FMR: Fast randomized algorithms for covariance matrix computations

Pierre Blanchard, Olivier Coulaud, Eric Darve, Alain Franc

To cite this version:

Pierre Blanchard, Olivier Coulaud, Eric Darve, Alain Franc. FMR: Fast randomized algorithms for covariance matrix computations. Platform for Advanced Scientific Computing (PASC), Jun 2016, Lausanne, Switzerland. 2016. hal-01334747

HAL Id: hal-01334747
https://hal.archives-ouvertes.fr/hal-01334747
Submitted on 23 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Fast randomized algorithms for
covariance matrix computations

Pierre Blanchard 1 · Olivier Coulaud 1 · Eric Darve 2 · Alain Franc 1

ABSTRACT

We present an open-source library implementing fast algorithms for covariance matrices computations, e.g., randomized low-rank approximations (LRA) and fast multipole matrix multiplication (FMM). This library can be used to approximate square roots of low-rank covariance matrices in SVD form using randomized LRA, instead of the standard O(N^3) cost. Low-rank covariance matrices given as kernels, e.g., Gaussian decay, evaluated on 3D grids can be decomposed in O(N) operations using the FMM. The performance of the library is illustrated on two examples:

- Generation of Gaussian Random Fields (GRF) on large spatial grids
- MultiDimensional Scaling (MDS) for the classification of species.

RANDOM PROJECTION BASED LRA

Randomized SVD is a random projection-based LRA algorithms made popular by Halko et al. [4], which returns an approximate SVD of a (symmetric) matrix $C \in \mathbb{R}^{N \times N}$ given a prespecified numerical rank $r$ in $O(N^2 \times r)$ operations.

- Form an approximate basis $Q \in \mathbb{R}^{N \times r}$ for the range of $C$.
- Form a sketched version of $C$ using a random projection, i.e., application of $Q$ to a $N \times r$-by-$r$ Gaussian random matrix $\Omega$.

$$Y = \Omega^T C \Omega$$

- Then, orthogonalize $Y$ by means of an QR Decomposition

$$\tilde{Q} \tilde{R} = Y$$

Thus, we get a low-rank representation of $C$ in the form $C \approx \tilde{Q} \tilde{S} \tilde{Q}^T$ with Frobenius/spectral error bounds that hold with high probability.

- Factorize $C_r$ in SVD form: $C_r = U S V^T$.
- We start by assembling the small $r$-by-$r$ matrix $B = \Omega^T C \Omega$.
- Then, perform a small SVD: $B = U \Sigma V^T$.
- Form $\tilde{U} = \tilde{Q} U$ and $\tilde{\Sigma} = \tilde{Q} \Sigma$.

If $C$ is positive semi-definite, then $C \approx C_r = AA^T$.

$\tilde{C}_r = 2 \times r \times r$ -- The method offers many advantages:

- Easily implemented and parallelized.
- Easily extended to Cholesky, Interpolative Decomposition.
- Cost dominated by matrix multiplication, i.e., $O(C \times N^2)$.

However, $C$ should fulfill the following conditions:

- be low-rank ($r << N$).
- have a fast decreasing spectrum $\{\lambda_i(C)\}$.

EFFICIENT GENERATION OF GRF

Aim: This project aims at promoting new highly efficient FMM algorithms to perform resource demanding computations in geostatistics.

Correlation kernels: A Gaussian Random Field $Y \sim \mathcal{N}(0, C)$ is a multivariate Gaussian random variable with mean 0 and covariance $C \in \mathbb{R}^{N \times N}$.

The covariance can be prescribed as a kernel matrix, i.e.,

$$C = \{k(x_i, x_j)\}_{x_i, x_j \in X}$$

where $k_{ij} = \|x_i - x_j\|^2$ denotes the distances between points of an arbitrary grid and $k$ is a correlation kernel such as $k_{ij} = \exp(-|x_i - x_j|^2)$ (Exponential decay) or $k_{ij} = \exp(-|x_i - x_j|^p)$ (Gaussian decay).

The length scale $\ell$ characterizes the decreasing speed of the correlation.

Square-root algorithms: Covariance matrices are split by definition of correlation kernels. Hence, $C$ admits the following representation:

$$C = A A^T$$

where the matrix factor $A \in \mathbb{R}^{N \times N}$ is often called a square root of $C$. Methods for generating Gaussian Random Fields usually differ by the way $A$ is prescribed:

- standard matrix decompositions ($O(N^3)$)
- fast parallel (in shared and distributed memory) algorithms on the unit sphere with $O(N \log N)$ for equispaced grids
- the turning bands method (approximate)

Most of them become computationally prohibitive for large $N$, i.e., $N$ over a few thousands.

Fast Multipole Acceleration of the matrix multiplications involved in the randomized SVD provides an algorithm for approximating $C$ in $O(N \times r)$ operations with many benefits:

- matrix-free method with a $O(r \times N)$ memory footprint.
- hierarchical methods handle heterogeneous grids more efficiently.

However, the extra error involved by the FMM has to be monitored.

The $k$-structure should apply well to $C_r$.

TAXONOMY VIA MULTIDIMENSIONAL SCALING (MDS)

Aim: This project aims at developing new strategies for the classification of species that benefit from the massive amount of data provided by New Generation Sequencing (NGS) techniques.

Metric: MDS aims at reconstructing a cloud of points $X$ in a low-dimensional feature space, e.g., $X \in \mathbb{R}^{N \times m}$, from a given distance/disimilarity matrix $D \in \mathbb{R}^{N \times N}$ (Smith-Webster scores of local alignment). The algorithm [2] consists in:

- Assembling a covariance/similarity matrix as $C_{ij} = \{x_i, x_j\} = \frac{1}{2} (D_{ij} - \bar{D}_i \bar{D}_j - \sum_k D_{ik} \frac{1}{n_k} \sum_{j \neq k} D_{jk})$.
- Computing the SVD of $C$, i.e., $C = U \Sigma U^T$
- Forming $X = U \Sigma_k^{1/2}$ (LS minimizer)

Perspectives: Develop automatic procedure for community inventories.

- Analyze clustering, concentration of reads, . . .
- Improve visualization tools and methods.

Enhance algorithm and numerical analysis.

- Compare with existing approaches based on random column selection.
- Improve storage and running time by partitioning data sets and compressing covariance matrices.

OTHER FEATURES OF THE LIBRARY

Sources: are available online as part of the open-source package FMR. They can be downloaded for free at the following address https://github.inria.fr/projects/fmr

Dependences: FMR relies on:

- MCL for dense linear algebra and FFT
- Scotch or ClusteringLib for partitioning.

Features:

- routines for generating Gaussian Random Fields based on standard LRA, Cholesky Decomposition, SVD or FFT for regular grids.
- randomized LRA: RandSVD and Ixystrox method with uniform or leverage-score based sampling.
- a variety of correlation kernels: Matern, Spherical model, Oseen-Gauss.
- a Python interface for MDS using Randomized SVD or Ixystrox.
- a Matlab interface for Ensemble Kalman Filtering.

REFERENCES


[5] FastLA. This work was partially supported by the associate team FastLA (Inria, Stanford University & Lawrence Berkeley National Laboratory).

This work was partially supported by the associate team FastLA (Inria, Stanford University & Lawrence Berkeley National Laboratory).