Nonnegative Low Rank Matrix Approximation for Nonnegative Matrices

Guang-Jing Song* Michael K. Ng†

December 17, 2019

Abstract

This paper describes a new algorithm for computing Nonnegative Low Rank Matrix (NLRM) approximation for nonnegative matrices. Our approach is completely different from classical nonnegative matrix factorization (NMF) which has been studied for more than twenty five years. For a given nonnegative matrix, the usual NMF approach is to determine two nonnegative low rank matrices such that the distance between their product and the given nonnegative matrix is as small as possible. However, the proposed NLRM approach is to determine a nonnegative low rank matrix such that the distance between such matrix and the given nonnegative matrix is as small as possible. There are two advantages. (i) The minimized distance by the proposed NLRM method can be smaller than that by the NMF method, and it implies that the proposed NLRM method can obtain a better low rank matrix approximation. (ii) Our low rank matrix admits a matrix singular value decomposition automatically which provides a significant index based on singular values that can be used to identify important singular basis vectors, while this information cannot be obtained in the classical NMF. The proposed NLRM approximation algorithm was derived using the alternating projection on the low rank matrix manifold and the non-negativity property. Experimental results are presented to demonstrate the above mentioned advantages of the proposed NLRM method compared the NMF method.

1 Introduction

Nonnegative data matrices appear in many data analysis applications. For instance, in image analysis, image pixel values are nonnegative and the associated nonnegative image data matrices can be formed for clustering and recognition [5, 9, 10, 11, 15, 16, 19, 24, 25, 30, 34]. In text mining, the frequencies of terms in documents...
are nonnegative and the resulted nonnegative term-to-document data matrices can be constructed for clustering [3, 22, 27, 32]. In bioinformatics, nonnegative gene expression values are studied and nonnegative gene expression data matrices are generated for diseases and genes classification [6, 8, 11, 17, 18, 26, 31]. Nonnegative Low Rank Matrix (NLRM) approximation for nonnegative matrices play a key role in all these applications. Its main purpose is to identify a latent feature space for objects representation. The classification, clustering or recognition analysis can be done by using these latent features. Lee and Seung [19] proposed and developed Nonnegative Matrix Factorization (NMF) algorithms, and demonstrated that NMF has part-based representation which can be used for intuitive perception interpretation.

1.1 Related Work

NMF has emerged in 1994 by Paatero and Tapper [35] for performing environmental data analysis. The purpose of NMF is to decompose an input $m$-by-$n$ nonnegative matrix $A \in \mathbb{R}^{m \times n}_+$ into $m$-by-$r$ nonnegative matrix $B \in \mathbb{R}^{m \times r}_+$ and $r$-by-$n$ nonnegative matrix $C \in \mathbb{R}^{r \times n}_+$:

$$A \approx BC,$$

and

$$\min_{B,C \geq 0} \| A - BC \|_F^2,$$  \hspace{1cm} (1)

where $B, C \geq 0$ means that each entry of $B$ and $C$ is nonnegative, $\| \cdot \|_F$ is the Frobenius norm of a matrix, and $r$ (the low rank value) is smaller than $m$ and $n$. For simplicity, we assume that $m \geq n$. Several researchers have proposed and developed algorithms for determining such nonnegative matrix factorization in the literature. For instance, Lee and Seung [19, 20] proposed to solve NMF by using the multiplicative update algorithm by finding both $B$ and $C$ iteratively. Also Yuan and Oja [33] considered and studied a projective nonnegative matrix factorization and proposed the following minimization problem:

$$\min_{B \geq 0} \| A - BB^T A \|_F^2,$$

where $B^T$ is the transpose of $B$. In the optimization problem, it is required to find a projection matrix $BB^T$ such that the difference between the given nonnegative matrix $A$ and its projection $BB^T A$ is as small as possible.

We remark that there can be many possible solutions in (1). In practice, it is necessary to impose additional constraints for finding NMF. In some applications, orthogonality, sparsity and/or smoothness constraints on $B$ and/or $C$ are incorporated in (1). Because of these constraint formulations, many optimization techniques have been designed to solve these minimization problems. For example, the multiplicative update algorithms [7, 8, 10] are revised to deal with tackle these constraint minimization problems.
1.2 The Contribution

The proposed Nonnegative Low Rank Matrix (NLRM) approximation method is completely different from classical NMF. Here NLRM approximation is to find a nonnegative low rank matrix \( X \) such that \( X \approx A \) such that their difference is as small as possible. Mathematically, it can be formulated as the following optimization problem

\[
\min_{\text{rank}(X)=r,X \geq 0} \| A - X \|_F^2. \tag{2}
\]

There are two advantages in the proposed NLRM method.

- It is obvious in (1) that when \( B \) and \( C \) are nonnegative, then the resulting matrix \( BC \) is also nonnegative. But these constraints are more restricted that that required in (2). Instead of using NMF in (1), we study NLRM in (2). The distance \( \| A - X \|_F^2 \) by the proposed NLRM method can be smaller than \( \min_{B,C \geq 0} \| A - BC \|_F^2 \) by the NMF method. It implies that the proposed NLRM method can obtain a better low rank matrix approximation.

- The proposed NLRM approximation admits a matrix singular value decomposition, i.e.,

\[
X = U \Sigma V^T, \tag{3}
\]

where \( U \) is an \( m \)-by-\( n \) matrix, \( \Sigma \) is an \( n \)-by-\( n \) diagonal matrix, and \( V^T \) is also an \( n \)-by-\( n \) matrix. The columns of \( U \) are called the left singular vectors of the singular value decomposition \( \{ u_i \}_{i=1}^m \). These left singular vectors form an orthonormal basis system in \( \mathbb{R}^{m \times m} \) such that \( u_i^T u_j = 1 \) if \( i = j \), otherwise 0. The rows of \( V^T \) refer to the elements of the right singular vectors of the singular value decomposition \( \{ v_i \}_{i=1}^n \). These right singular vectors also form an orthonormal basis system in \( \mathbb{R}^{n \times n} \) such that \( v_i^T v_j = 1 \) if \( i = j \), otherwise 0. The diagonal elements of \( \Sigma = \text{diag}(\sigma_1, \sigma_2, \cdots , \sigma_n) \) are called the singular values. As \( X \) is a rank \( r \) matrix, we have \( \sigma_i \geq 0 \) for \( 1 \leq i \leq r \) and \( \sigma_i = 0 \) for \( r + 1 \leq i \leq n \). The ordering of the singular values follows the descending order, i.e., \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \). We remark that both \( U \) and \( V \) are not necessary to be nonnegative, but the resulting matrix \( X = U \Sigma V^T \) must be non-negative. Therefore, this decomposition is different from principal component analysis as there is such requirement in principal component analysis. According to the singular value decomposition of \( X \), the proposed method can identify important singular basis vectors based on singular values. In the classical NMF, this information cannot be obtained directly.

Our experimental results are presented to demonstrate the above mentioned advantages of the proposed NLRM method compared the NMF method.

The paper is organized as follows. In Section 2, we present our algorithm and show the convergence. In Section 3, numerical results are presented to demonstrate the proposed algorithm. Finally, some concluding remarks are given in Section 4.
2 The Optimization on Manifolds

Constrained optimization is quite well studied as an area of research, and many powerful methods are proposed to solve the general problems in that area. In some special cases, the constrain sets possess particularly rich geometric properties, i.e., they are manifolds in the meaning of classical differential geometry. Then some constrained optimization problems can be rewritten as optimizing a real-valued function \( f(x) \) on a manifold \( M \):

\[
\min_{x \in M} f(x). \tag{4}
\]

Here, \( M \) can be the Stiefel manifold, the Grassmann manifold and the fixed rank matrices manifold and so on. In order to better understand manifolds and some related definitions, e.g., charts, atlases and tangent spaces, we refer to [21] and the references therein. In general, the dimensions that some classical constrained techniques work are much bigger than the corresponding manifold (see e.g., [1] [28]).

2.1 The Algorithm

Alternating projection method is popular in searching a point in the intersection of convex sets because of its simplicity and intuitive appeal. Its basic idea is iteratively projecting a point one set and then the other. In contrast to the well known cases, this paper concerns with the extensions of convex sets to non-convex sets. Here, one set is the \( m \times n \) fixed rank matrices manifold

\[
M_r := \{ X \in \mathbb{R}^{m \times n}, \text{rank}(X) = r \}, \tag{5}
\]

and the other one is the convex set of \( m \times n \) nonnegative matrices

\[
M_n := \{ X \in \mathbb{R}^{m \times n}, X_{i,j} \geq 0, i = 1, \ldots, m, j = 1, \ldots, n \}. \tag{6}
\]

In order to introduce the main algorithm, we need to define two projections that project the given matrix onto \( M_r \) and \( M_n \), respectively. By the Eckart-Young-Mirsky theorem [12], the projection onto fixed rank matrix set \( M_r \) can be expressed as follows:

\[
\pi_1(X) = \sum_{i=1}^{r} \sigma_i(X) u_i(X) v_i^T(X), \tag{7}
\]

where \( \sigma_i(X) \) is the \( i \)-th singular values of \( X \), and their corresponding left and right singular vectors: \( u_i(X) \) and \( v_i(X) \). The projection onto the nonnegative matrix set \( M_n \) can be expressed as follows:

\[
\pi_2(X) = \begin{cases} 
X_{ij}, & \text{if } X_{ij} \geq 0, \\
0, & \text{if } X_{ij} < 0.
\end{cases} \tag{8}
\]
In particular, since the manifold $M_r$ is not convex, the projection mapping $\pi_1(X)$ can no longer be single valued (for example when the $r$-th singular value has multiplicity higher than 1). Then the algorithm of alternating projection on $M_r$ and $M_n$ can be given as Algorithm 1. The framework of this algorithm is the same as the general case, while the only difference is that the projections are respectively chosen as $\pi_1$ and $\pi_2$ given in (7) and (8).

**Algorithm 1 Alternating Projections On Manifolds**

**Input:** Given a nonnegative matrix $A \in \mathbb{R}^{m \times n}$ this algorithm computes nearest rank-$r$ nonnegative matrix.

1: Initialize $X_0 = A$;
2: for $k = 1, 2, ...$
3: $Y_{k+1} = \pi_1(X_k)$;
4: $X_{k+1} = \pi_2(Y_{k+1})$;
5: end

**Output:** $X_k$ when the stopping criterion is satisfied.

Different from the convex sets case, the intersection of $M_r$ and $M_n$ decides whether the sequence generated by Algorithm 1 converges or not. In order to show the convergence of Algorithm 1 we compute the dimension of the intersection of $M_r$ and $M_n$.

**Theorem 1.** Let $M_r$ and $M_n$ be defined as (5) - (6). Then

$$M_{rn} := M_r \cap M_n = \{X \in \mathbb{R}^{m \times n}, \text{rank}(X) = r, X_{ij} \geq 0, i = 1, ..., m, j = 1, ..., n\}$$

is a smooth manifold with dimension $(m + n)r - r^2$.

The proof of Theorem 1 can be found in Supplementary. Moreover, we need to define the angle $\alpha(A)$ of $A \in M_{rn}$ where

$$\alpha(A) = \cos^{-1}(\sigma(A)) \quad \text{and} \quad \sigma(A) = \lim_{\xi \to 0} \sup_{B_1 \in F_1^\xi(A), B_2 \in F_2^\xi(A)} \left\{ \frac{\langle B_1 - A, B_2 - A \rangle}{\|B_1 - A\|_F \|B_2 - A\|_F} \right\},$$

with

$$F_1^\xi(A) = \{B_1 \mid B_1 \in M_r \backslash A, \|B_1 - A\|_F \leq \xi, B_1 - A \perp T_{M_r \cap M_n}(A)\},$$

$$F_2^\xi(A) = \{B_2 \mid B_2 \in M_n \backslash A, \|B_2 - A\|_F \leq \xi, B_2 - A \perp T_{M_r \cap M_n}(A)\},$$

and $T_{M_r \cap M_n}(A)$ is the tangent space of $M_r \cap M_n$ at point $A$ (the definition of the tangent space can be found in Supplementary). The angle is calculated based on the two points belonging $M_1$ and $M_2$ respectively.

A point $A$ in $M_{rn}$ is nontangential if $\alpha(A)$ has a positive angle $2$, i.e., $0 \leq \sigma(A) < 1$. It is interesting to note that if there is one nontangential point, majority of points are also nontangential because of the manifold smoothness. We show that $M_{rn}$ contains a non-empty set $M_{rn}^\prime$ of nontangential points. The proof can be found in Supplementary.
Theorem 2. $\mathcal{M}_{p_{n}}^{\mu} \neq \emptyset$.

By using this property, Andersson and Carlsson [2] have shown the following theorem.

Theorem 3. Let $\mathcal{M}_r$, $\mathcal{M}_n$ and $\mathcal{M}_{rn}$ be given as (3), (6) and (9), the projections onto $\mathcal{M}_r$ and $\mathcal{M}_{rn}$ be given as (7)-(8), respectively. Denote $\pi$ as the projection onto $\mathcal{M}_{rn}$. Suppose that $P \in \mathcal{M}_{p_{n}}^{\mu}$ is a non-tangential intersection point, then for any given $\epsilon > 0$ and $1 > c > \sigma(P)$, there exist a $\xi > 0$ such that for any $A \in \text{Ball}(P, \xi)$ (the ball neighborhood of $P$ with radius $\xi$ contains the given nonnegative matrix $A$) the sequence $\{X_k\}_{k=0}^{\infty}$ generated by the alternating projection algorithm initializing from given $A$ satisfies the following results:

(1) the sequence converges to a point $X_\infty \in \mathcal{M}_{rn}$,
(2) $\|X_\infty - \pi(A)\|_F \leq \epsilon \|A - \pi(A)\|_F$,
(3) $\|X_\infty - X_k\|_F \leq \text{const} \cdot c^k \|A - \pi(A)\|_F$,

where $\pi(A) \equiv \arg \min_{\text{rank}(X) = r, X \geq 0} \|A - X\|_F^2$ (the optimized solution).

According to Theorem 3, we know Algorithm 1 can converge and find a matrix that can be sufficiently close the best nonnegative approximation to $\|A - \pi(A)\|_F$.

3 Numerical Results and Discussion

The errors or the residuals of the proposed NLRM alternating projection algorithm were evaluated with three experiments. In the first experiment, we randomly generated $m$-by-$r$ nonnegative matrices $B$ and $r$-by-$n$ nonnegative matrices $C$ where their matrix entries follow a uniform distribution in between 0 and 1. The given non-negative matrix $A$ is computed by the multiplication of $B$ with $C$. We employed Algorithm 1 to test the relative error $\|A - X_c\|_F/\|A\|_F$ of the proposed NLRM algorithm and other existing NMF algorithms: A-MU [13], HALS [4], A-HALS [13], APG [23] and A-APG [23]. Here $X_c$ are the computed solutions by different algorithms. In the second experiment, we randomly generated $m$-by-$n$ nonnegative matrices $A$ where their matrix entries follow a uniform distribution in between 0 and 1. The low rank minimizer is unknown in this setting. However, the residual between the given nonnegative matrix and the computed solution $\|A - X_c\|_F$ can still be calculated. In the third experiment, we considered the CBCL face database [36]. In the face database, there are $m = 2429$ facial images, each consisting of $n = 19 \times 19 = 361$ pixels, and constituting a face image matrix $A \in \mathbb{R}_{+}^{361 \times 2429}$. We tested several values of $r$ for nonnegative low rank minimization, compared the proposed NLRM algorithm with other existing NMF algorithms. The relative residual $\|A - X_c\|_F/\|A\|_F$ was employed to show that the performance of the proposed NLRM algorithm.
In the first experiment, Table 1 shows the relative errors \( \|A - X_c\|_F / \|A\|_F \) of the computed solution \( X_c \) from Algorithm 1 and the other existing NMF algorithms for synthetic data sets. Because the input data matrix \( A \) is exactly a nonnegative rank \( r \) matrix, the proposed algorithm NLRM can provide exact recovery results. However, there is no guarantee that other existing NMF algorithms can determine exact nonnegative low rank factorization.

In the second experiment, the input data matrix \( A \) is just nonnegative. The corresponding nonnegative low rank solution is not known. Table 2 shows that shows the residuals \( \|A - X_c\|_F \) of the computed solution \( X_c \) from Algorithm 1 and the other existing NMF algorithms. We see from the table that the residuals by all the methods decrease when \( r \) increases. Moreover, the residuals by the proposed algorithm are smaller than those by other methods, in particular, when \( r \) is large.

In the third experiment, the input face data matrix \( A \) is 361-by-2429 where each column refers to a face image of size 19-by-19 arranged in chronological order. There are 2429 face images in total. We tested different values of \( r \), and computed the relative residuals \( \|A - X_c\|_F / \|A\|_F \) of the computed solution \( X_c \) from Algorithm 1 and the other existing NMF algorithms. We see in Table 3 that the relative residuals by the proposed algorithm is smaller than those by other methods, in particular, when \( r \) is large. On the other hand, our proposed method can provide a significant index based on singular values. In Figure 1, we displayed the \( r \) singular values of the computed solutions for \( r = 20, 40, 60, 80, 100 \). It is interesting to note that for each \( r \), their distributions behave similarly. When \( r \) increases, the ratio between the sum of the \( r \) singular values and the sum of all the singular value of \( A \) also increases. For example, when \( r = 20 \), the ratio is 55.01%; when \( r = 80 \), the ratio is 77.83%. This phenomena is similar to the low rank approximation, but the main difference is that it is also nonnegative low rank approximation. According to the magnitude of \( r \) singular values \( \{\sigma_i\}_{i=1}^r \), we can identify importance of singular basis vectors \( \{u_i\}_{i=1}^r \). In Figure 2, we showed the case for \( r = 40 \) as an example. It is clear in the figure that different face components are obtained and their importance are given in the face image data matrix according to the magnitude of \( r \) singular values. We remark that the above mentioned information cannot be obtained in the classical NMF directly.

4 Concluding Remarks

In this paper, we proposed and developed a new algorithm for computing NLRM approximation for nonnegative matrices. The new method is different from classical nonnegative matrix factorization method. We have shown the convergence of the proposed algorithm based on the results in manifold. Moreover, we have demonstrated numerical results that the minimized distance by the proposed NLRM method can be smaller than that by the NMF method. According to the ordering of singular values, the proposed method identifies important singular basis vectors, while this information cannot be obtained in the classical NMF.
Table 1: The relative errors by different algorithms for synthetic data matrices.

| sizes $(m,n)$ | rank $r$ | NLRM      | A-MU      | HALS      | A-HALS     | APG       | A-APG     |
|--------------|----------|-----------|-----------|-----------|------------|-----------|-----------|
| (100,80)     | 5        | **4.30e-15** | 0.0066 | 0.0067 | 0.0067 | 0.0097 | 0.0010 |
|              | 10       | **1.11e-15** | 0.0043 | 0.0044 | 0.0044 | 0.0083 | 0.0039 |
|              | 20       | **5.03e-16** | 0.0094 | 0.0066 | 0.0066 | 0.0074 | 0.0041 |
|              | 50       | **1.52e-15** | 0.0108 | 0.0031 | 0.0031 | 0.0066 | 0.0007 |
| (200,160)    | 5        | **5.41e-16** | 0.0051 | 0.0071 | 0.0071 | 0.0076 | 0.0006 |
|              | 10       | **6.61e-16** | 0.0083 | 0.0070 | 0.0067 | 0.0105 | 0.0051 |
|              | 50       | **8.63e-15** | 0.0124 | 0.0060 | 0.0054 | 0.0099 | 0.0063 |
|              | 100      | **1.13e-15** | 0.0123 | 0.0023 | 0.0022 | 0.0078 | 0.0075 |
| (500,400)    | 5        | **3.06e-15** | 0.0075 | 0.0123 | 0.0092 | 0.0129 | 0.0003 |
|              | 10       | **1.59e-16** | 0.0090 | 0.0087 | 0.0096 | 0.0123 | 0.0028 |
|              | 50       | **3.00e-16** | 0.0080 | 0.0092 | 0.0076 | 0.0126 | 0.0038 |
|              | 200      | **1.20e-15** | 0.0142 | 0.0022 | 0.0021 | 0.0102 | 0.0114 |

Table 2: The residuals computed by different algorithms for synthetic data matrices.

| sizes $(m,n)$ | rank $r$ | NLRM | A-MU | HALS | A-HALS | APG | A-APG |
|--------------|----------|------|------|------|--------|-----|-------|
| (100,80)     | 5        | **23.46** | 23.50 | 23.49 | 23.49 | 23.52 | 23.49 |
|              | 10       | **21.10** | 21.42 | 21.35 | 21.35 | 21.43 | 21.35 |
|              | 20       | **16.82** | 18.11 | 17.91 | 17.91 | 18.13 | 17.98 |
|              | 50       | **6.88**  | 11.59 | 10.94 | 10.94 | 11.93 | 11.03 |
| (200,160)    | 5        | **49.13** | 49.16 | 49.15 | 49.15 | 49.17 | 49.15 |
|              | 10       | **46.76** | 47.00 | 46.96 | 46.96 | 47.04 | 46.96 |
|              | 50       | **29.83** | 35.23 | 34.56 | 34.56 | 35.50 | 34.71 |
|              | 100      | **13.75** | 26.51 | 25.13 | 25.10 | 27.33 | 25.73 |
| (500,400)    | 5        | **126.48** | 126.50 | 126.51 | 126.51 | 126.52 | 126.49 |
|              | 10       | **123.96** | 124.08 | 124.09 | 124.09 | 124.19 | 124.06 |
|              | 50       | **104.64** | 109.80 | 109.22 | 109.22 | 110.26 | 109.32 |
|              | 200      | **48.82**  | 82.67 | 79.64 | 79.73 | 85.40 | 81.61 |

References

[1] Absil, P. A., Mahony, R. & Sepulchre, R. *Optimization algorithms on matrix manifolds* (Princeton University Press, 2009).

[2] Andersson, F. & Carlsson, M. Alternating projections on nontangential manifolds. *Constructive approximation*, 38, 489-525 (2013).

[3] Berry, M. W. & Kogan, J. *Text Mining: Applications and Theory* (West Sussex, PO19 8SQ, UK: John Wiley and Sons, 2010).

[4] Cichocki, A., Zdunek, R. & Amari, S. Hierarchical ALS algorithms for nonnegative matrix and 3D tensor factorization. *International Conference on Indepen-
Table 3: The relative residuals computed by different algorithms for face data matrices.

| rank | NLRM   | A-MU   | HALS   | A-HALS | APG    | A-APG  |
|------|--------|--------|--------|--------|--------|--------|
| 5    | 0.1825 | 0.1835 | 0.1836 | 0.1835 | 0.1835 | 0.1838 |
| 10   | 0.1494 | 0.1534 | 0.1538 | 0.1526 | 0.1531 | 0.1536 |
| 20   | 0.1170 | 0.1245 | 0.1236 | 0.1232 | 0.1241 | 0.1246 |
| 30   | 0.0972 | 0.1073 | 0.1048 | 0.1048 | 0.1109 | 0.1087 |
| 40   | 0.0839 | 0.0959 | 0.0925 | 0.0931 | 0.1021 | 0.0992 |
| 50   | 0.0744 | 0.0881 | 0.0853 | 0.0851 | 0.0959 | 0.0942 |
| 60   | 0.0654 | 0.0840 | 0.0776 | 0.0777 | 0.0949 | 0.0928 |
| 70   | 0.0587 | 0.0801 | 0.0728 | 0.0731 | 0.1000 | 0.0920 |
| 80   | 0.0529 | 0.0766 | 0.0676 | 0.0683 | 0.0898 | 0.0934 |
| 90   | 0.0480 | 0.0738 | 0.0652 | 0.0656 | 0.1018 | 0.0980 |
| 100  | 0.0438 | 0.0727 | 0.0616 | 0.0630 | 0.1020 | 0.1011 |

Figure 1: The singular values distribution of NLAM with \( r = 20, 40, 60, 80, 100 \) for face data set. The percentage refers to the ratio between the sum of the first \( r \) singular values and the sum of all the singular values of \( A \).

[4] “Non-negative sparse representation based on block NMF” by Chen, M., Chen, W. S., Chen, B. & Pan, B. In Chinese Conference on Biometric Recognition. Springer, Cham, 26-33 (2013).

[5] Cho, H., Dhillon, I. S., Guan, Y. & Sra, S. In Proc. 4th SIAM International Conference on Data Mining (SDM) Florida, 114-125 (2004).
Figure 2: The first 40 singular vectors \( \{u_i\}_{i=1}^{40} \) corresponding to bases of face images.

[7] Choi, S. Algorithms for orthogonal nonnegative matrix factorization. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* IEEE, 1828-1832 (2008).

[8] Cichocki, A., Zdunek, R., Phan, A. H. & Amari, S. I. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation* (John Wiley and Sons, 2009).

[9] Ding, C., He, X. & Simon, H. D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proc. SIAM International Conference on Data Mining (SDM’05)*, 606-610 (2005).

[10] Ding, C., Li, T., Peng, W. & Park, H. Orthogonal nonnegative matrix trifactorizations for clustering. In *KDD06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*. New York, NY, USA, ACM Press, 126-135 (2006).

[11] Gao, Y. & Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* 21(21), 3970-3975 (2005).

[12] Golub, G. H. & Van Loan, C. F. *Matrix computations* (vol. 3, JHU Press, 2012).

[13] Gillis, N. & Glineur, F. Accelerated multiplicative updates and hierarchical ALS algorithms for nonnegative matrix factorization. *Neural computation* 24(4), 1085-1108 (2012).
[14] Guillamet, D. & Vitria, J. Non-negative matrix factorization for face recognition. In *Topics in artificial intelligence*. Springer Berlin Heidelberg, 336-344 (2002).

[15] Guillamet, D., Vitria, J. & Schiele, B. Introducing a weighted nonnegative matrix factorization for image classification. *Pattern Recognition Letters* 24(14), 2447-2454 (2003).

[16] Jing, L. P., Zhang, C. & Ng, M. K. SNMFCA: supervised NMF-based image classification and annotation. *IEEE Transactions on Image Processing* 21(11), 4508-4521 (2012).

[17] Kim, P. M. & Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Research* 13, 1706-1718 (2003).

[18] Kim, H. & Park, H. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. *Bioinformatics* 23(12), 1495-1502, 2007.

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

[20] Lee, D. D. & Seung, H. S. Algorithms for Nonnegative Matrix Factorization. *MIT Press* 13, (2001).

[21] Lee, J. M. *Smooth manifolds*, in *Introduction to Smooth Manifolds* (Springer, 2013).

[22] Li, T. & Ding, C. The relationships among various nonnegative matrix factorization methods for clustering. In *Proc. 6th International Conference on Data Mining (ICDM06)*. Washington, DC, USA, IEEE Computer Society, 362-371 (2006).

[23] Lin, C. J. Projected gradient methods for nonnegative matrix factorization. *Neural computation* 19(10), 2756-2779 (2007).

[24] Liu, H., Wu, Z., Cai, D. & Huang, T. S. Constrained nonnegative matrix factorization for image representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34(7), 1299-1311 (2012).

[25] Liu, Y., Jing, L. P. & Ng, M. K. Robust and non-negative collective matrix factorization for text-to-image transfer learning. *IEEE Transactions on Image Processing* 24(12), 4701-4714 (2015).

[26] Pascual-Montano, A., Carmona-Saez, P., Chagoyen, M., Tirado, F., Carazo, J. M. & Pacual-Marqui, R. bioNMF: A versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics* 7(1) (2006).
The supplementary information consists of the theoretical proofs of Theorems 1-3. Some basic definitions, propositions of algebraic geometry and differential geometry are needed to prove the main results. We only provide some results here and for details we refer to [1, 21] and references therein.

In order to show the convergence of Algorithm 1 we need to prove the intersection of $M_r$ given as (5) and $M_n$ given as (6) is a manifold first. It is well known
that $\mathcal{M}_r$ is a manifold and $\mathcal{M}_n$ is a manifold with boundary, respectively. In many real applications, some manifolds are actually real algebraic varieties which are defined as the vanishing of a set of polynomials on $\mathbb{R}^n$, thus some algebraic geometry methods can help us to study the above problem. Note that all complex varieties are real varieties, but not conversely. For a given real algebraic variety $V \in \mathbb{R}^n$, if we identify $\mathbb{R}^n$ as a subset of $\mathbb{C}^n$ and denote $\mathbb{I}_R(V)$ as the set of real polynomials that vanish on $V$, then $V$ has a related complex variety given by its Zariski closure

$$V_{\text{Zar}} = \{ z \in \mathbb{C}^n : p(z) = 0, \ \forall p \in \mathbb{I}_R(V) \},$$

which is defined as the subset in $\mathbb{C}^n$ of common zeros to all polynomials that vanish on $V$. Moreover, H. Whitney in [29] showed that a real algebraic variety $V$ can be decomposed as $V = \bigcup_{j=0}^{m} V_j$ where each $V_j$ is either void or a $\mathbb{C}(\infty)$-manifold with dimension $j$. If $V_m \neq \emptyset$, then $m$ equals the algebraic dimension of $V_{\text{Zar}}$. Each $V_j$ contains at most a finite number of connected components. This result shows that the main part of a variety is a manifold.

Varieties have singular points which is different from manifolds. Note that a point $A \in V$ is nonsingular if it is nonsingular in the sense of algebraic geometry as an element of $V_{\text{Zar}}$. This result changes to be much simpler, if we restrict on the irreducible variety. Here a algebraic variety $V$ is said to be irreducible if there does not exist non-trivial real algebraic varieties $V_1$ and $V_2$, such that $V = V_1 \cup V_2$. Denote $\nabla$ as the gradient operator and set $\mathcal{N}_V(z) = \{ \nabla p(z) : p \in \mathbb{I}_R(V) \}$. Suppose that $V \in \mathbb{R}^n$ is an irreducible real algebraic variety of dimension $m$, then $\dim \mathcal{N}_V(z) \leq n - m$ and $z \in V$ is non-singular if and only if $\dim \mathcal{N}_V(z) = n - m$. In practice, it is not easy to check a given variety is irreducible or not, thus some additional tools are needed. The following statements can help us to make a decision and will be used many times in the sequel. A real algebraic variety $V$ is irreducible if $V$ is connected and can be covered with analytic patches or if a dense subset of $V$ can be given as the image of one real analytic function $\phi$. Then an irreducible variety $V$ has dimension $d$ if an open subset of $V$ is the image of a bijective real analytic map defined on a subset of $\mathbb{R}^d$.

With the above tools in hand we can prove the following results.

**Proof of Theorem 1** The proof can be divided into two parts. Firstly, we need to prove the following set

$$V_{rn} = \{ X \in \mathbb{R}^{m \times n} : \text{rank}(X) \leq r, X_{i,j} \geq 0, \ i = 1, \ldots, m, j = 1, \ldots, n \}$$

(10)
is an irreducible variety with dimension $(m + n)r - r^2$. Then we need to prove $V_{rn}^{ns}$, i.e., the set of all the non-singular points in $V_{rn}$, equals all the nonnegative matrices with rank equal to $r$.

In the first step, we mainly prove $V_{rn}$ is an irreducible variety with dimension $(m + n)r - r^2$. Denote $\mathcal{K}$ as set of $m \times n$ matrices over $\mathbb{R}$. If all elements of a matrix $A \in \mathcal{K}$ are considered as variables, $\mathcal{K}$ is a linear manifold with dimension $mn$. The same statement is also satisfied for the nonnegative matrix set $\mathcal{M}_n$. Denote $\mathcal{V}_r :=$
\{X \in \mathbb{R}^{m \times n}, \text{rank}(X) \leq r\}. Recall that \( B \in \mathcal{K} \) has rank greater that \( r \) if and only if one can find a nonzero \((r+1) \times (r+1)\) invertible minor. The determinant of each such minor is a polynomial, and \( \mathcal{V}_r \) is clear the variety obtained from the collection of such polynomials. Thus \( \mathcal{V}_r \) is a real algebraic variety. The same statement is also true for \( \mathcal{V}_{rn} \), which is derived by adding the nonnegative constrain conditions to those defining \( \mathcal{V}_r \). Moreover, for \( A \in \mathcal{M}_r \), it follows from the singular value decomposition theory (SVD) given in [12] that there exist two column orthogonal matrices \( U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r} \) and a diagonal matrix \( \Sigma \in \mathbb{R}^{r \times r} \) such that

\[ A = U_{m \times r} \cdot \Sigma_{r \times r} \cdot V_{n \times r}^T. \quad (11) \]

In the other direction, any matrix written as \((11)\) has rank less or equal to \( r \), thus \( \mathcal{V}_r \) can be covered by a real polynomial which is saying that \( \mathcal{V}_r \) is irreducible. Choose \( \Gamma \) as a subset of \( \mathcal{V}_r \) such that the singular values are different, then the matrices \( U, V \) and \( \Sigma \) in \((11)\) are unique. The sets of \( m \times r, r \times r \) and \( r \times n \) matrices appear in \((11)\) contain \((nr - \frac{r(r+1)}{2})\), \( r \), and \((nr - \frac{r(r+1)}{2})\) independent variables, respectively, then we can identify the sets of such matrices with \( \mathbb{R}^{nr - \frac{r(r+1)}{2}}, \mathbb{R}^r \) and \( \mathbb{R}^{nr - \frac{r(r+1)}{2}} \).

Denote the inverses of these identifications by

\[ t_1 : \mathbb{R}^{nr - \frac{r(r+1)}{2}} \rightarrow \mathbb{R}^{m \times r}; \quad t_2 : \mathbb{R}^r \rightarrow \mathbb{R}^{r \times r}; \quad t_3 : \mathbb{R}^{nr - \frac{r(r+1)}{2}} \rightarrow \mathbb{R}^{n \times r}; \quad (12) \]

and denote \( \Omega \subset \mathbb{R}^{(m+n)r-r^2} \) as the open set corresponding to those matrices having the same structure as \( \Gamma \). Define \( \phi : \Omega \rightarrow \mathcal{V}_{rn} \) by

\[ \phi(y_1, y_2, y_3) = t_1(y_1) \times t_2(y_2) \times (t_3(y_3))^T. \quad (13) \]

It is easy to see that the polynomial projection \( \phi \) is bijective correspondence with an open set including \( B \). Thus \( \mathcal{V}_r \) can be covered with one real polynomial, and is irreducible with dimension \((m + n)r - r^2\) as desired. We next turn our attention to \( \mathcal{V}_{rn} \). We need to show it can be covered by analytic patches. Suppose that \( \pi : \{1, \ldots, m\} \rightarrow \{1, \ldots, r\} \), and consider all \( U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r}, \) where \( U_{j, \pi_j} = x_j \) is an undetermined variable whereas all other values are fixed. Denote the ith row of \( U \) by \( U_i \), the jth row of \( V \) by \( V_j \) and let \( \Sigma = \text{diag}\{\sigma_1, \ldots, \sigma_r\} \) be a fixed diagonal matrix. Then \( U_i \Sigma V_j^T \geq 0 \) is a linear inequality with \( x_j \) as unknown, which may have infinitely many real solutions. Suppose the remaining values of \( U \) and \( V \) are such that this inequality has a positive solution for all \( 1 \leq i \leq m, 1 \leq j \leq n \). Denote the corresponding matrices by \( \hat{U} \) and \( \hat{V} \) and note that it is a real analytic function if we consider the remaining values of \( \hat{U} \) and \( \hat{V} \) as variables. Moreover, the variables and the values on the diagonal of \( \Sigma \) are \((m + n)r - r^2 - m \) in number. Then it can be identified with points \( y \) in an open subset of \( \mathbb{R}^{(m+n)r-r^2} \). Let \( \Omega \) be a particular connected component of this open set, and consider \( U, \Sigma \) and \( V \) as functions of \( y \) on \( \Omega \) as

\[ \psi_x,\Omega(y) = U(y) \Sigma(y)(V(y))^T, \quad y \in \Omega. \quad (14) \]

Denote \( \mathbb{I} \) as the set of all possible \( \pi \) and \( \Omega \), then \( B \in \mathcal{V}_{rn} \) is in the image of at least one of these maps \( \psi_i, i \in \mathbb{I} \), which is saying that \( \mathcal{V}_{rn} \) can be covered by the analytic
functions $\{\psi_i\}_{i \in I}$. We also need to prove $V_{r,n}$ is connected. For any $B \in V_{r,n}$, it is sufficient to show $B$ is path connected with the matrix $1$ (with all elements equal to $1$). Let $B$ be fixed. Then all the singular values of $B$ are nonnegative and ordered decreasingly. Suppose that $\sigma_1 = 1$, and then pick $\pi$ such that $\pi(j) = r$ for all $j$ and choose $\Omega$ such that the representation (11) can be expressed as (14). Now if $\sigma_2 \neq 1$ we can continuously moved until it is not without leaving $\Omega$. Then the values of $y$ corresponding to the first and second column of $\hat{U}$ and $\hat{V}$ can be continuously moved until all elements of the first column of $\hat{U}$ and $\hat{V}$ are positive, respectively. At this point, we can reduce all values of $\hat{U}$ and $\hat{V}$ except the first column to zero, increase the first value of each row whenever necessary to stay in $\Omega$. Then we can move $y$ so that the values in the first column become the same. Finally, we can let these values increase simultaneously until they reach $1$. Then the matrix $1$ is derived, which is saying that $V_{r,n}$ is connected and hence it is irreducible. In this case, choose an open set of $V_{r,n}$ as $\Lambda$ with $U$, $\Sigma$ and $V$ contain, respectively, $(mr - \frac{r(r+1)}{2})$, $r$, and $(nr - \frac{r(r+1)}{2})$ independent variables, with the difference that the last element in each row of $U$ is a variable satisfy $U_i \Sigma V_j^T \geq 0$. Thus they identify the set of matrices with $\mathbb{R}^{nr - \frac{r(r+1)}{2}} \times \mathbb{R}^r$ and $\mathbb{R}^{nr - \frac{r(r+1)}{2}}$. Denote the inverses of those identifications by

$$
t_1 : \mathbb{R}^{mr - \frac{r(r+1)}{2}} \to \mathbb{R}^{m \times r}; \ t_2 : \mathbb{R}^r \to \mathbb{R}^{r \times r}; \ t_3 : \mathbb{R}^{nr - \frac{r(r+1)}{2}} \to \mathbb{R}^{n \times r}; \quad (15)$$

and denote $\Omega \subset \mathbb{R}^{(m+n)r - r^2}$ as the open set corresponding to those matrices having the same structure as $\Lambda$. Define $\phi : \Omega \to V_{r,n}$ by

$$
\phi(y_1, y_2, y_3) = t_1(y_1) \times t_2(y_2) \times (t_3(y_3))^T. \quad (16)
$$

It is easy to see that $\phi$ is bijective correspondence with an open set including $B$, and $\phi$ is a polynomial. Thus $V_{r,n}$ can be covered with one real polynomial. It follows that $V_{r,n}$ is irreducible and possessing dimension $(m + n)r - r^2$ as desired.

In the second step, we mainly prove $V_{r,n}^{ns}$ equals all the nonnegative matrices with rank equal to $r$. Recall that $V_{r,n}$ is an irreducible real algebraic variety with dimension $(m + n)r - r^2$, then we need to prove

$$
\dim \mathcal{N}_{V_{r,n}}(A) = mn - (m + n)r + r^2, \quad (17)
$$

if and only if rank$(A) = r$. It is sufficient to show that $\dim \mathcal{N}_{V_{r,n}}(A) \leq mn - (m + n)r + r^2$ is strict when rank$(A) < r$ and the reverse inequality holds when rank$(A) = r$. Define $\omega : \mathbb{R}^{mn} \to \mathcal{K}$. If $p$ is a polynomial on $\mathcal{K}$, then write $\nabla p$ in stead of the right form $\omega(\nabla (p \circ \omega))$. Now for a given polynomial $p \in \mathbb{I}(V_{r,n})$, and two unitary matrices $U$ and $V$, $q_{U,V}(\phi) = p(U \circ V^T)$, which is also in $\mathbb{I}(V_{r,n})$. Let $A$ be a fixed nonnegative matrix with rank$(A) = k \leq r$. By the above proof we can produce two unitary matrices $U$ and $V$ such that

$$
UAV^T = \text{diag}\{\sigma_1, ..., \sigma_j, 0, ..., 0\} = E_j,
$$

where the first $j$ diagonal values of $E$ are positive numbers and $0$ otherwise. In particular $\nabla q_{U,V}(A) = U \nabla p(UAV^T)V^T = U \nabla p(E_j)V^T$, implying that $\dim \mathcal{N}_{V_{r,n}}(A) =
dim \mathcal{N}_{\mathcal{V}_n}(E_j). Note that the \( R_{r+1,r+1} \) subdeterminants of \( \mathcal{K} \) form polynomials in \( \mathcal{I}(\mathcal{V}_r_n) \), and their derivatives at \( E_j \) can be derived. Then

\[ \dim \mathcal{N}_{\mathcal{V}_n}(E_j) \geq mn - (m + n)r + r^2, \]  

(18)

which proves that any rank \( r \) element of \( \mathcal{V}_r_n \) is nonsingular. Moreover, if \( j < r \), consider a fixed \( u \in \mathbb{R}^m \) and \( v \in \mathbb{R}^n \), define the map \( \vartheta_{u,v} : \mathbb{R}_+ \rightarrow \mathcal{V}_r_n \) via \( \vartheta_{u,v}(x) = \bigcap \mathcal{V}_n \cap \mathcal{M}_r \). After that it is not difficult to derive \( T_{\mathcal{M}_r} \) with boundary look exactly the same as those on a manifold. Some details can be found in \([1,21]\). After that it is not difficult to derive \( T_{\mathcal{M}_r} \) with boundary look exactly the same as those on a manifold. Some details can be found in \([1,21]\).

Proof of Theorem 2
Suppose that the angle \( \alpha(A) \) of \( A \in \mathcal{M}_r \cap \mathcal{M}_n \) is well defined, then \( A \) is tangential if \( \alpha(A) = 0 \), and is nontangential if \( \alpha(A) > 0 \). Moreover, \( A \in \mathcal{M}_r \cap \mathcal{M}_n \) is nontangential if and only if

\[ T_{\mathcal{M}_r}(A) \cap T_{\mathcal{M}_n}(A) = T_{\mathcal{M}_r \cap \mathcal{M}_n}(A), \]  

(20)

where \( T_{\mathcal{M}_r}(A) \) and \( T_{\mathcal{M}_n}(A) \) denote the differential geometry tangent spaces of manifolds \( \mathcal{M}_r \) and \( \mathcal{M}_n \) at point \( A \), respectively. Denote \( \mathcal{M}_{r_n}^{nt} \subseteq \mathcal{M}_{r_n} \) as the set of all nontangential points of \( \mathcal{M}_r \cap \mathcal{M}_n \). Recall that the tangent space of \( \mathcal{M}_r \) at \( A = U_m \times \Sigma_{r \times r} \cdot V_n^{T} \) can be expressed as

\[ T_{\mathcal{M}_r}(A) = \left\{ [U, U_\perp] \left( \begin{array}{cc} \mathbb{R}^{r \times r} & \mathbb{R}^{r \times (n-r)} \\ \mathbb{R}^{(m-r) \times r} & 0^{(m-r) \times (n-r)} \end{array} \right) [V, V_\perp]^T \right\}, \]

where \( U_\perp \in \mathbb{R}^{m \times (m-r)} \) and \( V_\perp \in \mathbb{R}^{(n-r) \times n} \) stand for the orthogonal completions of \( U \) and \( V \), respectively. In particular, \( \mathcal{M}_n \) is a manifold with boundary so we need to introduce the tangent space of the points on the boundary. Assume that \( P \) is a boundary point of \( \mathcal{M}_n \), there are many method to define the tangent space to \( \mathcal{M}_n \) at \( P \), the standard choice is to define \( T_{\mathcal{M}_n}(p) \) to be an \( mn \)-dimensional vector space. This may or may not seem like the most geometrically intuitive choice, but it has the advantage of making most of the definitions of geometric objects on a manifold with boundary look exactly the same as those on a manifold. Some details can be found in \([11,21]\). After that it is not difficult to derive \( T_{\mathcal{M}_n}(A) = \text{Span}\{E_{ij}\} \) with \( E_{ij} = 1 \) or 0, for all \( i = 1, \ldots, m \) and \( j = 1, \ldots, n \). Thus

\[ \dim (T_{\mathcal{M}_r}(A) \cap T_{\mathcal{M}_n}(A)) \leq (m + n)r - r^2. \]  

(21)

Recall Theorem 1 \( \mathcal{M}_{r_n} \) is a smooth manifold with dimension \((m+n)r-r^2\), which is homeomorphism to its tangent space \( T_{\mathcal{M}_{r_n}}(A) \). Thus,

\[ \dim (T_{\mathcal{M}_r}(A) \cap T_{\mathcal{M}_n}(A)) \leq \dim (T_{\mathcal{M}_{r_n}}(A)), \]

(22)
which is sufficient to say $\mathcal{M}_{rrn}^n$ is not empty. □

Since $\mathcal{V}_{rn}$ is an irreducible variety with dimension $(m+n)r - r^2$, $\mathcal{V}_{rn} \setminus \mathcal{M}_{rrn}^n$ is a real algebraic variety of dimension strictly less than $(m+n)r - r^2$. This result tell us the majority of points are nontangential if one is. Moreover, (22) is satisfied, then by Theorem 5.1 in [2], we can get Theorem 3.