Fast Approximation of Covariance Functions Using a Hierarchical Matrices Approach

A. Gorshechnikova[a], C. Gaetan[b]

[a]University of Padova, Padova, Italy
[b]Ca' Foscari University of Venice, Venice, Italy

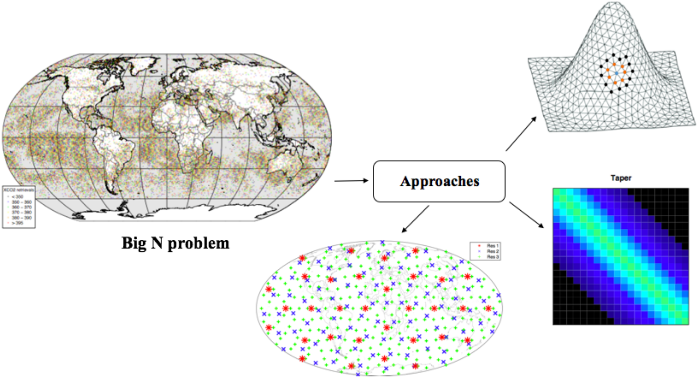
November 8, 2018

Table of Contents

Table of Contents

"Big N problem"

Methods proposed include: Covariance Tapering[1]; Low-rank approximations: Predictive Processes[2], Fixed Rank Kriging[3]; Gaussian Markov Random Field[4]...



[1] Furrer, Genton, and Nychka 2006.
[2] Banerjee et al. 2008.
[3] Cressie and Johannesson 2008.
[4] Lindgren, Rue, and Lindström 2011.

Formulation of the problem

- Consider a single realization $Z = (Z(x_1), \ldots, Z(x_n))'$ from a spatial random field.
- $Z(x)$ is zero mean Gaussian field. The likelihood is written as

$$L(\theta) = -\frac{n}{2}\log 2\pi - \frac{1}{2}\log|G(x,y)| - \frac{1}{2}Z^T G(x,y)^{-1} Z$$

where $G(x,y)$ is the covariance matrix.

Operation	Complexity
Matrix-vector multiplication	$O(n^2)$
Matrix inversion	$O(n^3)$

Exact computation of the likelihood requires computational complexity of order $O(n^3)$

Stochastic Partial Differential Equation (SPDE)

- Consider linear operator equation $Lu = f$ in $\Omega \subset \mathbb{R}^d$, where L is a boundedly invertible elliptic differential operator of order $r \in \mathbb{R}^d$
- The covariance function of a **Matérn field**

$$G(x,y) = \frac{1}{\Gamma(\lambda + d/2)(4\pi)^{d/2}\kappa^{2\lambda}2^{\lambda-1}}(\kappa||x-y||)^{\lambda}K_{\lambda}(\kappa||x-y||)$$

is the **Green's function** of the differential operator $L_{\lambda}^2 = (\kappa^2 - \Delta)^{\lambda+d/2}$ of the linear fractional SPDE[5]

$$(\kappa^2 - \Delta)^{\lambda+d/2} Z = W, \quad \kappa > 0, \quad \lambda > 0, \quad (\lambda + d/2) \in \mathbb{R}$$

with Laplacian Δ, a smoothness parameter λ and a spatial Gaussian white noise $W = \{W(x)\}$ with unit variance.

[5] Fasshauer 2012.

Green's function

- For all $x, y \in \Omega$, the Green's function $G(x, y)$ satisfies

$$LG(\cdot, y) = \delta_y$$

where δ_y is the Dirac distribution at $y \in \Omega$, and subject to the boundary conditions. Thus

$$u(x) = (L^{-1}f)(x) = \int_\Omega G(x, y) f(y) dy$$

- If $G(x, y)$ is analytic away from the diagonal, it allows for the separable approximation by $\mathcal{H}$-methods[6]

[6] Hackbusch 2015.

Table of Contents

Hierarchical Matrices Approach

New method based on the low-rank k approximation of $G(x, y)$ that results in $O(n \log n)$ order of computation[7]

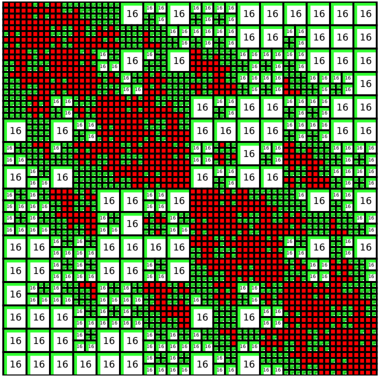
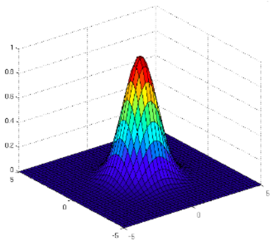


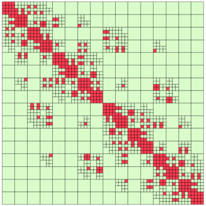
Figure 1: Hierarchical matrix representation with rank $k = 16$ and $n = 16641$

[7] Hackbusch 2015.

General framework

Hierarchical ($\mathcal{H}$)-matrix method is based on the following components:





1. **Analytical component**: local, separable approximation of covariance function $G(x, y)$ (Green's function)
2. **Linear algebra**: singular value decompositions to organise the local matrix data
3. **Discrete structures**: a suitable partition in submatrices for data compression and the ability to perform matrix operations in a linear cost.

Degenerate functions

Problem: Treat matrices G resulting from a covariance function $G(x, y)$

$$G_{ij} = G(x_i, y_j), \quad G_{ij} = \int\int G(x,y)\,\phi_i(x)dx\,\phi_j(y)dy, \quad x \in D_x, y \in D_y$$

Goal: Find a low-rank approximation $G \approx AB^T$

$$G_{ij} \approx \sum_{\nu=1}^{k} A_{i\nu} B_{j\nu}$$

Approach: Use a degenerate approximation of the covariance function

$$G^k(x,y) \approx \sum_{\nu=1}^{k} a_\nu(x) b_\nu(y) \tag{1}$$

$$G_{ij} \approx \sum_{\nu=1}^{k} \underbrace{\int a_\nu(x)\phi_i(x)dx}_{A_{i\nu}} \underbrace{\int b_\nu(y)\phi_j(y)dy}_{B_{j\nu}}$$

Admissibility condition

- Two domains $D_x, D_y \subset \mathbb{R}^2$ are η-**admissible** if for some fixed $\eta > 0$

$$\min\{\mathrm{diam}(D_x), \mathrm{diam}(D_y)\} \leq \eta \, \mathrm{dist}(D_x, D_y) \tag{2}$$

$$\mathrm{dist}(D_x, D_y) = \inf\{|x-y| : x \in D_x, y \in D_y\}$$

$$\mathrm{diam}(D_x) = \sup\{|x-y| : x \in D_x, y \in D_x\}$$

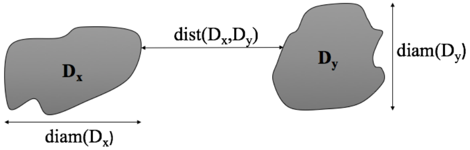


Figure 2: admissibility condition

Asymptotic smoothness condition

- Define multi-index $\alpha = (\alpha_1, \ldots, \alpha_d) \in \mathrm{N}_0^d$ with $|\alpha| = \alpha_1 + \cdots + \alpha_d$ and $\partial^\alpha = \partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d}$
- The covariance function $G(x, y) \in C^\infty$ is **asymptotically smooth** if there are constants $C, \sigma \in \mathbb{R}_{>0}$ satisfying

$$|\partial_x^\alpha \partial_y^\beta G(x, y)| \leq C(\alpha, \beta)|x - y|^{-|\alpha|-|\beta|-\sigma} \tag{3}$$

  for all $x, y \in \mathbb{R}^d$.
- The **Matérn covariance** satisfies the smoothness condition (3) (e.g. proof based on the symbols calculus for pseudodifferential operators[8])

[8] Dölz, Harbrecht, and Schwab 2017.

Admissibility condition in the space-time domain

- The standard asymptotic smoothness condition (3) is constructed with Green's functions for elliptic operators and leads to the standard admissibility condition
- For the space-time covariance function, a new **space-time admissibility condition** can be derived provided that the covariance is approximated on Cartesian products of spatial bounding boxes S and temporal intervals $[0, T]$, e.g., interpolation on $S \times [0, T]$ for $S \subseteq \mathbb{R}^2$ and $[0, T] \subseteq \mathbb{R}^+$ converges exponentially at a bounded rate
- This new condition can be used to construct a block tree with space-time clusters

Question: How do we find degenerate approximations?

Separable expansion of the covariance

Taylor expansion around $\hat{x}$ is a tool to obtain approximating polynomials

$$G(x,y) \approx G^k(x,y) = \sum_{|\nu|\le k} \underbrace{\frac{(x-\hat{x})^\nu}{\nu}}_{a_\nu(x)} \underbrace{\frac{\partial^\nu G}{\partial x}(\hat{x},y)}_{b_\nu(y)}$$

Error estimate for analytic functions[9]

$$|G(x,y) - G^k(x,y)| \le \left(\frac{\mathrm{diam}(D_x)}{\mathrm{dist}(D_x, D_y)}\right)^{k+1} \quad \text{for all } x \in D_x, y \in D_y$$

Disadvantages: Evaluation of the derivatives
Idea: Use Lagrange interpolation instead of Taylor expansion (see Appendix)

[9] Hackbusch 2015.

Table of Contents

$\mathcal{H}$-Approximation of Covariance Matrix

The $\mathcal{H}$-matrix technique is used to approximate the Gaussian likelihood function. The $\mathcal{H}$-approximation of the exact log-likelihood $L(\theta)$ is defined by $\tilde{L}(\theta, k)$:

$$\tilde{L}(\theta, k) = -\frac{n}{2}\log 2\pi - \sum_{i=1}^{n} \log\{\tilde{\Lambda}_{ii}(\theta, k)\} - \frac{1}{2}u(\theta)^T u(\theta)$$

where $\tilde{\Lambda}(\theta, k)$ is an $\mathcal{H}$-matrix approximation[10] of the Cholesky factor $\Lambda(\theta)$ with the maximal rank k and $u(\theta)$ is the solution of the linear system $\tilde{\Lambda}(\theta, k)u(\theta) = Z$

Operation	$\mathcal{H}$-Complexity	Complexity
MV multiplication	$O(kn\, log^2 n)$	$O(n^2)$
Cholesky decomposition	$O(kn\, log^2 n)$	$O(n^3)$
MLE cost[11]	$O(\#I \cdot kn \log^2 n)$	$O(\#I \cdot n^3)$

[10] Litvinenko et al. 2017.

[11] $\#I$ is the number of interations

Application in the spatial context

- We make an inference on the spatial process $\{S(x) : x \in D \subset \mathbb{R}^2\}$ which is assumed to have linear mean structure

$$S(x) = t(x)'\alpha + Z(x), \quad x \in D$$

  where $Z(x)$ is the Gaussian Random Field with the Matérn covariance function, $t(x) = (t_1(x), \ldots, t_p(x))$ is the vector process of p known covariates and coefficients $\alpha = (\alpha_1, \ldots, \alpha_p)$ are unknown.
- **Fixed Rank Kriging**: $\eta(x) = S(x) + \epsilon(x)$, where $\epsilon(x)$ is a spatial white noise with diagonal covariance matrix $\mathrm{Var}(\epsilon)$
- We detrend the data, i.e. $\tilde{Z}(x) = S(x) - t(x)'\hat{\alpha}$ using the OLS estimate for α. Then we estimate the parameter θ of the covariance function $G(x, y, \theta)$.

Kriging prediction with $\mathcal{H}$-matrices

With the estimated ML parameters $\hat{\theta}$ of the covariance function the prediction

$$\hat{S}_{\mathcal{H}}(x_0) = t(x_0)'\hat{\alpha} + G_{\mathcal{H}}(x_0, \hat{\theta})^T G_{\mathcal{H}}^{-1}(\hat{\theta})(S - T\hat{\alpha}),$$

$$\hat{\alpha} = (T' G_{\mathcal{H}}^{-1} T)^{-1} T' G_{\mathcal{H}}^{-1} S$$

where $G_{\mathcal{H}}$ is the approximated in the $\mathcal{H}$-format covariance and the covariance vector $G_{\mathcal{H}}(x_0, \hat{\theta}) = [G(x_0, x_1), \ldots, G(x_0, x_n)]'$ is taken between the sites $x_1, \ldots, x_n$ and a prediction location x_0.

Therefore, taking n data locations and m prediction locations, we obtain the computational cost of order $O(mn\, log\, n)$ compared to $O(mn^3)$

Comparison study with the simulated dataset

- We perform the experiments with simulated data to recover the true values of the parameters of the Matérn covariance $(\kappa, \sigma^2) = (1.0, 2.4)$ with a cross validation taking $k = \{2000, 6000, 10000\}$ points at random and predicting them. The procedure is repeated $M = 30$ times.
- The model-q root-mean-squared prediction error $(RMSPE)$ for the m-th simulation

$$RMSPE_q(m) = \sqrt{\sum_{s \in D} \left(\hat{S}(x,m) - S(x,m)\right)^2}, \quad m = 1, \ldots, M$$

where $q = FRK, HLM$ and D is the domain of the prediction locations

Sample size	*Method*	*RMSPE*	*Likelihood*	*Time(lik)*	*Time(kr)*	$\hat{\sigma}^2$	$\hat{\kappa}$
10000	*HM*	**1.158397**	**-12872.22**	15	**1.02**	2.456	1.003
	FRK	1.4135	-14377.21	**11.83**	2.74	-	-
30000	*HM*	**1.045781**	**-34060.76**	34.63	9.16	2.473	1.033
	FRK	1.361863	-42270.61	**17.89**	**8.11**	-	-
100000	*HM*	**1.098694**	**-95640.89**	**148.7**	**45.2**	2.462	1.082
	FRK	1.370223	-139004.2	170.7	56.84	-	-

Comparison study with the real dataset

- A study of tropospheric CO_2 $n = 43059$ measurements retrieved from the Atmospheric InfraRed Sounder (*AIRS*) between 1^{st} and 3^{rd} of May 2003.
- We compare $\mathcal{H}$*-method* with *FRK* and assess the utility of the methods on a validation dataset that we hold out

Method	*RMSPE*	*Time (lik),m*	*Time (kr),m*
HM	3.12	150.57	**18.32**
FRK	**3.09**	**119.87**	31.82

Table of Contents

Any suggestions are welcome...

1. Prove the exponential convergence of the expansion error for the spatio-temporal covariance function
2. Reformulate the admissibility condition in the space-time domain $S \times [0, T]$
3. Implement the procedures for the cluster routines: construct a block tree with space-time clusters

Thanks for your attention!

Appendix: Lagrange interpolation

- One-dimensional example $x \in \mathbb{R}$: smooth $G(x,y)$ is approximated by a separable expansion with rank k as long as x and y are well separated[12]
- A separable approximation[13] $G^k(x,y)$ of $G(x,y)$ is given by the Lagrange polynomial $L_j(x)$ in x which interpolates $G(x,y)$ at the points $\hat{x}_j \in [-1,1]$

$$G^k(x,y) = \sum_{j=1}^{k} L_j(x) G(\hat{x}_j, y)$$

with k-order Chebyshev interpolation points

$$\hat{x}_j = \cos\left(\frac{2j-1}{2k}\pi\right), \quad j \in \{1,\ldots,k\}$$

[12] Iske, Borne, and Wende 2017.

[13] The affine map from $[-1,1]$ onto $[a,b]$ yields Chebyshev nodes $\frac{a+b}{2} + \frac{b-a}{2}\hat{x}_j$

References I

Banerjee, Sudipto et al. (2008). "Gaussian predictive process models for large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.4, pp. 825–848.

Cressie, Noel and Gardar Johannesson (2008). "Fixed rank kriging for very large spatial data sets". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70.1, pp. 209–226.

Dölz, Jürgen, Helmut Harbrecht, and Ch Schwab (2017). "Covariance regularity and H-matrix approximation for rough random fields". In: *Numerische Mathematik* 135.4, pp. 1045–1071.

Fasshauer, Gregory E (2012). "Green's functions: Taking another look at kernel approximation, radial basis functions, and splines". In: *Approximation Theory XIII: San Antonio 2010*, pp. 37–63.

Furrer, Reinhard, Marc G Genton, and Douglas Nychka (2006). "Covariance tapering for interpolation of large spatial datasets". In: *Journal of Computational and Graphical Statistics* 15.3, pp. 502–523.

Hackbusch, Wolfgang (2015). *Hierarchical matrices: algorithms and analysis*. Vol. 49. Springer.

References II

Iske, Armin, Sabine Le Borne, and Michael Wende (2017). "Hierarchical Matrix Approximation for Kernel-Based Scattered Data Interpolation". In: *SIAM Journal on Scientific Computing* 39.5, A2287–A2316.

Lindgren, Finn, Håvard Rue, and Johan Lindström (2011). "An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73.4, pp. 423–498.

Litvinenko, Alexander et al. (2017). "Likelihood Approximation With Hierarchical Matrices For Large Spatial Datasets". In: *arXiv preprint arXiv:1709.04419*.