King Abdullah University of Science and Technology KAUST

Parallel Hierarchical Matrix Technique to Approximate Large Covariance Matrices, Likelihood Functions and Parameter Identification

Alexander Litvinenko[1],
(joint work with V. Berikov[2], M. Genton[3], D. Keyes[3], R. Kriemann[4], and Y. Sun[3])
SIAM CSE 2021

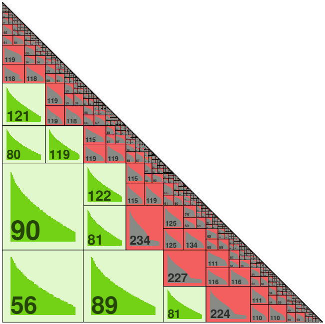
[1]RWTH Aachen, Germany, [2]Novosibirsk Staate University, [3]KAUST, Saudi Arabia, [4]MPI for Mathematics in the Sciences in Leipzig, Germany

Analysis of large datasets and prediction

Goal: analyse a large dataset

Need to use a dense cov. matrix C of size $2,000,000 \times 2,000,000$

0.422416	0.30634	0.483978
0.327703	0.818662	0.255766
0.391363	0.253407	0.809925
0.860548	0.231373	0.398615
0.533181	0.83534	0.445115
0.888258	0.357244	1.629436
0.301518	0.934554	-1.120526
0.064668	0.764294	-0.424035
0.301305	0.903365	-0.805926
0.310582	0.984848	-1.040024
0.247669	0.162615	0.03276
0.67071	0.881453	1.714354
0.300404	0.118761	1.21913
0.707469	0.925216	1.262887
0.551263	0.817304	0.655349



We show how to:

1. reduce storage cost from 32TB to 16 GB
2. approximate Cholesky factorisation of C, its determinant, inverse in 8 minutes on modern desktop computer.
3. make prediction

The structure of the talk

1. Motivation: improve statistical models, data analysis, prediction
2. Identification of unknown parameters via maximizing Gaussian log-likelihood (MLE)
3. Tools: Hierarchical matrices [Hackbusch 1999]
4. Matérn covariance function, joint Gaussian log-likelihood
5. Error analysis
6. Prediction at new locations
7. Comparison with machine learning methods

Given:
Let $\mathrm{s}_1, \ldots, \mathrm{s}_n$ be locations.
$\mathrm{Z} = \{Z(\mathrm{s}_1), \ldots, Z(\mathrm{s}_n)\}^\top$, where $Z(\mathrm{s})$ is a Gaussian random field indexed by a spatial location $\mathrm{s} \in \mathbb{R}^d$, $d \geq 1$.

Assumption: Z has mean zero and stationary parametric covariance function, e.g. Matérn:

$$\mathrm{C}(\boldsymbol{\theta}) = \frac{2\sigma^2}{\Gamma(\nu)} \left(\frac{r}{2\ell}\right)^\nu K_\nu\left(\frac{r}{\ell}\right) + \tau^2 \mathrm{I}, \quad \boldsymbol{\theta} = (\sigma^2, \nu, \ell, \tau^2).$$

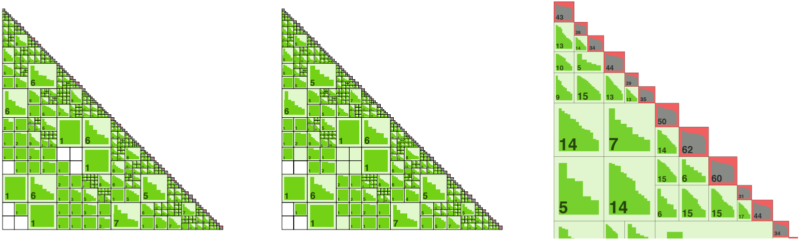
To identify: unknown parameters $\boldsymbol{\theta} := (\sigma^2, \nu, \ell, \tau^2)$.

Statistical inference about $\boldsymbol{\theta}$ is then based on the Gaussian log-likelihood function:

$$\mathcal{L}(\mathrm{C}(\boldsymbol{\theta})) = \mathcal{L}(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathrm{C}(\boldsymbol{\theta})| - \frac{1}{2}\mathrm{Z}^\top \mathrm{C}(\boldsymbol{\theta})^{-1}\mathrm{Z}, \quad (1)$$

where the covariance matrix $\mathrm{C}(\boldsymbol{\theta})$ has entries $C(\mathrm{s}_i - \mathrm{s}_j; \boldsymbol{\theta})$, $i, j = 1, \ldots, n$.
The maximum likelihood estimator of $\boldsymbol{\theta}$ is the value $\widehat{\boldsymbol{\theta}}$ that maximizes (1).

$\mathcal{H}$-matrix approximations of C, L and C^{-1}



$\mathcal{H}$-matrix approximations of the exponential covariance matrix (left), its hierarchical Cholesky factor $\tilde{L}$ (middle), and the zoomed upper-left corner of the matrix (right), $n = 4000$, $\ell = 0.09$, $\nu = 0.5$, $\sigma^2 = 1$.

What will change after $\mathcal{H}$-matrix approximation?

Approximate C by $C^{\mathcal{H}}$

1. How the eigenvalues of C and $C^{\mathcal{H}}$ differ ?
2. How det(C) differs from det($C^{\mathcal{H}}$) ?
3. How L differs from $L^{\mathcal{H}}$? [Mario Bebendorf et al]
4. How C^{-1} differs from $(C^{\mathcal{H}})^{-1}$? [Mario Bebendorf et al]
5. How $\tilde{\mathcal{L}}(\theta, k)$ differs from $\mathcal{L}(\theta)$?
6. What is optimal $\mathcal{H}$-matrix rank?
7. How $\theta^{\mathcal{H}}$ differs from θ?

For theory, estimates for the rank and accuracy see works of Bebendorf, Grasedyck, Le Borne, Hackbusch,...

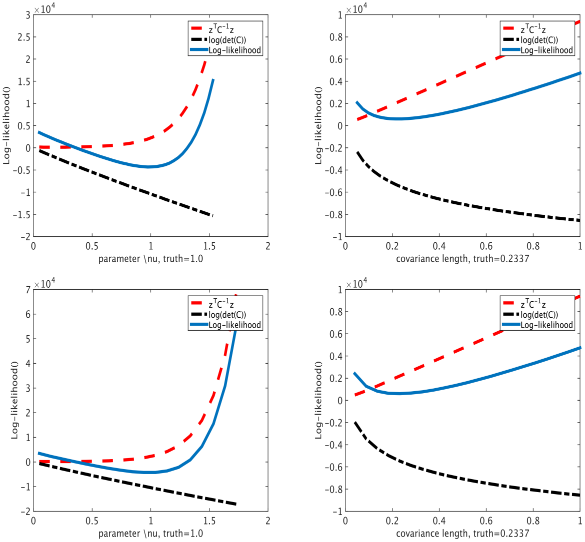
Details of the parameter identification

To maximize the log-likelihood function we use the Brent's method (combining bisection method, secant method and inverse quadratic interpolation) or any other.

1. $\mathrm{C}(\boldsymbol{\theta}) \approx \widetilde{\mathrm{C}}(\boldsymbol{\theta}, \varepsilon)$.
2. $\widetilde{\mathrm{C}}(\boldsymbol{\theta}, k) = \widetilde{\mathrm{L}}(\boldsymbol{\theta}, k)\widetilde{\mathrm{L}}(\boldsymbol{\theta}, k)^T$
3. $\mathrm{Z}^T\widetilde{\mathrm{C}}^{-1}\mathrm{Z} = \mathrm{Z}^T(\widetilde{\mathrm{L}}\widetilde{\mathrm{L}}^T)^{-1}\mathrm{Z} = \mathrm{v}^T \cdot \mathrm{v}$, where v is a solution of $\widetilde{\mathrm{L}}(\boldsymbol{\theta}, k)\mathrm{v}(\boldsymbol{\theta}) := \mathrm{Z}$.

$$\log\det\{\widetilde{\mathrm{C}}\} = \log\det\{\widetilde{\mathrm{L}}\widetilde{\mathrm{L}}^T\} = \log\det\{\prod_{i=1}^{n}\lambda_i^2\} = 2\sum_{i=1}^{n}\log\lambda_i,$$

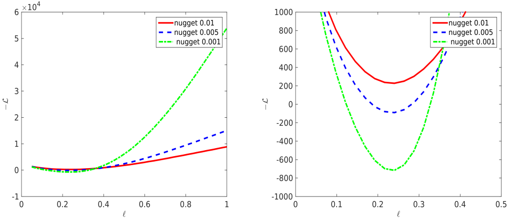
$$\widetilde{\mathcal{L}}(\boldsymbol{\theta}, k) = -\frac{n}{2}\log(2\pi) - \sum_{i=1}^{n}\log\{\widetilde{\mathrm{L}}_{ii}(\theta, k)\} - \frac{1}{2}(\mathrm{v}(\boldsymbol{\theta})^T \cdot \mathrm{v}(\boldsymbol{\theta})). \quad (2)$$



Dependence of log-likelihood ingredients on parameters, $n = 4225$. $k = 8$ in the first row and $k = 16$ in the second.

Remark: stabilization with nugget

To avoid instability in computing Cholesky, we add: $\widetilde{\mathrm{C}}_m = \widetilde{\mathrm{C}} + \tau^2 \mathrm{I}$.
Let λ_i be eigenvalues of $\widetilde{\mathrm{C}}$, then eigenvalues of $\widetilde{\mathrm{C}}_m$ will be $\lambda_i + \tau^2$,
$\log\det(\widetilde{\mathrm{C}}_m) = \log \prod_{i=1}^{n}(\lambda_i + \tau^2) = \sum_{i=1}^{n} \log(\lambda_i + \tau^2)$.



(left) Dependence of the negative log-likelihood on parameter ℓ with nuggets $\{0.01, 0.005, 0.001\}$ for the Gaussian covariance. (right) Zoom of the left figure near minimum; $n = 2000$ random points from moisture example, rank $k = 14$, $\tau^2 = 1$, $\nu = 0.5$.

Theorem (1)

Let $\widetilde{\mathsf{C}}$ be an $\mathcal{H}$-matrix approximation of matrix $\mathsf{C} \in \mathbb{R}^{n \times n}$ such that

$$\rho(\widetilde{\mathsf{C}}^{-1}\mathsf{C} - \mathsf{I}) \leq \varepsilon < 1.$$

Then

$$|log \det \mathsf{C} - log \det \widetilde{\mathsf{C}}| \leq -n log(1 - \varepsilon), \tag{3}$$

Proof: See [Ballani, Kressner 14] and [Ipsen 05].
Remark: factor n is pessimistic and is not really observed numerically.

Error in the log-likelihood

Theorem (2)

Let $\widetilde{C} \approx C \in \mathbb{R}^{n\times n}$ and Z be a vector, $\|Z\| \leq c_0$ and $\|C^{-1}\| \leq c_1$. Let $\rho(\widetilde{C}^{-1}C - I) \leq \varepsilon < 1$. Then it holds

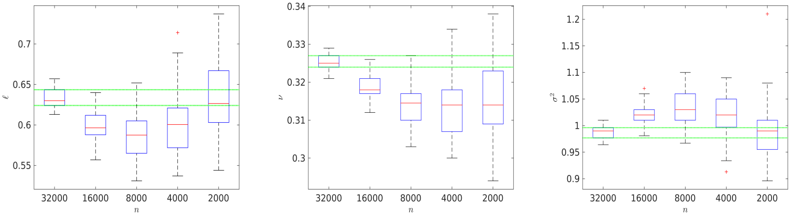
$$
\begin{aligned}
|\widetilde{\mathcal{L}}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta})| &= \frac{1}{2}(log|C| - log|\widetilde{C}|) + \frac{1}{2}|Z^{\mathrm{T}}\left(C^{-1} - \widetilde{C}^{-1}\right)Z| \\
&\leq -\frac{1}{2}\cdot nlog(1-\varepsilon) + \frac{1}{2}|Z^{\mathrm{T}}\left(C^{-1}C - \widetilde{C}^{-1}C\right)C^{-1}Z| \\
&\leq -\frac{1}{2}\cdot nlog(1-\varepsilon) + \frac{1}{2}c_0^2\cdot c_1\cdot\varepsilon.
\end{aligned}
$$

ε accuracy in each sub-block, $n = 16641$, $\nu = 0.5$, c.r.=compression ratio.

ε	$\lvert \log\lvert C \rvert - \log\lvert \tilde{C} \rvert \rvert$	$\lvert \frac{\log\lvert C \rvert - \log\lvert \tilde{C} \rvert}{\log\lvert \tilde{C} \rvert} \rvert$	$\lVert C - \tilde{C} \rVert_F$	$\frac{\lVert C - \tilde{C} \rVert_2}{\lVert C \rVert_2}$	$\lVert I - (\tilde{L}\tilde{L}^\top)^{-1} C \rVert_2$	c.r. in %
$\ell = 0.0334$						
1e-1	3.2e-4	1.2e-4	7.0e-3	7.6e-3	2.9	9.16
1e-2	1.6e-6	6.0e-7	1.0e-3	6.7e-4	9.9e-2	9.4
1e-4	1.8e-9	7.0e-10	1.0e-5	7.3e-6	2.0e-3	10.2
1e-8	4.7e-13	1.8e-13	1.3e-9	6e-10	2.1e-7	12.7
$\ell = 0.2337$						
1e-4	9.8e-5	1.5e-5	8.1e-5	1.4e-5	2.5e-1	9.5
1e-8	1.45e-9	2.3e-10	1.1e-8	1.5e-9	4e-5	11.3

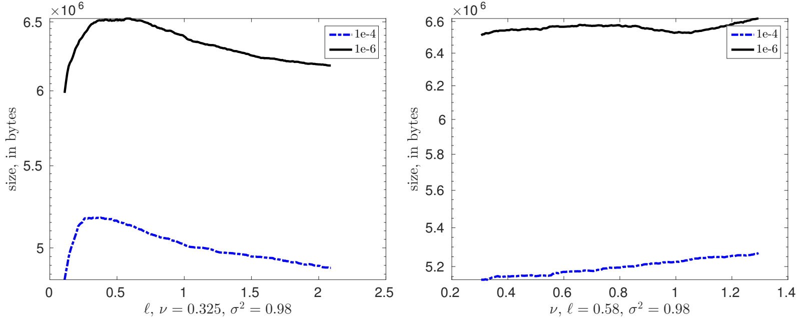
$\log\lvert C \rvert = 2.63$ for $\ell = 0.0334$ and $\log\lvert C \rvert = 6.36$ for $\ell = 0.2337$.

Boxplots for unknown parameters



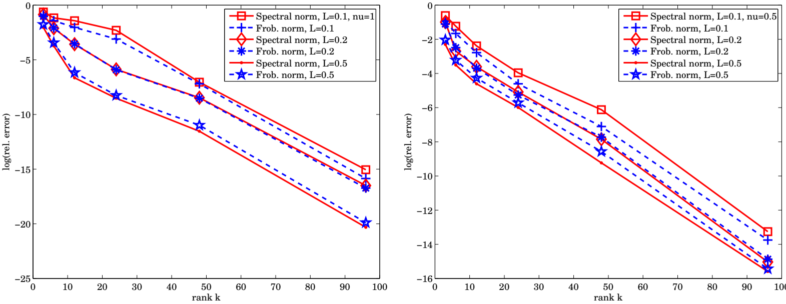
Moisture data example. Boxplots for the 100 estimates of (ℓ, ν, σ_2), respectively, when $n = 32K, 16K, 8K, 4K, 2K$. $\mathcal{H}$-matrix with a fixed rank $k = 11$. Green horizontal lines denote 25% and 75% quantiles for $n = 32K$.

How much memory is needed?



(left) Dependence of the matrix size on the covariance length ℓ, and (right) the smoothness ν for two different $\mathcal{H}$-accuracies $\varepsilon = \{10^{-4}, 10^{-6}\}$

Error convergence



Convergence of the $\mathcal{H}$-matrix approximation errors for covariance lengths $\{0.1, 0.2, 0.5\}$; (left) $\nu = 1$ and (right) $\nu = 0.5$, computational domain $[0,1]^2$.

How log-likelihood depends on n?

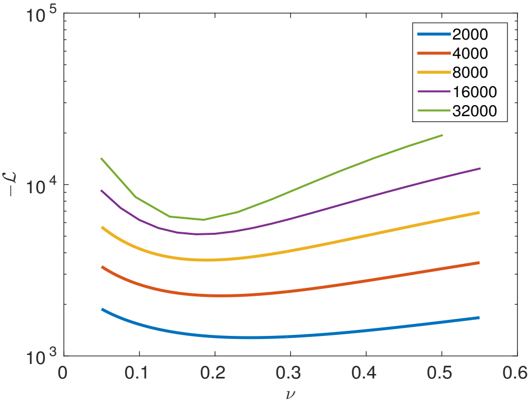


Figure: Dependence of negative log-likelihood function on different number of locations $n = \{2000, 4000, 8000, 16000, 32000\}$ in log-scale.

Maximal # cores is 40, $\nu = 0.325$, $\ell = 0.64$, $\sigma^2 = 0.98$

n	$\tilde{C}$			$\tilde{L}\tilde{L}^\top$		
	time sec.	size MB	kB/dof	time sec.	size MB	$\\|I - (\tilde{L}\tilde{L}^\top)^{-1}\tilde{C}\\|_2$
32.000	3.3	162	5.1	2.4	172.7	$2.4 \cdot 10^{-3}$
128.000	13.3	776	6.1	13.9	881.2	$1.1 \cdot 10^{-2}$
512.000	52.8	3420	6.7	77.6	4150	$3.5 \cdot 10^{-2}$
2.000.000	229	14790	7.4	473	18970	$1.4 \cdot 10^{-1}$

Dell Station, 20×2 cores, 128 GB RAM, bought in 2013 for 10.000 USD.

Let $Z = (Z_1, Z_2)^\top$ has mean zero and a stationary covariance, Z_1 - known, Z_2 unknown.

$$C = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22}, \end{bmatrix}$$

We compute predicted values

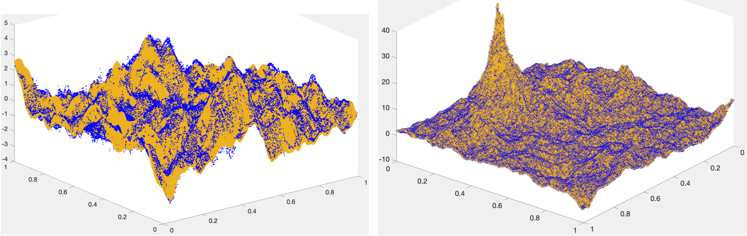
$$Z_2 = C_{21} C_{11}^{-1} Z_1$$

Z_2 has the conditional distribution with the mean value $C_{21} C_{11}^{-1} Z_1$ and the covariance matrix $C_{22} - C_{21} C_{11}^{-1} C_{12}$.

We participated in **2021 KAUST Competition on Spatial Statistics for Large Datasets**.

You can download the datasets and look the final results here
`https://cemse.kaust.edu.sa/stsds/2021-kaust-competition-spatial-statistics-large-datasets`

Prediction

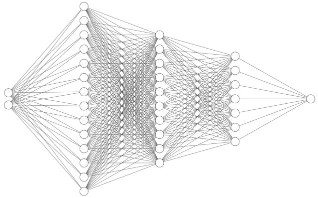


Prediction for two datasets. The yellow points at 900.000 locations were used for training and the blue points were predicted at 100.000 new locations.

k-nearest neighbours (kNN): For each point x find its k nearest neighbors $x_1, \ldots, x_k$, and set: $\hat{y}(x) = \frac{1}{k} \sum_{i=1}^{k} y_i$.

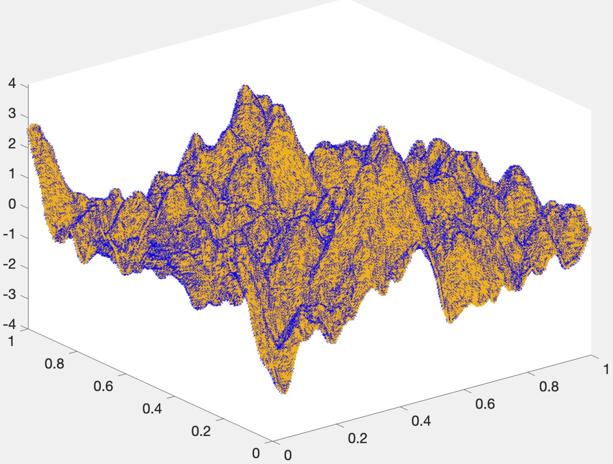
Random Forest (RF): a large number of decision (or regression) trees are generated independently on random sub-samples of data. The final decision for x is calculated over the ensemble of trees by averaging the predicted outcomes.

Deep Neural Network (DNN): fully connected neural network



Input layer consists of two neurons (input feature dimensionality), and output layer consists of one neuron (predicted feature dimensionality).

Prediction by kNN



Prediction obtained by the kNN method. The yellow points at 900.000 locations were used for training and the blue points were predicted at 100.000 new locations. One can see a very well alignment of both.

Conclusion

- With $\mathcal{H}$-matrices you can approximate Matérn covariance matrices, Gaussian log-likelihoods, identify unknown parameters and make predictions
- MLE estimate and predictions depend on $\mathcal{H}$-matrix accuracy
- parameter identification problem has multiple solutions
- Investigated dependence of $\mathcal{H}$-matrix approximation error on the estimated parameters
- Each of ML methods needs fine-tuning stage to optimize its hyperparameters or architecture.

Open questions and TODOs

- The Gaussian log-likelihood function has some drawbacks for very large matrices
- How to skip/avoid redundant data?
- A good starting point for optimization is needed
- a “preconditioner” (a simple cov. matrix) is needed
- $\mathcal{H}$-matrices become expensive for large number of parameters to be identified
- error estimates are needed

Literature

All tests are reproducible

https://github.com/litvinen/large_random_fields.git

1. A. Litvinenko, R. Kriemann, M.G. Genton, Y. Sun, D.E. Keyes, HLIBCov: Parallel hierarchical matrix approximation of large covariance matrices and likelihoods with applications in parameter identification, MethodsX 7, 100600, 2020
2. A. Litvinenko, Y. Sun, M.G. Genton, D.E. Keyes, Likelihood approximation with hierarchical matrices for large spatial datasets, Computational Statistics & Data Analysis 137, 115-132, 2019
3. A. Litvinenko, R. Kriemann, V. Berikov, Identification of unknown parameters and prediction with hierarchical matrices, look arXiv and ResearchGate, March 2021
4. M. Espig, W. Hackbusch, A. Litvinenko, H.G. Matthies, E. Zander, Iterative algorithms for the post-processing of high-dimensional data, Journal of Computational Physics 410, 109396, 2020

5. A. Litvinenko, D. Keyes, V. Khoromskaia, B.N. Khoromskij, H.G. Matthies, Tucker tensor analysis of Matérn functions in spatial statistics, Computational Methods in Applied Mathematics 19 (1), 101-122, 2019
6. B. N. Khoromskij, A. Litvinenko, H.G. Matthies, Application of hierarchical matrices for computing the Karhunen-Loéve expansion, Computing 84 (1-2), 49-67, 31, 2009
7. W. Nowak, A. Litvinenko, Kriging and spatial design accelerated by orders of magnitude: Combining low-rank covariance approximations with FFT-techniques, Mathematical Geosciences 45 (4), 411-435, 2013
8. M. Espig, W. Hackbusch, A. Litvinenko, H.G. Matthies, E. Zander, Efficient analysis of high dimensional data in tensor formats, Sparse Grids and Applications, 31-56, Springer, Berlin, 2013

Acknowledgement

V. Berikov was supported by the state contract of the Sobolev Institute of Mathematics (project no 0314-2019-0015), and by RFBR grant 19-29-01175.
A. Litvinenko was supported by funding from the Alexander von Humboldt Foundation.