An efficient algorithm for non-negative matrix factorization

Mai-Quyen Pham and Thierry Chonavel

I. Introduction

II. Problem formulation

In this framework we are interested in the minimization of the following quadratic cost function

$$F : \mathbb{R}_+^{m\times p} \times \mathbb{R}_+^{p\times n} \to \mathbb{R}_+ : (\mathbf{X}, \mathbf{H}) \mapsto \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\mathbf{H}\|^2 \tag{1}$$

where $\|\cdot\|$ denotes the Frobenius norm and $\mathbf{Y} \in \mathbb{R}_+^{m\times n}$ is a given non-negative matrix. Solving this problem allows us to reduce the rank of the given matrix $\mathbf{Y}$, for this reason p is often chosen to be smaller than n and m [2], [3]. Several applications ... []

To end this section, let us introduce the following notations: $\odot$ and $\oslash$ will denote the Hadamard (elementwise) product and division of two matrices. $\mathrm{diag}(\mathbf{v})$ will represent the diagonal matrix with diagonal entries defined from vector $\mathbf{v}$ and $\mathrm{vec}(\mathbf{M})$ the vector formed by the successive columns of matrix $\mathbf{M}$.

III. Proposed method

We propose to use the Majorize-Minimize (MM) subspace algorithm [1] which is based on the idea of constructing at the current iteration a convex quadratic majorizing approximation function (also called an auxiliary function [3]) of the cost function.

Definition 1. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function and $\mathbf{x} \in \mathbb{R}^n$. Let us define, for every $\mathbf{x}' \in \mathbb{R}^n$,*

$$g(\mathbf{x}, \mathbf{x}') = f(\mathbf{x}) + (\mathbf{x} - \mathbf{x}')^\top \nabla f(\mathbf{x}) + \frac{1}{2}\|\mathbf{x} - \mathbf{x}'\|^2_{\mathbf{A}_\mathbf{x}} \tag{2}$$

Mai-Quyen Pham and Thierry Chonavel are with IMT Atlantique; UMR CNRS 6285 Lab-STICC (e-mail: firstname.lastname@imt-atlantique.fr).

$$
\begin{aligned}
G_{\mathbf{H}}(\mathbf{X},\mathbf{X}') &= F(\mathbf{X}',\mathbf{H}) + \mathbb{1}_m^\top \left((\mathbf{X}-\mathbf{X}) \odot \nabla_{\mathbf{X}} F(\mathbf{X}',\mathbf{H})\right) \mathbb{1}_p + \tfrac{1}{2}\|\text{vec}\,(\mathbf{X}-\mathbf{X}')\,\|^2_{\mathbf{A}_{\mathbf{H}}(\mathbf{X}')} \\
G_{\mathbf{X}}(\mathbf{H},\mathbf{H}') &= F(\mathbf{X},\mathbf{H}') + \mathbb{1}_p^\top \left((\mathbf{H}-\mathbf{H}') \odot \nabla_{\mathbf{X}} F(\mathbf{X},\mathbf{H}')\right) \mathbb{1}_n + \tfrac{1}{2}\|\text{vec}\,(\mathbf{H}-\mathbf{H}')\,\|^2_{\mathbf{A}_{\mathbf{X}}(\mathbf{H}')}
\end{aligned}
\tag{3}
$$

where $\mathbf{A}_\mathbf{x} \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix and $\|\cdot\|^2_{\mathbf{A}_\mathbf{x}}$ denotes the weighted Euclidean norm induced by matrix $\mathbf{A}_\mathbf{x}$, that is, $\forall \mathbf{z} \in \mathbb{R}^n,\ \|\mathbf{z}\|^2_{\mathbf{A}_\mathbf{x}} = \mathbf{z}^\top \mathbf{A}_\mathbf{x}\mathbf{z}$. Then, $\mathbf{A}_\mathbf{x}$ satisfies the majoration condition for f at $\mathbf{x}$ if $g(\mathbf{x},\cdot)$ is a quadratic marojant function of f at $\mathbf{x}$, that is, for every $\mathbf{x} \in \mathbb{R}^n,\ f(\mathbf{x}) \leq g(\mathbf{x},\mathbf{x}')$.

Based on this definition, we introduce the following functions: where $\nabla_{\mathbf{X}} F(\mathbf{X}',\mathbf{H}) \in \mathbb{R}^{n\times p}$ and $\nabla_{\mathbf{H}} F(\mathbf{X},\mathbf{H}') \in \mathbb{R}^{p\times m}$ denote the partial gradients of F with respect to the variables $\mathbf{X}$ and $\mathbf{H}$, computed at $(\mathbf{X}',\mathbf{H})$ and $(\mathbf{X},\mathbf{H}')$ respectively.

The positive matrices $\mathbf{A}_{\mathbf{H}}(\mathbf{X}')$ and $\mathbf{A}_{\mathbf{X}}(\mathbf{H}')$ are chosen in such way that $G_{\mathbf{H}}(\mathbf{X},\mathbf{X}')$ and $G_{\mathbf{X}}(\mathbf{H},\mathbf{H}')$ represent majorizing functions of $F(\mathbf{X},\mathbf{H})$ for any $\mathbf{X}'$ and $\mathbf{H}'$ respectively.

Then, as in [2], [3], we consider the following iterative alternanting procedure to minimize F:

- initialize $\mathbf{X}^0 \in_+^{m\times p}$ and $\mathbf{H}^0 \in_+^{p\times n}$
- for $k = 0, 1, 2, \ldots$ calculate

$$
\begin{aligned}
\mathbf{X}^{k+1} &= \arg\min\nolimits_{\mathbf{X}} G_{\mathbf{H}^k}(\mathbf{X},\mathbf{X}^k) \\
\mathbf{H}^{k+1} &= \arg\min\nolimits_{\mathbf{H}} G_{\mathbf{X}^{k+1}}(\mathbf{H},\mathbf{H}^k).
\end{aligned}
\tag{4}
$$

Now we turn our attention to the choice of the matrices $\mathbf{A}_{\mathbf{H}^k}(\mathbf{X})$ and $\mathbf{A}_{\mathbf{X}}^{k+1}(Hm)$. We want above functions $\mathbf{X} \to G_{\mathbf{H}^k}(\mathbf{X},\mathbf{X}^k)$ and $\mathbf{H} \to G_{\mathbf{X}^{k+1}}(\mathbf{H},\mathbf{H}^k)$ to be quadratic majorant functions of F. Adapting to the matrix notations used here the definitions proposed in [2], [3] we get

$$
\begin{aligned}
\mathbf{A}_{\mathbf{H}^k}(\mathbf{X}^k) &= \text{diag}\left[\text{vec}\left((\mathbf{X}^k\mathbf{H}^k\mathbf{H}^{k\top}) \oslash \mathbf{X}^k\right)\right] \\
\mathbf{A}_{\mathbf{X}^{k+1}}(\mathbf{H}^k) &= \text{diag}\left[\text{vec}\left((\mathbf{X}^{k+1\top}\mathbf{X}^{k+1}\mathbf{H}^k) \oslash \mathbf{H}^k\right)\right]
\end{aligned}
\tag{5}
$$

They have shown that these choices yield majorizing functions of F for $G_{\mathbf{H}}$ and $G_{\mathbf{X}}$ and convergence of iterations (4) is guaranteed.

The purpose of our work is to improve the speed of convergence of the MM algorithm thanks to an improved choice of matrices $\mathbf{A}_{\mathbf{H}}$ and $\mathbf{A}_{\mathbf{X}}$. To this end, we replace matrices $\mathbf{A}_{\mathbf{H}^k}(\mathbf{X}^k)$ and $\mathbf{A}_{\mathbf{X}^{k+1}}(\mathbf{H}^k)$ by matrices $\mathbf{A}_{\mathbf{H}^k}(\mathbf{U}^k)$ and $\mathbf{A}_{\mathbf{X}^{k+1}}(\mathbf{V}^k)$ where $\mathbf{U}$ and $\mathbf{V}$ are positive matrices. We propose to select these matrices as solutions of the following convex optimization problems:

$$
\left\{
\begin{aligned}
\mathbf{U}^k &= \arg\min\nolimits_{\mathbf{U}\geq 0} \textstyle\sum_{i,j} \frac{\left(\mathbf{X}^k\mathbf{H}^k\mathbf{H}^{k\top}\right)_{i,j}}{\mathbf{U}_{i,j}} \quad \text{s.t. } \|\mathbf{U}\|_1 = \alpha \\
\\
\mathbf{V}^k &= \arg\min\nolimits_{\mathbf{V}\geq 0} \textstyle\sum_{i,j} \frac{\left(\mathbf{X}^{k+1\top}\mathbf{X}^{k+1}\mathbf{H}^k\right)_{i,j}}{\mathbf{V}_{i,j}} \quad \text{s.t. } \|\mathbf{V}\|_1 = \beta
\end{aligned}
\right.
\tag{6}
$$

Particular solutions of this problem are given as follows:

Proposition 1. *Letting* $\mathbf{B} = \mathbf{X}^k \mathbf{H}^k \mathbf{H}^{k\top}$*, a particular solutions of problem (6) for* $\mathbf{U}^k$ *is given by*

$$\mathbf{U}^k_{i,j} = \begin{cases} \mathbf{X}^k_{i,j} & \textit{if } (i,j) \notin I \\ \tilde{\alpha} \frac{\sqrt{\mathbf{B}_{i,j}}}{\sum\limits_{(i,j)\in I} \sqrt{\mathbf{B}_{i,j}}} & \textit{otherwise} \end{cases} \tag{7}$$

where $I = \{(i,j) \in (1:m) \times (1:n);\ \mathbf{B}_{i,j} \neq 0\}$ *and* $\tilde{\alpha} = \alpha - \sum\limits_{(i,j)\notin I} \mathbf{X}^k_{i,j}$.

In the same way, letting $\mathbf{C} = \mathbf{X}^{k+1\top} \mathbf{X}^{k+1} \mathbf{H}^k$*, a particular solutions of problem (6) for* $\mathbf{V}^k$ *is given by ...*

Let us remark that $\mathbf{A}_{\mathbf{H}^k}(\mathbf{U})$ and $\mathbf{A}_{\mathbf{X}^{k+1}}(\mathbf{V})$ are invariant up to a scalling factor of $\mathbf{U}$ and $\mathbf{V}$ and thus the choice of α and β is not very critical [CHECK ON SIMULATIONS]

Proof. The Laplacian for the optimization of $\mathbf{U}$ is given by

$$L(\mathbf{U}) = \sum_{ij} \frac{\mathbf{B}_{ij}}{\mathbf{U}_{ij}} + \lambda(\sum_{ij} \mathbf{U}_{ij} - \alpha) - \sum_{ij} \mu_{ij} \mathbf{U}_{ij}. \tag{8}$$

For KKT conditions to be satisfied for (7) we must have $\frac{\mathbf{B}_{ij}}{\mathbf{U}^2_{ij}} = \lambda - \mu_{ij}$ and $\mu_{ij}\mathbf{U}_{ij} = 0$. Then, setting $\mu_{ij} = 0$ if $\mathbf{B}_{ij} \neq 0$ or $\mathbf{X}_{ij} \neq 0$ and $\mu_{ij} = \lambda$ otherwise, with $\lambda = \left(\frac{\sum_{(i,j)\in I} \sqrt{\mathbf{B}_{i,j}}}{\tilde{\alpha}} \right)^2$, it is clear that (7) satisfies the KKT conditions.

...

$\square$

Note that alternative constraints such that $\|\mathbf{U}\mathbf{H}^k\|_1 = \|\mathbf{Y}\|_1$ or $\|\mathbf{X}^{k+1}\mathbf{V}\|_1 = \|\mathbf{Y}\|_1$ can be considered instead of $\|\mathbf{U}\|_1 = \alpha$ or $\|\mathbf{V}\|_1 = \beta$. However, similar simulation results have been obtained with this choice.

IV. Convergence rate analysis

V. Experimental resuls

We choose to factorize one random positive matrix $\mathbf{Y}$ of size 50×50 into the product of two positive matrices $\mathbf{X}$ of size 50×20 and $\mathbf{H}$ of size 20×50. So that the rank of the matrix $\mathbf{V}$ is reduced.

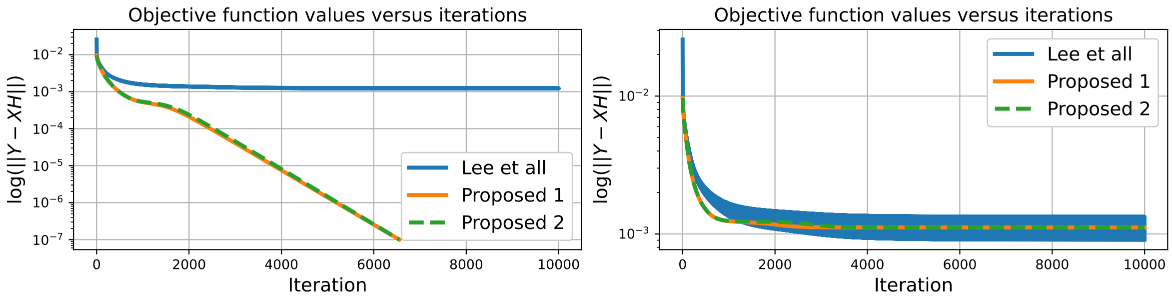

Fig. 1: Objective function values versus iterations for different majorizing functions (Lee et ell, proposed 1 and proposed 2), for data without noise (left), and data with SNR of 35 dB.

Noise level		SNR (dB)	without noise	55	45	35
		Error	0	2.00e-4	6.00e-4	2.00e-3
Lee *et all*		SNR (db)	38	38	38	37
		Error	1.23e-3	1.04e-3	0.66e-3	0.88e-3
		Nb Iteration	10000	10000	10000	10000
Proposed method	proposed 1	SNR (dB)	120	59	50	39
		Error	1.00e-7	1.10e-4	0.33e-3	1.10e-3
		Nb Iteration	6553	10000	10000	10000
	proposed 2	SNR (dB)	120	59	50	39
		Error	1.00e-7	1.10e-4	0.33e-3	1.10e-3
		Nb Iteration	6556	10000	10000	10000

Fig. 2: Comparison between Lee *et all* and proposed method. These results are averaged over 100 random noise realizations.

VI. Conclusions

References

[1] David R. Hunter and Kenneth Lange. A tutorial on mm algorithms. *American Statistician*, 58(1):30–37, February 2004.

[2] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 1999.

[3] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*. MIT Press, 2001.