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). The library can be used to approximate square roots of low-rank covariance matrices in $O(N^3)$ operations from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$, from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$, from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$. They can be downloaded for free at the following address https://gforge.inria.fr/projects/fmr

Randomized approach: the randomization procedure is based on the approach used to approximate square roots of low-rank covariance matrices in $O(N^3)$ operations from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$, from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$, from a given distance/dissimilarity matrix $D \in \mathbb{R}_{++}^{N \times N}$. The library is available in different languages, including C, C++, Python, R and Matlab interfaces. It supports parallel and distributed computation on heterogeneous grids. It is designed to be easily extended to Cholesky, Interpolative Decomposition, Randomized SVD, etc.

Randomized Procedure Based LRA

Randomized procedures for covariance matrices computations, e.g., randomized low-rank approximations (LRA) and Fast Multipole Matrix Multiplication (FMM). The library can be used to approximate square roots of low-rank covariance matrices in $O(N^3)$ operations 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 $O(N^r)$ 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.

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 prescribed numerical rank $r$ in $O(N^3 \times r)$ operations:

- Form an approximate basis $Q \in \mathbb{R}^{N \times r}$ for the range of $C$.
- Form a sketched version of $C$ using Gaussian random projection, i.e., a matrix $C_s$.
- Apply $C_s$ to a $N \times r$ Gaussian random matrix $\Omega$.
- $Y \equiv C \Omega$.
- Then, orthogonalize $Y$ by means of a QR Decomposition.

Thus, we get a low-rank approximation of $C$ in the form of $C_\approx = QQ^T C$ with Frobenius/spectral error bounds that hold with high probability.

- Factorize $C_s$ in SVD form:
  $C_s = U \Sigma V^T$.
- We start by assembling the small $r \times r$ matrix $B = Q^T C Q$.
- Then, perform a small SVD, i.e., $B = U \Sigma V^T$.
- Form $U \Sigma Q$ and $\Sigma = U \Sigma$.
- If $C$ is positive semi-definite, then $C_\approx = U \Sigma U^T$.

The method offers many advantages:
- Easily implemented and parallelized,
- Easily extended to Cholesky, Interpolative Decomposition...
- Cost dominated by matrix multiplication, i.e., $O(C N^3)$. However, $C$ should fulfill the following conditions;
- be low-rank ($r \ll N$),
- have a fast decreasing spectrum ($\sigma(C) \approx [\sigma(C)^{1/2} / C] \ll N$).

Efficient Generation of GRF

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

Correlation kernels A Gaussian Random Field $Y \sim p(0,C)$ is a multi-variate 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., $C = (k(x_i,x_j))_{i,j=1}^N$.

where $r \equiv ||x_i - x_j||_2$ denotes the distances between points of an arbitrary grid and $k$ is a kernel function such as:

$$k_{\text{ExpDec}}(r) = e^{-r/\ell} \quad \text{(Exponential decay)}$$

$$k_{\text{GaussDec}}(r) = e^{-r^2/\sigma^2} \quad \text{(Gaussian decay)}$$

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 ($D \in \mathbb{H}^N$),
- random uniform embedding ($D \in \mathbb{X}^N$) for equispaced grids,
- the turning bands method (approximate),
- Most of them become computationally prohibitive for large $N$, i.e., $N$ over a few thousands.

TAXONY VIA MULTI-DIMENSIONAL SCALING (MDS)

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.

Metro: MDS aims at reconstructing a cloud of points $X$ in a low-dimensional feature space, e.g., $X \in \mathbb{R}^{N \times 3}$, from a given distance/dissimilarity matrix $D \in \mathbb{R}^{N \times N}$ (Smith-Watson scores of local alignment). The algorithm [2] consists in:

- Assembling a covariance/similarity matrix as $C_{ij} = \hat{d}(x_i,x_j) = \frac{1}{2} (D_{ij} - \frac{1}{N} \sum_{k=1}^N D_{ik} - \frac{1}{N} \sum_{k=1}^N D_{kj} + \frac{1}{N^2} \sum_{k=1}^N D_{kk})$.
- Computing the SVD of $C$, i.e, $C = U \Sigma U^T$.
- Forming $X = C^{1/2} = U \Sigma^{1/2} U^T$ (LS minimizer).

Perspectives

- Develop automatic procedures 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.

FUNDINGS

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

REFERENCES


FIGURE 3: MDS using Gaussian Random Projection on 60K reads from a data set of maize ITS for green plants. Mark families (colors) can clearly be distinguished. However, most points contain more than one species and it is necessary to render further axes to have a complete view of the whole data set.