Large Scale Kriging: A High Performance Multi-Level Computational Mathematics Approach

Julio E. Castrillón Candás[1], Marc G. Genton[2], Xiaoyu Wang[1], Rio Yokota[3]

[1] Department of Mathematics and Statistics, Boston University
[2] Division of Applied Mathematics and Computational Science, King Abdullah University of Science and Technology.
[3] Global Scientific Information and Computing Center, Tokyo Inst. of Tech.

December 15, 2019

Kriging: Problem Setup

Consider the following model random field Z:

$$Z(\mathbf{s}) = \underbrace{\mathbf{m}(\mathbf{s})^{\mathrm{T}}\boldsymbol{\beta}}_{\textit{Deterministic}} + \underbrace{\varepsilon(\mathbf{s})}_{\textit{Random}}, \qquad \mathbf{s} \in \mathbb{R}^d, \tag{1}$$

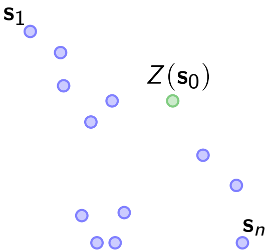
where

- $\mathbf{m}(\mathbf{s}) := [m_1(\mathbf{s}), \ldots, m_p(\mathbf{s})]^{\mathrm{T}}$ is a known function vector with respect to the location $\mathbf{s} := [s_1, \ldots, s_d]^{\mathrm{T}}$
- $\boldsymbol{\beta} := [\beta_1, \ldots, \beta_p]^{\mathrm{T}}$ is an unknown vector of coefficients
- ε is a stationary zero mean Gaussian random field with parametric covariance function $C(\mathbf{s}, \mathbf{s}'; \boldsymbol{\theta}) = \mathrm{cov}\{\varepsilon(\mathbf{s}), \varepsilon(\mathbf{s}')\}$ having an unknown vector $\boldsymbol{\theta} := [\theta_1, \ldots, \theta_w]^{\mathrm{T}}$ of parameters

Problem Setup

We observe $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))^{\mathrm{T}}$ at $\mathbb{S} := \{\mathbf{s}_1, \dots, \mathbf{s}_n\}$, and wish to:

- (**Estimation**) estimate the unknown vectors $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ by maximizing the log-likelihood function
- (**Prediction**) predict the Best Linear Unbiased Predictor (BLUP) $Z(\mathbf{s}_0)$ at a new location $\mathbf{s}_0 \in \mathbb{R}^d$



- Vectorial form

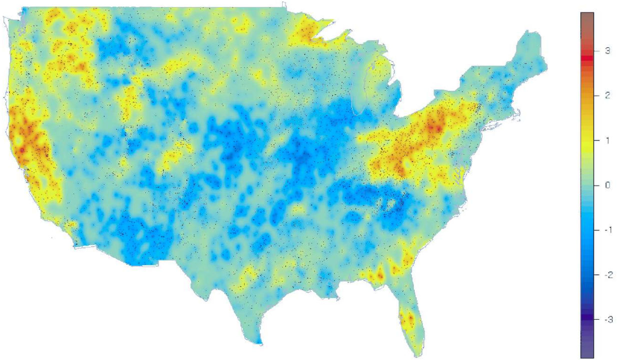
$$\mathbf{Z} = \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2}$$

where $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\mathbf{0}, \mathbf{C}(\boldsymbol{\theta}))$, and

$$\mathbf{M} := \begin{bmatrix} m_1(\mathbf{s}_1) & m_2(\mathbf{s}_1) & \dots & m_p(\mathbf{s}_1) \\ \vdots & \vdots & \vdots & \vdots \\ m_1(\mathbf{s}_n) & m_2(\mathbf{s}_n) & \dots & m_p(\mathbf{s}_1) \end{bmatrix}$$

Applications

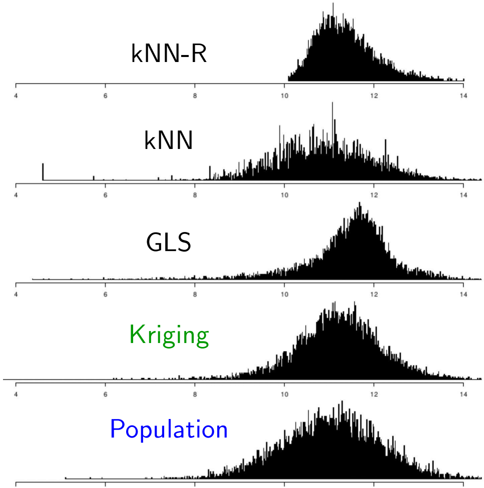
Environmental science (Kriging): Estimation of missing data from high resolution satellite data and monitoring networks for temperature, rainfall, pressure, air quality, etc.



Estimation and prediction of precipitation anomaly field with 5,906 observations.

Applications

Missing Data Hospital Datasets 90,000/10,000 (Training / Validation)



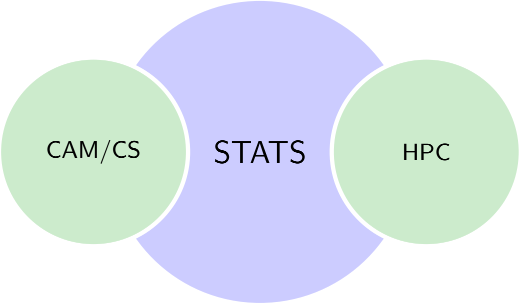
Challenges

- Machine Learning / Regression challenge: Estimate noise model with good accuracy

$$Z(\mathbf{s}) = \mathbf{m}(\mathbf{s})^{\mathrm{T}} \underbrace{\beta}_{\text{Unknown}} + \underbrace{\varepsilon(\mathbf{s})}_{\text{Unknown}}, \qquad \mathbf{s} \in \mathbb{R}^d$$

(a) Challenge: Data points n and dimensions d are large and suffers from the Curse of Dimensionality.
(b) Challenge: Usually numerically unstable

Leverage Computational Applied Mathematics (CAM), Computer Science (CS) and High Performance Computing (HPC)



Estimation

The log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{C}(\boldsymbol{\theta})\} - \frac{1}{2}(\mathbf{Z} - \mathbf{M}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\boldsymbol{\beta}),$$

that can be profiled by the generalized least squares with

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}.$$

Maximum Likelihood Estimation

$\mathrm{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^w}\, \ell(\hat{\beta}, \boldsymbol{\theta})$

Problems:

Estimation

The log-likelihood function is

$$\ell(\boldsymbol{\beta}, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{C}(\boldsymbol{\theta})\} - \frac{1}{2}(\mathbf{Z} - \mathbf{M}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\boldsymbol{\beta}),$$

that can be profiled by the generalized least squares with

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}.$$

Maximum Likelihood Estimation

$\operatorname{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^w} \ell(\hat{\beta}, \boldsymbol{\theta})$

Problems:

- $\mathbf{C}(\boldsymbol{\theta})$: Large scale, ill-conditioned

  Bad condition numbers are death – Vladimir Rokhlin

- $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}$ are coupled, Curse of dimensionality with respect to d

Prediction

The minimization of $\mathrm{E}[\{Z(\mathbf{s}_0) - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}\}^2]$ under the constraint $\mathbf{M}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{m}(\mathbf{s}_0)$ yields the optimal unbiased estimate

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z},\\
\hat{\gamma}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\boldsymbol{\beta}}),\\
\hat{Z}(\mathbf{s}_0) &= \mathbf{m}(\mathbf{s}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{c}(\boldsymbol{\theta})^{\mathrm{T}}\hat{\gamma}(\boldsymbol{\theta}),
\end{aligned}$$

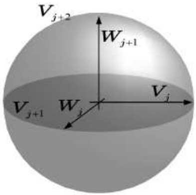
where $\mathbf{c}(\boldsymbol{\theta}) = \mathrm{cov}\{\mathbf{Z}, Z(\mathbf{s}_0)\} \in \mathbb{R}^n$.

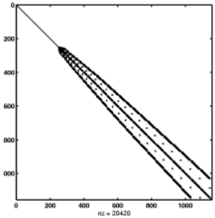
Problems:

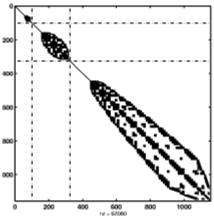
- $\mathbf{C}(\boldsymbol{\theta})$: Large scale, ill-conditioned (Still dead).
- $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\beta}}$ are coupled
- Curse of dimensionality with respect to d

Different Approaches

- Current techniques: FFT, Hierarchical Matrices, Heuristics, etc: Limited to small dimensions (2D and 3D), grid like geometries or $\mathbf{C}(\boldsymbol{\theta})$ relatively well conditioned and fast decay [FGN06, KSN08, SLG12, ACW12, SS14, LSGK16].
- Promising approach: Pivoted Cholesky decomposition [LM15]
- We exploit techniques from CAM and Partial Differential Equations





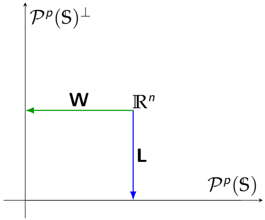


Multi-level Approach

- Proposed Solution: Suppose $\mathcal{P}^p(\mathbb{S}) := span\{\mathbf{M}(:,1),\ldots,\mathbf{M}(:,p)\}$ and we have an orthonormal basis of $\mathbb{R}^n$ such that

$$\mathbb{R}^n \rightarrow \mathcal{P}^p(\mathbb{S}) \oplus \mathcal{P}^p(\mathbb{S})^{\perp}$$

- Build the operators $\mathbf{L} : \mathbb{R}^n \rightarrow \mathcal{P}^p(\mathbb{S})$ and $\mathbf{W} : \mathbb{R}^n \rightarrow \mathcal{P}^p(\mathbb{S})^{\perp}$



- By applying the operator $\mathbf{W}$ to (2) we obtain

$$\mathbf{Z}_W := \mathbf{W}\mathbf{Z} = \mathbf{W}(\mathbf{M}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \mathbf{W}\boldsymbol{\varepsilon}$$

i) First Consequence: $\boldsymbol{\beta}$ is gone. The estimation problem is decoupled (but also the prediction problem!)

Multi-level Estimation

New log-likelihood becomes

$$\ell_W(\mathbf{C}_W(\boldsymbol{\theta}), \mathbf{Z}_W) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{C}_W(\boldsymbol{\theta})\} - \frac{1}{2}\mathbf{Z}_W^{\mathrm{T}}\mathbf{C}_W(\boldsymbol{\theta})^{-1}\mathbf{Z}_W,$$

where $\mathbf{C}_W(\boldsymbol{\theta}) := \mathbf{W}\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}}$ and $\mathbf{Z}_W \sim \mathcal{N}_{n-p}(\mathbf{0}, \mathbf{W}\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}})$

ii) Second Consequence:

If covariance function is smooth entries of $\mathbf{C}_W(\boldsymbol{\theta})$ will decay fast

iii) Third Consequence:

Theorem: If $\kappa(\mathbf{A})$ is the condition number of $\mathbf{A}$ then

$$\kappa(\mathbf{C}_W(\boldsymbol{\theta})) \leq \kappa(\mathbf{C}(\boldsymbol{\theta}))$$

Difference between dead and alive

Multi-level Estimation

- Multi-level log-likelihood

$$\ell_W(\mathbf{C}_W(\boldsymbol{\theta}), \mathbf{Z}_W) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\det\{\mathbf{C}_W(\boldsymbol{\theta})\} - \frac{1}{2}\mathbf{Z}_W^{\mathrm{T}}\mathbf{C}_W(\boldsymbol{\theta})^{-1}\mathbf{Z}_W$$

- Solve decoupled optimization

Maximum Likelihood Estimation

$$\operatorname{argmax}_{\boldsymbol{\theta}\in\mathbb{R}^w} \ell_W(\mathbf{C}_W(\boldsymbol{\theta}), \mathbf{Z}_W)$$

- In practice a sparse version of $\mathbf{C}_W(\boldsymbol{\theta})$ is used instead

Multi-level Prediction

- Prediction: Min. of $\mathrm{E}[\{Z(\mathbf{s}_0) - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}\}^2]$ with constraint: $\mathbf{M}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{m}(\mathbf{s}_0)$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}, \\
\hat{\gamma}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\boldsymbol{\beta}}), \\
\hat{Z}(\mathbf{s}_0) &= \mathbf{m}(\mathbf{s}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{c}(\boldsymbol{\theta})^{\mathrm{T}}\hat{\gamma}(\boldsymbol{\theta})
\end{aligned} \tag{3}$$

Multi-level Prediction

- Prediction: Min. of $\mathrm{E}[\{Z(\mathbf{s}_0) - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}\}^2]$ with constraint: $\mathbf{M}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{m}(\mathbf{s}_0)$

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}, \\ \hat{\gamma}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\boldsymbol{\beta}}), \\ \hat{Z}(\mathbf{s}_0) &= \mathbf{m}(\mathbf{s}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{c}(\boldsymbol{\theta})^{\mathrm{T}}\hat{\gamma}(\boldsymbol{\theta}) \end{aligned} \tag{3}$$

- Alternative equivalent formulation for solving the estimate $\hat{\mathbf{Z}}(\mathbf{s}_0)$:

$$\begin{pmatrix} \mathbf{C}(\boldsymbol{\theta}) & \mathbf{M} \\ \mathbf{M}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{0} \end{pmatrix} \tag{4}$$

Multi-level Prediction

- Prediction: Min. of $\mathrm{E}[\{Z(\mathbf{s}_0) - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}\}^2]$ with constraint: $\mathbf{M}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{m}(\mathbf{s}_0)$

$$\begin{aligned}
\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}, \\
\hat{\gamma}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\boldsymbol{\beta}}), \\
\hat{Z}(\mathbf{s}_0) &= \mathbf{m}(\mathbf{s}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{c}(\boldsymbol{\theta})^{\mathrm{T}}\hat{\gamma}(\boldsymbol{\theta})
\end{aligned} \tag{3}$$

- Alternative equivalent formulation for solving the estimate $\hat{\mathbf{Z}}(\mathbf{s}_0)$:

$$\begin{pmatrix} \mathbf{C}(\boldsymbol{\theta}) & \mathbf{M} \\ \mathbf{M}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{0} \end{pmatrix} \tag{4}$$

- Critical Observation: Constrained problem $(\mathbf{M}^{\mathrm{T}}\hat{\gamma} = \mathbf{0})$

Use transformation of $\mathbb{R}^n$ onto $\mathcal{P}^p(\mathbb{S}) \oplus \mathcal{P}^p(\mathbb{S})^{\perp}$

Multi-level Prediction

- Prediction: Min. of $\mathrm{E}[\{Z(\mathbf{s}_0) - \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{Z}\}^2]$ with constraint: $\mathbf{M}^{\mathrm{T}}\boldsymbol{\lambda} = \mathbf{m}(\mathbf{s}_0)$

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) &= (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z}, \\ \hat{\gamma}(\boldsymbol{\theta}) &= \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\boldsymbol{\beta}}), \\ \hat{Z}(\mathbf{s}_0) &= \mathbf{m}(\mathbf{s}_0)^{\mathrm{T}}\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) + \mathbf{c}(\boldsymbol{\theta})^{\mathrm{T}}\hat{\gamma}(\boldsymbol{\theta}) \end{aligned} \tag{3}$$

- Alternative equivalent formulation for solving the estimate $\hat{\mathbf{Z}}(\mathbf{s}_0)$:

$$\begin{pmatrix} \mathbf{C}(\boldsymbol{\theta}) & \mathbf{M} \\ \mathbf{M}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \hat{\gamma} \\ \hat{\boldsymbol{\beta}} \end{pmatrix} = \begin{pmatrix} \mathbf{Z} \\ \mathbf{0} \end{pmatrix} \tag{4}$$

- Critical Observation: Constrained problem $(\mathbf{M}^{\mathrm{T}}\hat{\gamma} = \mathbf{0})$

Use transformation of $\mathbb{R}^n$ onto $\mathcal{P}^p(\mathbb{S}) \oplus \mathcal{P}^p(\mathbb{S})^{\perp}$

- Thus $\hat{\gamma} \in \mathcal{P}^p(\mathbb{S})^{\perp} \Rightarrow \hat{\gamma} = \mathbf{W}^{\mathrm{T}}\hat{\gamma}_W$ for some $\hat{\gamma}_W \in \mathbb{R}^{n-p}$

Multi-level Prediction

- Write $\mathbf{C}(\boldsymbol{\theta})\hat{\gamma} + \mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{Z}$ as

$$\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}}\hat{\gamma}_W + \mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{Z}$$

Multi-level Prediction

- Write $\mathbf{C}(\boldsymbol{\theta})\hat{\gamma} + \mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{Z}$ as

$$\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}}\hat{\gamma}_W + \mathbf{M}\hat{\boldsymbol{\beta}} = \mathbf{Z}$$

- Applying $\mathbf{W}$ we have

$$\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W = \mathbf{Z}_W$$

  - Solved efficiently with Preconditioned Conjugate Method (PCG) and Kernel Independent Fast Multi-Pole Method (KIFMM).
  - Obtain $\hat{\gamma}$ from γ_W (i.e. $\hat{\gamma} = \mathbf{W}\hat{\gamma}_W$).
  - $\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma})$

Multi-level Prediction

- Write $\mathbf{C}(\boldsymbol{\theta})\hat{\gamma} + \mathbf{M}\hat{\beta} = \mathbf{Z}$ as
$$\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}}\hat{\gamma}_W + \mathbf{M}\hat{\beta} = \mathbf{Z}$$

- Applying $\mathbf{W}$ we have
$$\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W = \mathbf{Z}_W$$
  - Solved efficiently with Preconditioned Conjugate Method (PCG) and Kernel Independent Fast Multi-Pole Method (KIFMM).
  - Obtain $\hat{\gamma}$ from γ_W (i.e. $\hat{\gamma} = \mathbf{W}\hat{\gamma}_W$).
  - $\hat{\beta}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma})$

$$\begin{array}{l} \hat{\beta}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})^{-1}\mathbf{Z} \\ \hat{\gamma}(\boldsymbol{\theta}) = \mathbf{C}(\boldsymbol{\theta})^{-1}(\mathbf{Z} - \mathbf{M}\hat{\beta}) \end{array} \Longleftrightarrow \begin{array}{l} \mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W = \mathbf{Z}_W, \hat{\gamma} = \mathbf{W}\hat{\gamma}_W \\ \hat{\beta}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma}) \end{array}$$

- Decoupled problem, better conditioned, equivalent, *but much easier to solve numerically* (speed and stability).

Multi-level Approach

How do we build **L** and **W** such that:

- **L** and **W** have at most $\mathcal{O}(n \log n)$ non zero coefficients ?
- $\mathbf{C}_W(\boldsymbol{\theta})$ is stable ?
- $\mathbf{C}_W(\boldsymbol{\theta})$ can be made sparse ?

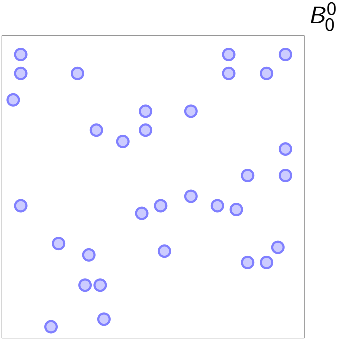
Multi-level Approach

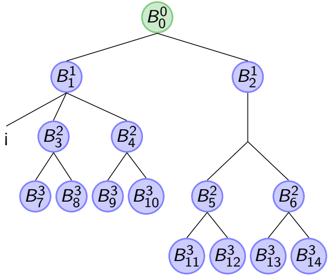
How do we build **L** and **W** such that:

- **L** and **W** have at most $\mathcal{O}(n \log n)$ non zero coefficients ?
- $\mathbf{C}_W(\boldsymbol{\theta})$ is stable ?
- $\mathbf{C}_W(\boldsymbol{\theta})$ can be made sparse ?

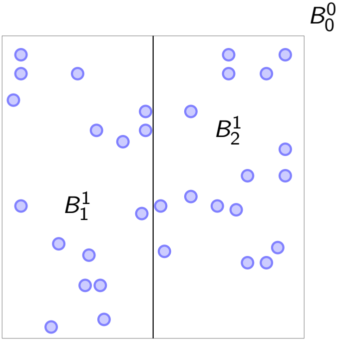
- First Step: Decompose the domain in a series of multiresolution cells $\{B_0^0, B_0^1, B_1^1, \ldots, B_k^t\}$.
  - Binary Tree Partition: KD tree, Random Projection Tree, etc
  - Depth of tree: t levels

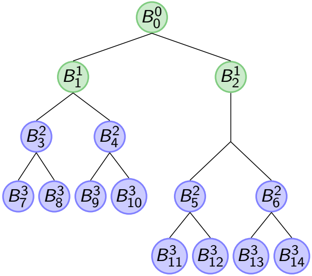
Multi-level Domain Decomposition



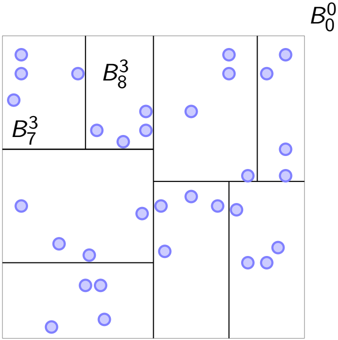


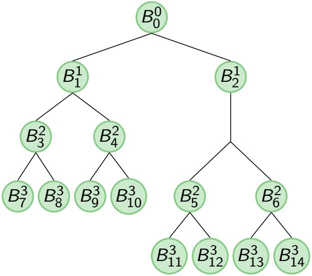
Multi-level Domain Decomposition





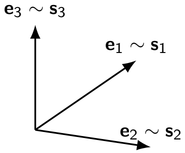
Multi-level Domain Decomposition





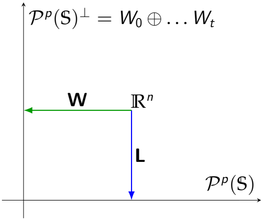
Multi-level Vector Spaces

- For each $\mathbf{s}_i$ node, $i = 1, \dots, n$, assign a unit vector $\mathbf{e}_i$



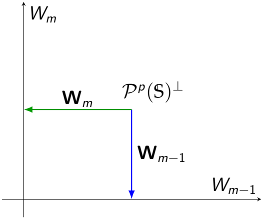
- Transform $\mathbb{R}^n := span\{\mathbf{e}_1, \dots, \mathbf{e}_n\}$ to multi-level spaces with levels $i = 0, 1, \dots, t$

$$\mathbb{R}^n \to \overbrace{\mathcal{P}(\mathbb{S})}^{\mathbf{L}} \oplus \underbrace{W_0 \oplus \dots W_t}_{\mathbf{W}}$$

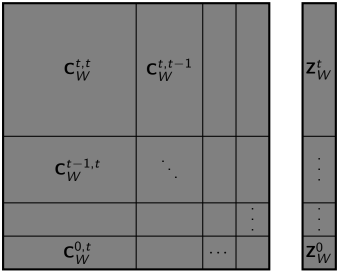


Multi-level Covariance

- Decompose $\mathbf{W}$ as $\mathbf{W} = [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$



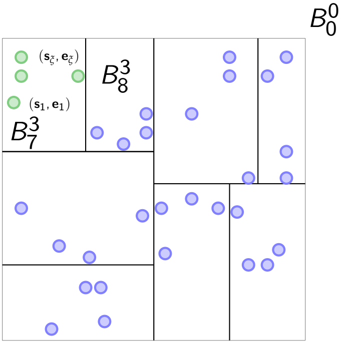
- Form covariance matrix: For all $i, j = -1, \ldots, t$

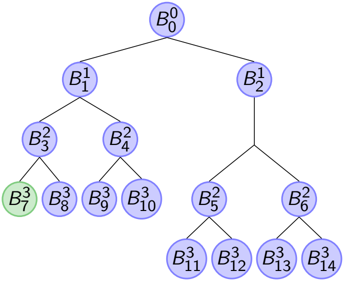


$$\mathbf{C}_W(\boldsymbol{\theta}) := \mathbf{W}\mathbf{C}(\boldsymbol{\theta})\mathbf{W}^{\mathrm{T}}$$
$$\mathbf{C}_W^{i,j} := \mathbf{W}_i^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})\mathbf{W}_j$$
$$\mathbf{Z}_W^i := \mathbf{W}_i\mathbf{Z}$$

Leaf Cell

- Let $q := t$. Initial basis construction from leaf cell at B_k^q
- Local indexing of nodes: $\{\mathbf{s}_1, \ldots, \mathbf{s}_\xi\} \leftrightarrow \{\mathbf{e}_1, \ldots, \mathbf{e}_\xi\}$





Multilevel Basis Construction at Leaf Cell

1. Let $\mathbf{V} := [\mathbf{e}_1, \ldots, \mathbf{e}_\xi]$ and $V := span\{\mathbf{e}_1, \ldots, \mathbf{e}_\xi\}$, a new basis for V is construct with a linear combination:

$$\phi_j^{q,k} := \sum_{i=1}^{\xi} c_{i,j}^{q,k} \mathbf{e}_i, \quad j = 1, \ldots, a_{q,k}$$

$$\psi_j^{q,k} := \sum_{i=1}^{\xi} d_{i,j}^{q,k} \mathbf{e}_i, \quad j = a_{q,k} + 1, \ldots, \xi$$

where $c_{i,j}^{q,k}$, $d_{i,j}^{q,k} \in \mathbb{R}$ and for some $a_{q,k} \in \mathbb{Z}^+$

Multilevel Basis Construction at Leaf Cell

1. Let $\mathbf{V} := [\mathbf{e}_1, \ldots, \mathbf{e}_\xi]$ and $V := span\{\mathbf{e}_1, \ldots, \mathbf{e}_\xi\}$, a new basis for V is construct with a linear combination:

$$\phi_j^{q,k} := \sum_{i=1}^{\xi} c_{i,j}^{q,k} \mathbf{e}_i, \quad j = 1, \ldots, a_{q,k}$$

$$\psi_j^{q,k} := \sum_{i=1}^{\xi} d_{i,j}^{q,k} \mathbf{e}_i, \quad j = a_{q,k} + 1, \ldots, \xi$$

where $c_{i,j}^{q,k}$, $d_{i,j}^{q,k} \in \mathbb{R}$ and for some $a_{q,k} \in \mathbb{Z}^+$

2. We desire that the new vector $\psi_j^{q,k}$ to be orthogonal to $\mathcal{P}^p(\mathbb{S})$:

$$\sum_{l=1}^{N} r[l] \psi_j^{q,k}[l] = 0, \tag{5}$$

for all $r \in \mathcal{P}^p(\mathbb{S})$. If $\Psi^{q,k} := [\psi_{a_{q,k}+1}, \ldots, \psi_\xi]$ then $\mathbf{M}^{\mathrm{T}} \Psi^{q,k} = 0$

3. Form the matrix

$$\mathbf{M}_{\xi,p}^{q,k} := \mathbf{M}^{\mathrm{T}}\mathbf{V}$$

3. Form the matrix

$$\mathbf{M}_{\xi,p}^{q,k} := \mathbf{M}^{\mathrm{T}}\mathbf{V}$$

4. Compute the SVD

$$\mathbf{M}_{\xi,p}^{q,k} \to \mathbf{U}_{\xi,p}^{q,k}\mathbf{D}_{\xi,p}^{q,k}(\mathbf{V}_{\xi,p}^{q,k})^{\mathrm{T}}$$

3. Form the matrix

$$\mathbf{M}_{\xi,p}^{q,k} := \mathbf{M}^{\mathrm{T}}\mathbf{V}$$

4. Compute the SVD

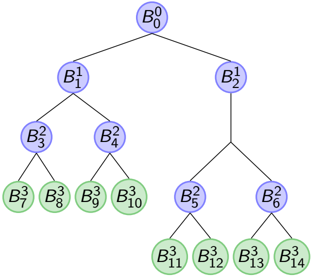
$$\mathbf{M}_{\xi,p}^{q,k} \to \mathbf{U}_{\xi,p}^{q,k}\mathbf{D}_{\xi,p}^{q,k}(\mathbf{V}_{\xi,p}^{q,k})^{\mathrm{T}}$$

5. We then pick

$$\left[\begin{array}{ccc|ccc} c_{0,1} & \dots & c_{a,1} & d_{a_{q,k}+1,1} & \dots & d_{\xi,1} \\ c_{0,2} & \dots & c_{a,2} & d_{a_{q,k}+1,2} & \dots & d_{\xi,2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ c_{0,\xi} & \dots & c_{a,\xi} & d_{a_{q,k}+1,\xi} & \dots & d_{\xi,\xi} \end{array}\right] := (\mathbf{V}_{\xi,p}^{q,k})^{\mathrm{T}}, \tag{6}$$

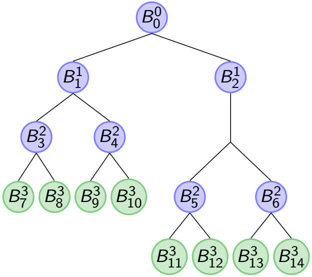
where $a_{q,k} = \dim\mathcal{R}(\mathbf{M}_{\xi,p}^{q,k})$

6. Let $\mathbf{V}^{q,k} := [\phi_1^{q,k}, \ldots, \phi_{a_{q,k}}^{q,k}]$ and $\mathbf{W}^{q,k} := [\psi_{a_{q,k}+1}^{q,k}, \ldots, \psi_{\xi}^{q,k}]$.

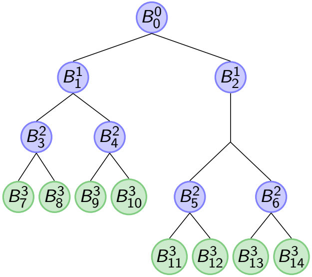


6. Let $\mathbf{V}^{q,k} := [\phi_1^{q,k}, \ldots, \phi_{a_{q,k}}^{q,k}]$ and $\mathbf{W}^{q,k} := [\psi_{a_{q,k}+1}^{q,k}, \ldots, \psi_{\xi}^{q,k}]$.
7. Let $V^{q,k} := span\{\phi_1^{q,k}, \ldots, \phi_{a_{q,k}}^{q,k}\}$ and $W^{q,k} := span\{\psi_{a_{q,k}+1}^{q,k}, \ldots, \psi_{\xi}^{q,k}\}$ then the following properties holds:

- $V = V^{q,k} \oplus W^{q,k}$
- $W^{q,k} \perp \mathcal{P}^p(\mathbb{S})$.



6. Let $\mathbf{V}^{q,k} := [\phi_1^{q,k}, \ldots, \phi_{a_{q,k}}^{q,k}]$ and $\mathbf{W}^{q,k} := [\psi_{a_{q,k}+1}^{q,k}, \ldots, \psi_{\xi}^{q,k}]$.
7. Let $V^{q,k} := span\{\phi_1^{q,k}, \ldots, \phi_{a_{q,k}}^{q,k}\}$ and $W^{q,k} := span\{\psi_{a_{q,k}+1}^{q,k}, \ldots, \psi_{\xi}^{q,k}\}$ then the following properties holds:
   - $V = V^{q,k} \oplus W^{q,k}$
   - $W^{q,k} \perp \mathcal{P}^p(\mathbb{S})$.
8. Repeat 1 - 6 for all non empty cells B_k^q at level q and let $W_q := \oplus_k W^{q,k}$, $V_q := \oplus_k V^{q,k}$ and $\mathbf{W}^q := [\mathbf{W}^{q,k}, \ldots,]$



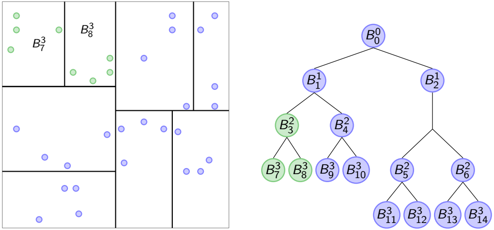
Multi-level Basis Non-leaf Cell

9. Next level:

   i) $q = q - 1$

   ii) For each sibling nodes $(B^{q+1,k}, B^{q+1,k+1})$:

      - Let $\mathbf{V} := [\mathbf{V}^{q+1,k}, \mathbf{V}^{q+1,k+1}]$ and repeat 1 - 6



10. When $q = -1$ stop.

11. $\mathbf{L} := (\mathbf{V}^{0,0})^{\mathrm{T}}$ and $\mathbf{W} := [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$.

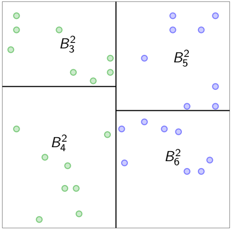
Multi-level Basis Non-leaf Cell

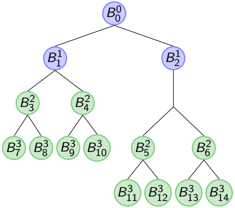
9. Next level:

   i) $q = q - 1$

   ii) For each sibling nodes $(B^{q+1,k}, B^{q+1,k+1})$:

      - Let $\mathbf{V} := [\mathbf{V}^{q+1,k}, \mathbf{V}^{q+1,k+1}]$ and repeat 1 - 6





10. When $q = -1$ stop.

11. $\mathbf{L} := (\mathbf{V}^{0,0})^{\mathrm{T}}$ and $\mathbf{W} := [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$.

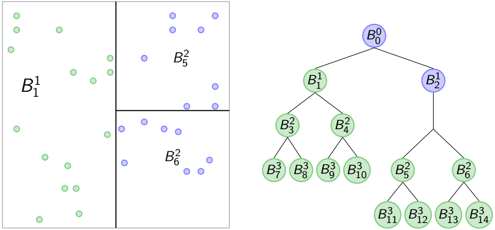
Multi-level Basis Non-leaf Cell

9. Next level:

   i) $q = q - 1$

   ii) For each sibling nodes $(B^{q+1,k}, B^{q+1,k+1})$:

      - Let $\mathbf{V} := [\mathbf{V}^{q+1,k}, \mathbf{V}^{q+1,k+1}]$ and repeat 1 - 6



10. When $q = -1$ stop.

11. $\mathbf{L} := (\mathbf{V}^{0,0})^{\mathrm{T}}$ and $\mathbf{W} := [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$.

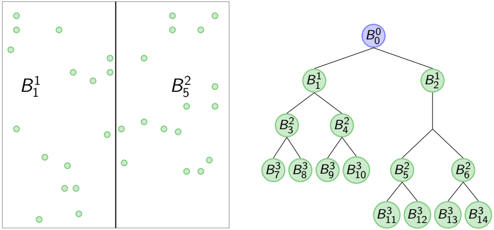
Multi-level Basis Non-leaf Cell

9. Next level:

   i) $q = q - 1$

   ii) For each sibling nodes $(B^{q+1,k}, B^{q+1,k+1})$:

      - Let $\mathbf{V} := [\mathbf{V}^{q+1,k}, \mathbf{V}^{q+1,k+1}]$ and repeat 1 - 6



10. When $q = -1$ stop.

11. $\mathbf{L} := (\mathbf{V}^{0,0})^{\mathrm{T}}$ and $\mathbf{W} := [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$.

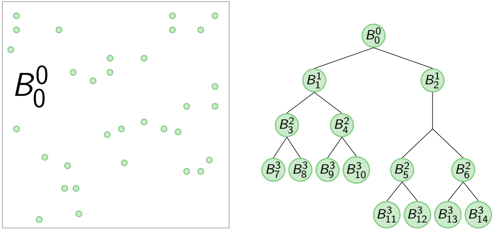
Multi-level Basis Non-leaf Cell

9. Next level:

   i) $q = q - 1$

   ii) For each sibling nodes $(B^{q+1,k}, B^{q+1,k+1})$:

      - Let $\mathbf{V} := [\mathbf{V}^{q+1,k}, \mathbf{V}^{q+1,k+1}]$ and repeat 1 - 6



10. When $q = -1$ stop.

11. $\mathbf{L} := (\mathbf{V}^{0,0})^{\mathrm{T}}$ and $\mathbf{W} := [\mathbf{W}_t, \ldots, \mathbf{W}_0]^T$.

Multi-level Basis Properties

Theorem 1.

We have now decomposed $\mathbb{R}^n$ *as*

$$\mathbb{R}^n \rightarrow \overbrace{\mathcal{P}(\mathbb{S})}^{\mathbf{L}} \oplus \underbrace{W_0 \oplus \ldots W_t}_{\mathbf{W}}$$

Theorem 2.

The complexity cost of the multi-level is bounded by $\mathcal{O}(nt)$

Theorem 3.

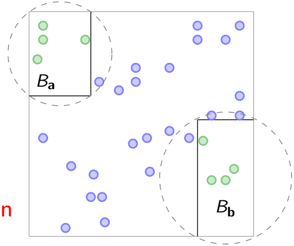
The multi-level basis vectors of $\mathcal{P}(\mathbb{S}) \oplus W_0 \oplus \ldots W_t$ *form an orthonormal set*

Matrix Coefficient Decay

Theorem 4 (Total Degree).

Let $B_{\mathbf{a}}$ be the smallest ball in $\mathbb{R}^d$ with radii $r_{\mathbf{a}}$ centered around the midpoint $\mathbf{a} \in \mathbb{R}^d$ of the cell B^i_k such that $B^i_k \subset B_{\mathbf{a}}$. Similarly, let $B_{\mathbf{b}}$ be the smallest ball in $\mathbb{R}^d$ with radii $r_{\mathbf{b}} \in \mathbb{R}^d$ centered around the midpoint $\mathbf{b}$ of the cell B^j_l such that $B^j_l \subset B_{\mathbf{b}}$. If $\boldsymbol{\psi}^{i,k}_{\tilde{k}} \in W_i$ and $\boldsymbol{\psi}^{j,l}_{\tilde{l}} \in W_j$ then for $i, j = 0, \ldots, t$:

$$|(\boldsymbol{\psi}^{i,k}_{\tilde{k}})^{\mathrm{T}} \mathbf{C}(\boldsymbol{\theta}) \boldsymbol{\psi}^{j,l}_{\tilde{l}}| \leq \sum_{|\alpha|=w+1} \sum_{|\beta|=w+1} \frac{r_a^\alpha}{\alpha!} \frac{r_b^\beta}{\beta!} \sup_{\mathbf{x} \in B_{\mathbf{a}}, \mathbf{y} \in B_{\mathbf{b}}} |D^\alpha_{\mathbf{x}} D^\beta_{\mathbf{y}} \phi(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta})|$$

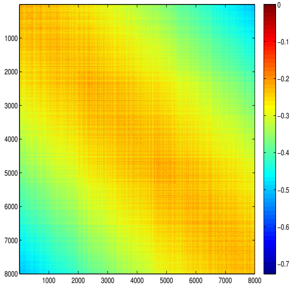


Warning: Derivative information hard to obtain

Example

- Let $\mathbb{S}$ a collection of the observation locations ($n = 8,000$) that are sampled from a uniform distribution on the unit cube $[0,1]^3$ with $\phi(r;\boldsymbol{\theta}) := \exp(-r)$
- $\mathbf{m}(\mathbf{s})$ set to cubic polynomials in $\mathbb{R}^3$ i.e. $p = 20$

(a) $\log_{10}|\mathbf{C}(\boldsymbol{\theta})|$

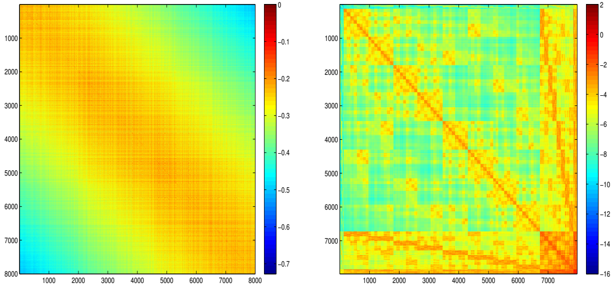


(b) $\log_{10}|\mathbf{C}_W(\boldsymbol{\theta})|$

Example

- Let $\mathbb{S}$ a collection of the observation locations ($n = 8,000$) that are sampled from a uniform distribution on the unit cube $[0,1]^3$ with $\phi(r;\boldsymbol{\theta}) := \exp(-r)$
- $\mathbf{m}(\mathbf{s})$ set to cubic polynomials in $\mathbb{R}^3$ i.e. $p = 20$

(a) $\log_{10}|\mathbf{C}(\boldsymbol{\theta})|$ (b) $\log_{10}|\mathbf{C}_W(\boldsymbol{\theta})|$

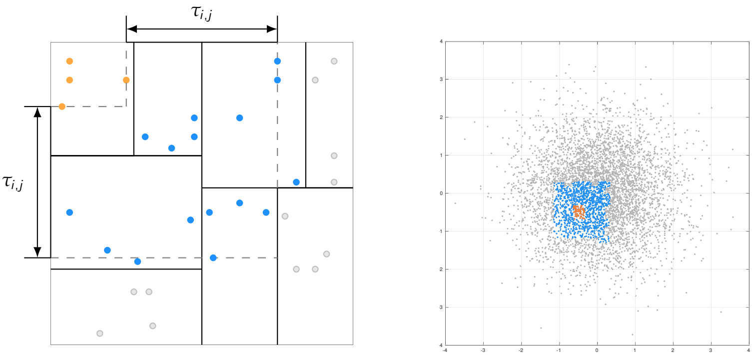


- Use sparse matrix $\tilde{\mathbf{C}}_W$ instead of $\mathbf{C}_W$

Sparse matrix $\tilde{\mathbf{C}}_W$

- Most of the entries of $\mathbf{C}_W$ are small and not necessary to compute.
- Distance criterion parameter $\tau \geq 0$; $\tau_{i,j} := \tau 2^{t-(i+j)/2}$.

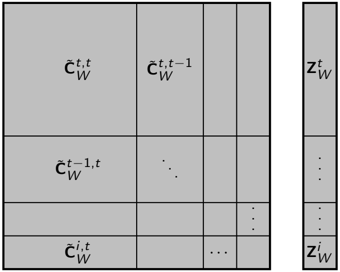
Entry $(\boldsymbol{\psi}_{\tilde{k}}^{i,k})^{\mathrm{T}} \mathbf{C}(\boldsymbol{\theta}) \boldsymbol{\psi}_{\tilde{l}}^{j,l}$ is computed if axiswise $\mathit{dist}(\boldsymbol{\psi}_{\tilde{k}}^{i,k}, \boldsymbol{\psi}_{\tilde{l}}^{j,l}) \leq \tau_{i,j}$

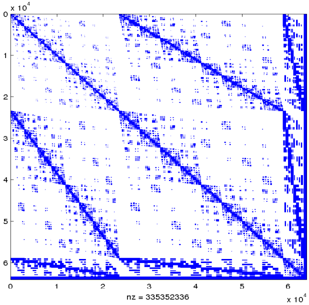


(a) (b)

Estimation with $\tilde{\mathbf{C}}_W$

- For *Estimation* a reduced matrix of $\tilde{\mathbf{C}}_W$ is used
- Reduced sparse covariance matrix: For all $k, l = i, \ldots, t$ form the reduced matrix $\tilde{\mathbf{C}}_{W,i,\ldots,t}$
- Sparse Cholesky decomposition used to compute likelihoods





Polynomial Approximation in High Dimensions

i) Let $\mathbf{p} := [p_1, \ldots, p_d] \in \mathbb{N}_0^d$, $w \in \mathbb{R}$ and $\Lambda(w) \subset \mathbb{N}_0^d$ be an index set that will determine the indices of polynomial basis functions and thus the size of p.

ii) Restrict the number of polynomials along each dimension by using the set of monomials contained in

$$\mathcal{Q}_{\Lambda(w)} := \left\{ s_1^{p_1} s_2^{p_2} \ldots s_d^{p_d} \text{ with } \mathbf{p} \in \Lambda(w) \right\}$$

i.e. $\mathbf{m}(\mathbf{s})$ is built from the monomials in $\mathcal{Q}_{\Lambda(w)}$.

iii) Based on sparse grid representations of high dimensional functions.

Polynomial Approximation in High Dimensions

Approx. space	Index Set: $\Lambda(w)$
Tensor Prod. (TP)	$\Lambda(w) \equiv \{\mathbf{p} \in \mathbb{N}_+^d : \ \max_{i=1}^d p_i \leq w\}$
Total Degree (TD)	$\Lambda(w) \equiv \{\mathbf{p} \in \mathbb{N}_+^d : \ \sum_{i=1}^d p_i \leq w\}$
Smolyak (SM)	$\Lambda(w) \equiv \{\mathbf{p} \in \mathbb{N}_+^d : \ \sum_{i=1}^d f(p_i) \leq w\}$ $f(b) = \begin{cases} 0, \ b = 0 \\ 1, \ b = 1 \\ \lceil \log_2(b) \rceil, \ b \geq 2 \end{cases}$
Hyperbolic Cross (HC)	$\Lambda(w) \equiv \{\mathbf{p} \in \mathbb{N}_+^d : \ \prod_{i=1}^d (p_i + 1) \leq w\}$

Table: Index set of different polynomial approximation choices.

Polynomial Approximation in High Dimensions

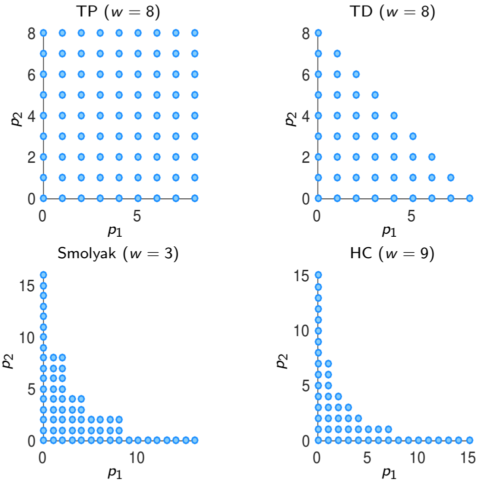


Figure: Tensor Product (TP), Total Degree (TD) , Smolyak (SM) and Hyperbolic Cross (HC) index sets for $d = 2$

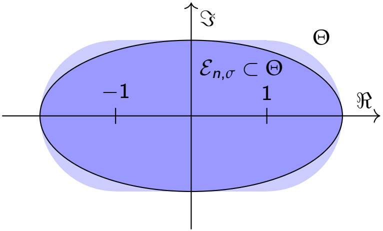
Matrix Coefficient Decay

- Do not use derivative information. Too difficult and/or expensive to obtain. Decay based on analytic extensions of covariance function
- Let $\Gamma := [-1, 1]^d$ and assume w.l.o.g. that $\phi(\mathbf{x}, \mathbf{y}; \theta) : \Gamma \times \Gamma \to \mathbb{R}$

Matrix Coefficient Decay

- Do not use derivative information. Too difficult and/or expensive to obtain. Decay based on analytic extensions of covariance function
- Let $\Gamma := [-1,1]^d$ and assume w.l.o.g. that $\phi(\mathbf{x},\mathbf{y};\theta) : \Gamma \times \Gamma \to \mathbb{R}$
- Form Bernstein polyellipse $\mathcal{E}_\sigma := \Pi_{n=1}^d \mathcal{E}_{n,\sigma} \subset \mathbb{C}^d$, $\sigma > 0$, where

$$\mathcal{E}_{n,\sigma} = \left\{ z \in \mathbf{C};\ \Re(z) = \frac{e^\sigma + e^{-\sigma}}{2} cos(\theta);\ \Im(z) = \frac{e^\sigma - e^{-\sigma}}{2} sin(\theta), \theta \in [0, 2\pi) \right\},$$



- For Matérn covariance function we can show that $\phi(\mathbf{x},\mathbf{y};\boldsymbol{\theta})$ admits an analytic extension on $\mathcal{E}_\sigma \times \mathcal{E}_\sigma$ and uniformly bounded $|\phi| \leq \tilde{M} < \infty$

(Total Degree) Matrix Coefficient Decay

Theorem 5.

i) *Let* $0 < \delta < 1$ *and* $\hat{\sigma} := \sigma(1-\delta)$

ii) $\boldsymbol{\psi}_{\tilde{k}}^{i,k} \in \mathcal{P}^p(\mathbb{S})^{\perp}$, *with* n_m *non-zero entries and* $\boldsymbol{\psi}_{\tilde{l}}^{j,l} \in \mathcal{P}^p(\mathbb{S})^{\perp}$, *with* n_q *non-zero entries*

If $\eta(d,w) \geq \left(\frac{2d}{\kappa(d)}\right)^d$, $\kappa(d) := (d!)^{\frac{1}{d}}$, *then*

$$|(\boldsymbol{\psi}_{\tilde{k}}^{i,k})^{\mathrm{T}}\mathbf{C}(\boldsymbol{\theta})\boldsymbol{\psi}_{\tilde{l}}^{j,l}| \leq \sqrt{n_m n_q}\left(\frac{C(\tilde{M},\sigma)^d e^{d-\sigma(1-\delta)+1}\hat{\sigma}d}{(\sigma\delta)^d}\right)^2 \left(\frac{e^{\hat{\sigma}}}{1-e^{-\hat{\sigma}}}\right)^{2d} \exp\left(-\frac{2d}{e}\hat{\sigma}\eta^{\frac{1}{d}}\right)\eta^{2\left(\frac{d-1}{d}\right)}$$

Key observations:

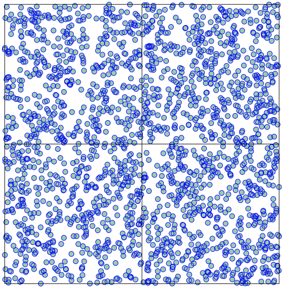
- *Sub-exponential Decay*
- *All coefficients are known*
- *Similar bound obtained for Smolyak*

Implementation

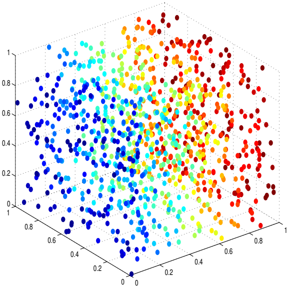
- Implemented with g++ GNU, Intel MKL, and Matlab
- Results performed on Intel i7-3770 CPU @ 3.40GHz / 32 Gb / Linux
- For $d = 2, 3$ problems a Kernel Independent Fast Multipole Method (KIFMM) is used
- KIFMM accuracy to medium (10^{-6} to 10^{-8}) or high ($\geq 10^{-8}$) with one core
- For $d > 3$ Direct method (no KIFMM), but 4 cores are used instead

Numerical Results

Data Set #1 & #2 The sets of observation locations $\mathbf{S}_1^d, \ldots, \mathbf{S}_{10}^d$ vary from 1×10^3, 2×10^3, 4×10^3, $\ldots$, 512×10^3, $\mathbf{S}_l^d \subset \mathbf{S}_{l+1}^d$ for $l = 1, \ldots, 9$. Observations locations are sampled from a uniform distribution over $[0, 1]^d$



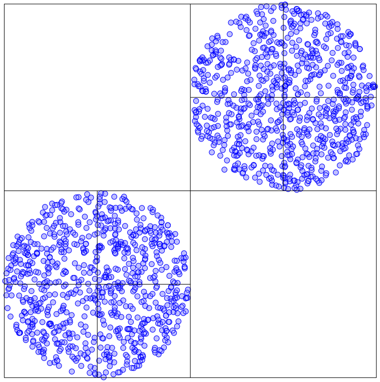
Data Set #1 ($d = 2$)



Data Set #2 ($d = 3$)

Numerical Results

Data Set #3: We take the data set generated by $\mathbf{S}_9^d$ for $d = 2$ (256,000 observations points) and carve out two disks located at (1/4,1/4) and (3/4,3/4) with radii 1/4. This generates 100,637 observation points.

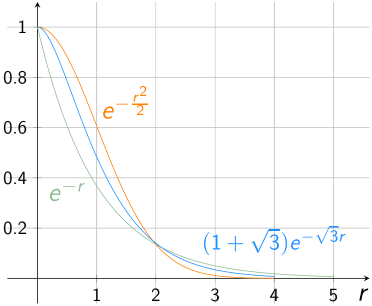


Matérn Covariance function

$$\phi(r;\boldsymbol{\theta}) := \frac{1}{\Gamma(\nu)2^{\nu-1}}\left(\sqrt{2\nu}\frac{r}{\rho}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu}\frac{r}{\rho}\right),$$

$\Gamma(\nu)$: gamma function, K_ν : modified Bessel function of the second kind and $\boldsymbol{\theta} := (\nu, \rho)$ where $\nu, \rho \in \mathbb{R}^+$

- $r := \|\mathbf{s} - \mathbf{s}'\|_2$ (distance function)
- $\nu = 0.5 \Rightarrow \phi(r;\boldsymbol{\theta}) = \exp\left(-\frac{r}{\rho}\right)$
- $\nu = 1.5 \Rightarrow \phi(r;\boldsymbol{\theta}) = \left(1 + \frac{\sqrt{3}r}{\rho}\right)\exp\left(-\frac{\sqrt{3}r}{\rho}\right)$
- $\nu \to \infty \Rightarrow \phi(r;\boldsymbol{\theta}) \to \exp\left(-\frac{r^2}{2\rho^2}\right)$



Condition Numbers

- Data Set #1 (2D): Condition numbers of $\mathbf{C}_W$ and $\mathbf{C}$ with TD design matrix.

n	w	p	d	ν	ρ	$\kappa(\mathbf{C}_W)$	$\kappa(\mathbf{C})$
4000	20	231	2	3/4	1	4.71×10^4	7.46×10^8
8000	20	231	2	3/4	1	2.47×10^5	3.85×10^9
16000	20	231	2	3/4	1	8.21×10^5	1.29×10^{10}
32000	20	231	2	3/4	1	$\boxed{2.45 \times 10^7}$	$\boxed{3.86 \times 10^{11}}$

- Condition numbers of $\mathbf{C}_W$, $\mathbf{C}$ and $\mathbf{R} := \begin{pmatrix} \mathbf{C}(\boldsymbol{\theta}) & \mathbf{M} \\ \mathbf{M}^{\mathrm{T}} & \mathbf{0} \end{pmatrix}$

n	w	p	d	ν	ρ	$\kappa(\mathbf{C}_W)$	$\kappa(\mathbf{C})$	$\kappa(\mathbf{R})$
4000	14	120	2	3/4	1/6	1.12×10^5	9.30×10^6	5.93×10^{20}
8000	14	120	2	3/4	1/6	5.99×10^5	4.77×10^7	1.51×10^{20}
16000	14	120	2	3/4	1/6	1.99×10^6	1.59×10^8	3.14×10^{19}
4000	14	120	2	3/4	1	1.27×10^5	7.59×10^8	4.43×10^{20}
8000	14	120	2	3/4	1	6.81×10^5	3.85×10^9	7.59×10^{20}
16000	14	120	2	3/4	1	$\boxed{2.27 \times 10^6}$	$\boxed{1.29 \times 10^{10}}$	$\boxed{2.12 \times 10^{20}}$
4000	14	120	2	3/4	100	1.27×10^5	1.09×10^{12}	4.58×10^{16}
8000	14	120	2	3/4	100	6.83×10^5	5.61×10^{12}	9.76×10^{16}
16000	14	120	2	3/4	100	$\boxed{2.28 \times 10^6}$	$\boxed{1.88 \times 10^{13}}$	$\boxed{1.25 \times 10^{17}}$

Estimation Numerical Results

- Data Set #1 (2D), $\nu = 0.75$, $\rho = 1/6$
- Total Degree Design Matrix with degree w
- Generate realizations of $\mathbf{Z}_5^2$, $\mathbf{Z}_6^2$ and $\mathbf{Z}_7^2$ from the Gaussian random field $\mathbf{Z}(\mathbf{s})$ model. (with $\mathbf{S}_5^2$ ($n = 32,000$), $\mathbf{S}_6^2$ ($n = 64,000$) and $\mathbf{S}_7^2$ ($n = 128,000$) respectively.
- Compute $\mathbf{Z}_{W,5} := \mathbf{W}\mathbf{Z}_6^2$, $\mathbf{Z}_{W,6} := \mathbf{W}\mathbf{Z}_6^2$ and $\mathbf{Z}_{W,7} := \mathbf{W}\mathbf{Z}_7^2$.
- Solve

  $$\hat{\boldsymbol{\theta}}_j := \operatorname*{argmin}_{\tilde{\boldsymbol{\theta}}} \ell_W(\tilde{\mathbf{C}}_{W,i,\dots,t}(\tilde{\theta}), \mathbf{Z}_{W,j}^i)$$

  for $j = 5, 6, 7$.
- $i = 3, \dots, t$ levels

Estimation Numerical Results

Data set #1 (2D,TD)

n	w	i	p	$\hat{\nu}-\nu$	$\hat{\rho}-\rho$	$nz(\mathbf{G})(\%)$	size($\tilde{\mathbf{C}}_W^{i,\dots,t}$)	$t_{cons}(s)$	$t_{chol}(s)$
64,000	6	6	28	-0.0759	0.0333	8.9	23	14	0
64,000	6	5	28	0.0182	-0.0132	1.7	35328	40	1
64,000	6	4	28	-0.0043	0.0046	4.5	56832	230	11
64,000	6	3	28	-0.0049	0.0048	10.7	62208	961	65
64,000	5	6	21	0.0071	-0.0146	0.4	810	13	0
64,000	5	5	21	0.0037	-0.0027	1.7	42496	43	2
64,000	5	4	21	-0.0030	0.0048	3.7	58624	220	12
64,000	5	3	21	-0.0048	0.0046	7.0	62656	750	32

n	w	i	p	$\hat{\nu}-\nu$	$\hat{\rho}-\rho$	$nz(\mathbf{G})(\%)$	size($\tilde{\mathbf{C}}_W^{i,\dots,t}$)	$t_{cons}(s)$	$t_{chol}(s)$
128,000	6	6	28	0.0010	-0.0011	0.3	17179	75	0
128,000	6	5	28	0.0025	-0.0020	2.1	99328	350	13
128,000	6	4	28	-0.0002	0.0005	4.0	120832	1200	70
128,000	5	6	21	-0.0010	0.0015	0.5	42154	80	0
128,000	5	5	21	0.0004	-0.0002	1.9	106496	300	14
128,000	5	4	21	-0.0016	0.0017	3.3	122624	1000	50

Estimation Numerical Results

100 realizations from Data Set #1 with $\nu = .75$, $\rho = 1/6$ and TD

n	w	i	p	t	$\mathbb{E}_M[\hat{\nu}-\nu]$	$\mathbb{E}_M[\hat{\rho}-\rho]$	$std_M[\hat{\nu}]$	$std_M[\hat{\rho}]$
32,000	4	5	15	5	-6.0×10^{-4}	1.0×10^{-3}	1.3×10^{-2}	1.0×10^{-2}
32,000	4	4	15	5	-7.2×10^{-4}	7.0×10^{-4}	5.9×10^{-3}	4.5×10^{-3}
32,000	4	3	15	5	-7.0×10^{-4}	6.0×10^{-4}	5.6×10^{-3}	4.0×10^{-3}
64,000	4	6	15	6	6.8×10^{-4}	1.1×10^{-3}	2.0×10^{-2}	1.9×10^{-2}
64,000	4	5	15	6	7.4×10^{-4}	-5.6×10^{-4}	5.4×10^{-3}	4.6×10^{-3}
64,000	4	4	15	6	2.5×10^{-4}	-1.7×10^{-4}	3.9×10^{-3}	3.3×10^{-3}
128,000	6	6	28	6	-1.3×10^{-3}	1.5×10^{-3}	8.3×10^{-3}	7.7×10^{-3}
128,000	6	5	28	6	-6.2×10^{-4}	6.5×10^{-4}	3.7×10^{-3}	3.3×10^{-3}

Disk problem: Data set #3 with $\nu = 1.25$, $\rho = 1/6$:

n	w	i	p	$\hat{\nu}-\nu$	$\hat{\rho}-\rho$	$nz(\mathbf{G})(\%)$	size$(\tilde{\mathbf{C}}_W^{i,\ldots,t})$	$t_{cons}(s)$	$t_{chol}(s)$
100,637	6	6	66	0.0548	-0.0237	0.5	2613	60	0
100,637	6	5	66	-0.0031	0.0020	3.1	72231	600	12

e

Prediction (Kriging) Numerical Results

- First step: (Preconditioned Conjugate Gradient)

$$\mathbf{D}_W^{-1}(\boldsymbol{\theta})\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W(\boldsymbol{\theta}) = \mathbf{D}_W^{-1}(\boldsymbol{\theta})\mathbf{Z}_W$$

  - Preconditioner: $\mathbf{D}_W(\boldsymbol{\theta}) := \mathrm{diag}(\mathbf{C}_W(\boldsymbol{\theta}))$,

- Second step: Recover $\hat{\gamma}$ from $\hat{\gamma}_W$ and compute

$$\hat{\beta}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma})$$

under the following:

i) Data Set #2 (3D), $\boldsymbol{\theta}_a^3 = (\nu, \rho) = (3/4, 1/6)$, $\boldsymbol{\theta}_b^3 = (5/4, 1/6)$

ii) Observation locations: $\mathbf{S}_1^3, \dots, \mathbf{S}_{10}^3$ ($n = 1,000, \dots, 512,000$)

iii) ε_{PCG} set such that the *unpreconditioned* system $\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W = \mathbf{Z}_W$ relative residual error is 10^{-3}

iv) Each matrix vector product computed with KIFMM

v) Total Degree

Prediction Numerical Results (3D)

(a) $\boldsymbol{\theta}_a = (\nu, \rho) = (3/4, 1/6)$, $d = 3$, $w = 3$ $(p = 20)$

n	itr($\mathbf{C}_W$)	itr($\mathbf{C}$)	ε_{PCG}	Diag. (s)	Itr (s)	Total (s)
16,000	166	1296	1.02×10^{-4}	80	113	193
32,000	247	3065	9.88×10^{-5}	215	321	536
$\boxed{64{,}000}$	$\boxed{372}$	$\boxed{5517}$	$\boxed{1.00 \times 10^{-4}}$	$\boxed{665}$	$\boxed{1226}$	$\boxed{1891}$
128,000	547	-	4.84×10^{-5}	2060	3237	5397
256,000	847	-	5.00×10^{-5}	5775	9885	15660
$\boxed{512{,}000}$	$\boxed{1129}$	$\boxed{-}$	$\boxed{3.74 \times 10^{-5}}$	$\boxed{17896}$	$\boxed{33116}$	$\boxed{51012}$

(b) $\boldsymbol{\theta}_b = (\nu, \rho) = (5/4, 1/6)$, $d = 3$, $w = 3$ $(p = 20)$

n	itr($\mathbf{C}_W$)	itr($\mathbf{C}$)	ε_{PCG}	Diag. (s)	Itr (s)	Total (s)
16,000	500	5953	6.27×10^{-5}	138	580	718
32,000	827	17029	7.29×10^{-5}	346	1574	1920
64,000	1567	37018	4.45×10^{-5}	910	6474	7384
128,000	2381	-	2.23×10^{-5}	3974	25052	29026
256,000	4299	-	2.61×10^{-5}	10322	72374	82696

Observation: At least about 200 times faster than traditional method for $n = 64,000$.

Prediction Numerical Results (2D)

- First step: (Preconditioned CG)

$$\mathbf{D}_W^{-1}(\boldsymbol{\theta})\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W(\boldsymbol{\theta}) = \mathbf{D}_W^{-1}(\boldsymbol{\theta})\mathbf{Z}_W$$

  - Preconditioner: $\mathbf{D}_W(\boldsymbol{\theta}) := \mathrm{diag}(\mathbf{C}_W(\boldsymbol{\theta}))$

- Second step: Recover $\hat{\gamma}$ from $\hat{\gamma}_W$ and compute

$$\hat{\beta}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma})$$

under the following:

i) Data Set #1 (2D), $\boldsymbol{\theta}_a^2 = (1/2, 1/6)$, $\boldsymbol{\theta}_b^2 = (1, 1/6)$

ii) Observation locations: $\mathbf{S}_1^2, \dots, \mathbf{S}_{10}^2$ ($n = 1,000, \dots, 512,000$)

iii) ε_{PCG} set such that the *unpreconditioned* system $\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W = \mathbf{Z}_W$ relative residual error is 10^{-2}

iv) Each matrix vector product computed with KIFMM

v) Total Degree

Prediction Numerical Results

(a) $\theta_a^2 = (\nu, \rho) = (1/2, 1/6)$, $d = 2$, $w = 3$ ($p = 10$)

n	itr($\mathbf{C}_W$)	itr($\mathbf{C}$)	ε_{PCG}	Diag. (s)	Itr (s)	Total (s)
16,000	330	3603	2.39×10^{-3}	246	115	361
32,000	333	5429	1.39×10^{-3}	750	251	1001
64,000	455	8152	1.32×10^{-3}	1947	589	2536
128,000	564	-	7.10×10^{-4}	5570	1577	7147
256,000	619	-	9.78×10^{-4}	15266	3065	18331
512,000	1230	-	4.50×10^{-4}	42254	13101	55355

Prediction Numerical Results

(b) $\boldsymbol{\theta}_b^2 = (\nu, \rho) = (1, 1/6)$, $d = 2$, $w = 14$ ($p = 120$)

n	itr($\mathbf{C}_W$)	itr($\mathbf{C}$)	ε_{PCG}	Diag. (s)	Itr (s)	Total (s)
16,000	2710	$> 100,000$	1.90×10^{-3}	553	1844	2397
32,000	4261	-	1.43×10^{-3}	1522	5713	7235
64,000	8801	-	1.00×10^{-4}	5022	23785	28807
128,000	14405	-	7.28×10^{-4}	12587	75937	88524

Observation: $\mathbf{C}(\boldsymbol{\theta})$ ill-conditioned thus the iterative solver stagnates.

Observation: $\mathbf{C}_W(\boldsymbol{\theta})$ solves prediction problem to 10^{-2} relative residual error.

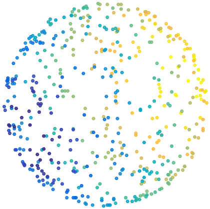
Extension to higher dimensions

- **Random n-sphere data set:** The set of nested random observations $\mathbf{S}_0^d \subset \cdots \subset \mathbf{S}_{10}^d$ varies from 1,000, 2000, 4000 to 128,000 knots generated on the n-sphere $\mathbf{S}_{d-1} := \{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$.

- **Gaussian data set:** Observations values $\mathbf{Z}_0^d$, $\mathbf{Z}_1^d, \ldots \mathbf{Z}_5^d$ formed from $\mathbf{S}_0^d, \mathbf{S}_1^d, \ldots \mathbf{S}_5^d$ locations with Matérn covariance parameters (ν, ρ).

- d up to 50 dimensions. Hyperbolic Cross design matrix.

- For matrix vector products we use direct method. To our knowledge there does not exist a fast summation method.



Estimation Results

- Sparsity and numerical stability of multilevel covariance matrix.
- Excellent condition numbers

Sparse $\tilde{\mathbf{C}}_{W,i,\ldots,n}$, $\nu = 1.25$, $\rho = 1$, $\boxed{\kappa(\mathbf{C}) = 4.91 \times 10^4}$

n	d	p	t	i	$\kappa(\mathbf{C}_W)$	$\kappa(\tilde{\mathbf{C}}_{W,i,\ldots,t})$	Size($\mathbf{C}_{W,i,\ldots,t}$)	τ	Sparsity
32000	50	1426	3	3	3.36	1.58	20,592	1×10^{-6}	6%
32000	50	1426	3	2	3.36	2.09	26,296	1×10^{-6}	10%
32000	50	1426	3	1	3.36	2.63	29,148	1×10^{-6}	14%
32000	50	1426	3	0	3.36	3.09	30,574	1×10^{-6}	18%
64000	50	1426	4	4	-	1.62	41,184	1×10^{-7}	3.1%
64000	50	1426	4	3	-	1.87	52,592	1×10^{-7}	4.3%
64000	50	1426	4	2	-	2.29	58,296	1×10^{-7}	5.9%
64000	50	1426	4	1	-	3.10	61,148	1×10^{-7}	7.6%

- Total Degree, KD tree, n-Sphere, $d = 10$, $M = 100$, $\nu = 1.25$, $\rho = 1$, $\tau = 10^{-7}$

n	w	t	i	$\mathbb{E}_M[\hat{\nu} - \nu]$	$\mathbb{E}_M[\hat{\rho} - \rho]$	$std_M[\hat{\nu}]$	$std_M[\hat{\rho}]$
64000	2	9	9	-5.86e-03	5.21e-03	4.85e-02	3.15e-02
64000	2	9	8	-3.37e-02	1.93e-02	3.50e-02	2.46e-02
64000	2	9	7	-1.19e-01	6.92e-02	2.88e-02	2.51e-02
128000	2	10	10	-2.70e-03	2.74e-03	3.76e-02	2.44e-02
128000	2	10	9	-2.00e-02	1.20e-02	2.47e-02	1.80e-02
128000	2	10	8	-8.40e-02	5.03e-02	2.21e-02	1.89e-02

Prediction Numerical Results

- First step: (Conjugate Gradient)

$$\mathbf{C}_W(\boldsymbol{\theta})\hat{\gamma}_W(\boldsymbol{\theta}) = \mathbf{Z}_W$$

- Second step: Recover γ from γ_W and compute

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta}) = (\mathbf{M}^{\mathrm{T}}\mathbf{M})^{-1}\mathbf{M}^{\mathrm{T}}(\mathbf{Z} - \mathbf{C}(\boldsymbol{\theta})\hat{\gamma})$$

- Relative error set to 10^{-6}, $d = 50$ dimensions
- Hyperbolic Cross (HC) design matrix

(a) $\boldsymbol{\theta} = (\nu, \rho) = (5/4, 1)$, $d = 50$, HC, $w = 4$ ($p = 1376$), No preconditioner, Direct

N	$\kappa(\mathbf{C})$	itr($\mathbf{C}_W$)	MB (s)	D_W (s)	Itr (s)	Total (s)
16,000	2.3×10^4	6	45	-	350	399
32,000	4.9×10^4	7	116	-	1660	1276
64,000	-	9	279	-	8370	8649
128,000	-	11	645	-	41006	41651

(b) $\boldsymbol{\theta} = (\nu, \rho) = (5/4, 5)$, $d = 50$, HC, $w = 4$ ($p = 1376$), No preconditioner, Direct

N	$\kappa(\mathbf{C})$	itr($\mathbf{C}_W$)	MB (s)	D_W (s)	Itr (s)	Total (s)
16,000	2.8×10^6	7	46	-	406	452
32,000	$\boxed{6.4 \times 10^6}$	9	115	-	2083	2198
64,000	-	11	272	-	10219	10491
128,000	-	14	643	-	51870	52513

Hospital Data Benchmarks

- Comparison with PMM, PPD, BEM, DA methods from current missing data packages (Mi, Amelia, Norm).
- Totchg $\sim$ los + npr + ndx + age

Total Charge Imputation ($n = 10^5$)

Methods	rMSE	MAPE	lnQ
PMM	0.864	1.235	1.00
PPD	0.869	3.378	1.779
BEM	0.869	3.317	1.745
DA	0.867	3.449	1.787
Kriging	0.535	0.861	0.492

Total Charge (with Transformation)

Methods	rMSE	MAPE	lnQ
PMM	0.802	1.102	0.888
PPD	0.967	1.117	0.924
BEM	1.092	1.171	0.943
DA	0.968	1.192	0.935
Kriging	0.545	0.653	0.418

Last Comments

- High dimensional large ill-conditioned statistical problems can now be solved efficiently and accurately
  - Multi-level method that decouples the estimation and prediction problems
  - Multilevel covariance matrix exhibits fast decay and well conditioned
  - Numerically stable
- Future work
  - Develop a fast summation method in high dimensions
  - Extension to non-stationary covariance function

Acknowledgements

- George Biros, Victor Eijkhout, David Keyes, Alexander Litvinenko, Sudarshan Raghunathan, Raul Tempone, and Lexing Ying.

References I

Mihai Anitescu, Jie Chen, and Lei Wang, *A matrix-free approach for solving the parametric gaussian process maximum likelihood problem*, SIAM Journal on Scientific Computing **34** (2012), no. 1, 240–262.

Satish Balay, Jed Brown, Kris Buschelman, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang, *PETSc users manual*, Tech. Report ANL-95/11 - Revision 3.4, Argonne National Laboratory, 2013.

Satish Balay, Jed Brown, Kris Buschelman, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Lois Curfman McInnes, Barry F. Smith, and Hong Zhang, *PETSc Web Page*, 2013, http://www.mcs.anl.gov/petsc.

Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith, *Efficient management of parallelism in object oriented numerical software libraries*, Modern Software Tools in Scientific Computing (E. Arge, A. M. Bruaset, and H. P. Langtangen, eds.), Birkhäuser Press, 1997, pp. 163–202.

References II

- Julio E. Castrillón-Candás, Marc G. Genton, and Rio Yokota, *Multi-level restricted maximum likelihood covariance estimation and kriging for large non-gridded spatial datasets*, Spatial Statistics (2015), –.
- Julio E. Castrillón-Candás, Jun Li, and Victor Eijkhout, *A discrete adapted hierarchical basis solver for radial basis function interpolation*, BIT Numerical Mathematics **53** (2013), no. 1, 57–86 (English).
- R. Furrer, M. G. Genton, and D. Nychka, *Covariance tapering for interpolation of large spatial datasets*, Journal of Computational and Graphical Statistics **15** (2006), no. 3, 502–523.
- Cari G. Kaufman, Mark J. Schervish, and Douglas W. Nychka, *Covariance tapering for likelihood-based estimation in large spatial datasets*, Journal of the American Statistical Association **103** (2008), no. 484, 1545–1555.
- Dishi Liu and Hermann G. Matthies, *Pivoted cholesky decomposition by cross approximation for efficient solution of kernel systems*, Arxiv (2015).

References III

- Alexander Litvinenko, Ying Sun, Marc G. Genton, and David E. Keyes, *Likelihood approximation with hierarchical matrices for large spatial datasets*, Soon to be published (2016).
- Y. Sun, B. Li, and M. G. Genton, *Geostatistics for large datasets*, Space-Time Processes and Challenges Related to Environmental Problems (M. Porcu, J. M. Montero, and M. Schlather, eds.), Springer, 2012, pp. 55–77.
- Y. Sun and M. L. Stein, *Statistically and computationally efficient estimating equations for large spatial datasets*, Journal of Computational and Graphical Statistics, (2014).
- L. Ying, G. Biros, and D. Zorin, *A kernel-independent adaptive fast multipole method in two and three dimensions*, Journal of Computational Physics **196** (2004), no. 2, 591–626.