NONNEGATIVE MATRIX FACTORIZATION USING ADMM: ALGORITHM AND CONVERGENCE ANALYSIS

Davood Hajinezhad[1], *Tsung-Hui Chang*[2], *Xiangfeng Wang*[3], *Qingjiang Shi*[4], *Mingyi Hong*[1]

[1] Dept. of IMSE, Iowa State Univ., Ames, IA, 50011, USA

[2] School of Sci. & Eng. Chinese Univ. of Hong Kong, Shenzhen Shenzhen, China.

[3] Software Eng. Institute East China Normal University, Shanghai, China

[4] Dept. of Elect. Eng. Zhejiang Sci-Tech Univ., Hangzhou, China

ABSTRACT

The nonnegative matrix factorization (NMF) has been a popular model for a wide range of signal processing and machine learning problems. It is usually formulated as a nonconvex cost minimization problem. This work settles the convergence issue of a popular algorithm based on the alternating direction method of multipliers proposed in Boyd *et al* 2011. We show that the algorithm converges globally to the set of KKT solutions whenever certain penalty parameter ρ satisfies $\rho > 1$. We further extend the algorithm and its analysis to the problem where the observation matrix contains missing values. Numerical experiments on real and synthetic data sets demonstrate the effectiveness of the algorithms under investigation.

Index Terms— Nonnegative Matrix Factorization, ADMM, Convergence Analysis, Nonconvex Optimization.

1. INTRODUCTION

The well-known NMF problem extracts from an observation matrix $M \in \mathbb{R}^{N\times Q}$ two nonnegative factors $X \in \mathbb{R}^{N\times K}$, and $Y \in \mathbb{R}^{K\times Q}$. A popular nonconvex formulation for NMF is given by [1]:

$$\min f(X,Y) = \frac{1}{2}\|XY - M\|_F^2, \quad \text{s.t. } X \geq 0,\ Y \geq 0. \quad (1.1)$$

When M contains missing values, one would like to find the nonnegative factors that complete the matrix. Such nonnegative matrix factorization/completion (NMFC) problem can be formulated as [2]

$$\min \hat{f}(X,Y) = \frac{1}{2}\|\mathcal{P}_\Omega\left(XY - M\right)\|_F^2, \text{ s.t. } X \geq 0,\ Y \geq 0 \quad (1.2)$$

where $\mathcal{P}_\Omega$ denotes the projection on to the index set Ω which contains the known entries. The seminal work of Lee and Seung [1] has motivated a variety of applications of NMF/NMFC, such as text mining [3], pattern discovery [4], bioinformatics [5], as well as clustering [6]; for a recent survey, see [7].

Many efficient algorithms have been proposed for NMF/NMFC. For example the multiplicative update proposed in [1] alternates between solving certain surrogate functions for X and Y, respectively. The convergence of this algorithm is analyzed in [8], but in practice it often converges slowly [7, 9]. The alternating nonnegative least square (ANLS) is another class of useful algorithms, which includes the projected gradient descent method [10], the block principal pivoting method [11], and an algorithm proposed in [12]. Recently, the alternating direction method of multipliers (ADMM) has become a popular framework for NMF. In a highly cited survey [13, Chapter 9.2], Boyd *et al* proposed one of the first ADMM algorithms to solve the nonconvex NMF problem (1.1). At roughly the same time many variants have been developed, each demonstrating encouraging numerical performance; see [2, 9, 14, 15]. Unfortunately, there is a significant gap between the algorithms' good practical performance and our understanding for such behavior – to the best of our knowledge the theoretical convergence of such ADMM based method is still open[1]. It is the aim of this work to partially close this gap.

This work settles the convergence issue of the nonconvex ADMM proposed in [13]. We show that as long as certain penalty parameter is chosen greater than 1, the algorithm globally converges to the set of KKT points of problem (1.1). This result provides theoretical justification for the good practical performance observed for this algorithm. Further, we develop an extension to the aforementioned algorithm for the NFMC problem (1.2). We expect that our analysis technique will serve as the basis for analyzing a much wider range of ADMM based methods for nonconvex matrix factorization.

2. THE ALGORITHM

We begin with reviewing the algorithm proposed in [13, Chapter 9.2] for NMF. Consider the following reformulation of (1.1)

$$\min_{X,Y,Z} \frac{1}{2}\|Z - M\|_F^2, \text{ s.t. } X, Y \geq 0,\ Z = XY, \quad (2.3)$$

where a new variable $Z \in \mathbb{R}^{N\times Q}$ is introduced. The augmented Lagrangian for the above problem is given by

$$L_\rho(X,Y,Z;\Lambda) = \frac{1}{2}\|Z - M\|_F^2 + \langle \Lambda, Z - XY\rangle + \frac{\rho}{2}\|Z - XY\|_F^2.$$

where $\Lambda \in \mathbb{R}^{N\times Q}$ is the dual variable. In [13], an ADMM based algorithm is proposed to solve the nonconvex NMF problem (1.1). The algorithm alternates between updating Y and (X, Z), followed by the update of the dual variable Λ; see the following table[2].

Algorithm 1. ADMM for Problem (1.1)
Initialize: X^0, Z^0, Λ^0
Repeat: Let $r = r + 1$; update Y, (X, Z) and Λ alternatingly by:

$$Y^{r+1} = \arg\min_{Y\geq 0} \frac{\rho}{2}\left\|Z^r - X^r Y + \frac{\Lambda^r}{\rho}\right\|_F^2, \quad (2.4a)$$

$$(X,Z)^{r+1} = \arg\min_{X\geq 0, Z} \frac{1}{2}\|Z - M\|_F^2 + \frac{\rho}{2}\left\|Z - XY^{r+1} + \frac{\Lambda^r}{\rho}\right\|_F^2, \quad (2.4b)$$

$$\Lambda^{r+1} = \Lambda^r + \rho\left(Z^{r+1} - X^{r+1}Y^{r+1}\right). \quad (2.4c)$$

Until Convergence.

[1] A few works such as [2] have analyze the convergence of nonconvex ADMM based on some *nonstandard* assumptions that the successive difference of the iterates goes to zero. These assumptions are made on the algorithm iterates and hence are impossible to verify *a priori*.

[2] Note that this is precisely the algorithm developed in [13, Chapter 9.2], except that we have exchanged the order of Y and (X,Z) update.

This algorithm has very good practical performance and the code can be easily parallelized for high dimensional problems [13]. Unfortunately, despite its good performance, there has been no rigorous convergence analysis available. This is also the case for many ADMM based NMF algorithms that follow suite. The main challenge in analyzing its convergence lies in the *nonconvexity* and *nonseparability* in (1.1)– most of the known convergence analysis for ADMM is only applicable to convex separable problems (see, e.g., [13,16,17]), hence are not applicable in our context. Recently [18] analyzes the convergence of ADMM for a special nonconvex *separable* global consensus problem, but the analysis again does not apply here (the nonconvexity and nonseparability of the NMF/NMFC problem arise from its bi-convex structure, which cannot be handled by [18]).

Algorithm 1 can be easily extended to handle the NMFC problem. Consider the following equivalent reformulation of (1.2)

$$\begin{aligned} \min_{X,Y,Z,W} \quad & \frac{1}{2}\|Z-W\|_F^2 \\ \text{s.t.} \quad & X \geq 0,\ Y \geq 0, \\ & Z = XY,\ \mathcal{P}_\Omega\left(W - M\right) = 0, \end{aligned} \tag{2.5}$$

where new variables $Z \in \mathbb{R}^{N\times Q}$ and $W \in \mathbb{R}^{N\times Q}$ are introduced. The augmented Lagrangian for the above problem is given by

$$\hat{L}_{\hat{\rho}}(X,Y,Z,W;\Lambda) = \frac{1}{2}\|Z-W\|_F^2 + \langle \Lambda, Z-XY\rangle + \frac{\hat{\rho}}{2}\|Z-XY\|_F^2.$$

By grouping the variables into (Y,W) and (X,Z), a direct application of the conventional ADMM yields the following algorithm.

Algorithm 2. ADMM for Problem (1.2)
Initialize: X^0, Z^0, Λ^0
Repeat: Let $r = r+1$; update (Y,W), (X,Z) and Λ alternatingly by:

$$(Y,W)^{r+1} = \underset{Y\geq 0, \mathcal{P}_\Omega(W-M)=0}{\arg\min} \frac{1}{2}\|Z^r - W\|_F^2 + \frac{\hat{\rho}}{2}\|Z^r - X^rY + \Lambda^r/\hat{\rho}\|_F^2 \tag{2.6a}$$

$$(X,Z)^{r+1} = \underset{X\geq 0, Z}{\arg\min} \frac{1}{2}\|Z - W^{r+1}\|_F^2 + \frac{\hat{\rho}}{2}\|Z - XY^{r+1} + \Lambda^r/\hat{\rho}\|_F^2 \tag{2.6b}$$

$$\Lambda^{r+1} = \Lambda^r + \hat{\rho}\left(Z^{r+1} - X^{r+1}Y^{r+1}\right). \tag{2.6c}$$

Until Convergence.

It is easy to observe that Algorithm 2 reduces to Algorithm 1 if M is a full matrix (in which case the solution of (2.6a) yields $W^r = M,\ \forall\, r$). Also note that the subproblems (2.6a) and (2.6b) are both convex therefore can be solved by general purpose solvers such as CVX [19]. Later in the simulation section we will discuss a particular efficient implementation for solving these subproblems.

3. CONVERGENCE ANALYSIS

We begin analyzing the convergence of Algorithm 1 and 2. Due to its generality we will focus on proving the latter algorithm, and make comments on the former whenever necessary. To highlight, we provide below the main steps of the analysis:

1. Bound the size of the successive difference of the multipliers by that of the successive difference of the primal variables.
2. Show that the augmented Lagrangian is lower bounded and decreasing.
3. Combine the previous two steps and show convergence.

We note that our analysis differs from that of [18], because we have to deal with the following two challenging issues: *1)* The *nonconvexity* in the constraint $Z = XY$; *2)* The absence of the Lipschitzian gradient for $L_\rho(X,Y,Z;\Lambda)$ and $\hat{L}_{\hat{\rho}}(X,Y,Z,W;\Lambda)$ [3].

The following lemma represents the first step of the proof.

Lemma 3.1 *We have the following estimates of* $\|\Lambda^{r+1} - \Lambda^r\|$*:*

1. *For Algorithm 1, the following is true*

$$\|\Lambda^{r+1} - \Lambda^r\|_F^2 \leq \|Z^{r+1} - Z^r\|_F^2 \tag{3.7}$$

2. *For Algorithm 2, the following is true*

$$\|\Lambda^{r+1} - \Lambda^r\|_F^2 \leq 2(\|Z^{r+1} - Z^r\|_F^2 + \|W^{r+1} - W^r\|_F^2) \tag{3.8}$$

Proof. The optimality condition of (2.6b) for variable Z is given by

$$Z^{r+1} - W^{r+1} + \hat{\rho}\left(Z^{r+1} - X^{r+1}Y^{r+1} + \Lambda^r/\hat{\rho}\right) = 0. \tag{3.9a}$$

This condition combined with (2.6c) can be equivalently written as

$$W^{r+1} - Z^{r+1} = \Lambda^{r+1}. \tag{3.10}$$

Then it is easy to see that we have

$$\|\Lambda^{r+1} - \Lambda^r\|_F^2 \leq 2(\|Z^{r+1} - Z^r\|_F^2 + \|W^{r+1} - W^r\|_F^2). \tag{3.11}$$

The lemma is proved. **Q.E.D.**

Our second step shows that the augmented Lagrangian functions are lower bounded and decreasing, provided that the penalty parameters ρ and $\hat{\rho}$ are chosen sufficiently large. Note that the analysis below explicitly makes use of the property of the nonconvex quadratic function of (1.1), therefore it is different from what is presented in [18]. For notational simplicity, define $S^r := \{X^r, Y^r, Z^r, W^r\}$.

Lemma 3.2 *We have the following estimates for the decent of the augmented Lagrangian function*

1. *For Algorithm 1, if* $\rho > 1$*, then for some* $c_1, c_2, c_3 > 0$*,*

$$\begin{aligned} & L_\rho(X^{r+1}, Y^{r+1}, Z^{r+1}; \Lambda^{r+1}) - L_\rho(X^r, Y^r, Z^r; \Lambda^r) \\ & \leq -c_1\|Z^{t+1} - Z^t\|_F^2 - c_2\|X^r(Y^{r+1} - Y^r)\|_F^2 \\ & - c_3\|(X^{r+1} - X^r)Y^{r+1}\|_F^2. \end{aligned}$$

Further, we have that $L_\rho(X^r, Y^r, Z^r; \Lambda^r) \geq 0$*.*

2. *For Algorithm 2, if* $\rho > 4$*, then for some* $\hat{c}_1, \hat{c}_2, \hat{c}_3, \hat{c}_4 > 0$*,*

$$\begin{aligned} & \hat{L}_{\hat{\rho}}(S^{r+1}; \Lambda^{r+1}) - \hat{L}_{\hat{\rho}}(S^r; \Lambda^r) \\ & \leq -\hat{c}_1\|Z^{t+1} - Z^t\|_F^2 - \hat{c}_2\|W^{t+1} - W^t\|_F^2 \\ & - \hat{c}_3\|X^r(Y^{r+1} - Y^r)\|_F^2 - \hat{c}_4\|(X^{r+1} - X^r)Y^{r+1}\|_F^2. \end{aligned}$$

Further, we have $\hat{L}_{\hat{\rho}}(S^r; \Lambda^r) \geq 0$*.*

Proof. First let us examine the (Y,W)-step (2.6a). We have

$$\begin{aligned} & \hat{L}_{\hat{\rho}}(X^r, Y^{r+1}, Z^r, W^{r+1}; \Lambda^r) - \hat{L}_{\hat{\rho}}(S^r; \Lambda^r) \\ & = \langle W^{r+1} - Z^r, W^{r+1} - W^r\rangle - \frac{1}{2}\|W^{r+1} - W^r\|_F^2 \\ & + \left\langle \hat{\rho}\left(X^rY^{r+1} - \Lambda^r/\hat{\rho} - Z^r\right), X^r(Y^{r+1} - Y^r)\right\rangle \\ & - \frac{\hat{\rho}}{2}\|X^r(Y^{r+1} - Y^r)\|_F^2 \\ & \leq -\frac{1}{2}\|W^{r+1} - W^r\|_F^2 - \frac{\hat{\rho}}{2}\|X^r(Y^{r+1} - Y^r)\|_F^2 \end{aligned} \tag{3.12}$$

[3] The gradient of $L(X,Y,Z,W;\Lambda)$ with respect to either X or Y is not Lipschitz continuous, because the potential unboundedness of X and Y.

where the first equality comes from the fact that the second order Taylor expansion for a quadratic function is exact. Note that here the expansion is performed on the variable X^rY. The last inequality is due to the optimality condition of problem (2.6a).

Next let us examine the (X, Z)-step (2.6b). Similarly, we have

$$\begin{aligned}
&\hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^r) - \hat{L}_{\hat{\rho}}(X^r, Y^{r+1}, Z^r, W^{r+1};\Lambda^r)\\
&= \left\langle (Z^{r+1} - W^{r+1}) + \hat{\rho}\left(Z^{r+1} - X^{r+1}Y^{r+1} + \Lambda^r/\hat{\rho}\right), Z^{r+1} - Z^r\right\rangle\\
&\quad - \left\langle \hat{\rho}(Z^{r+1} - X^{r+1}Y^{r+1} + \Lambda^r/\hat{\rho})(Y^{r+1})^T, X^{r+1} - X^r\right\rangle\\
&- \frac{1+\hat{\rho}}{2}\|Z^{r+1} - Z^r\|_F^2 - \frac{\hat{\rho}}{2}\|(X^{r+1} - X^r)Y^{r+1}\|_F^2\\
&\le -\frac{1+\hat{\rho}}{2}\|Z^{r+1} - Z^r\|_F^2 - \frac{\hat{\rho}}{2}\|(X^{r+1} - X^r)Y^{r+1}\|_F^2. \qquad (3.13)
\end{aligned}$$

Utilizing the above two inequalities, we can bound the successive difference of the augmented Lagrangian by

$$\begin{aligned}
&\hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^{r+1}) - \hat{L}_{\hat{\rho}}(S^r;\Lambda^r)\\
=&\hat{L}_{\hat{\rho}}(X^r, Y^{r+1}, Z^r, W^{r+1};\Lambda^r) - \hat{L}_{\hat{\rho}}(S^r;\Lambda^r)\\
&+ \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^r) - \hat{L}_{\hat{\rho}}(X^r, Y^{r+1}, Z^r, W^{r+1};\Lambda^r)\\
&+ \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^{r+1}) - \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^r)\\
\overset{(3.12),(2.6c)}{\le}& -\frac{1}{2}\|W^{r+1} - W^r\|_F^2 - \frac{\rho}{2}\|X^r(Y^{r+1} - Y^r)\|_F^2\\
&+ \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^r) - \hat{L}_{\hat{\rho}}(X^r, Y^{r+1}, Z^r, W^{r+1};\Lambda^r)\\
&+ \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^{r+1}) - \hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^r)\\
\overset{(3.13)}{\le}& -\frac{\hat{\rho}+1}{2}\|Z^{r+1} - Z^r\|_F^2 - \frac{1}{2}\|W^{r+1} - W^r\|_F^2\\
&- \frac{\hat{\rho}}{2}\|X^r(Y^{r+1} - Y^r)\|_F^2 - \frac{\hat{\rho}}{2}\|(X^{r+1} - X^r)Y^{r+1}\|_F^2\\
&+ \frac{1}{\hat{\rho}}\|\Lambda^{r+1} - \Lambda^r\|_F^2. \qquad (3.14)
\end{aligned}$$

Utilizing (3.11), we have

$$\begin{aligned}
&\hat{L}_{\hat{\rho}}(S^{r+1};\Lambda^{r+1}) - \hat{L}_{\hat{\rho}}(S^r;\Lambda^r)\\
&\le -\left(\frac{\hat{\rho}+1}{2} - \frac{2}{\hat{\rho}}\right)\|Z^{r+1} - Z^r\|_F^2 - (\frac{1}{2} - \frac{2}{\hat{\rho}})\|W^{r+1} - W^r\|_F^2\\
&- \frac{\hat{\rho}}{2}\|X^r(Y^{r+1} - Y^r)\|_F^2 - \frac{\hat{\rho}}{2}\|(X^{r+1} - X^r)Y^{r+1}\|_F^2 \qquad (3.15)
\end{aligned}$$

Therefore if $\hat{\rho} > 4$, the augnmented Lagrangian is decreasing.

Next we show the lower boundedness of the augmented Lagrangian. We have

$$\begin{aligned}
&\hat{L}_{\hat{\rho}}(S^r;\Lambda^r)\\
&= \frac{1}{2}\|Z^r - W^r\|_F^2 + \langle\Lambda^r, Z^r - X^rY^r\rangle + \frac{\hat{\rho}}{2}\|Z^r - X^rY^r\|_F^2\\
&\overset{(3.10)}{=} \frac{1}{2}\|Z^r - W^r\|_F^2 + \langle W^r - Z^r, Z^r - X^rY^r\rangle + \frac{\hat{\rho}}{2}\|Z^r - X^rY^r\|_F^2\\
&= \frac{\hat{\rho}-1}{2}\|Z^r - X^rY^r\|_F^2 + \frac{1}{2}\|W^r - Z^r + Z^r - X^rY^r\|_F^2\\
&= \frac{\hat{\rho}-1}{2}\|Z^r - X^rY^r\|_F^2 + \frac{1}{2}\|W^r - X^rY^r\|_F^2 \ge 0. \qquad (3.16)
\end{aligned}$$

The claim is proved. We mention that similar analysis can be done for Algorithm 1, however in that case the range of the penalty parameter can be made even wider ($\rho > 1$ instead of $\hat{\rho} > 4$). **Q.E.D.**

Our final step collects all the results we have so far to show convergence. The following is our main result.

Theorem 3.1 *We have the following convergence results:*

1. *For Algorithm 1, if $\rho > 1$, then the primal gap is satisfied in the limit, i.e.,*

$$\lim_{r\to\infty} \|X^{r+1}Y^{r+1} - Z^{r+1}\|_F \to 0. \qquad (3.17)$$

   Further, every limit point of the iterates (X^r, Y^r) is a KKT point of the problem (1.1)*.*

2. *For Algorithm 2: If $\hat{\rho} > 4$, then the primal gap is satisfied in the limit, i.e.,*

$$\lim_{r\to\infty} \|X^{r+1}Y^{r+1} - Z^{r+1}\|_F \to 0. \qquad (3.18)$$

   Further, every limit point of the iterates (X^r, Y^r) is a KKT point of the original problem (1.2)*.*

Proof. We focus on analyzing the second claim. When $\hat{\rho} > 4$, by (3.15) we have

$$\begin{aligned}
&Z^{r+1} - Z^r \to 0, \quad X^r(Y^{r+1} - Y^r) \to 0,\\
&(X^{r+1} - X^r)Y^{r+1} \to 0, \quad W^{r+1} - W^r \to 0, \qquad (3.19)
\end{aligned}$$

By Lemma 3.1, we have: $\Lambda^{r+1} - \Lambda^r \to 0$, which further implies

$$X^{r+1}Y^{r+1} - Z^{r+1} \to 0. \qquad (3.20)$$

Once the constraint violation is shown to go to zero, the rest of the proof simply involves in checking the KKT solution of problem (1.2). Due to space limitation we will not show them here. **Q.E.D.**

4. NUMERICAL RESULTS

In this section we compare the performance of Algorithms 1-2 with some existing methods for NMF/NMFC. Our experiments are performed using Matlab 2013a on a PC with 8GB memory and Intel Core i5-4690 CPU.

4.1. Procedures for solving the subproblems

To efficiently implement Algorithm 1 and 2, we use a procedure inspired by the recent work [12] to solve the convex subproblems (2.4a), (2.4b) and (2.6a), (2.6b). Below we outline the procedure for solving (2.6a). Procedures for the rest of the subproblems can be developed similarly.

First, we reformulate the subproblem (2.6a) for updating (Y, W) by introducing a new variable $\hat{Y}$:

$$\begin{aligned}
\min_{Y,W,\hat{Y}}\ & \frac{1}{2}\|Z^r - W\|_F^2 + \frac{\hat{\rho}}{2}\|Z^r - X^rY + \Lambda^r/\hat{\rho}\|_F^2\\
\text{s.t.}\ \ & Y = \hat{Y}, \quad \hat{Y} \ge 0\\
& \mathcal{P}_\Omega(W - M) = 0.
\end{aligned}$$

The augmented Lagrangian for the above problem is given by

$$\begin{aligned}
\tilde{L}(Y, W, \hat{Y}; \hat{U}) = &\frac{1}{2}\|Z^r - W\|_F^2 + \frac{\hat{\rho}}{2}\|Z^r - X^rY + \Lambda^r/\hat{\rho}\|_F^2\\
&+ \langle\hat{U}, Y - \hat{Y}\rangle + \frac{\alpha}{2}\|Y - \hat{Y}\|_F^2
\end{aligned}$$

The ADMM steps are summarized in the following table.

Algorithm 3. ADMM for Problem (2.6a)
Input: $M, Z^r, X^r, \Lambda^r, \hat{U}, \Omega, \alpha, \rho$
Initialize: Y^0, compute $P = (\hat{\rho}(X^r)^T X^r + \alpha I)^{-1}$
Repeat

$$S1: W \leftarrow Z^r + \mathcal{P}_\Omega(M - Z^r)$$
$$S2: \hat{Y} \leftarrow \max(0, Y + \hat{U}/\alpha)$$
$$S3: Y \leftarrow P\left(\hat{\rho}(Z^r + \Lambda^r/\hat{\rho}) + (\hat{Y} - \hat{U}/\alpha)\right)$$
$$S4: \hat{U} \leftarrow \hat{U} + \alpha(Y - \hat{Y})$$

Until Convergence.
Output: $\hat{Y}, W, \hat{U}$

We note that the algorithm involves performing the inversion of a $K \times K$ matrix $\hat{\rho}(X^r)^T X^r + \alpha I$ once, an operation that is relatively easy because in most practical NMF/NMFC problems we have $K \ll \min\{Q, N\}$.

A similar procedure can be developed for solving the subproblem (2.6b), by introducing a variable $\hat{X}$ to handle the constraint $X \geq 0$. The associated augmented Lagrangian is given by

$$\bar{L}(Z, X, \hat{X}; \hat{V}) = \frac{1}{2}\|Z - W^{r+1}\|_F^2 + \frac{\hat{\rho}}{2}\|Z - XY^{r+1} + \Lambda^r/\hat{\rho}\|_F^2 + \langle \hat{V}, X - \hat{X}\rangle + \frac{\beta}{2}\|X - \hat{X}\|_F^2$$

We omit the details due to space limitation.

4.2. The Performance of Algorithm 1

In this subsection we compare the performance of Algorithm 1 with the following algorithms for solving (1.1): *1)* the multiplicative updating rule (MULT) [1]; *2)* the AO-ADMM proposed recently by Huang *et al* [12]; *3)* The AO-BPP proposed by Kim *et al* [11]; *4)* the ADM method proposed by Xu *et al* [2].

We choose $\rho = 1.1$, which satisfies the condition given in Theorem (3.1). Furthermore, we choose $\alpha = \|W\|_F$, and $\beta = 1$, for the subproblems. The stopping criteria for Algorithm 1 is given by: $L_\rho(S^{r+1}; \Lambda^{r+1}) - L_\rho(S^r; \Lambda^r) \leq 10^{-4}$, $\frac{\|M - XY\|_F}{\|M\|_F} \leq 10^{-4}$. First we randomly generate M which satisfies $M = WH + N$, where $W \in \mathbb{R}^{N \times R}$, and $H \in \mathbb{R}^{R \times Q}$ are random matrices with iid entries generated from $\mathsf{Uniform}(0, 1)$; $N \in \mathbb{R}^{N \times Q}$ is a random matrix with iid entries generated from $\mathcal{N}(0, 0.01^2)$. Different algorithms are compared in terms of the quality of solutions as well as their convergence speed. In this experiment we set the problem size $N = Q = 5000$, $R = 1000$, and $K = 200$. The results which are averaged over 50 independent trials are summarized in Table 1. As we can observe, Algorithm 1 outperforms other algorithms for this specific test problem in terms of both absolute error of the final solution as well as the run time. We further demonstrate the convergence speed of different algorithms by testing on problems of different sizes. We set $N = Q \in \{100, 200, \cdots, 1000\}$, $K = Q/2$, and $R = Q/10$. The results are shown in Fig. 1. The x-axis displays the size of the problem (N, Q), and y-axis displays running time. The results are averaged over 50 independent trials.

Table 1. The performance of Algorithm 1 on synthetic data sets

Method	$\|Y - WH\|_F$	Run Time (s)	Iterations
Algorithm 1	**10176**	**80**	**22**
AO_ADMM [12]	10187	129	39
ADM [2]	10191	500	242
AO_BPP [11]	10179	134	23
MULT [1]	19311	497	500

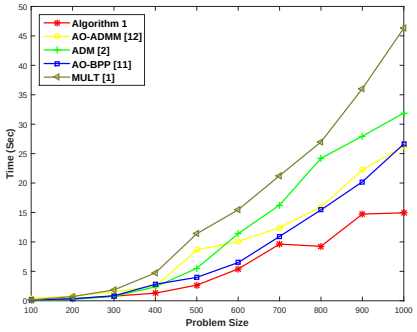

Fig. 1. The convergence speed of different algorithms

From Figure 1 it can be seen that Algorithm 1 converges faster compared with other algorithms, especially for the large-size problems. Also note that the behavior of the objective values of different algorithms are similar as in the previous case (omitted due to space limitation).

Next we demonstrate the performance of various algorithms on the ORL face data set [20]. In this case the data matrix M is a 10304×400 matrix, each column of which is a picture with 112×92 pixels. We apply different algorithms on this data set to get a nonnegative basis matrix X and a nonnegative coefficient matrix Y such that $M \approx XY$. The results are given in Table 2. We observe that Algorithm 1 enjoys a slight advantage over the rest of the methods.

Table 2. The performance of Algorithm 1 on ORL face data set

Method	$\|Y - WH\|_F$	Run Time (s)	Iterations
Algorithm 1	**90**	**185**	109
AO_ADMM [12]	91	202	46
ADM [2]	646	301	500
AO_BPP [11]	94	202	**23**
MULT [1]	104	246	500

4.3. The Performance of Algorithm 2

In this subsection we compare the performance of Algorithm 2 with the algorithm in [2], both of which deal with problems with missing values in the observation. The data matrices we use are generated similarly as in the previous section, except that the entries of M are sampled uniformly according to different sample rates (percentages of known entries). Specifically we set $N = Q = 3000$, $R = 1000$, $K = 300$ and the comparison results are reported on Table 3. From Table 3 we can see that Algorithm 2 has significantly better performance than the ADM proposed in [2].

Table 3. The performance of Algorithm 2 on synthetic data sets

Algorithm	Algorithm 2			ADM [2]		
Sample Rate	25%	50%	75%	25%	50%	75%
Absolute Error	**2875**	**3679**	**4827**	2922	3712	4881
Run Time (S)	**110**	**88**	**54**	139	92	131
Iteration	**310**	**75**	**40**	406	230	249

Acknowledgement: The authors would like to thank Stephen Boyd from Stanford and Kejun Huang from University of Minnesota for helpful comments and discussions.

5. REFERENCES

[1] D.D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, pp. 556–562. MIT Press, 2001.

[2] Y. Xu, W. Yin, Z. Wen, and Y. Zhang, "An alternating direction algorithm for matrix completion with nonnegative factors," *Journal of Frontiers of Mathematics in China, Special Issues on Computational Mathematics*, pp. 365–384, 2011.

[3] V. P. Pauca, F. Shahnaz, M. W. Berry, and R. J. Plemmons, *Text Mining using Non-Negative Matrix Factorizations*, chapter 45, pp. 452–456.

[4] J-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Metagenes and molecular pattern discovery using matrix factorization," *Proceedings of the National Academy of Sciences*, vol. 101, no. 12, pp. 4164–4169, 2004.

[5] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinformatics*, vol. 23, no. 12, pp. 1495–1502, 2007.

[6] A. Caner Turkmen, "A review of nonnegative matrix factorization methods for clustering," 2015, Preprint, available at arXiv:1507.03194v2.

[7] N. Gillis, "The why and how of nonnegative matrix factorization," 2015, Book Chapter available at arxiv: 1401.5226v2.

[8] E..F Gonzalez and Y. Zhang, "Accelerating the leeseung algorithm for non-negative matrix factorization," *Department of Computational and Applied Mathematics, Rice University, Technical report*, 2005.

[9] D. L. Sun and C. Fevotte, "Alternating direction method of multipliers for non-negative matrix factorization with the beta-divergence," in *the Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2014.

[10] C. H. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Comput.*, vol. 19, no. 10, pp. 2756–2779, Oct. 2007.

[11] J. Kim and H. Park, "Fast nonnegative matrix factorization: An active-set-like method and comparisons," *SIAM J. Sci. Comput.*, vol. 33, no. 6, pp. 3261–3281, Nov. 2011.

[12] K. Huang, N.D. Sidiropoulos, and A.P.Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *arXiv preprint arXiv:1506.04209*, 2015.

[13] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[14] R. Zdunek, "Alternating direction method for approximating smooth feature vectors in nonnegative matrix factorization," in *Machine Learning for Signal Processing (MLSP), 2014 IEEE International Workshop on*, Sept 2014, pp. 1–6.

[15] D. Song, D. A. Meyer, and Martin M. R. Min, "Fast nonnegative matrix factorization with rank-one admm," *NIPS 2014 Workshop on Optimization for Machine Learning (OPT2014)*, 2014.

[16] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena-Scientific, second edition, 1999.

[17] J. Eckstein and D. P. Bertsekas, "On the Douglas-Rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, no. 1, pp. 293–318, 1992.

[18] M. Hong, Z.-Q. Luo, and M. Razaviyayn, "Convergence analysis of alternating direction method of multipliers for a family of nonconvex problems," 2014, technical report, University of Minnesota.

[19] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 1.21," Apr. 2011.

[20] AT&T laboratories, "Cambridge orl database of faces," Available at: http://www.uk.research.att.com/facedatabase.html.