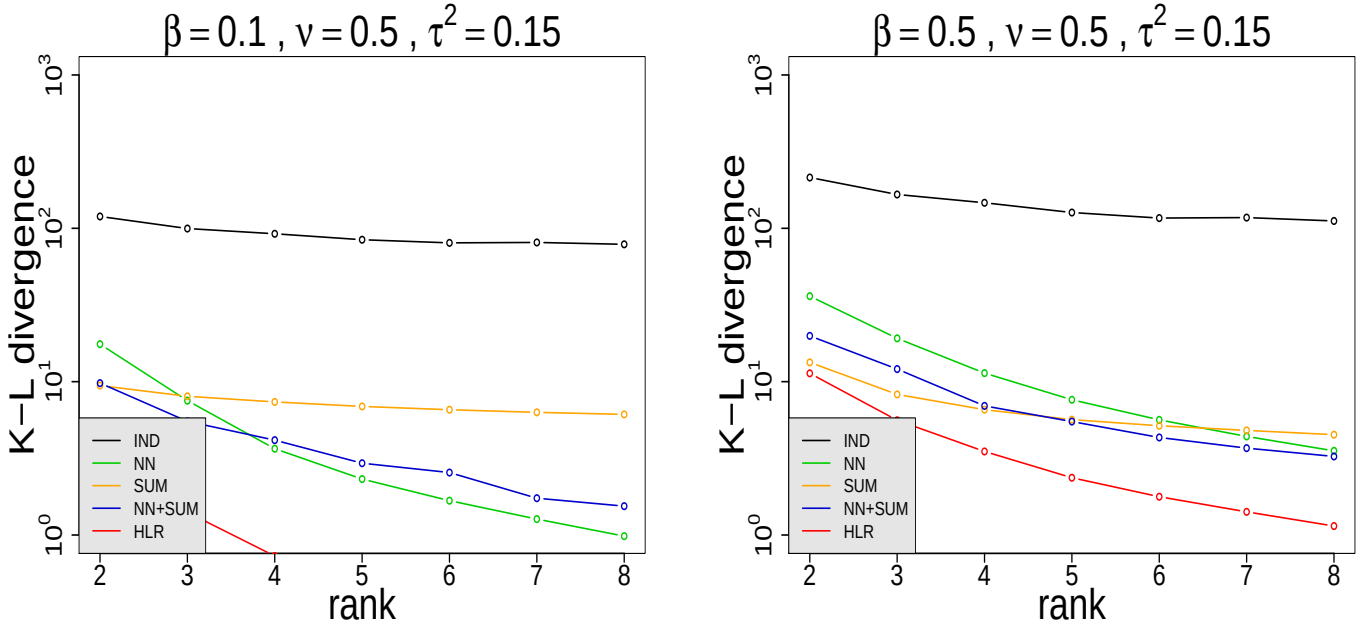
- Matérn covariance function has parameters: scale $\phi = 1$, range β, smoothness ν, and nugget τ^2.
- Compare different approximations by Kullback-Leibler divergence, where r/n is from 0.2% to 0.9%, and $m = 2$.

$$D_{\text{K-L}}(N_e \| N_a) = \frac{1}{2}\left\{\text{tr}(\Sigma_a^{-1}\Sigma_e) + \log(|\Sigma_a|) - \log(|\Sigma_e|) - n\right\},$$

where N_e and N_a are the exact and the approximated Gaussian distributions, respectively.

Dependence Level

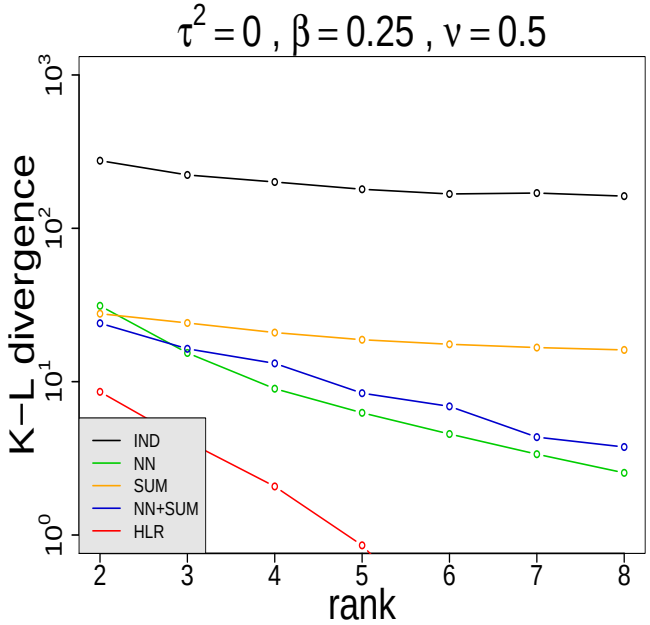
- Exponential covariance with a nugget effect and $\beta = 0.1, 0.5$.

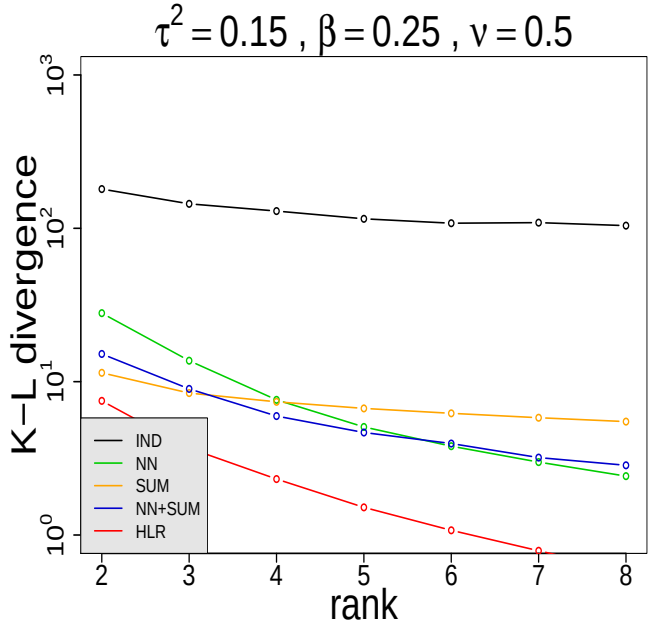


- It shows how **HLR**, **SUM**, and **NN+SUM** improve the **NN** approximation by conditioning on more neighbors.

Noise Level

- Exponential covariance with and without a nugget effect.

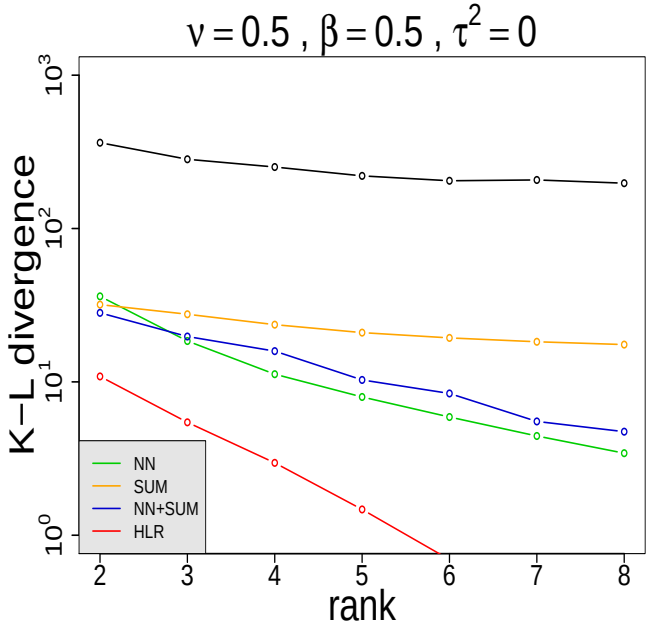


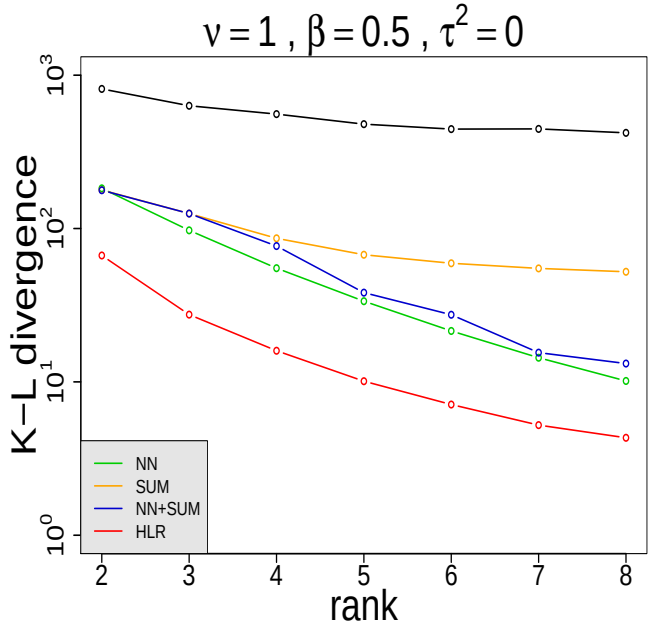


- It suggests that **HLR** is the best, and **SUM** and **NN+SUM** are useful when the process is noisy and r is small.

Smoothness Level

- Exponential and Whittle covariance models without nugget.

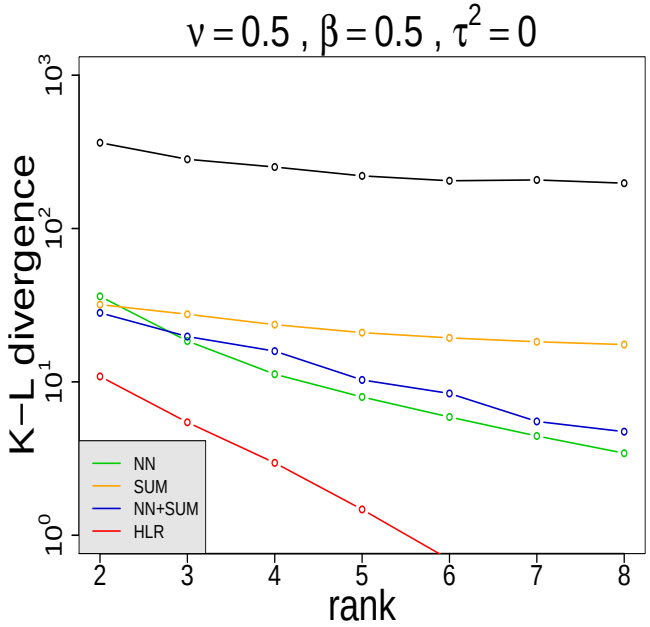


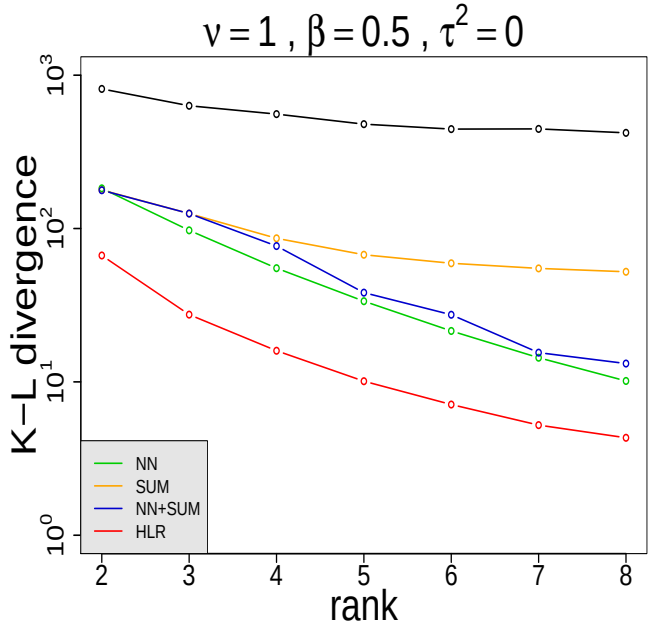


- **HLR** shows significant improvement, but the improvement of **SUM** and **NN+SUM** over **NN** is negligible.

Smoothness Level

- Exponential and Whittle covariance models without nugget.



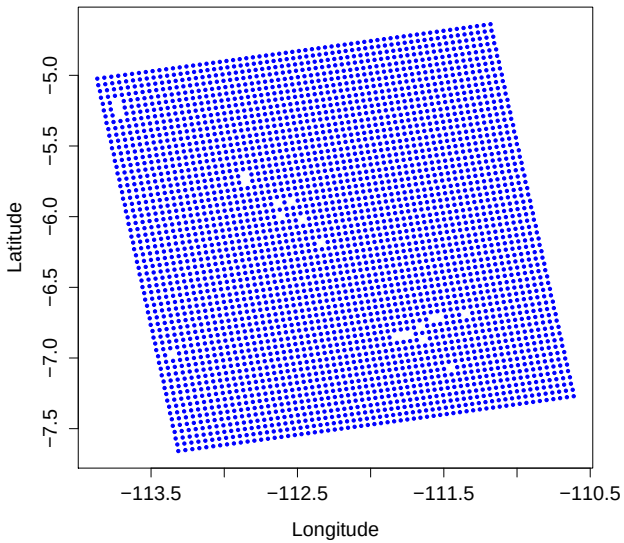


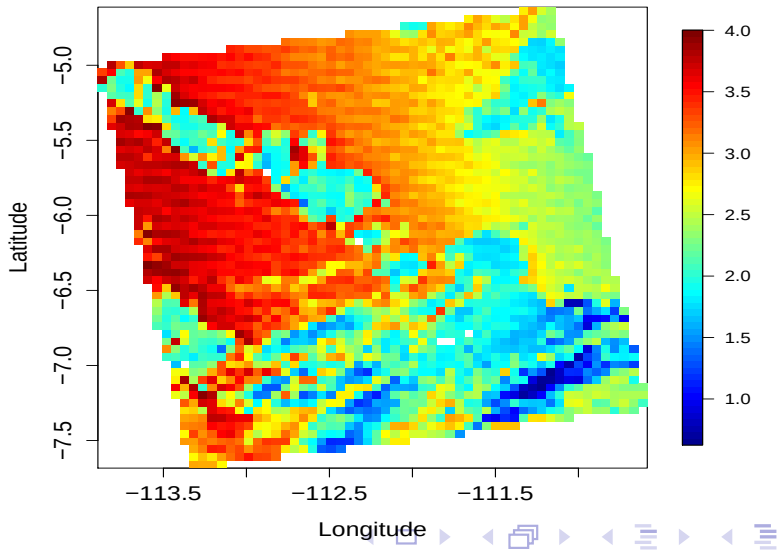
- **HLR** shows significant improvement, but the improvement of **SUM** and **NN+SUM** over **NN** is negligible.

Application: High Resolution Satellite Data

Fit a GP model to explore the spatial variability of water vapor

- Precipitable water product from MODIS at 1 km $\times$ 1 km spatial resolution
- Study region: one region of the Southeast Pacific Ocean
- Datasets: **89,479** measurements with 521 missing values

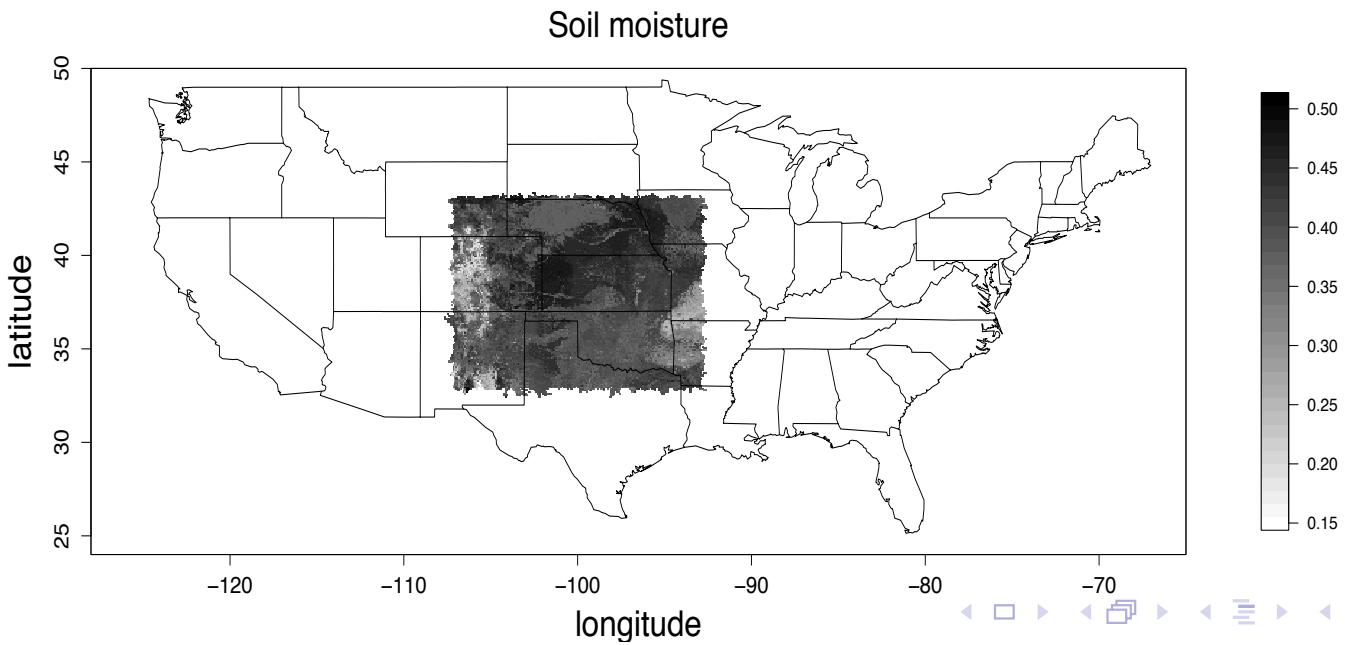


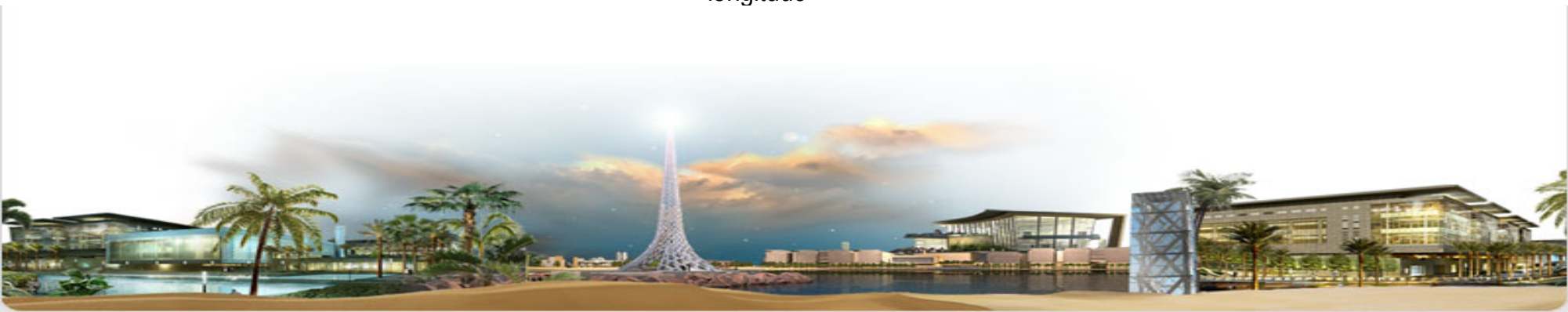


Application: Computer Model Output

Soil moisture at the top layer of the Mississippi basin

- Randomly choose **2 million** irregularly spaced observations
- Fit a GP model with a Matérn covariance (HLR and NN)
- The fitted models are used to predict the left out observations





Discussion

Fit a Gaussian process model to large and irregularly spaced sets of observations

- New **unbiased** estimating equations:
  - **Sparse** approximation of Σ^{-1}: sparse inverse Cholesky
  - Require **one solve**: iterative methods and preconditioning
  - Main **computation**: matrix-vector multiplication
- The proposed **HLR** method:
  - Provides a **unified** framework for likelihood approximation
  - The approximated covariance is **not** low rank
  - It is not necessary to fix the rank r for different hierarchies
  - Suitable for **parallel computing**

References

- Huang, H. and Sun, Y. (2016), "Hierarchical low rank approximation of likelihoods for large spatial datasets," manuscript.
- Sun, Y. and Stein, M. L. (2016), "Statistically and computationally efficient estimating equations for large spatial datasets," *Journal of Computational and Graphical Statistics*, 25, 187-208.
- Sun, Y., Li, B. and Genton, M. G. (2012), "Geostatistics for large datasets," in Advances And Challenges In Space-time Modelling Of Natural Events, J. M. Montero, E. Porcu, M. Schlather (eds), Springer, Vol. 207, Chapter 3, 55-77.