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 (FMF). 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 FMF. 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 algorithm made popular by Halko et al. [6], which returns an approximate SVD of a (symmetric) matrix \(C \in \mathbb{R}^{N \times N}\) given a prescribed numerical rank \(r \in O(N^2 \times r)\) operations:

- Form an approximate basis \(B \in \mathbb{R}^{N \times N}\) for the range of \(C\).
- Form a sketched version of \(C\) using Gaussian random projection, i.e., application of \(C\) to a \(N \times r\)-by-
  \(N\)-Gaussian random matrix \(W\).
- Then, orthogonalize \(Y\) by means of a QR Decomposition
  \(B = Q \cdots R\).
- Thus, we get a low-rank representation of \(C\) in the form
  \[C \approx QQ^T\]

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

- Factorize \(C\) in SVD form: \(C = U \Sigma U^T\).
- We start by assembling the small \(r\)-by-
  \(r\) matrix
  \[B = \Sigma^{1/2} W\]

- Then, perform a small SVD: \(B = U \Sigma U^T\).
- Form \(U \Sigma U^T\) and \(\Sigma = B\).
- If \(C\) is positive semi-definite, then \(C \approx C = AA^T\)

where \(\Sigma\) should fulfill the following conditions:

- be low-rank (\(r \ll N\))
- have a fast decreasing spectrum \(\langle \sigma/C \rangle = \langle C \rangle / \langle C \rangle \ll N\).

However, \(C\) should fulfill the following conditions:

- be low-rank (\(r \ll N\))
- have a fast decreasing spectrum \(\langle \sigma/C \rangle = \langle C \rangle / \langle C \rangle \ll N\).

EFFICIENT GENERATION OF GRF

Aim: This project aims at proposing high efficient FMF algorithms to perform resource demanding computations in geostatistics.

Correlation kernels: A Gaussian Random Field \(Y \sim \nu(\mu, C)\) is a multi-variate Gaussian random variable with mean \(\mu\) and covariance \(C \in \mathbb{R}^{N \times N}\). The covariance can be prescribed as a kernel matrix, i.e.,

\[C \approx \frac{4G(x,y), x \in \mathcal{X}, y \in \mathcal{X}}{4}\]

where \(G_{ij} = \|x_i - x_j\|\) denote the distances between points of an arbitrary grid and \(\mathcal{K}\) is a kernel function such as:

\[\mathcal{K}(x, y) = e^{-||x-y||^2}\] (Exponential decay)
\[\mathcal{K}(x, y) = e^{-||x-y||^2/\sigma^2}\] (Gaussian decay)

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

Square-root algorithm: Covariance matrices are split by definition of correlation kernels. Hence, \(C\) admits the following representation:

\[C = A \cdot A^T\]

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 \(\text{Cholesky}\)
- circulant embedding \(\mathcal{C}(N \times N)\) for equispaced grids
- the turning bands method (approximate)

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

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

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

- matrix-free method with \(A(\times N)\) memory footprint
- hierarchical methods handle heterogeneous grids more efficiently
- the extra error involved by the FMF has to be monitored
- \(A\)-structure should apply well to \(C\).

TAXONY 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., \(X \in \mathbb{R}^{N \times N}\), from a given distance/dissimilarity matrix \(D \in \mathbb{R}^{N \times N}\) (Smith-Watson scores of local alignment). The algorithm [2] consists in:

- Constructing a covariance/similarity matrix as:

\[C_{ij} = \langle x_i, x_j \rangle = \frac{1}{N} \sum_1^N x_{ij}\]

- Computing the SVD of \(C\), i.e.,

\[C = U \Sigma U^T\]

- Forming \(X = C^{1/2} = U \Sigma^{1/2}\) (LS minimizer)

References:


FUNDINGS

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

OTHER FEATURES OF THE LIBRARY

The library is available online as part of the open-source package FMR. It can be downloaded for free at the following address https://fgrs.inria.fr/projects/fmr

Dependancies: FMF relies on
- ScalFMM [5] for performing fast multipole matrix multiplication in parallel (in shared and distributed memory)
- Mkl 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 Iyxri methods with uniform or leverage score-based sampling
- a variety of correlation kernels: Matérn, Spherical model, Oseen-Gauss
- a Python interface for MDS using Randomized SVD or Iyxri
- a Matlab interface for Ensemble Kalman Filtering

Sources: