Journées SMAI MODE 2018

Stochastic Subsampling for Factorizing Huge Matrices

Arthur Mensch
Inria Parietal

Julien Mairal
Inria Thoth

Bertrand Thirion
Inria Parietal

Gaël Varoquaux
Inria Parietal

Abstract. We present a matrix-factorization algorithm that scales to input matrices with both huge number of rows and columns. Learned factors may be sparse or dense and/or non-negative, which makes our algorithm suitable for dictionary learning, sparse component analysis, and non-negative matrix factorization. Our algorithm streams matrix columns while subsampling them to iteratively learn the matrix factors. At each iteration, the row dimension of a new sample is reduced by subsampling, resulting in lower time complexity compared to a simple streaming algorithm. Our method comes with convergence guarantees to reach a stationary point of the matrix-factorization problem. We demonstrate its efficiency on massive functional Magnetic Resonance Imaging data (2 TB), and on patches extracted from hyperspectral images (103 GB). For both problems, which involve different penalties on rows and columns, we obtain significant speed-ups compared to state-of-the-art algorithms.

Keywords : Stochastic optimization – matrix factorization – random methods.

1 Context and goals

Matrix factorization is a flexible approach to uncover latent factors in low-rank or sparse models. With sparse factors, it is used in dictionary learning, and has proven very effective for denoising and visual feature encoding in signal and computer vision [see *e.g.,* 4]. When the data admit a low-rank structure, matrix factorization has proven very powerful for various tasks such as matrix completion [10, 1], word embedding [9, 2], or network models [13]. Matrix factorization techniques can be tackled with stochastic optimization: matrix decompositions are learned by observing a single matrix column (or row) at each iteration. Those techniques have been successful in handling matrices with a large number of rows but a reasonable number of columns, *e.g.,* in computer vision [5]. However, stochastic algorithms for matrix factorization were unable to deal efficiently with matrices that are large in both dimensions. In two successive works [7, 6], we presented a matrix-factorization algorithm that scales to input matrices with both huge number of rows and columns. Learned factors may be sparse

or dense and/or non-negative, which makes our algorithm suitable for dictionary learning, sparse component analysis, and non-negative matrix factorization. Our algorithm, called *subsampled online matrix factorization* (SOMF) is faster than state-of-the-art algorithms by an order of magnitude on large real-world datasets (hyperspectral images, large fMRI data). It leverages random sampling with stochastic optimization to learn sparse and dense factors more efficiently. More precisely, it streams matrix columns while subsampling them to iteratively learn the matrix factors. At each iteration, the row dimension of a new sample is reduced by subsampling, resulting in lower time complexity compared to a simple streaming algorithm.

2 Problem setting and algorithm

In our setting, the goal of matrix factorization is to decompose a matrix $\mathbf{X} \in \mathbb{R}^{p \times n}$ — typically n signals of dimension p — as a product of two smaller matrices:

$$\mathbf{X} \approx \mathbf{D}\mathbf{A} \quad \text{with} \quad \mathbf{D} \in \mathbb{R}^{p \times k} \text{ and } \mathbf{A} \in \mathbb{R}^{k \times n},$$

with potential sparsity or structure requirements on $\mathbf{D}$ and $\mathbf{A}$. In signal processing, sparsity is often enforced on the code $\mathbf{A}$, in a problem called *dictionary learning* [8]. In such a case, the matrix $\mathbf{D}$ is called the "dictionary" and $\mathbf{A}$ the sparse code. Learning the factorization is performed by minimizing a quadratic data-fitting term, with constraints and/or penalties over the code and the dictionary:

$$\min_{\mathbf{D} \in \mathcal{C}} \Big(\bar{f}(\mathbf{D}) \triangleq \min_{\mathbf{A} \in \mathbb{R}^{k \times n}} \quad \sum_{i=1}^{n} \frac{1}{2} \big\| \mathbf{x}^{(i)} - \mathbf{D}\boldsymbol{\alpha}^{(i)} \big\|_2^2 + \lambda\, \Omega(\boldsymbol{\alpha}^{(i)}) \Big), \tag{1}$$

where $\mathbf{A} \triangleq [\boldsymbol{\alpha}^{(1)}, \dots, \boldsymbol{\alpha}^{(n)}]$, $\mathcal{C}$ is a column-wise separable convex set of $\mathbb{R}^{p \times k}$ and $\Omega : \mathbb{R}^p \to \mathbb{R}$ is a penalty over the code. In our work, Ω is the elastic-net penalty [14] and $\mathcal{C}$ enforces that each columns of $\mathbf{D}$ lies in the elastic-net ball: we may thus enforce sparsity on either $\mathbf{A}$ or $\mathbf{D}$; we may optionally add positivity constraints on any of these terms, to perform non-negative matrix factorization.
Equation (1) can be solved using alternated minimization on $\mathbf{D}$ and $\mathbf{A}$, which is guaranteed to converge toward a critical point of the objective. However this method is not scalable to problem with large number of samples n or large number of features p, as each iteration requires to go through all data $\mathbf{X}$. As our use-cases requires to factorize terabyte matrices with $n \sim 10^6$ and $p \sim 10^5$, we design the SOMF algorithm, which

- handles numerous data (large n) by updating $\mathbf{D}$ as we stream the columns of matrix $\mathbf{X}$, following the principles of *online matrix factorization* [5];
- handles high dimensional data (large p) by randomly reducing the dimension of the stream $(\mathbf{x}_t)_t$ using subsampling masks $(\mathbf{M}_t)_t$. This introduces noise in parameter updates but improves single-iteration complexity, resulting in large speed-ups compared to vanilla online matrix factorization.

With some simplification, at each iteration, our algorithm draws a column $\mathbf{x}_t$ from $\mathbf{X}$ and a random diagonal masking matrix $\mathbf{M}_t$, that selects a subspace $\mathbf{P}_t(\mathbb{R}^p)$. An approximate code $\boldsymbol{\alpha}_t$ is computed from $\mathbf{M}_t\mathbf{x}_t$ and $\mathbf{M}_t\mathbf{D}$:

$$\boldsymbol{\alpha}_t = \min_{\boldsymbol{\alpha} \in \mathbb{R}^k} \frac{1}{2} \|\mathbf{M}_t(\mathbf{x}_t - \mathbf{D}_{t-1}^\top \boldsymbol{\alpha})\|_2^2 + \lambda \Omega(\boldsymbol{\alpha}). \tag{2}$$

We use this code $\boldsymbol{\alpha}_t$ to update a quadratic surrogate function $\bar{g}_t$ of dictionary $\mathbf{D}$, which is then minimized over the set $\mathcal{C} \cap \mathbf{P}_t(\mathbb{R}^p)$. Defining $\bar{\mathbf{C}}_t = \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{\alpha}_s \boldsymbol{\alpha}_s^\top$ and $\bar{\mathbf{B}}_t = \frac{1}{t}\sum_{s=1}^{t} \boldsymbol{\alpha}_s \mathbf{x}_s^\top$,

$$\mathbf{D}_t = \underset{\substack{\mathbf{D}\in\mathcal{C} \\ \mathbf{P}_t^\perp \mathbf{D} = \mathbf{P}_t^\perp \mathbf{D}_{t-1}}}{\operatorname{argmin}} \frac{1}{2}\operatorname{Tr}\left(\mathbf{D}^\top \mathbf{D} \bar{\mathbf{C}}_t\right) - \operatorname{Tr}\left(\mathbf{D}^\top \bar{\mathbf{B}}_t\right), \tag{3}$$

before moving to a new samples $\mathbf{x}_{t+1}$ and a *new* mask $\mathbf{M}_{t+1}$ at iteration

3 Convergence analysis and empirical results

Extending the framework of stochastic majorization-minimization [3] to handle approximations in the majorization and minimization phase of the algorithm, we can show the following asymptotic convergence guarantee, for a slightly modified version of (2) and under some non-restrive assumptions — we refer the reader to the published papers for more details.

*Proposition 1 (*SOMF *convergence).* $\bar{f}(\mathbf{D}_t)$ converges with probability one and every limit point $\mathbf{D}_\infty$ of $(\mathbf{D}_t)_t$ is a stationary point of $\bar{f}$: for all $\mathbf{D} \in \mathcal{C}$

$$\nabla \bar{f}(\mathbf{D}_\infty, \mathbf{D} - \mathbf{D}_\infty) \geq 0. \tag{4}$$

We measured the performance of SOMF on the largest available resting-state functional MRI dataset HCP [11] — brain activation images collected over time, from which to extract representative sparse brain components — and on sets of hyperspectral image patches from the AVIRIS project [12]. We observed speed-ups of an order of magnitude, with subsampling ratio (average rank of $\mathbf{M}_t$) that can be increased up to $r = 12$, as illustrated in Figure 1. Qualitatively, We outline the dictionary atoms connex components obtained from fMRI decomposition in Figure 2 using SOMF and OMF: qualitatively, we obtain well defined dictionaries $10\times$ faster.

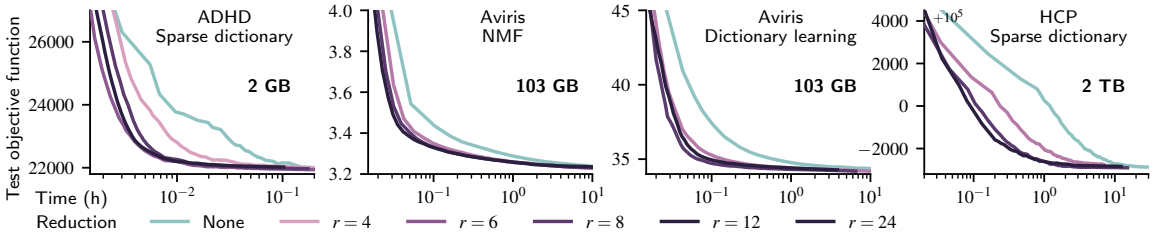

Figure 1: Subsampling provides significant speed-ups on all fMRI and hyperspectral datasets.

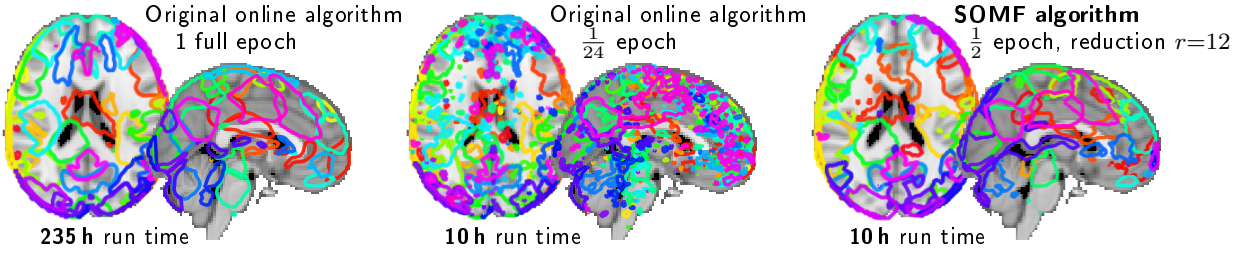

Figure 2: Outlines of each columns of $\mathbf{D}$. The dictionary converges much faster using SOMF.

References

[1] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[2] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, pages 2177–2185, 2014.

[3] Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *Adv. Neural Inform. Process. Syst.*, pages 2283–2291, 2013.

[4] Julien Mairal. Sparse Modeling for Image and Vision Processing. *Foundations and Trends in Computer Graphics and Vision*, 8(2-3):85–283, 2014.

[5] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *J. Machine Learning Research*, 11:19–60, 2010.

[6] A. Mensch, J. Mairal, B. Thirion, and G. Varoquaux. Stochastic subsampling for factorizing huge matrices. *IEEE Transactions on Signal Processing*, 66(1):113, 128, 2018.

[7] Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Dictionary learning for massive matrix factorization. In *Proc. ICML*, pages 1737–1746, 2016.

[8] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Res.*, 37(23):3311–3325, 1997.

[9] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global Vectors for Word Representation. In *Proc. Conf. EMNLP*, volume 14, pages 1532–43, 2014.

[10] Nathan Srebro, Jason Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems*, pages 1329–1336, 2004.

[11] David C. Van Essen et al. The WU-Minn Human Connectome Project: An overview. *NeuroImage*, 80:62–79, 2013.

[12] Gregg Vane. First results from the airborne visible/infrared imaging spectrometer (AVIRIS). In *Ann. Tech. Symp. Int. Soc. Optics Photonics*, pages 166–175, 1987.

[13] Yin Zhang, Matthew Roughan, Walter Willinger, and Lili Qiu. Spatio-Temporal Compressive Sensing and Internet Traffic Matrices. 2009.

[14] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Series B Stat. Methodol.*, 67(2):301–320, 2005.