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.science/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 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 O(N^2 × r) operations, given a prespecified numerical rank r in O(N^2 × r) operations.

- Form an approximate basis Q ∈ R^N×r for the range of C.
- Form a projected version of C using Gaussian randomization, i.e., application of C to a N-by-r Gaussian random matrix Ω.

Y = Ω^T C Ω

- Then, orthonormalize Y by means of a QR Decomposition
QR = Y

- Thus, we get a low-rank representation of C in the form
C ≈ C_r = Σ_r Q_r Σ_r^T

with Frobenius/spectral error bounds that hold with high probability.

- Factorize C in SVD form: C = UΣV^T.
  - We start by assembling the small r-by-r matrix
  B = Q^T C Q.
  - Then, perform a small SVD: B = UΣ V^T.
  - Form U = UΣ and Σ = Σ_r.

If C is positive semi-definite, then C ≈ C_r = Σ_r Σ_r^T

A = Σ_r^T U_r Σ_r^T U_r

The method offers several advantages:

- Easily implemented and parallelized,
- easily extended to Cholesky, Iterative Decomposition...,
- cost dominated by matrix multiplication, i.e. O(N^2).

However, C_r should fulfill the following conditions:

- be low-rank (r << N),
- have a fast decreasing spectrum (ab(0) = ||C||_F/||C||_2 << N).

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 ∼ N(0, C) is a multivariate Gaussian random variable with mean 0 and covariance C ∈ R^N×N. The covariance can be prescribed as a kernel matrix, i.e.,

C = (k(x_i,x_j))_{i,j=1}^N

where \( k(x_i,x_j) \) denotes the distances between points of an arbitrary grid and \( k \) is a correlation kernel such as

\[ k_{Exp}(r) = e^{-r^2} \quad \text{(Exponential decay)} \]

\[ k_{Gauss}(r) = e^{-r^2/2} \quad \text{(Gaussian decay)} \]

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

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

\[ C = A A^T \]

where the matrix factor A ∈ R^N×r is often called a square root of C. Methods for generating Gaussian Random Fields usually differ by the way A is precomputed:

- standard matrix decompositions (Cholesky),
- circulant embedding (Covariances matrices for equispaced grids),
- turning bands method (approximate).

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

TAXONOMY VIA MULTIDIMENSIONAL SCALING (MDS)

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

Randomized approach: Dehban and Deutsch [3] used the FastSVD in order to precompute A in low-rank form in O(N^2 × r) operations and thus efficiently generate realizations of Gaussian Random Fields at a O(N × r) cost. This approach still requires C_r to be fully assembled.

Fast Multipole Acceleration of the matrix multiplications involved in the randomized SVD provides an algorithm for approximating A in O(N^2 × r) with much benefits:

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

However, the extra error involved by the FMM has to be monitored. The H^2-structure should apply well to C_r.

References:


FUNDINGS

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