ISSN 1749-3889 (print), 1749-3897 (online)
International Journal of Nonlinear Science
Vol.33(2022) No.2/3,pp.127-138

An Approximate Augmented Lagrangian Method for Symmetric Nonnegative Matrix Factorization

Yongjin Liang*
School of Mathematical Science, Jiangsu University, Zhenjiang, Jiangsu 212013, P.R. China
(Received March 1 2022, accepted May 12 2022)

Abstract: Nonnegative matrix factorization is a powerful tool for data dimensionality reduction and has found important applications in pattern recognition, data mining, etc. In this paper, we study symmetric nonnegative matrix factorization (SymNMF) and develop an approximate augmented Lagrangian method. We show that the proposed method converges to a stationary point of SymNMF. Experiments on synthetic data for clustering demonstrate that the proposed method is noticeably efficient and achieves competitive performance compared with existing methods.

Keywords: Symmetric nonnegative matrix factorization; Augmented Lagrangian method; Proximal alternating minimizaition method; Clustering

1 Introduction

Dimensionality reduction is a fundamental task for high dimensional data analysis. Nonnegative matrix factorization (NMF) [14], [5]is a powerful tool for data dimensionality reduction by taking a constrained low-rank approximation. NMF has found important applications in a great variety of contexts such as face recognition [10], data mining [6], text clustering [16], etc.

NMF aims to find two nonnegative low-rank matrices $U \in \mathbb{R}^{m\times r}$ and $V \in \mathbb{R}^{n\times r}$ to approximate given matrix $A \in \mathbb{R}^{m\times n}$ with $r \ll \min(m,n)$, such that $A \approx UV^\top$. The common used NMF is given by

$$\min_{U\in\mathbb{R}^{m\times r},\, V\in\mathbb{R}^{n\times r}} \frac{1}{2}\|A - UV^\top\|_F^2, \quad \text{s.t. } U \geq 0,\ V \geq 0,$$

where $X \geq 0$ means that all entries of X are nonnegative. Due to the nonnegative constraint, NMF offers the interpretability that the clustering assignment of each data point can be easily obtained by choosing the largest entry in the corresponding row of V [8].

When NMF is used for clustering propose, matrix $A \in \mathbb{R}^{n\times n}$ usually refers to the similarity matrix defined based on the inherent cluster structure, and thus is symmetric. To preserve the symmetry of the low-rank approximation, the factors U and V should be identical, which leads to the following symmetric nonnegative matrix factorization (SymNMF) model:

$$\text{(SymNMF)} \qquad \min_{X\in\mathbb{R}^{n\times r},\, X\geq 0} f(X) \triangleq \frac{1}{2}\|A - XX^\top\|_F^2. \tag{1}$$

Compared to NMF, SymNMF is more flexible in terms of choosing similarities in the data points.

The objective function of SymNMF model (1) is a nonconvex fourth-order polynomial. Optimization algorithms guarantee only the stationary of the limit points, so one only looks for a local minimum. Existing methods for solving SymNMF can be divided into two categories: methods working directly on the original problem (1) and nonsymmetric relaxation methods working on relaxation forms of the problem (1) [7]. The first kind of methods include projected gradient method, Newton-like method [9]. The projected gradient method suffers from extremely slow convergence, while the Newton-like method is efficient only for small-scale problems. The main idea of nonsymmetric relaxation

*Corresponding author. *E-mail address*: yjliang@stmail.ujs.edu.cn

methods [12, 19] is to relax the symmetry requirement $A \approx XX^\top$ by $A \approx XY^\top$ and enforce $X = Y$ gradually. A common nonsymmetric relaxation is of the form

$$\min_{X,Y\in\mathbb{R}^{n\times r}, X,Y\geq 0} \frac{1}{2}\|A - XY^\top\|_F^2 + \frac{\rho}{2}\|X - Y\|_F^2, \tag{2}$$

where $\rho > 0$ is a penalty parameter to control the violation of the symmetry. The alternating algorithm is widely used for solving the above problem, which leads to two nonnegative least square (NLS) subproblems with respect to variables X and Y, respectively. Many nonsymmetric relaxations follow this scheme, such as SymANLS [19] and SymHALS [19]. Zhu et al. [19] proved that when ρ is greater than a certain threshold, any critical point of (2) matching a critical point of SymNMF. In [12], authors proved that the Karush-Kuhn-Tucker (KKT) points of SymNMF and the KKT points of the following problem

$$\min_{Y\in\mathbb{R}_+^{n\times r}} \frac{1}{2}\|A - XY^\top\|_F^2 \quad \text{s.t. } X = Y, \ \ \|Y_{i,:}\|_2^2 \leq \hat{\tau}, \quad i = 1, \cdots, n, \tag{3}$$

have a one-to-one correspondence when the penalty parameter which defined in the algorithm is greater than a certain threshold, where $Y_{i,:}$ denotes the i-th row of Y, $\hat{\tau} > 0$ is some given constant associated with A. Then a nonconvex splitting method (NS-SNMF) is proposed to solve (3).

In this paper, we focus on SymNMF and address it by an approximate augmented Lagrangian method (AALM). The augmented Lagrangian (AL) method (or its variant) is a good candidate for solving constrained nonlinear programming problems and has been well-studied in literatures [1, 13, 18]. Notice that Problem (1) can be rewritten as

$$\min_{X,Y} \frac{1}{2}\|A - XY^\top\|_F^2 + \delta_{\mathbb{R}_+^{n\times r}}(Y), \quad \text{s.t. } X - Y = 0, \tag{4}$$

where $\mathbb{R}_+^{n\times r}$ denotes the set of all n-by-r nonnegative matrices and $\delta_{\mathbb{R}_+^{n\times r}}(\cdot)$ denotes the indictor function on $\mathbb{R}_+^{n\times r}$, i.e., $\delta_{\mathbb{R}_+^{n\times r}}(X) = 0$ if $X \in \mathbb{R}_+^{n\times r}$, otherwise, $\delta_{\mathbb{R}_+^{n\times r}}(X) = +\infty$. Denote $W = [X^\top,\ Y^\top]^\top \in \mathbb{R}^{2n\times r}$, $B = [I_n,\ 0_n] \in \mathbb{R}^{n\times 2n}$ and $C = [0_n,\ I_n] \in \mathbb{R}^{n\times 2n}$. Notice that $\delta_{\mathbb{R}_+^{n\times r}}$ is an extended real valued function and the subdifferential of $\delta_{\mathbb{R}_+^{n\times r}}$, which defined by $\partial\delta(X) \triangleq \{Z \mid 0 \geq \langle Z, Y - X\rangle, \forall Y\}$ is unbounded. The convergence proof gaven in [13] cannot be directly applied to SymNMF. In this paper, we will show that the limit point indeed is a stationary point of SymNMF and discuss details of solving the involved subproblems by the proximal alternating minimization (PAM) method [2] to the variable $(X^\top, Y^\top)$ and prove that the sequence generated by the PAM method is indeed convergence to the critical point of the AL subproblem.

We use the following notations in this paper: 1) $X \geq 0$ means all $X_{i,j} \geq 0$. We use $\Pi_+[Z] = \arg\min_{X\geq 0}\|X - Z\|_F^2$ to denote the orthogonal projection of Z onto $\mathbb{R}_+^{n\times r}$; 2) $\|X\|_\infty = \max_{i,j}\{|\ X_{i,j}\ |\}$; 3) $\mathcal{S}_+^n$ denotes the set of n-by-n symmetric semi-definite positive matrices; 4) We use I_n and 0_n to denote n-by-n identity matrix and zero matrix, respectively; 5) For $D \in \mathcal{S}_+^n$ and $Z \in \mathbb{R}^{n\times n}$, $\|Z\|_D^2 = \mathrm{tr}(Z^\top DZ)$, where $\mathrm{tr}(X) = \sum_i X_{i,i}$ represents the matrix trace. Then the Frobenius norm is defined with D equal to the identity matrix I_n; 6) Matrix inner product is defined as $\langle X, Y\rangle_D = \mathrm{tr}(X^\top DY)$. We omit the subscript D if the inner product is defined with $D = I_n$; 7) For nonsmooth convex function $\delta(X)$, the subdifferential is given by $\partial\delta(X) \triangleq \{Z \mid \delta(Y) \geq \delta(X) + \langle Z, Y - X\rangle,\ \forall Y\}$; 8) We say function f is coercive if $\lim_{\|X\|\to\infty} f(X) = \infty$; 9) We say a point is a KKT point if there exists Lagrangian multiplier such that the first-order optimality condition hold.

The paper is organized as follows. In Section 2, we present the approximate augmented Lagrangian method for solving SymNMF, together with the convergence analysis. In Section 3, we give details on how to solve the involved subproblem by the PAM method. We demonstrate the efficiency of our proposed method in Section 4 by numerical experiments using synthetic data. Finally, some conclusions follow in Section 5.

2 The Approximate Augmented Lagrangian Method

In this section, we give details on how to solve SymNMF by the approximate augmented Lagrangian method.

2.1 Algorithm

The augmented Lagrangian function of Problem (4) is of the form

$$\widehat{\mathcal{L}}_\rho(X, Y; \Lambda) = \frac{1}{2}\|A - XY^\top\|_F^2 + \langle\Lambda, X - Y\rangle + \frac{\rho}{2}\|X - Y\|_F^2,$$

where $(X, Y, \Lambda) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{n\times r}_{+} \times \mathbb{R}^{n\times r}$, Λ is the Lagrangian multiplier and $\rho > 0$ is the penalty parameter. For any fixed Λ^k and ρ_k, the general update step of the AL method employed on Problem (4) takes the form

$$\begin{cases} (X^{k+1}, Y^{k+1}) \approx \arg\min_{X,Y}\{\widehat{\mathcal{L}}_{\rho_k}(X, Y; \Lambda^k) + \delta_{\mathbb{R}^{n\times r}_{+}}(Y)\}, \\ \Lambda^{k+1} = \Lambda^k + \rho_k(X^{k+1} - Y^{k+1}). \end{cases}$$

Then update ρ_{k+1} according to the violation of equality $X = Y$. (X^{k+1}, Y^{k+1}) is an approximate minimizer of the AL subproblem. Let $(X^{\text{feas}}, Y^{\text{feas}}) \in \mathbb{R}^{n\times r} \times \mathbb{R}^{n\times r}_{+}$ and $X^{\text{feas}} = Y^{\text{feas}}$ be any feasible point of Problem (4). We give an approximate AL method to address Problem (4) in Algorithm 1, which is similar to the one that proposed in [13].

Algorithm 1 (**AALM**: Approximate Augmented Lagrangian Method for (4))

Require: $\{\epsilon_k\}_{k\in\mathbb{N}} \downarrow 0$, $\tau \in [0, 1)$, $\nu > 0$, $\mu > 1$, $k = 1$, $\rho_0 > 0$. $X^0, Y^0 \in \mathbb{R}^{n\times r}_{+}$, $\Lambda^0 \in \mathbb{R}^{n\times r}$, $\Upsilon \geq \max\{\frac{1}{2}\|A - X^{\text{feas}}(Y^{\text{feas}})^\top\|_F^2, \widehat{L}_{\rho_0}(X^0, Y^0; \Lambda^0)\}$.

Ensure: A sequence $\{(X^k, Y^k, \Lambda^k)\}_{k\in\mathbb{N}}$.

1: **while** stopping criterion is not satisfied **do**

2: **Step 1.** For given ρ_{k-1}, Λ^{k-1}, compute (X^k, Y^k) such that $Y^k \in \mathbb{R}^{n\times r}_{+}$,

$$\widehat{L}_{\rho_{k-1}}(X^k, Y^k; \Lambda^{k-1}) \leq \Upsilon \tag{5}$$

and there exists an $\xi^k \in \partial\check{L}_{\rho_{k-1}}(X^k, Y^k; \Lambda^{k-1})$ satisfying

$$\|\xi^k\|_\infty < \epsilon_{k-1}. \tag{6}$$

3: **Step 2.** Update the Lagrangian multiplier

$$\Lambda^k = \Lambda^{k-1} + \rho_{k-1}(X^k - Y^k).$$

4: **Step 3.** Update the penalty parameter

$$\rho_k = \begin{cases} \rho_{k-1} & \text{if } \|X^k - Y^k\|_\infty \leq \tau\|X^{k-1} - Y^{k-1}\|_\infty, \\ \max\{\mu\rho_{k-1}, \|\Lambda^k\|_\infty^{1+\nu}\} & \text{otherwise.} \end{cases}$$

5: **Step 4.** Update $k \leftarrow k + 1$;

6: **end while**

7: **return** $\{(X^k, Y^k; \Lambda^k)\}_{k\in\mathbb{N}}$.

Remark 1 *Here are some comments on Algorithm 1.*

(i) Different from method presented in [13], we update penalty ρ_k based on the violation of equality constraints to control the growth of ρ_k.

(ii) Theoretically, Λ^0 can be set as any n-by-r matrix.

(iii) By subdifferentiability property [15], the condition (6) *is equivalent to that there exists*

$$(\hat{\xi}_X^k, \hat{\xi}_Y^k) \in (\nabla_X\widehat{\mathcal{L}}_{\rho_{k-1}}(X^k, Y^k; \Lambda^{k-1}), \nabla_Y\widehat{\mathcal{L}}(X^k, Y^k; \Lambda^{k-1}) + \partial\delta_{\mathbb{R}^{n\times r}_{+}}(Y))$$

such that $\max\{\|\hat{\xi}_X^k\|_\infty, \|\hat{\xi}_Y^k\|_\infty\} < \epsilon_{k-1}$ *and* $\epsilon_k \downarrow 0$ *as* $k \to +\infty$.

2.2 Convergence

Suppose that (X^*, Y^*) is a local solution of (4). Then there is a Lagrangian multiplier $\Lambda^* \in \mathbb{R}^{n\times r}$ and $W^* \in \partial\delta_{\mathbb{R}^{n\times r}_{+}}(Y^*)$, such that the following conditions are satisfied at (X^*, Y^*):

$$\begin{cases} Y^* \geq 0; \\ X^* - Y^* = 0; \\ (X^*(Y^*)^\top - A)Y^* + \Lambda^* = 0; \\ W^* + (Y^*(X^*)^\top - A)X^* - \Lambda^* = 0. \end{cases} \tag{7}$$

We say (X^*, Y^*) is a KKT point of (4) if (7) holds. We say $X^* \in \mathbb{R}_+^{n\times r}$ is a stationary point of (1) if X^* satisfies

$$2(A - X^*(X^*)^\top)X^* \in \mathcal{N}_{\mathbb{R}_+^{n\times r}}(X^*),$$

where $\mathcal{N}_{\mathbb{R}_+^{n\times r}}(X^*)$ is the normal cone of $\mathbb{R}_+^{n\times r}$ at point X^*. Suppose Step 1 of Algorithm 1 is well defined, we give the following convergence result regarding Algorithm 1.

Theorem 2 *Let $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ be a sequence generated by Algorithm 1. Suppose $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ is bounded. Then any limit point (X^*, Y^*) of $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ is a KKT point of Problem* (4)*. Moreover, X^* is a stationary point of Problem* (1)*.*

Proof. Under the boundedness assumption of $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ and the closeness of $\mathbb{R}_+^{n\times r}$, there exists an index set $\mathcal{K} \subset \mathbb{N}$ and (X^*, Y^*), such that

$$\lim_{\mathcal{K}\ni k\to\infty} X^k = X^* \quad \text{and} \quad \lim_{\mathcal{K}\ni k\to\infty} Y^k = Y^* \in \mathbb{R}_+^{n\times r}.$$

If $\{\rho_k\}$ is bounded, then by the updating rule of ρ_k in Algorithm 1, there exists an $k_0 \in \mathbb{N}$ such that

$$\|X^k - Y^k\|_\infty \le \tau\|X^{k-1} - Y^{k-1}\|_\infty \quad \forall k \ge k_0.$$

Thus

$$X^* - Y^* = 0.$$

If $\{\rho_k\}$ is unbounded, in view of (5), we have

$$\frac{1}{2}\|A - X^k(Y^k)^\top\|_F^2 + \langle\Lambda^{k-1}, X^k - Y^k\rangle + \frac{\rho_{k-1}}{2}\|X^k - Y^k\|_F^2 \le \Upsilon, \quad \forall k.$$

It follows that

$$\|X^k - Y^k\|_F^2 \le \frac{2}{\rho_{k-1}}\left[\Upsilon - \frac{1}{2}\|A - X^k(Y^k)^\top\|_F^2 - \langle\Lambda^{k-1}, X^k - Y^k\rangle\right].$$

According to the update rule of ρ_k in Algorithm 1, $\frac{\|\Lambda^k\|_\infty}{\rho_k} \to 0$ as $\rho_k \to 0$. Let $k \in \mathcal{K}$ goes to infinity on both sides of the above inequality, under the boundedness assumption on $\{(X^k, Y^k)\}$, we obtain that

$$X^* - Y^* = 0.$$

Therefore, in both cases, we show that (X^*, Y^*) is a feasible point of Problem (4).

Next, we show that there exist $\Lambda^* \in \mathbb{R}^{n\times r}$, such that $(X^*, Y^*; \Lambda^*)$ satisfies (7). Notice that (6) implies that (X^k, Y^k) satisfies

$$\xi_X^k \triangleq (X^k(Y^k)^\top - A)Y^k + \Lambda^{k-1} + \rho_{k-1}(X^k - Y^k) \quad \text{with } \|\xi_X^k\|_\infty < \epsilon_{k-1},$$

$$\xi_Y^k \triangleq W^k + (Y^k(X^k)^\top - A)X^k - \Lambda^{k-1} + \rho_{k-1}(Y^k - X^k) \quad \text{with } \|\xi_Y^k\|_\infty < \epsilon_{k-1},$$

for some $W^k \in \partial\delta_{\mathbb{R}_+^{n\times r}}(Y^k)$, and $\epsilon_k \downarrow 0$ as $k \to \infty$.

Combing with the update rule of Λ^k, we have

$$\xi_X^k \triangleq (X^k(Y^k)^\top - A)Y^k + \Lambda^k \quad \text{with } \|\xi_X^k\|_\infty < \epsilon_{k-1}, \tag{8}$$

$$\xi_Y^k \triangleq W^k + (Y^k(X^k)^\top - A)X^k - \Lambda^k \quad \text{with } \|\xi_Y^k\|_\infty < \epsilon_{k-1}. \tag{9}$$

By passing to a subsequence on $\mathcal{K}$ if necessary, together with $X^* = Y^*$ and the definition of $\partial\delta_{\mathbb{R}_+^{n\times r}}$, we obtain from (8) and (9) that the second and third equality in (7) hold, which implies that (X^*, Y^*) is a KKT point of Problem (4). Moreover,

$$2(A - X^*(X^*)^\top)X^* \in \partial\delta_{\mathbb{R}_+^{n\times r}}(X^*) = \mathcal{N}_{\mathbb{R}_+^{n\times r}}(X^*),$$

where the equality holds due to the convexity of $\mathbb{R}_+^{n\times r}$. We have thus proved the theorem. ■

3 PAM Method for AL Subproblems

In this section, we discuss how to solve the AL subproblem

$$\min_{X,Y}\{\widehat{\mathcal{L}}_{\rho_k}(X,Y;\Lambda^k)+\delta_{\mathbb{R}_+^{n\times r}}(Y)\}. \tag{10}$$

To get a smaller Lipschitz constant of the AL subproblem, we consider the following scaled form of $\widehat{\mathcal{L}}_\rho$. Define

$$\widetilde{L}_{\rho_k}(X,Y;\Lambda^k)\triangleq\frac{1}{\rho}\widehat{\mathcal{L}}_{\rho_k}(X,Y;\Lambda^k)=\frac{1}{2\rho_k}\|A-XY^\top\|_F^2+\langle\Lambda_k/\rho_k,X-Y\rangle+\frac{1}{2}\|X-Y\|_F^2.$$

For the sake of brevity, we omit Λ^k and the subscript of ρ_k to write $\widetilde{L}_{\rho_k}(X,Y;\bar{\Lambda}^k)$ as $\widetilde{L}_\rho(X,Y)$. Denote $\check{L}_\rho(X,Y):=\widetilde{L}_\rho(X,Y)+\delta_{\mathbb{R}_+^{n\times r}}(Y)$. Then $\check{L}_\rho(X,Y)$ satisfies the following properties:

i) $\delta_{\mathbb{R}_+^{n\times r}}(Y)$ is proper closed, convex and continuous over $\mathbb{R}_+^{n\times r}$;

ii) $\widetilde{L}_\rho$ is closed and differentiable over $\mathbb{R}^{n\times r}\times\mathbb{R}_+^{n\times r}$;

iii) for each $(\overline{X},\overline{Y})$, $\min_X\left\{\check{L}_\rho(X,\overline{Y})\right\}$ or/and $\min_Y\left\{\check{L}_\rho(\overline{X},Y)\right\}$ has a unique minimizer since $\mathbb{R}_+^{n\times r}$ is convex and $\widetilde{L}_\rho(X,\overline{Y})$ is 1-strongly convex w.r.t. X and Y for any given Λ, respectively;

iv) for any $\alpha\in\mathbb{R}$, the level set $\{(X,Y)\mid\check{L}_\rho(X,Y)\le\alpha\}$ is bounded.

Next, we show how to solve the AL subproblem by the PAM method. For given $\bar{\Lambda}$, ρ and the current iteration point (X^-,Y^-), the iterative of PAM update (X,Y) by

$$\begin{cases}X^+=\arg\min_X\left\{\check{L}_\rho(X,Y^-)+\frac{1}{2}\|X-X^-\|_{D_1}^2\right\};\\ Y^+=\arg\min_Y\left\{\check{L}_\rho(X^+,Y)+\frac{1}{2}\|Y-Y^-\|_{D_2}^2\right\},\end{cases}$$

where $D_1,\ D_2\in\mathcal{S}_+^n$. In the following, we give details for updating $(X^\top,Y^\top)$ (rather than (X,Y)) based on the fact that for any $X\in\mathbb{R}^{n\times r}$ and $\widetilde{D}\in\mathcal{S}_+^r$, $\|X\|_F=\|X^\top\|_F$, $\|X^\top\|_{\widetilde{D}}^2=\|X\|_{D(X)}^2$ with $D(X)=\frac{\|X^\top\|_{\widetilde{D}}^2}{\|X\|_F^2}I_n$.

3.1 PAM for Updating $(X^\top,Y^\top)$

For ease of notation, define

$$\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)=\frac{1}{2\rho}\|YX^\top-A^\top\|_F^2+\langle\Lambda^\top/\rho,X^\top-Y^\top\rangle+\frac{1}{2}\|X^\top-Y^\top\|_F^2.$$

Then the following properties hold for $\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)$.

Proposition 3 *The following statements hold.*

i) $\nabla_{X^\top}\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)$ *and* $\nabla_{Y^\top}\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)$ *are Lipschitz continuous, i.e., there exist* L_X *and* L_Y*, such that*

$$\|\nabla_{X^\top}\widetilde{\mathcal{L}}_\rho(X^\top,Y_1^\top)-\nabla_{X^\top}\widetilde{\mathcal{L}}_\rho(X^\top,Y_2^\top)\|_F\le L_X\|Y_1-Y_2\|_F,\quad\forall Y_1^\top,Y_2^\top;$$

$$\|\nabla_{Y^\top}\widetilde{\mathcal{L}}_\rho(X_1^\top,Y^\top)-\nabla_{Y^\top}\widetilde{\mathcal{L}}_\rho(X_2^\top,Y^\top)\|\le L_Y\|X_1-X_2\|_F,\quad\forall X_1^\top,X_2^\top,$$

where $L_X=1+\frac{1}{\rho}\|Y^\top Y\|_F$ *and* $L_Y=1+\frac{1}{\rho}\|X^\top X\|_F$.

ii) For any $Z\in\mathbb{R}^{r\times n}$,

$$\nabla_{X^\top}^2\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)[Z]=\frac{1}{\rho}Y^\top YZ+Z;\quad\nabla_{Y^\top}^2\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)[Z]=\frac{1}{\rho}X^\top XZ+Z.$$

In this subsection, for the sake of brevity, we omit the superscripts of iteration, and use X^- and X^+ to denote the previous and the current iteration, respectively. To update $X^\top$ and $Y^\top$, we rewrite the X-subproblem as

$$\begin{aligned}\left(X^\top\right)^+&=\arg\min_{X^\top}\widetilde{\mathcal{L}}_\rho(X^\top,Y^\top)+\frac{1}{2}\|X^\top-\left(X^-\right)^\top\|_{\widetilde{D}_1}^2\\&=\arg\min_Z\frac{1}{2}\mathrm{tr}(Z^\top(\frac{1}{\rho}Y^\top Y+I_r+\widetilde{D}_1)Z)-\mathrm{tr}(Z^\top(\frac{1}{\rho}Y^\top A^\top+(Y-\Lambda/\rho)^\top+\widetilde{D}_1(X^-)^\top)).\end{aligned}$$

When $\widetilde{D}_1$ is set to be $\widetilde{D}_1 = \frac{\gamma_1}{\rho} I_r - \frac{1}{\rho} Y^\top Y$, it follows that

$$\left(X^\top\right)^+ = \arg\min_Z \frac{1+\gamma_1/\rho}{2} \|Z - \frac{Y^\top A^\top + \rho Y^\top - \Lambda^\top + \gamma_1 (X^-)^\top - Y^\top Y (X^-)^\top}{\rho + \gamma_1}\|_F^2,$$

which yields that

$$X^+ = \frac{1}{\rho+\gamma_1}\left(AY + \rho Y - \Lambda + \gamma_1 X^- - X^- Y^\top Y\right). \tag{11}$$

Notice that $\widetilde{\mathcal{L}}_\rho(X^\top, Y^\top)$ is quadratic w.r.t. $X^\top$. It follows from Proposition 3 ii) that

$$\left[\nabla^2_{X^\top} \widetilde{L}_\rho((X^-)^\top, Y^\top) + \widetilde{D}_1\right][Z] = (1+\gamma_1/\rho) Z.$$

Thus, we have

$$\begin{aligned}
&\widetilde{\mathcal{L}}_\rho(X^\top, Y^\top) + \frac{1}{2}\|X^\top - \left(X^-\right)^\top\|^2_{\widetilde{D}_1} \\
=&\widetilde{\mathcal{L}}_\rho((X^-)^\top, Y^\top) + \langle X^\top - (X^-)^\top, \nabla_{X^\top} \widetilde{\mathcal{L}}_\rho((X^-)^\top, Y^\top)\rangle + \frac{1+\gamma_1/\rho}{2}\|X^\top - (X^-)^\top\|_F^2.
\end{aligned} \tag{12}$$

Hence, the update formula (11) identifies with

$$(X^+)^\top = (X^-)^\top - \frac{1}{1+\gamma_1/\rho} \nabla_{X^\top} \widetilde{\mathcal{L}}_\rho((X^-)^\top, Y^\top).$$

Combining with the sufficient decrease Lemma [4], it holds that

$$\widetilde{\mathcal{L}}_\rho((X^+)^\top, Y^\top) \leq \widetilde{\mathcal{L}}_\rho((X^-)^\top, Y^\top) - (1 + \frac{\gamma_1}{\rho} - \frac{L_X}{2})\|(X^+)^\top - (X^-)^\top\|_F^2. \tag{13}$$

Therefore, to ensure sufficient descent, it is sufficient to set $1 + \frac{\gamma_1}{\rho} > \frac{L_X}{2}$, i.e.,

$$\gamma_1 > \rho(\frac{L_X}{2} - 1) = \frac{\|Y^\top Y\|_F - \rho}{2}. \tag{14}$$

Similarly, the Y-subproblem can be rewritten as

$$\begin{aligned}
\left(Y^\top\right)^+ =& \arg\min_{Y^\top \in \mathbb{R}_+^{r\times n}} \widetilde{\mathcal{L}}_\rho((X^\top)^+, Y^\top) + \frac{1}{2}\|Y^\top - (Y^-)^\top\|^2_{\widetilde{D}_2} \\
=& \arg\min_{Z \in \mathbb{R}_+^{r\times n}} \frac{1}{2}\mathrm{tr}(Z^\top(\frac{1}{\rho}(X^+)^\top X^+ + I_r + \widetilde{D}_2)Z) - \mathrm{tr}(Z^\top(\frac{1}{\rho}(X^+)^\top A + (X^+ + \Lambda/\rho)^\top + \widetilde{D}_2 (Y^-)^\top)).
\end{aligned}$$

When $\widetilde{D}_2$ is set to be $\widetilde{D}_2 = \frac{\gamma_2}{\rho} I - \frac{1}{\rho} X^\top X$, it follows that

$$\left(Y^\top\right)^+ = \arg\min_{Z\in\mathbb{R}_+^{r\times n}} \|Z - \frac{(X^+)^\top A + \rho (X^+)^\top + \Lambda^\top + \gamma_2 (Y^-)^\top - (X^+)^\top X^+ (Y^-)^\top}{\rho + \gamma_2}\|_F^2$$

which yields that

$$Y^+ = \frac{1}{\rho+\gamma_2}\Pi_+[\left(A^\top X^+ + \rho X^+ + \Lambda + \gamma_2 Y^- - Y^-(X^+)^\top X^+\right)]. \tag{15}$$

Similar to the analysis about X^+, it holds that

$$\begin{aligned}
&\widetilde{\mathcal{L}}_\rho((X^+)^\top, Y^\top) + \frac{1}{2}\|Y^\top - \left(Y^-\right)^\top\|^2_{\widetilde{D}_2} \\
=&\widetilde{L}_\rho((X^+)^\top, (Y^-)^\top) + \frac{1+\gamma_2/\rho}{2}\|Y^\top - (Y^-)^\top\|_F^2 + \langle Y^\top - (Y^-)^\top, \nabla_Y \widetilde{L}_\rho((X^+)^\top, (Y^-)^\top)\rangle.
\end{aligned} \tag{16}$$

Hence, $(Y^+)^\top$ is consistent with

$$(Y^+)^\top = \mathrm{prox}_{\frac{\rho}{\gamma_2+\rho}\delta_{\mathbb{R}_+^{n\times r}}}\left((Y^-)^\top - \frac{1}{\gamma_2/\rho+1}\nabla_{Y^\top}\widetilde{\mathcal{L}}_\rho((X^+)^\top,(Y^-)^\top)\right),$$

where $\mathrm{prox}_{\lambda\delta_{\mathbb{R}_+^{n\times r}}}(U) = \arg\min_{Z\in\mathbb{R}_+^{n\times r}}\frac{1}{2\lambda}\|Z-U\|_F^2$ denotes the projection operator onto $\mathbb{R}_+^{n\times r}$. According to the sufficient decrease Lemma [4], we have

$$\widetilde{L}_\rho((X^+)^\top,(Y^+)^\top) \le \widetilde{L}_\rho((X^+)^\top,(Y^-)^\top) - \left(1+\frac{\gamma_2}{\rho}-\frac{L_Y}{2}\right)\|(Y^+)^\top-(Y^-)^\top\|_F^2. \tag{17}$$

Hence, to ensure sufficient descent, we can set $1+\frac{\gamma_2}{\rho} > \frac{L_Y}{2}$, i.e.,

$$\gamma_2 > \frac{1}{2}\|X^\top X\|_F - \frac{\rho}{2}. \tag{18}$$

Remark 4 (12) *and* (16) *illustrate that the update of variable* $(X^\top, Y^\top)$ *generated by the PAM iteration* (11) *and* (15) *is consistent with proximal alternating linearized minimization [3] iteration. We will discuss the convergence of iterative* (11) *and* (15) *based on this fact in the next subsection.*

Details about the PAM method for the AL subproblem are described in Algorithm 2.

Algorithm 2 (PAM for subproblem (6) with fixed k)

Require: $\gamma_{k,j}^1,\ \gamma_{k,j}^2 > 0$, $(X^{k,0}, Y^{k,0})$.
Ensure: A sequence (X^k, Y^k).
1: **while** $\|\xi^{k,j}\|_\infty \ge \frac{\epsilon_{k-1}}{\rho_{k-1}}$ **do**
2: Update $X^{k,j}$ by

$$X^{k,j} = \frac{AY^{k,j-1}+\rho_{k-1}Y^{k,j-1}-\Lambda^{k-1}+\gamma_1X^{k,j-1}-X^{k,j-1}(Y^{k,j-1})^\top Y^{k,j-1}}{\rho_{k-1}+\gamma_{k,j}^1}.$$

3: Update $Y^{k,j}$ by

$$Y^{k,j} = \mathcal{P}_{\mathbb{R}_+^{n\times r}}[\frac{A^\top X^{k,j}+\rho_{k-1}X^{k,j}+\Lambda^{k-1}+\gamma_2Y^{k,j-1}-Y^{k,j-1}(X^{k,j})^\top X^{k,j}}{\rho_{k-1}+\gamma_{k,j}^2}],$$

4: **end while**
5: **return** $X^k = X^{k,j}$ and $Y^k = Y^{k,j}$.

Remark 5 *In Algorithm 2, we generate* $\widehat{X}\in\mathbb{R}^{n\times r}$ *randomly, then project* $\widehat{X}$ *onto the convex set* $\{X\in\mathbb{R}_+^{n\times r} \mid \|X_{k,:}\|_2^2 \le 1.001\frac{A_{k,k}+\frac{1}{2}\sqrt{\sum_{i=1}^n(A_{i,k}+A_{k,i})^2}}{2},\ k=1,\cdots,n.\}$ *and set* $Y^{k,0}=X^{k,0}$ *when* $k=0$. *Otherwise, we set*

$$(X^{k,0},Y^{k,0}) = (X^{k-1},\Pi_+[Y^{k-1}-\alpha\nabla f(Y^{k-1})]),\quad k>1,$$

where $\alpha>0$ *satisfies the following line search condition*

$$f(Y^{k,0}) \le f(Y^{k-1}) - \delta\alpha\|\nabla f(Y^{k-1})\|_F^2. \tag{19}$$

In implementing, we set $\alpha^{(0)} = \dfrac{\|Y^{k-1}-Y^{k-2}\|_F^2}{|\ \mathrm{tr}((Y^{k-1}-Y^{k-2})^T(\nabla f(Y^{k-1})-\nabla f(Y^{k-2})))\ |}$ *as the initial gauss of the step size* α_k *(denoted by* $\alpha_k^{(0)}$*), then check* $\alpha_k = \eta\alpha_k^{(0)}$ *until the above inequality is established or 10 steps of iteration are reached. We use* $\eta=0.5$ *and* $\delta=10^{-4}$. *If* $\widehat{L}_{\rho_{k-1}}(X^{k,0},Y^{k,0};\Lambda^{k-1})>\Upsilon$, *then we set* $(X^{k,0},Y^{k,0}) = (X^{\mathrm{feas}},Y^{\mathrm{feas}})$ *as suggested in [13].*

To check the convergence condition (6), a candidation of $\xi^{k,j} = (\xi_X^{k,j}, \xi_Y^{k,j})$ could be

$$\xi_X^{k,j} = \nabla_X \mathcal{L}_{\rho_{k-1}}(X^{k,j}, Y^{k,j}; \Lambda^{k-1}) - \nabla_X \mathcal{L}_{\rho_{k-1}}(X^{k,j-1}, Y^{k,j-1}; \Lambda^{k-1}) + (X^{k,j-1} - X^{k,j}) D_1^{k,j-1}; \tag{20}$$

$$\xi_Y^{k,j} = \nabla_Y \mathcal{L}_{\rho_{k-1}}(X^{k,j}, Y^{k,j}; \Lambda^{k-1}) - \nabla_Y \mathcal{L}_{\rho_{k-1}}(X^{k,j}, Y^{k,j-1}; \Lambda^{k-1}) + (Y^{k,j-1} - Y^{k,j}) D_2^{k,j}, \tag{21}$$

where $D_1^{k,j-1} = \frac{\gamma_1}{\rho} I_r - \frac{1}{\rho}(Y^{k,j-1})^\top Y^{k,j-1}$ and $D_2^k = \frac{\gamma_2}{\rho} I_r - \frac{1}{\rho}(X^{k,j})^\top X^{k,j}$. In the next proposition, we show that $\xi^{k,j}$ indeed satisfies the convergence condition (6). Moreover, the sequence generated by Algorithm 1 is strongly convergence.

Proposition 6 *For each $k \geq 1$, let $\{(X^{k,j}, Y^{k,j})\}_{j\in\mathbb{N}}$ be the sequence generated by Algorithm 2 with*

$$\gamma_{k,j}^1 > \frac{\|(Y^{k,j-1})^\top (Y^{k,j-1})\|_F - \rho_k}{2} \quad \text{and} \quad \gamma_{k,j}^2 > \frac{\|(X^{k,j})^\top X^{k,j}\|_F - \rho_k}{2}.$$

Then

1) *$\xi^{k,j}$ defined in (20)-(21) satisfies $\xi^{k,j} \in \partial \check{L}_{\rho_{k-1}}(X^{k,j}, Y^{k,j}; \Lambda^{k-1})$. Moreover, $\|\xi^{k,j}\|_\infty \to 0$ as $j \to \infty$.*

2) *The sequence $\{X^{k,j}, Y^{k,j}\}_{j\in\mathbb{N}}$ has finite length, that is,*

$$\sum_{j=1}^{\infty} \|(X^{k,j+1}, Y^{k,j+1}) - (X^{k,j}, Y^{k,j})\| < \infty.$$

Moreover, $\{(X^{k,j}, Y^{k,j})\}_{j\in\mathbb{N}}$ converges to a critical point $(X^{k,*}, Y^{k,*})$ of function $\check{L}_{\rho_{k-1}}(X, Y; \Lambda^{k-1})$.

4 Numerical Experiments

In this section, we evaluate the effectiveness and efficiency of our proposed method on synthetic data. All numerical experiments are implemented in MATLAB R2020b running on a computer with an Intel(R) Core(TM) i7-9700 CPU @ 3.00GHz $\times$ 3.00GHz and 32GB of RAM.

In the following, we name our proposed algorithm (Algortihm 1 with Algorithm 2 for Setp 1) as AALM for short. We take $\tau = 0.99$, $\mu = 1.05$, $\nu = 0.01$ To solve the AL subproblem by Algorithm 2, for each k, we set

$$\gamma_1 = \tfrac{\theta}{2} \max(\|(Y^{k,j-1})^\top Y^{k,j-1}\|_F - \hat{\rho}, 1); \qquad \gamma_2 = \tfrac{\theta}{2} \max(\|(X^{k,j})^\top X^{k,j}\|_F - \hat{\rho}, 1) \tag{22}$$

with some $\theta > 1$ and $\hat{\rho} \geq 0$.

4.1 Comparison Methods and Stopping Criterion

We compare our proposed method with the nonconvex splitting method for SymNMF (NS-SNMF)[12], ANLS for SymNMF (SymANLS) [19] and HALS for SymNMF (SymHALS)* [19], Newton-like method PNewton† [9]. All of these methods expect for PNewton are nonsymmetric relaxization methods, where NS-SNMF works on problem (3) with $\hat{\tau} = \max_{1\leq i\leq n}\{\hat{\theta}_i\}$, where $\hat{\theta}_i = \frac{A_{i,i} + \frac{1}{2}\sqrt{\sum_{j=1}^m (A_{j,i}+A_{i,j})^2}}{2}$. SymANLS and SymHALS work on the nonsymmetric penalty problem (2).

We use the same initial point as described in Remark 5 for all comparison methods. We set parameters in the algorithm NS-SNMF according to the description of the paper [12], except for the parameter ϵ, which is used to update the penalty parameter, we set it as $1e-4$ to get better numerical results. For SymANLS and SymHALS, refer to paper [19] Corollary 1 and 2 to set the penalty parameter ρ.

We use $RPG \triangleq \frac{\|\mathrm{grad} f(Y^k)\|_F}{\|\mathrm{grad} f(Y^0)\|_F} \leq \varepsilon$ as the stopping criterions for each algorithm, where

$$\mathrm{grad} f(Y) = \begin{cases} \nabla f(Y)_{ij}, & \text{if } Y_{i,j} > 0; \\ \min\{\nabla f(Y)_{ij}, 0\}, & \text{if } Y_{i,j} = 0, \end{cases}$$

denotes the projected gradient of $f(Y)$ at Y. In addition, the iteration process is terminated when the number of steps reaches 5000 or the maximum running time exceeds 9000 seconds. For AALM, we also require the return value satisfies $\|X^k - Y^k\|_F < \varepsilon_1$. In order to reduce the iterative complexity of the subproblem, the termination iterative conditions of PAM are similar to (20)-(21).

*The codes for SymANLS and SymHALS algorithms can be downloaded from `http://mysite.du.edu//~zzhu61/Software.html`

†Matlab code is downloaded from `https://github.com/dakuang/symnmf`.

4.2 Synthetic Data

In this section, we perform on synthetic datasets. We first generate $\widetilde{X} \in \mathbb{R}^{n\times r}$ with entries sampled from a Bernoulli distribution of probability $2/r$. Then generate noise matrix whose entries follow an i.i.d. Gaussian distribution $N \sim \mathcal{N}(0,\sigma^2)$ and set $A = \mathrm{sym}\left(\max\{\widetilde{X}\widetilde{X}^\top + \mathrm{sym}(N)), 0\}\right)$ to ensure symmetry and nonnegativity, where N generated by MATLAB command $\sigma \times \mathrm{randn}(n,n)$. We will consider both $\sigma = 0$ and 0.3 and set $n = \{100, 500, 1000\}$ and $r = \{10, 20\}$ in this section.

For AALM, we take $\Lambda_0 = \frac{1}{r}[(Y_0(X_0)^T - A)X_0 - \rho_0(X_0 - Y_0)]$, $\varepsilon = 1e-7$ and $\varepsilon_1 = 1e-3$. For the inner solver of AALM, we set $\hat{\rho} = \rho_{k-1}$ and $\theta = 1.02$. We take $\epsilon_k = \max\{0.75^{k+1}, 1e-6\}$, or terminate the iteration when the number of steps reaches 100. Table 1 lists the parameter setting of ρ_0 associate with each pair of (n,σ).

Table 1: Penalty parameter ρ_0 for AALM.

σ	$\sigma = 0$			$\sigma = 0.3$		
n	100	500	1000	100	500	1000
ρ_0	40	200	500	100	500	500

Tables 2 and 3 report the average result on $\sigma = 0$ and $\sigma = 0.3$ over 10 tests of CPU Time (cputime), the relative projected gradient (RPG), the number of iterations (ITER), the relative square error ($\mathrm{RSE} = \frac{\|A - X_c X_c^\top\|_F^2}{\|A\|_F^2}$), the final function value ($\mathrm{FV} = \frac{1}{2}\|A - X_c X_c^\top\|_F^2$), and report the average value of $\mathrm{DXY} := \|X_c - Y_c\|_F$ for each nonsymmetric relaxation algorithm, where $X_c \geq 0$ and Y_c are the returned value that generated by each algorithm. It can be seen from Tables 2 and 3 that

a). Table 2, for $r = 10$: according to the column 'FV', AALM has better accuracy in capturing low-rank decomposition. According to the 'RPG' column, all algorithms achieve the desired approximate solution in most cases. According to the 'DXY' column, all nonsymmetric relaxation algorithms can reach good approximation in X_c and Y_c.

b). Table 2, for $r = 20$: As n increases, AALM still yields the lowest function value in less cputime compared to the other methods.

c). We note that the penalty methods SymANLS and SymHALS are sensitive to the penalty parameter ρ, while the Newton-like method PNewton takes more time. NS-SNMF can also obtain a better low-rank approximation, but in most cases the final objective function value is higher than that obtained by AALM.

d). Table 3: similar final objective function values obtained by each method except for $(n,r) = (100, 20)$. Similar to $\sigma = 0$, AALM can reach the approximate stationary point in a shorter time. All algorithms have good approximations in X_c and Y_c.

Table 2: Average results over 10 tests on $\sigma = 0$.

n	Alg.	r = 10						r = 20					
		cputime	RPG	ITER	RSE	FV	DXY	cputime	RPG	ITER	RSE	FV	DXY
100	AALM	0.02	6.75E-08	36	1.10E-14	3.67E-11	4.73E-06	0.03	7.03E-08	34	6.44E-13	1.25E-09	5.82E-06
	NS-SNMF	0.08	9.14E-08	98	2.70E-13	8.81E-10	8.61E-07	0.72	1.89E-06	811	4.23E-04	8.21E-01	2.38E-06
	SymANLS	0.22	9.56E-08	177	3.10E-13	1.01E-09	7.29E-07	1.53	7.31E-07	970	4.23E-04	8.22E-01	8.54E-06
	SymHALS	0.03	9.62E-08	192	3.07E-13	1.00E-09	7.04E-07	0.10	9.86E-08	475	1.98E-12	3.84E-09	1.08E-06
	PNewton	0.14	9.27E-08	258	1.15E-13	3.76E-10	-	0.54	2.06E-07	1059	8.48E-04	1.65E+00	-
500	AALM	0.06	6.55E-08	26	3.24E-14	2.62E-09	2.05E-05	0.09	6.45E-08	22	3.37E-13	1.11E-08	1.54E-04
	NS-SNMF	0.22	8.62E-08	67	5.03E-13	4.02E-08	5.17E-06	0.58	8.80E-08	116	3.74E-12	1.23E-07	8.42E-06
	SymANLS	0.89	9.28E-08	158	6.49E-13	5.18E-08	2.21E-06	2.92	9.57E-08	341	4.57E-12	1.50E-07	3.28E-06
	SymHALS	0.31	9.39E-08	172	6.35E-13	5.07E-08	2.17E-06	0.92	9.52E-08	427	4.50E-12	1.48E-07	3.24E-06
	PNewton	1.49	8.64E-08	138	2.44E-13	1.95E-08	-	3.38	9.22E-08	379	1.26E-12	4.15E-08	-
1000	AALM	0.15	5.17E-08	24	4.71E-14	1.40E-08	4.92E-05	0.21	7.02E-08	17	9.11E-13	1.10E-07	2.71E-04
	NS-SNMF	0.45	8.16E-08	58	5.14E-13	1.53E-07	9.76E-06	1.01	7.84E-08	78	5.06E-12	6.12E-07	1.72E-05
	SymANLS	2.26	9.15E-08	150	7.63E-13	2.27E-07	3.34E-06	6.14	9.59E-08	298	7.88E-12	9.53E-07	5.30E-06
	SymHALS	1.07	9.45E-08	161	8.26E-13	2.46E-07	3.37E-06	2.84	9.55E-08	380	8.04E-12	9.72E-07	5.26E-06
	PNewton	6.55	8.75E-08	115	2.31E-13	6.90E-08	-	8.48	8.23E-08	84	1.85E-12	2.23E-07	-

Figures 1 and 2 show the average convergence results of comparison methods in the logarithm of function value or function value relative square error as cputime changes over 10 tests. It can be seen that in the presence of precise low-rank decomposition, AALM in most cases achieves lower function values than other methods and achieves similar relative square errors in a shorter time. These observations are consistent with the results in Tables 2 and 3.

Table 3: Average results over 10 tests on $\sigma = 0.3$.

n	Alg.	r = 10						r = 20					
		cputime	RPG	ITER	RSE	FV	DXY	cputime	RPG	ITER	RSE	FV	DXY
100	AALM	0.02	9.87E-08	37	3.21E-02	1.09E+02	4.12E-07	0.14	9.96E-08	40	4.16E-02	8.53E+01	6.18E-07
	NS-SNMF	0.16	9.61E-08	173	3.21E-02	1.09E+02	3.85E-07	0.80	9.88E-08	689	4.16E-02	8.53E+01	3.28E-07
	SymANLS	0.49	9.76E-08	302	3.21E-02	1.09E+02	7.78E-07	2.37	9.87E-08	1157	4.14E-02	8.48E+01	1.11E-06
	SymHALS	0.04	9.78E-08	311	3.21E-02	1.09E+02	7.37E-07	0.27	9.90E-08	1386	4.19E-02	8.59E+01	1.09E-06
	PNewton	0.15	9.47E-08	547	3.21E-02	1.09E+02	-	0.59	9.67E-08	1520	4.14E-02	8.48E+01	-
500	AALM	0.09	9.77E-08	30	3.85E-02	3.22E+03	1.51E-06	0.28	9.87E-08	28	6.92E-02	2.51E+03	1.46E-06
	NS-SNMF	0.43	9.56E-08	141	3.85E-02	3.22E+03	1.76E-06	1.47	9.61E-08	221	6.92E-02	2.51E+03	2.85E-06
	SymANLS	1.79	9.79E-08	319	3.85E-02	3.22E+03	2.47E-06	7.04	9.89E-08	602	6.92E-02	2.51E+03	3.52E-06
	SymHALS	0.64	9.82E-08	372	3.85E-02	3.22E+03	2.36E-06	1.35	9.86E-08	649	6.92E-02	2.51E+03	3.46E-06
	PNewton	3.32	9.56E-08	771	3.85E-02	3.22E+03	-	3.84	9.28E-08	443	6.92E-02	2.51E+03	-
1000	AALM	0.20	9.70E-08	28	4.14E-02	1.30E+04	5.68E-06	0.33	9.78E-08	25	7.62E-02	1.02E+04	6.93E-06
	NS-SNMF	1.10	9.72E-08	151	4.14E-02	1.30E+04	2.79E-06	1.99	9.62E-08	148	7.62E-02	1.02E+04	5.55E-06
	SymANLS	5.21	9.84E-08	371	4.14E-02	1.30E+04	3.84E-06	16.14	9.88E-08	668	7.62E-02	1.02E+04	5.72E-06
	SymHALS	2.51	9.85E-08	379	4.14E-02	1.30E+04	3.70E-06	5.07	9.90E-08	672	7.62E-02	1.02E+04	5.67E-06
	PNewton	16.83	9.18E-08	796	4.14E-02	1.30E+04	-	23.38	9.44E-08	586	7.62E-02	1.02E+04	-

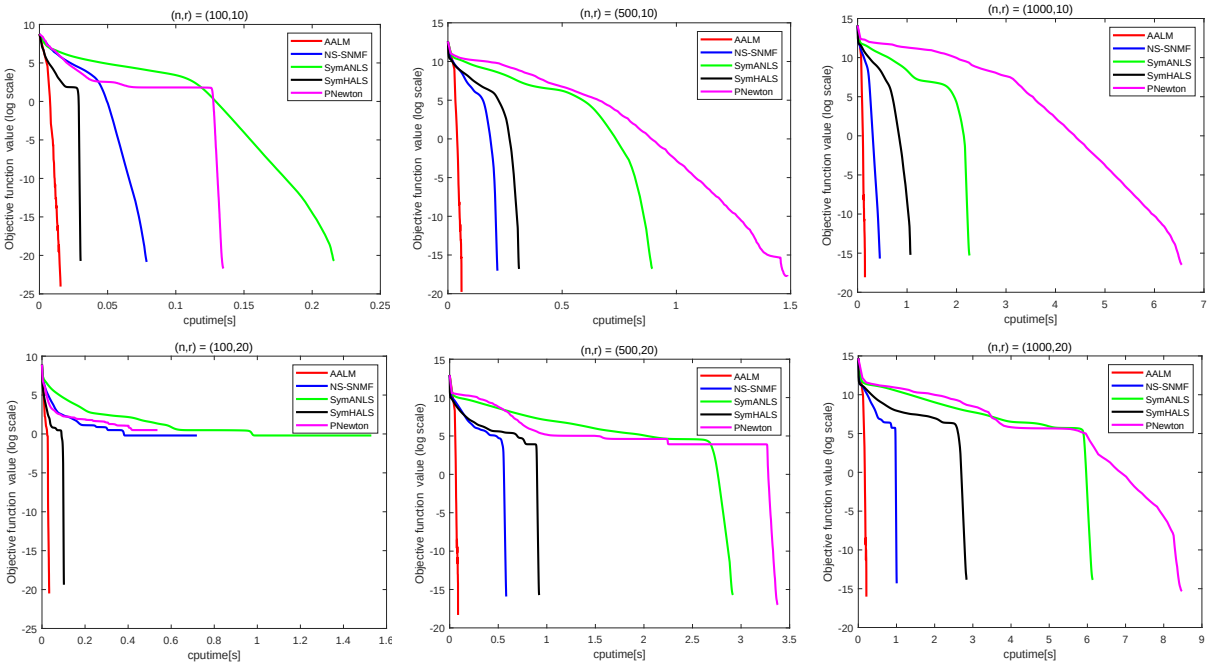

Figure 1: the logarithm of the objective function value (FV) as CPU time increase for Data I. Top $r = 10$: from left to right: $n = 100, 500, 1000$; Bottom $r = 20$: from left to right: $n = 100, 500, 1000$.

5 Conclusion

In this paper, we developed an approximate ALM scheme to solve SymNMF and address the AL subproblem by the PAM method. Faster iteration convergence to the stationary point can be observed on synthetic data.

References

[1] R. Andreani, E. G. Birgin, J. M. Martínez and M. L. Schuverdt. On augmented Lagrangian methods with general lower-level constraints. *SIAM Journal on Optimization*, 18(2008): 1286–1309

[2] H. Attouch, J. Bolte, P. Redont and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality. *Mathematics of operations research*, 35(2010):438–457.

[3] J. Bolte, S. Sabach and M. Teboulle. Proximal alternating linearized minimization for nonconvex and nonsmooth problems. *Mathematical Programming*, 146(2014):459–494.

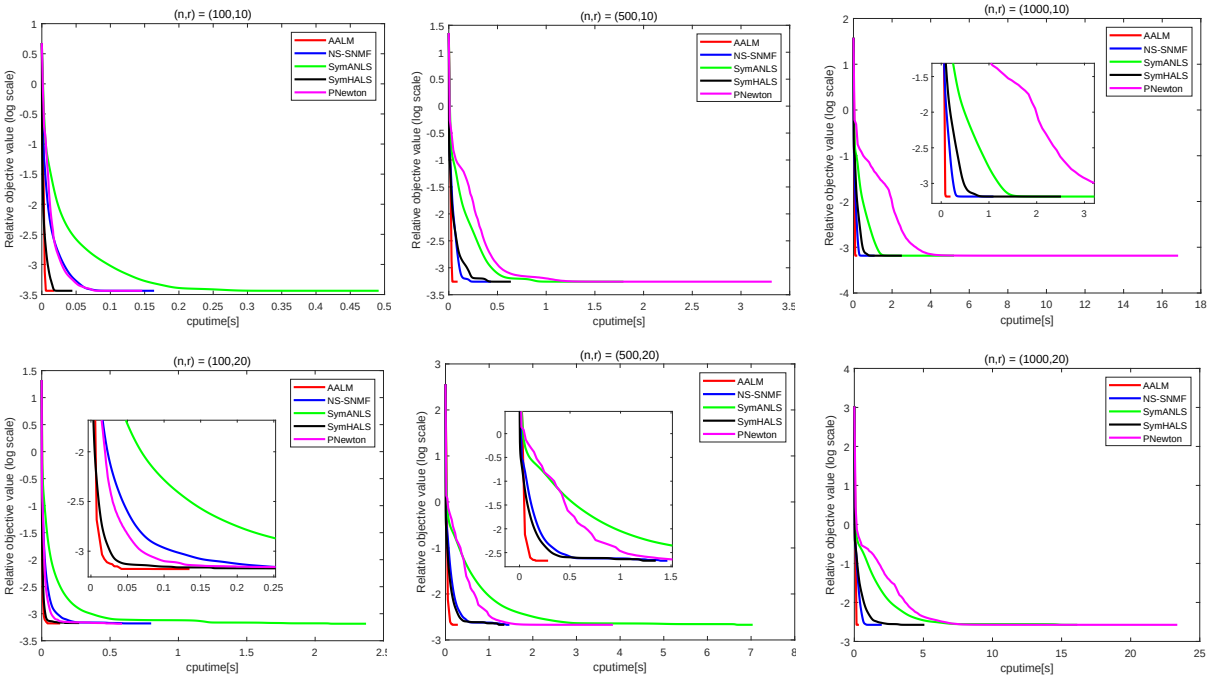

Figure 2: the logarithm of the relative square error (RSE) as CPU time increase for Data II. Top $r = 10$: from left to right: $n = 100, 500, 1000$; Bottom $r = 20$: from left to right: $n = 100, 500, 1000$.

[4] A. Beck. First-Order Methods in Optimization. Society for Industrial and Applied Mathematics, Philadelphia, PA. 2017.

[5] D. Chu, W. Shi, S. Eswar and H. Park. An alternating rank-k nonnegative least squares framework (ARkNLS) for nonnegative matrix factorization. *SIAM Journal on Matrix Analysis and Applications*, 42(2021):1451–1479.

[6] C. Ding, T. Li, W. Peng and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 126–135. Association for Computing Machinery, New York. 2006.

[7] R. A. Dragomir, A. D'Aspremont and J. Bolte. Quartic first-order methods for low-rank minimization *Journal of Optimization Theory and Applications*, 189(2021):341–363.

[8] D. Kuang, C. Ding and H. Park. Symmetric nonnegative matrix factorization for graph clustering. In Proceedings of the 2012 SIAM International Conference on Data Mining (SDM), 106–117. 2012.

[9] D. Kuang, S. Yun and H. Park. SymNMF: nonnegative low-rank approximation of a similarity matrix for graph clustering. *Journal of Global Optimization*, 62(2015):545–574.

[10] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(1999):788–791.

[11] J. Leskovec, K. J. Lang, A. Dasgupta and M. W. Mahoney. Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(2009):29–123.

[12] S. Lu, M. Hong and Z. Wang. A nonconvex splitting method for symmetric nonnegative matrix factorization: Convergence analysis and optimality. *IEEE Transactions on Signal Processing*, 65(2017):3120–3135.

[13] Z. Lu and Y. Zhang. An augmented Lagrangian approach for sparse principal component analysis. *Mathematical Programming*, 135(2012):149–193.

[14] P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(1994):111–126.

[15] R. T. Rockafellar and R. J. B. Wets. Variational Analysis. Grundlehren Der Mathematischen Wissenschaften. Springer-Verlag. 1998.

[16] F. Shahnaz, M. W. Berry, V. Pauca and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2006):373–386.

[17] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering In Proceedings of the 17th International Conference

on Neural Information Processing Systems, 1601–1608. MIT Press, Cambridge, MA, USA. 2004.

[18] H. Zhu, X. Zhang, D. Chu and L. Z. Liao. Nonconvex and nonsmooth optimization with generalized orthogonality constraints: An approximate augmented Lagrangian method. *Journal of Scientific Computing*, 72(2017):331–372.

[19] Z. Zhu, X. Li, K. Liu and Q. Li. Dropping symmetry for fast symmetric nonnegative matrix factorization. In Proceedings of the 32nd International Conference on Neural Information Processing Systems, 5160–5170. Curran Associates Inc., Red Hook, NY, USA. 2018.