FMR: Fast randomized algorithms for covariance matrix computations

Pierre Blanchard, Olivier Coulaud, Eric Darve, Alain Franc

FMR

Fast randomized algorithms for covariance matrix computations

Pierre Blanchard [1] - Olivier Coulaud [1] - Eric Darve [2] - Alain Franc [3]

[1]: HiePACS, Inria Bordeaux, 200, rue Vieille Tour 33405 Talence, France
name.surname@inria.fr

[2]: Mechanical Engineering, Stanford University, CA, USA
darve@stanford.edu

[3]: BioGeCo, Inra & PLEIADE, Inria Bordeaux, France
alain.franc@inria.fr

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 $\mathcal{O}(N^2)$ operations in SVD form using randomized LRA, instead of the standard $\mathcal{O}(N^3)$ cost.

Low-rank covariance matrices given as kernels, *e.g.*, Gaussian decay, evaluated on 3D grids can be decomposed in $\mathcal{O}(N)$ 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 $\mathbf{C} \in \mathbb{R}^{N\times N}$ given a prescribed numerical rank r in $\mathcal{O}(N^2 \times r)$ operations:

- **Form an approximate basis** $\mathbf{Q} \in \mathbb{R}^{N\times r}$ **for the range of** $\mathbf{C}$.
  - Form a sketched version of $\mathbf{C}$ using Gaussian random projection, *i.e.*, application of $\mathbf{C}$ to a N-by-r Gaussian random matrix $\mathbf{\Omega}$.
  $$\mathbf{Y} = \mathbf{C}\mathbf{\Omega}$$
  - Then, orthogonalize $\mathbf{Y}$ by mean of a QR Decomposition
  $$\mathbf{QR} = \mathbf{Y}$$
  - Thus, we get a low-rank representation of $\mathbf{C}$ in the form
  $$\mathbf{C} \approx \mathbf{C_r} = \mathbf{QQ^TCQQ^T}$$
  with Frobenius/spectral error bounds that hold with high probability.
- **Factorize** $\mathbf{C_r}$ **in SVD form:** $\mathbf{C_r} = \mathbf{U\Sigma U^T}$.
  - We start by assembling the small r-by-r matrix
  $$\mathbf{B} = \mathbf{Q^TCQ}$$
  - Then, perform a small SVD, *i.e.*, $\mathbf{B} = \mathbf{U_B\Sigma_B U_B^T}$.
  - Form $\mathbf{U} = \mathbf{QU_B}$ and $\mathbf{\Sigma} = \mathbf{\Sigma_B}$
- If $\mathbf{C}$ is positive semi-definite, then $\mathbf{C} \approx \mathbf{C_r} = \mathbf{AA^T}$
$$\mathbf{A} = \mathbf{QB^{1/2}} = \mathbf{QU_B\Sigma^{1/2}}$$

+/− The method offers many advantages:

- Easily implemented and parallelized,
- easily extended to Cholesky, Interpolative Decomposition ...
- cost dominated by matrix multiplication, *i.e.* $\mathcal{O}(r \times N^2)$.

However, C should fulfill the following conditions:

- be low-rank ($r << N$),
- have a fast decreasing spectrum ($st(C) = \|C\|_F^2/\|C\|_S^2 << N$),

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 $\mathbf{Y} \sim \mu(\mathbf{0}, \mathbf{C})$ is a multivariate Gaussian random variable with mean $\mathbf{0}$ and covariance $\mathbf{C} \in \mathbb{R}^{N\times N}$. The covariance can be prescribed as a kernel matrix, *i.e.*,

$$\mathbf{C} = \{k(r_{ij})\}_{i,j=1...N}$$

where $r_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ denotes the distances between points of an arbitrary grid and k is a correlation kernel such as

$$k_{1/2}(r) = e^{-r/\ell} \quad \text{(Exponential decay)}$$
$$k_{\infty}(r) = e^{-r^2/(2\ell^2)} \quad \text{(Gaussian decay)}$$

The length scale ℓ characterizes the decreasing speed of the correlation.

Square-root algorithms Covariance matrices are *spd* by definition of correlation kernels. Hence, $\mathbf{C}$ admits the following representation

$$\mathbf{C} = \mathbf{AA^T}$$

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

- standard matrix decompositions ($\mathcal{O}(N^3)$)
- circulant embedding ($\mathcal{O}(N \log N)$ for equispaced grids)
- the turning bands method (approximate)

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

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

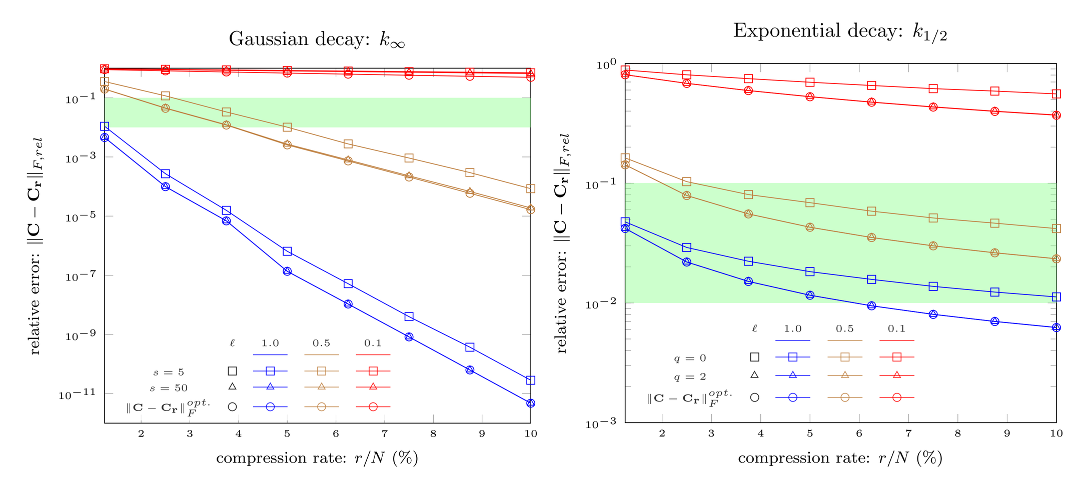

FIGURE 1: Accuracy of the fixed rank Randomized SVD *w.r.t.* the compression rate for $N = 2000$ points on the unit sphere and $r = 20\ldots200$. Left: The effects of q power iterations using a fixed oversampling of $s = 50$ on the exponential kernel. Right: The effects of the oversampling s alone ($q = 0$).

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

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

However,

- the **extra error** involved by the FMM has to be monitored.
- $\mathcal{H}^2$-structure should apply well to C.

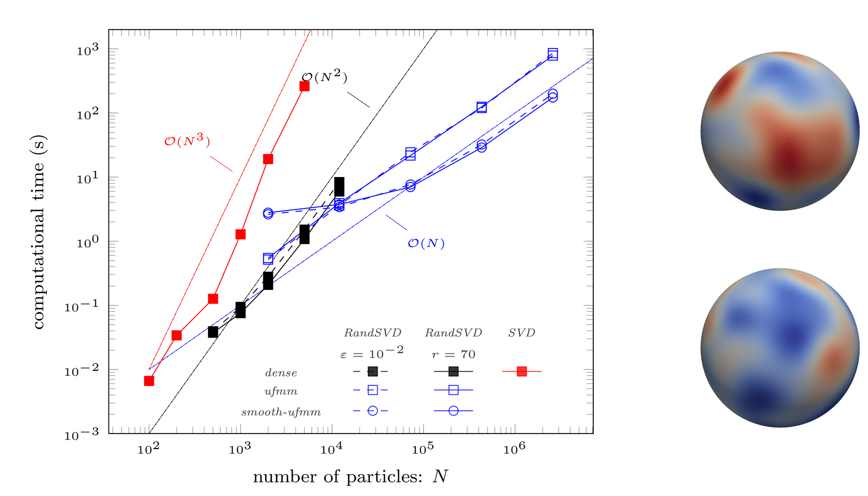

FIGURE 2: Left: Running time for computing a randomized SVD using either the fixed rank or fixed accuracy algorithm with $q = 0$ and $s = 10$. Matrices are multiplied either in a dense way (filled black squares) or by means of a fmm scheme, particles are randomly distributed on the unit sphere and the correlation is Gaussian (k_∞) with $\ell = 0.5$. Right: Realizations on the unit sphere with $72k$ points, k_∞ and $\ell = 0.25$ using FMM accelerated SVD (top) and smooth variant (bottom).

Perspectives Extension of the algorithm to other geoscientific application, such as **Data Assimilation**:

- Joint work with Stanford University: Amalia Kokkinaki, Peter Kitanidis (Civil Engineering) and Eric Darve.
- Efficient **Ensemble Kalman Filters (EnKF)**
  - Perturbation of measurements
  - Square Root EnKFs based on kernel methods
- Develop **Python, R and Matlab interfaces**

TAXONOMY VIA MULTIDIMENSIONAL SCALING (MDS)

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

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

- Assembling a covariance/similarity matrix as
$$C_{ij} = \langle x_i, x_j \rangle = -\frac{1}{2}\left(D_{ij}^2 - \frac{1}{n}\sum_i D_{ij}^2 - \frac{1}{n}\sum_j D_{ij}^2 + \frac{1}{n^2}\sum_{i,j} D_{ij}^2\right)$$
- Computing the SVD of C, *i.e.*, $\mathbf{C} = \mathbf{U\Sigma U^T}$
- Forming $\mathbf{X} = \mathbf{C^{1/2}} = U\Sigma^{1/2}$ (LS minimizer)

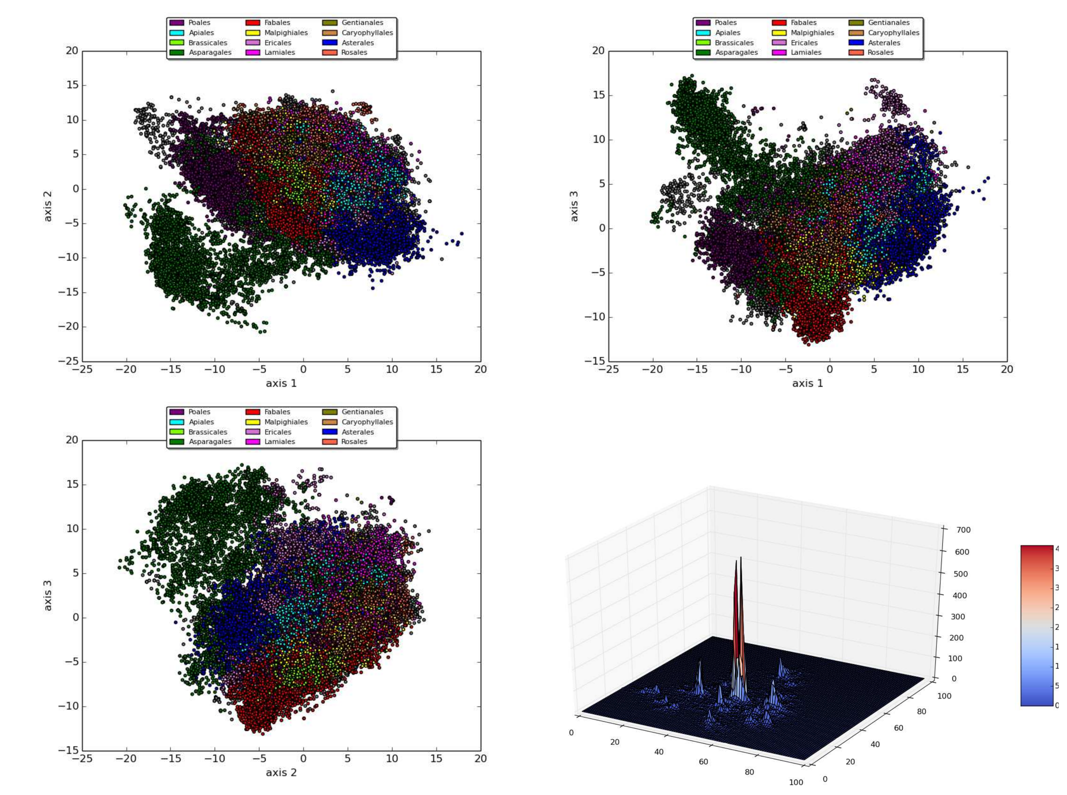

FIGURE 3: MDS using Gaussian Random Projection on 65k reads from a data set of marker ITS for green plants. Main families (colors) can clearly be distinguished. However, most pixels contain more than one specimen, and it is necessary to explore further axes to have a complete view of the whole data set. Bottom Right: Another 20k reads from a diatom commnity in Geneva lake, with genetic distances. Here, the z axis represents the number of reads which are stacked on the same pixel. Each spike represents a species, with high concentration of reads, the variations being either natural variability or sequencing errors. This permits an automatic procedure for community inventories.

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 **partitionning** 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://gforge.inria.fr/projects/fmr

Dependencies FMR relies on

- **ScalFMM** [1] for performing fast multipole matrix multiplication in parallel (in shared and distributed memory)
- **MKL** for dense linear algebra and FFT
- **Scotch** or **CClusteringLib** for partitionning

Features The package provides:

- routines for generating Gaussian Random Fields based on
  - standard LRA: Cholesky Decomposition, SVD or FFT for regular grids.
  - randomized LRA: `RandSVD` and `Nyström` method 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 `Nyström`
- a Matlab interface for Ensemble Kalman Filtering

REFERENCES

[1] Pierre Blanchard, Bérenger Bramas, Olivier Coulaud, Eric Darve, Laurent Dupuy, Arnaud Etcheverry, and Guillaume Sylvand. Scalfmm: A generic parallel fast multipole library. In *Computational Science and Engineering (CSE)*. SIAM, March 2015.

[2] T.F. Cox and M. A. A. Cox. *Multidimensional Scaling - Second edition*, volume 88 of *Monographs on Statistics and Applied Probability*. Chapman & al., 2001.

[3] Vahid Dehdari and Clayton V. Deutsch. Applications of randomized methods for decomposing and simulating from large covariance matrices. In *Geostatistics Oslo 2012*, volume 17 of *Quantitative Geology and Geostatistics*, pages 15–26. Springer Netherlands, 2012.

[4] Nathan Halko, Per-Gunnar Martinsson, and Joel A Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

FUNDINGS

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

① ② ③