Application of hierarchical matrices in spatial statistics

Applicazione di matrici gerarchiche nella statistica spaziale

Anastasiia Gorshechnikova and Carlo Gaetan

Abstract Large datasets with irregularly spatial (or spatio-temporal) locations are difficult to handle in many applications of Gaussian random fields such as maximum likelihood estimation (MLE) and prediction. We aim to approximate covariance functions in a format that facilitates the computation of MLE and prediction with very large datasets using a hierarchical matrix approach. We present a numerical study where we compare this approach with the covariance tapering method.

Abstract *Grandi dataset contenenti posizioni spazio-temporali disposte in maniera irregolare sono molto difficili da trattare in parecchie applicazioni dei campi aleatori gaussiani, quali la stima di massima verosimiglianza o la previsione. Il nostro obiettivo è di approssimare le funzioni di covarianza in un formato che faciliti il calcolo della stima di massima verosimiglianza e della previsione in caso di dataset molto grandi si basa sull'uso di matrici gerarchiche. Un esempio numerico in cui si confronta il metodo proposto con il 'tapering' viene presentato.*

Key words: computational methods, hierarchical matrices, large datasets, covariance matrices

1 Introduction

Large data sets are common in environmental sciences where data are often observed at a large number of spatial locations and at different temporal intervals. Therefore, computational and modeling challenges arise which were labeled by as

Anastasiia Gorshechnikova
Proximus, Boulevard du Roi Albert II 27 Brussels, Belgium, e-mail: agorshechnikova@gmail.com

Carlo Gaetan
Ca' Foscari University of Venice, Dorsoduro 3246 Venice, Italy, e-mail: gaetan@unive.it

"big N problem". The exact computation of the likelihood of a Gaussian Random Field (GRF) observed at N irregularly sited locations generally requires $O(N^3)$ floating point operations and $O(N^2)$ memory [3].

Consider a vector Z of N observations from a GRF $\{Z(x)\}$ defined over a domain indexed by x, where x denotes either a spatial $x : s \in \mathbb{R}^d$ or spatio-temporal domain of observations $x : (s,t) \in \mathbb{R}^d \times \mathbb{R}$. Without loss of generality we consider a zero-mean GRF. Considering parametric covariance function with the vector of the unknown p- dimensional parameters $\theta \in \Theta \subseteq \mathbb{R}^p$, the covariance function $c(x) := c(x;\theta)$ depends on unknown parameter θ. We make statistical inference with respect to θ based on the Gaussian log-likelihood

$$L(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|C_Z| - \frac{1}{2}Z^\top C_Z^{-1}Z \tag{1}$$

where N is the sample size and C_Z is the covariance matrix of Z. As can be seen from (1), to make an inference on the unknown parameter θ the exact computation of the log-likelihood requires a computation of the determinant of the covariance matrix $|C_Z|$ as well as its inverse C_Z^{-1} which both require $O(N^3)$ operations.

A similar computational burden is involved in evaluating the best linear unbiased prediction (BLUP), at an unobserved location x_0 defined as follows

$$Z(x_0) = c(x_0)^\top C_Z^{-1}Z, \tag{2}$$

where $c(x_0) = [c(x_0,x_1),\dots,c(x_0,x_N)]'$ is covariance vector formed based on a new location x_0 and $C_Z = C(x_i,x_j)$.

A comparison of current methods to tackle this computational problem is contained in [3]. For instance, in the covariance tapering approach [4] the covariance matrices are multiplied element-wise by a sparse correlation matrix which results in another positive definite function with a compact support, i.e. $C_T = C_Z \circ T(\delta)$, where $T(\delta)$ is a compactly supported correlation function which is identically zero whenever $||s-s'|| \geq \delta$ with $s,s' \in \mathbb{R}^d$ and taper (or cut-off distance) δ. Therefore

$$L(\theta) = -\frac{N}{2}\log 2\pi - \frac{1}{2}\log|C_T| - \frac{1}{2}Z^T C_T^{-1}Z \tag{3}$$

is the tapered likelihood, where $C_T = C_Z \circ T(\delta)$.

The covariance tapering method may not be effective in accounting for spatial dependence with long range dependence thereby sacrificing some precision. Also it is not straightforward how to choose the distance to taper off. In this work we present an approach based on the approximation of covariance functions by hierarchical matrices (or shortly $\mathscr{H}$-matrices). Focusing on the numerical analysis, the method of $\mathscr{H}$-matrix was exploited by [5] for MLE estimation. We extend this work by adapting the regularity conditions, performing kriging prediction on a simulated dataset and comparing this technique with covariance tapering in terms of both computational and statistical efficiencies.

2 Hierarchical matrices

The idea behind $\mathscr{H}$-matrices is to use a low-rank approximation of the blocks of a covariance matrix which are located far from the diagonal entries. To obtain the structure of a covariance matrix with the off-diagonal blocks $\tilde{C}^k_{\text{block}}$ approximated in a low-rank k format, an index $i \in I$ from the index set $I \subset \mathbb{N}$ is firstly assigned to each data location $x_i \in \mathbb{R}^d$. The hierarchical structure of a matrix is then obtained by partitioning the index set I into subsets or, equivalently, associated data locations x_i into clusters. This is required in order to obtain matrix blocks which further can be factored, such that a low-rank block $\tilde{C}^k_{\text{block}}$ is characterised by the rank $k << N$. These all are crucial steps required to compress data and perform matrix operations in a linear cost. We refer to [2] for the technical details of the $\mathscr{H}$-matrix method.

The matrix C resulting from a covariance function $c(\cdot)$ is not sparse. To find a data sparse representation of some blocks of the covariance matrix, their low-rank decomposition must be exploited. We refer to [1] for the description of analytical techniques to find a low-rank approximation of a block $\tilde{c}^k(x_i,x_j)$ of the covariance function $c(x_i,x_j)$. According to [2], to admit a low-rank representation it is necessary that the underlying functions satisfy so called ‘asymptotic smoothness condition’.

We define a d-dimensional multi-index notation $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ of non-negative integers. For the multi-index $\alpha \in \mathbb{N}_0^d$ sum of the components or absolute value can be written as $|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_d$ and higher-order partial derivatives as $\partial^\alpha = \partial_1^{\alpha_1} \partial_2^{\alpha_2} \ldots \partial_d^{\alpha_d}$, where $\partial_i^{\alpha_i} = \partial^{\alpha_i}/\partial x_i^{\alpha_i}$ of the dimension d. Let $X_i, X_j \subset \mathbb{R}^d$ be subsets such that the function $c(x_i,x_j)$ is defined and arbitrarily often differentiable for all spatial locations $x_i \in X_i$ and $x_j \in X_j$ with $x_i \neq x_j$ for $i, j = 1, \ldots, N$. Then the covariance function $c(x_i,x_j)$ is asymptotically smooth if there exist constants $p_1, p_2 \in \mathbb{R}^+$, such that for all multi-indices $\alpha \in N_0^d$, one has

$$|\partial_x^\alpha c(x_i,x_j)| \le p_1 |\alpha|! p_2^{|\alpha|} (||x_i - x_j||)^{-|\alpha|} \tag{4}$$

for all $x_i \neq x_j$.

The factor $p_2^{|\alpha|}$ allows for a change of the growth behaviour. The derivatives tend to 0 as $||x_i - x_j|| \to \infty$. The condition (4) is required to guarantee a fast decay of the eigenvalues of the underlying function which leads to an effective low-rank approximation $\tilde{C}^k_{\text{block}}$ of specific blocks of C, so that the error $|c(x_i,x_j) - \tilde{c}(x_i,x_j)|$ of a low-rank approximation of $c(x_i,x_j)$ converges exponentially fast. At first sight, the condition (4) seems to be restrictive. However, this condition is satisfied by some classes of spatial covariance functions such as Matérn and spatio-temporal covariance functions, see [1] for the details.

We aim to approximate a covariance matrix by the $\mathscr{H}$-method and perform a fast approximated Cholesky decomposition. We denote the $\mathscr{H}$-matrix approximation of the covariance matrix by $\tilde{C}$ and approximation of the Cholesky factor by $\tilde{\Lambda}$, so that $\tilde{C} = \tilde{\Lambda}\tilde{\Lambda}^T$. To be able to perform approximate Cholesky decomposition, the positive definiteness property of $\tilde{C}$ should be preserved. With the approximation by $\mathscr{H}$-matrices, the error can propagate and perturb the eigenvalues of the resulting matrix.

If the smallest eigenvalue is close to the origin, the result of these operations might become indefinite. To tackle this problem, we follow the suggestion of [5] to add a nugget value to the diagonal of $\tilde{C}$, albeit sacrificing approximation accuracy for the sake of positive definiteness. The $\mathscr{H}$-approximation of the exact log-likelihood $L(\theta)$ is defined by $\tilde{L}(\theta,k)$ with the maximal rank k

$$\tilde{L}(\theta,k) = -\frac{N}{2}\log 2\pi - \sum_{i=1}^{N}\log\tilde{\lambda}_i - \frac{1}{2}U^{\top}U, \tag{5}$$

where $U^TU = Z^T(\tilde{\Lambda}\tilde{\Lambda}^T)^{-1}Z = Z^TCZ$ which is composed of the matrix-vector multiplications with a log-linear cost and $\tilde{\lambda}_i$ are diagonal elements of $\tilde{\Lambda}$, such that $\log\det(C) = \log\det\tilde{\Lambda}\tilde{\Lambda}^T = \log\det\left(\prod_{i=1}^{N}\tilde{\lambda}_i^2\right) = 2\sum_{i=1}^{N}\log\tilde{\lambda}_i$.

As with the likelihood in (5), we substitute C_Z in (2) by the approximated by $\mathscr{H}$- covariance $\tilde{C}_Z$. Then a simple kriging prediction for a location x_0 using the estimated covariance function with $\hat{\theta}$ in (5) is $\tilde{Z}(x_0) = \tilde{c}(x_0)^{\top}\tilde{C}_Z^{-1}Z$, where $\tilde{c}(x_0) = [\tilde{c}(x_0,x_1),\ldots,\tilde{c}(x_0,x_N)]'$ is the $\mathscr{H}$-matrix approximation of the corresponding covariance vector. We note that as in (5) it is also based on the matrix-vector multiplications which leads to the log-linear cost computation due to the $\mathscr{H}$-matrices.

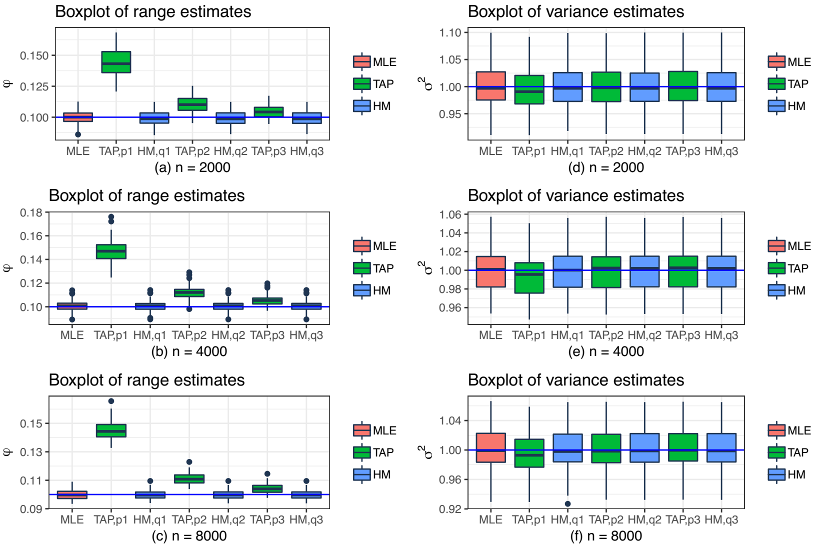
3 Numerical results

For the covariance tapering approach distant pairs of observations are modelled using a compactly supported covariance function. With the $\mathscr{H}$-matrices, off-diagonal elements of C_Z are defined through the low-rank factors. Because of the similarity of both methods, the main purpose of this section is to compare their performance based on computational and statistical efficiency. With the covariance tapering the 'score' function for θ based on (3) is biased. Since (5) also entails a biased score function, i.e $\frac{1}{2}\left(Z^T\tilde{C}_Z^{-1}C_i\tilde{C}_Z^{-1}Z - \mathrm{tr}(\tilde{C}_Z^{-1}C_{iZ})\right)$, we use (5) with $\mathscr{H}$-covariance $\tilde{C}_Z$ and (3) with tapered covariance C_T.

The simulation study is performed with the increasing domain asymptotics setup on the randomly perturbed grid of spatial locations by constructing a regular grid with increments 0.03 over $W_k = [0,2^{(k+2)/2}]\times[0,2^{(k+2)/2}],\quad k = 0,\ldots,2.$ and perturbing the regular grid points by adding a uniform random value on $[-0.01,0.01]$. With this setup, each data location is at least 0.01 units distant from its neighbours.

For the different sample size of $N_k = \{2000,4000,8000\}$ points with $k = 0,\ldots,2$ chosen without replacement, we simulate $L = 100$ realizations of zero-mean GRF with Matérn covariance with the true parameters $\theta = (\sigma^2,\varphi,\nu,\tau^2) = (1,0.1,0.5,0.1)$, where σ^2 is the marginal variance and τ^2 is the nugget parameter. We fix the smoothness parameter $\nu = 0.5$ (exponential covariance function) and $\tau^2 = 0.1$ is added to the diagonal in order to preserve the positive definiteness property. In addition, we scale the distance in (4) by the range parameter φ. This adjustment resulted in a computational efficiency that is doubled compared to the standard condition.

To check the predictive performance with the increasing N_k, we divide the simulated data into a training dataset chosen at random and a validation dataset containing the remaining 10%, i.e $M = \{200, 400, 800\}$ observations respectively. As practical range we set $\varphi = 0.1$ due to consistency of φ over the spatial domain to increasing domain framework. Because we keep distance as fixed, increasing k and consequently the number N_k of observations, the percentage of nonzero elements in the resulting tapered covariance matrix decreases. By varying the practical range $\delta = \{0.15, 0.3, 0.5\}$, the percentage of non-zero elements p in the tapered covariance matrix increases. For the $\mathscr{H}$-matrices we control the compression ratio q which is defined as the ratio between the sizes of a compressed (hierarchical matrix) $\tilde{C}$ and original matrix C. The *h2lib* library[1] was exploited for application of $\mathscr{H}$-matrices); for covariance tapering method *spam*[6] was used.



(a) n = 2000

(d) n = 2000

(b) n = 4000

(e) n = 4000

(c) n = 8000

(f) n = 8000

Fig. 1 Boxplots of sampled estimates (a)-(c) $\hat{\varphi}$ and (d)-(f) $\hat{\sigma}^2$ with the horizontal line of the true estimates $(\varphi = 0.1, \sigma^2 = 1)$ under the exact maximum likelihood estimation (MLE), covariance tapering (TAP) and $\mathscr{H}$-matrices (HM)

Figure 1 shows boxplots of the estimates of the φ and σ^2 parameters with the both methods, including the exact ML estimation. The horizontal line indicates the true values of the estimates $(\varphi = 0.1, \sigma^2 = 1)$. As taper δ decreases, the biases in the one-taper estimates increase. In contrast, we see negligible bias in the $\mathscr{H}$-matrices estimates. The difference in variance estimates with both methods is almost indistinguishable. In terms of computational efficiency, for example, for $n = 8000$ likelihood evaluation based on the $\mathscr{H}$-matrices with q_{max} required $t_{hm} = 2$ min compared to $t_{tap} = 7$ min by the covariance tapering approach with p_{max}. Thus, the application

[1] https://github.com/H2Lib/H2Lib, developed by Steffen Boerm and his group, Kiel, Germany

of $\mathscr{H}$-matrices approach for ML estimation results in computational efficiency as well as in a good statistical efficiency even with a small compression ratio q.

To compare the predictive performance of both methods we compute Root-Mean-Squared Prediction Error (RMSPE). The set of the predicted locations for each M is denoted as D_M^* with each new location $x_0 \in D_M^* \subset \mathbb{R}^d$. If $\tilde{Z}(x_0,l)$ denote the model-A predictor, where $Z(x_0,l)$ is the lth simulated process evaluated at a new location x_0 and A =TAP, HM, then the model-A predictor RMSPE for the lth simulation is $\text{RMSPE}_A(l) = \sqrt{\sum_{x_0 \in D_M^*} \left(\tilde{Z}(x_0,l) - Z(x_0,l)\right)^2}$, $l = 1,\ldots,L$. We then consider a measure of relative skill (RS), relative to HM, namely $\text{RS}(N) = \text{RMSPE}_{\text{HM}}(l)/\text{RMSPE}_{\text{TAP}}(l)$ for different $N = 2000, 4000, 8000$. As can be seen from the Figure 2 for different sample size N_k, density and correlation ratio p and q $RS(N) < 1$. Therefore, $\mathscr{H}$-matrices approach has a better predictive accuracy.

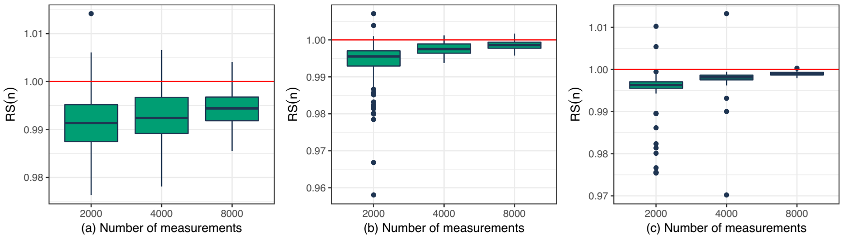


Fig. 2 Boxplots of RS(N) for (a) $N_1 = 2000 : p_1 = 0.2, q_1 = 0.3$, $N_2 = 4000 : p_1 = 0.15, q_1 = 0.25$, $N_3 = 8000 : p_1 = 0.1, q_1 = 0.2$, (b) $N_1 = 2000 : p_2 = 0.5, q_2 = 0.8$, $N_2 = 4000 : p_2 = 0.38, q_2 = 0.48$, $N_3 = 8000 : p_2 = 0.27, q_2 = 0.37$, (c) $N_1 = 2000 : p_3 = 1.5, q_3 = 1.9$, $N_2 = 4000 : p_3 = 1.33, q_3 = 1.56$, $N_3 = 8000 : p_3 = 1.12, q_3 = 1.31$

References

1. Gorshechnikova, A.: Likelihood Approximation and Prediction for Large Spatial and Spatio-temporal Datasets using H-matrix Approach. PhD-Thesis, University of Padua, Italy (2019)
2. Hackbusch, W.: Hierarchical matrices: algorithms and analysis. Vol. 49. Heidelberg: Springer (2015)
3. Heaton, M.J., Datta,A., Finley, A.O., Furrer, R., Guinness, J., Guhaniyogi, R., Zammit-Mangion, A. et al.: A case study competition among methods for analyzing large spatial data. Journal of Agricultural, Biological and Environmental Statistics, 24(3), 398–425 (2019).
4. Kaufman, C.G., Nychka, D.W., Schervish, M.J.: Covariance tapering for likelihood-based estimation in large spatial data sets. Journal of the American Statistical Association 103.484, 1545–1555. (2008)
5. Litvinenko, A., Genton, M.G., Keyes, D.E., Sun, Y.: Likelihood approximation with hierarchical matrices for large spatial datasets. Computational Statistics & Data Analysis 137, 115–132 (2019)
6. Furrer, R., Sain, S.R.: spam: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. Computational Statistics & Data Analysis 137, 36, 1–25. (2010)