Robust Bi-stochastic Graph Regularized Matrix Factorization for Data Clustering

Qi Wang, *Senior Member, IEEE,* Xiang He, Xu Jiang, and Xuelong Li, *Fellow, IEEE*

Abstract—Data clustering, which is to partition the given data into different groups, has attracted much attention. Recently various effective algorithms have been developed to tackle the task. Among these methods, non-negative matrix factorization (NMF) has been demonstrated to be a powerful tool. However, there are still some problems. First, the standard NMF is sensitive to noises and outliers. Although $\ell_{2,1}$ norm based NMF improves the robustness, it is still affected easily by large noises. Second, for most graph regularized NMF, the performance highly depends on the initial similarity graph. Third, many graph-based NMF models perform the graph construction and matrix factorization in two separated steps. Thus the learned graph structure may not be optimal. To overcome the above drawbacks, we propose a robust bi-stochastic graph regularized matrix factorization (RBSMF) framework for data clustering. Specifically, we present a general loss function, which is more robust than the commonly used L_2 and L_1 functions. Besides, instead of keeping the graph fixed, we learn an adaptive similarity graph. Furthermore, the graph updating and matrix factorization are processed simultaneously, which can make the learned graph more appropriate for clustering. Extensive experiments have shown the proposed RBSMF outperforms other state-of-the-art methods.

Index Terms—Matrix factorization, bi-stochastic graph, data clustering, robustness.

1 Introduction

Clustering, as one of the most fundamental and important tasks in machine learning and data mining fields [1], has been extensively studied for many years. Given some data samples, clustering aims to divide them into several different groups, such that there is high similarity for those samples within the same group. Because no label information of data is utilized, clustering can be considered as a special unsupervised classification. Therefore, the clustering performance mainly depends on the similarity relationships between data [2]. In the past decades, lots of approaches of clustering have been developed, such as k-means [3], hierarchical clustering [4], spectral clustering [5], subspace clustering [6], non-negative matrix factorization (NMF) [7], and so on.

As one of the most widely used clustering methods, NMF has drawn lots of attention in recent years. It was originally proposed in the seminal works [7] [8] as a kind of matrix factorization technique. Afterwards, Ding *et al.* [9] found the connection between NMF and k-means, and further proved NMF can be used as a clustering method. NMF aims to approximate the original matrix with two non-negative matrices. For data clustering, the two non-negative matrices are called clustering centroid matrix and clustering indicator matrix respectively. Such non-negative constraints make NMF easy to interpret the real-world data. Since there is only additive (no subtractive and combinative) operator, NMF can obtain a parts-based representation. So NMF has the ability to learn the parts of objects like human brain. Because of these advantages, NMF has been widely applied into many applications. For example, in [10], Luo *et al.* proposed the approach of non-negative latent factor analysis, which is specifically designed for handling a high-dimensional and sparse (HiDS) matrix. Its data density-oriented modeling and learning strategies enable its high-efficiency in both computation and storage [11]. Morevoer, it has good representative learning ability on an HiDS matrix, especially when its data density is extremely low [12]. In addition, NMF has also achieved huge success in the fields of face recognition [13], document clustering [14], image annotation [15], crowd analysis [16] and co-clustering [17].

Although NMF has acquired good performance in most tasks, there are still some limitations. One main problem is that the standard NMF utilizes the Frobenius norm (i.e., L_2 loss function) to define the objective function, which is widely considered unstable to noises and outliers. However, most real-world data contains noises and outliers in practical applications. Although L_2 loss function has some good mathematical properties and is usually applied in other tasks, it is not the best choice for the robustness. Kong *et al.* [18] has proven the standard NMF using Frobenius norm is most suitable for the data with the Gaussian noises (i.e., zero-mean normal distribution noises). However, most actual data does not satisfy the assumptions. In order to decrease the sensitivity for noises and outliers, Kong *et al.* [18] proposed a robust $\ell_{2,1}$-NMF, which replaces the Frobenius norm with $\ell_{2,1}$ norm to measure the reconstructed errors. Comparing with the standard NMF, $\ell_{2,1}$-NMF removes the square of errors such that a few outliers can not dominate the loss function. Besides, $\ell_{2,1}$-NMF can be better applied

- *This work was supported by the National Key R&D Program of China under Grant 2017YFB1002202, National Natural Science Foundation of China under Grant U1864204, 61773316, U1801262, and 61871470.*

- *Q. Wang, X. He, X. Jiang, and X. Li are with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China (e-mail: crabwq@gmail.com; xianghe@mail.nwpu.edu.cn; jx19961023@mail.nwpu.edu.cn; xuelong_li@nwpu.edu.cn).*

- *X. Li is the corresponding author.*

for the data with the Laplacian noises. Although $\ell_{2,1}$-NMF improves the robustness of NMF, it is still influenced by some very large noises. Only removing the square of errors is not enough.

Another main problem is that the local data structures are ignored, which makes the original NMF fails to preserve the data correlation of neighboring points. In order to cope with the problem, many graph-based NMF models [19] [20] [21] [22] [23] have been proposed to incorporate the essential manifold structures. Such methods assume that data samples with small distance have high probability of belonging to the same cluster. It is a reasonable manifold assumption and has derived many manifold learning methods [24] [25]. Most graph regularized NMF models first construct a similarity graph based on the distances of data samples. And then matrix factorization is performed by constraining the encoding matrix with the constructed graph. Therefore, graph-based NMF models strongly depend on the initial graph. If the quality of the input graph is low, the performance of NMF would be poor. Therefore, some problems come with these grapd-based NMF methods. Firstly, most methods usually apply some simple methods to construct the initial graph, such as 0-1 weighting and heat kernel weighting [19]. So the graph may be of low quality. Secondly, once the similarity graph is built for NMF, it would be fixed in the process of matrix factorization. So the graph may be not the optimal for NMF and would limit the performance for clustering task.

To address the above drawbacks, a robust bi-stochastic graph regularized matrix factorization (RBSMF) framework is proposed in this paper. First, we present a new robust $Hx(x)$, which is a general loss function and can be used in NMF to improve its robustness. Second, a better graph is learned by bi-stochastic matrix based on the input graph. Traditional graph construction methods (i.e., ε-ball graph, k-nn graph [5]) only employ the pairwise distances of the original data. Our graph learning strategy combines both the original data and its new encoding matrix to learn a better graph. Third, the proposed RBSMF simultaneously decomposes the given data matrix and learns an adaptive graph, which is not fixed one and helps to obtain better clustering indicator matrix. Finally, we summarize the main contributions of the paper as follows.

- We propose a simple and effective loss function $Hx(x)$ to measure the reconstruction errors of NMF. It has the following two main advantages: 1) By making the larger noises and outliers have the smaller weight on the whole loss, the proposed $Hx(x)$ has the better robustness than the widely used L_2 and L_1 loss functions. 2) It is a general loss function and can be extensively used for other reconstruction problems, such as linear regression, principal component analysis (PCA), and etc. Besides, we also derive an efficient optimization algorithm to solve the problems associated with $Hx(x)$.
- We present an adaptive bi-stochastic graph regularized NMF. Most graph construction methods only utilize the original data to get an input graph, which may be not appropriate. Combining the low-noise encoding matrix, the adaptive graph learning can learn a better high-quality graph by bi-stochastic matrix from an initial low-quality graph. Instead of fixing the input graph like most graph-based NMF, our method can update dynamically graph such that the local geometrical structures are fully exploited.
- The connection between iterative reweighted algorithm and NMF is built in this paper. We summarize a new important and valuable rule, i.e., for any given loss function $f(x)$ to measure the reconstruction errors of NMF, it can be uniformly and easily solved by iterative reweighted algorithm. But traditional ways of solving $f(x)$-NMF need to employ different specific algorithms, which may be complicated. The rule has been verified correctly by the original NMF and $\ell_{2,1}$-NMF. Also we employ the rule to successfully solve the proposed robust Hx-NMF.
- We develop a general robust and graph-based NMF framework for data clustering. The proposed framework mainly incorporates two powerful tools into the original NMF and greatly improves the clustering performance of NMF. The first tool is robust loss function. It makes the proposed RBSMF framework has less sensitivity to noises and outliers. The second tool is adaptive graph learning strategy. It helps to learn a high-quality graph to regularize NMF. Certainly, the general framework also can use other (not just the proposed RBSMF) robust functions and graph learning strategies to improve NMF.

The remainder of the paper is organized as follows. Section 2 introduces some related works about robust NMF and graph regularized NMF. Section 3 gives some brief reviews of iterative reweighted algorithm, NMF and $\ell_{2,1}$-NMF. Section 4 proposes a general loss function $Hx(x)$, the robust Hx-NMF, and the RBSMF framework. Section 5 derives an effective optimization algorithm to solve the proposed RBSMF framework. Section 6 shows extensive experiments on the proposed Hx-NMF and RBSMF. Finally, section 7 concludes the paper and gives some future works.

2 Related Work

In this paper, we mainly focus on two kinds of widely studied NMF, i.e., robust NMF and graph regularized NMF. Therefore, we will introduce some related works about them in this section.

2.1 Robust NMF

Due to using the least square error function, the original NMF is sensitive to noises and outliers. This leads to that NMF fails to accurately decompose most real-world data with some large noises in actual applications. To improve the robustness of NMF, various robust NMF methods have been developed. Hamza *et al.* [26] presented the $L_1 + L_2$ function to define the loss function of NMF. L_1+L_2 function combined the advantages of L_2 and L_1 functions, and became a more robust loss function. However, the solving algorithm of the proposed objective function was a general gradient descent method, which was very time-consuming. As was mentioned in section 1, Kong *et al.* [18] proposed the $\ell_{2,1}$-NMF which used $\ell_{2,1}$ norm to measure the loss.

Due to removing the square of reconstructed errors, $\ell_{2,1}$-NMF was more robust than the original NMF. In [27], a direct robust matrix factorization model was developed for anomaly dection. It applied two constraints (low-rank clean matrix and sparse noises) to improve the robustness of matrix factorization. Zhang *et al.* [28] presented an assumption that noises are sparse and used a matrix to extract the sparse noises. This method decomposed the original data into one sparse matrix, and two non-negative matrices. So the robustness was also improved. Moreover, Shen *et al.* [29] also proposed to encode the noises and outliers with an L_1 regularized sparse term. Meanwhile, an effective iterative solving algorithm was developed to obtain the desired clean factorization matrices. Considering that maximum correntropy criterion (MCC) has acquired some success to handle non-Gaussian noises and outliers, Peng *et al.* [30] proposed to employ MCC to measure the reconstructed errors of NMF. Besides, an L_1 sparse constraint on non-negative encoding matrix was also utilized to improve the clustering performance.

2.2 Graph Regularized NMF

As is well known to us, NMF decomposes the given data matrix into two non-negative matrices. But it only utilizes the global data information and ignores the local structures of neighboring data. This makes the ability of NMF to acquire discriminative data representation is not fully exploited and the clustering performance of NMF is also strongly restricted. To overcome the limitations, many relevant studies, which incorporate the manifold learning into NMF, have been developed in recent years. Cai *et al.* [19] proposed a novel graph regularized NMF (GNMF), which constructed an affinity graph to encode the low dimensional manifold for NMF. Thus the geometrical information of data was integrated into NMF and the data representation became more compact and discriminative. Considering that GNMF still used Frobenius norm based NMF, which was sensitive to noises and outliers, Huang *et al.* [20] presented a robust manifold regularized NMF (RMNMF), which replaced Frobenius norm with $\ell_{2,1}$ norm to measure the loss. The operation enhanced the robustness of NMF. Besides, RMNMF added an orthogonal constraint on clustering indicator matrix, which improved the clustering results of RMNMF. Furthermore, Li *et al.* [31] indicated that most useful information of data was hidden in the low-rank parts. So the work [31] combined low rank representation (LRR) and Laplacian graph to design a new non-negative low-rank matrix factorization. However, all the above graph regularized NMF methods used the given fixed graph, which may be not optimal for NMF. If the input graph was of low-quality, the performance of NMF would be poor. Afterwards, Huang *et al.* [32] developed a more effective NMF with adaptive neighbors (NMFAN), which learned an adaptive graph for NMF and better utilized the manifold structures of data. However, NMFAN had too many hyperparameters to be tuned, which would cost lots of time and was difficult to be applied in practical applications.

For better describing the proposed method, we summarize all notations in Table 1.

TABLE 1
Descriptions of all notations

notations	descriptions
$\Phi(\cdot)$	a general loss function
$(\cdot)_{Hx}$	the simplification of $\sum_i log(1+\|(\cdot)_i\|_2)$
$L_2(\cdot)$	L_2 loss function
$L_1(\cdot)$	L_1 loss function
$\psi(e_i)$	$d\Phi(e_i)/de_i$, influence function
$w(e_i)$	$d\psi(e_i)/de_i$, weight function
$\|\cdot\|_F$	Frobenius norm
$\|\cdot\|_2$	ℓ_2-norm of a vector
$\|\cdot\|_{2,1}$	$\ell_{2,1}$-norm
$Tr(\cdot)$	the trace of a matrix
$diag(\cdot)$	the column vector composed of diagonal elements of a square matrix
L_S	Laplacian matrix
μ, ρ	ALM parameters
α, β	regularization parameters

3 Preliminaries

In this section, we briefly review some important preliminaries. One of them is the iterative reweighted algorithm, which is usually used to solve the general reconstruction problem. Then we introduce two widely studied NMF, i.e., the standard NMF and the robust $\ell_{2,1}$-NMF. After reviewing the related preliminaries, we will propose a more robust NMF and build a connection between iterative reweighted algorithm and NMF with any loss functions in next section.

3.1 Iterative Reweighted Algorithm

Denote the reconstruction error of the *i*-th sample by e_i, which is the difference between the true value and reconstructed value of the *i*-th sample. The general reconstruction problems can be formulated as the following objective function [33]

$$\min \sum_i \Phi(e_i), \tag{1}$$

where $\Phi(\cdot)$ is a general loss function. It needs to satisfy the condition that $\Phi(e_i)$ is an increasing function with respect to $|e_i|$. Assume that there are m unknown variables, denoted by $\mathbf{p} = [p_1, p_2, ..., p_m]^T$, to be solved in problem (1). We can obtain the optimal solution by making the derivative of problem (1) be zero, which is

$$\sum_i \psi(e_i)\frac{\partial e_i}{\partial p_j} = 0, \quad j = 1, 2, ..., m, \tag{2}$$

where $\psi(e_i) = d\Phi(e_i)/de_i$ is called influence function. Furthermore, Eq. (2) can be rewritten as

$$\sum_i w(e_i)e_i\frac{\partial e_i}{\partial p_j} = 0, \quad j = 1, 2, ..., m, \tag{3}$$

where $w(e_i)$ is called weight function [33]. It is a very important function for the subsequent algorithms. The weight function $w(x)$ is defined as

$$w(x) = \frac{\psi(x)}{x} = \frac{\Phi'(x)}{x}. \tag{4}$$

It can be observed easily that Eq. (3) is also exactly the solution of the following iterative reweighted problem

$$\min \sum_i w(e_i^{k-1}) e_i^2, \tag{5}$$

where the superscript $k-1$ stands for the *(k-1)*-th iteration. Each iteration of solving problem (5) can be divided into two steps. Firstly, consider the weight $w(e_i^{k-1})$ as a constant, and then obtain the optimal solution according to the specific form of problem (5). Secondly, recompute the weight value of $w(e_i^{k-1})$ according to the current reconstructed error e_i^k and it would be used during the next iteration.

Finally, the detailed iterative reweighted algorithm for solving the general reconstructed problems like (1) is demonstrated as Algorithm 1.

Algorithm 1 iterative reweighted algorithm for solving problem (1)

Input: The loss function $\Phi(x)$.
Output: The optimal solution of problem (1).
Initialize: Compute the corresponding weight function $w(x)$ of the given loss function $\Phi(x)$. The original problem (1) is then converted as the iterative reweighted problem (5).
While not converged **do**

1) Fix $w(e_i^{k-1})$, and solve problem (5).
2) Recompute the weight value of $w(e_i^k)$.

End while

3.2 NMF and $\ell_{2,1}$-NMF

Denote the given data matrix by $X = [X_1, X_2, ..., X_n] \in \mathbb{R}^{m\times n}$, where each column represents a data sample. m and n represent the number of features and samples, respectively. NMF aims to find two non-negative matrices $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$, which can well reconstruct data matrix X as

$$X \approx UV^T. \tag{6}$$

A commonly used objective function, which uses Euclidean distance to measure the quality of reconstruction, can be seen as follows

$$\min_{U\geq 0, V\geq 0} \quad \|X - UV^T\|_F^2 = \sum_{i=1}^{n} \|(X - UV^T)_i\|_2^2, \tag{7}$$

where $\|\cdot\|_F$ is Frobenius norm. For the arbitrary matrix $M \in \mathbb{R}^{m\times n}$, the Frobenius norm of M is defined as $\|M\|_F = \sqrt{\sum_j^n \sum_i^m M_{i,j}^2}$. $\|(X-UV^T)_i\|_2$ is the reconstructed error of the *i*-th sample. Problem (7) is proven convex with respect to U only or V only. Therefore, it is impractical to obtain the globally optimal solution of problem (7). Fortunately, Lee *et al*. [8] found a simple iterative updating algorithm to acquire the locally optimal U and V. The detailed algorithm for solving problem (7) is given as

$$U_{ik} = U_{ik} \frac{(XV)_{ik}}{(UV^TV)_{ik}}, \tag{8}$$

$$V_{jk} = V_{jk} \frac{(X^TU)_{jk}}{(VU^TU)_{jk}}. \tag{9}$$

Because the standard NMF (7) uses the square of reconstructed errors to define the loss, some large noises and outliers would easily dominate the objective function. This makes the standard NMF very sensitive to noises and outliers. To improve the robustness of NMF, Kong *et al*. [18] proposed a more robust $\ell_{2,1}$-NMF, which replaced Frobenius norm with $\ell_{2,1}$ norm to measure the reconstructed errors. The objective function of $\ell_{2,1}$-NMF can be seen as follows

$$\min_{U\geq 0, V\geq 0} \quad \|X - UV^T\|_{2,1} = \sum_{i=1}^{n} \|(X - UV^T)_i\|_2, \tag{10}$$

where $\|\cdot\|_{2,1}$ denotes the $\ell_{2,1}$-norm. For the given matrix $M \in \mathbb{R}^{m\times n}$, its $\ell_{2,1}$-norm is defined as $\|M\|_{2,1} = \sum_j^n \sqrt{\sum_i^m M_{i,j}^2}$. It can be clearly observed that $\ell_{2,1}$-NMF uses the original reconstructed error $\|(X - UV^T)_i\|_2$, and removes the square. Therefore, $\ell_{2,1}$-NMF can better handle noises and outliers than the standard NMF. Moreover, Kong *et al*. [18] also provided an effective iterative updating algorithm as follows to solve $\ell_{2,1}$-NMF

$$U_{ik} = U_{ik} \frac{(XWV)_{ik}}{(UV^TWV)_{ik}}, \tag{11}$$

$$V_{jk} = V_{jk} \frac{(WX^TU)_{jk}}{(WVU^TU)_{jk}}, \tag{12}$$

$$W_i = 1/\|(X - UV^T)_i\|_2, \tag{13}$$

where W is a diagonal weight matrix with the diagonal element W_i. Experimental results [18] have shown that $\ell_{2,1}$-NMF can improve the robustness of NMF and achieve the better clustering performance.

4 The Proposed Hx-NMF and RBSMF

In this section, we firstly propose a general loss function $Hx(x)$, and then develop a more robust Hx-NMF using $Hx(x)$ function. After that, we derive a very efficient optimization method to solve the proposed Hx-NMF by iterative reweighted algorithm. Finally, combining the robust Hx-NMF and bi-stochastic graph learning strategy, we propose the RBSMF framework for data clustering.

4.1 Robust Loss Function Hx(x)

In this subsection, we propose a robust function $Hx(x)$ and its important properties are also given. Note that $Hx(x)$ is a general loss function which can be extensively applied in other reconstructed problems.

As we all know, there exists two widely used loss functions, L_2 and L_1 functions, and L_1 function is more robust than L_2 function. Here, we propose a general loss function, called $Hx(x)$, which is more robust than L_1 function. To be

specific, the three loss functions, their influence functions and weight functions are shown as follows

$$\begin{aligned} Hx(x) &= c * log(1 + \frac{|x|}{c}), \\ \psi_{Hx}(x) = c\frac{sgn(x)}{1 + \frac{|x|}{c}},\ & w_{Hx}(x) = \frac{c}{|x|(1 + \frac{|x|}{c})}, \end{aligned} \tag{14}$$

$$L_2(x) = \frac{x^2}{2},\ \psi_{L_2}(x) = x,\ w_{L_2}(x) = 1, \tag{15}$$

$$L_1(x) = |x|,\ \psi_{L_1}(x) = sgn(x),\ w_{L_1}(x) = \frac{1}{|x|}. \tag{16}$$

In our proposed loss function $Hx(x)$, $c > 0$ is a constant and we set $c = 1$ in subsequent analyses. It may have been noticed that the standard NMF and $\ell_{2,1}$-NMF employ L_2 and L_1 function respectively as their loss functions. Because the reconstructed error $\|(X - UV^T)_i\|_2$ is always greater than 0, the absolute value symbols ($|\cdot|$) of (14) and (16) can be removed.

The influence function $\psi(x)$ can reveal the influence of the reconstructed errors on the whole loss. For instance, for the L_2 function $L_2(x) = \frac{x^2}{2}$, its influence function is $\psi_{L_2}(x) = x$. This indicates the whole loss of L_2 function increases linearly with respect to the reconstructed errors. That is non-robust and sensitive to noises and outliers. Besides, the influence function of L_1 function is $\psi_{L_1}(x) = 1$ (due to $x > 0, sgn(x) = 1$). This implies that no matter how large the reconstructed errors are, they would has the same effect on the whole loss. From this perspective, we also prove that L_1 loss is more robust that L_2 loss. Similarly, for the proposed $Hx(x)$ loss function, its influence function is as Eq. (14). Due to $c = 1$ and $x > 0$, the influence function of $H(x)$ can be simplified into $\psi_{Hx}(x) = \frac{1}{1+x}$. It can be seen clearly that $\psi_{Hx}(x)$ is a strictly decreasing function, i.e., the larger the reconstructed error is, the less influence it has on the whole loss. Therefore, comparing with L_2 and L_1 functions, the proposed $Hx(x)$ is the most robust loss function. In order to better understand the superiority of the proposed $Hx(x)$, we draw the loss functions of $Hx(x)$, $L_2(x)$ and $L_1(x)$ in Fig. 1. In particular, we can see from Fig. 1 that: 1) $Hx(x) \leq L_1(x)$ whatever x-value is; 2) the value of $Hx(x)$ increases very slow. This indicates that $Hx(x)$ can decrease the influence of noises and outliers rather than enlarging the errors like $L_2(x)$, or keeping the errors unchanged like $L_1(x)$. Therefore, we also reach the same conclusion that the proposed $Hx(x)$ is the most robust loss function among the three functions according to Fig. 1.

4.2 Robust NMF Using Hx(x)

In this subsection, we propose a new robust NMF using the proposed robust loss fuction $Hx(x)$. It is more robust than the original NMF and $\ell_{2,1}$-NMF. Besides, the optimization algorithm is very efficient and has the almost same computational cost as the original NMF.

By considering $Hx(x)$ as the objective function, the new robust NMF model, called Hx-NMF, is formulated as

$$\min_{U\geq 0, V\geq 0} \quad \sum_{i=1}^{n} log(1 + \|(X - UV^T)_i\|_2). \tag{17}$$

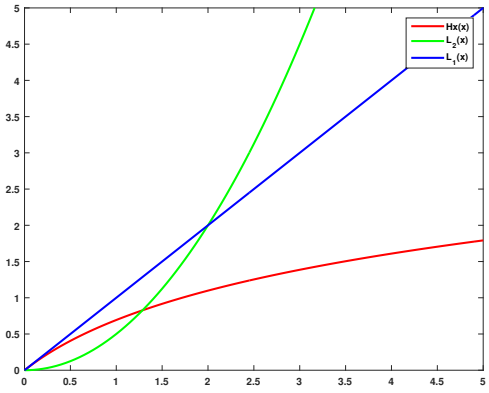


Fig. 1. Curves of the loss functions of $Hx(x)$, $L_2(x)$ and $L_1(x)$.

For simplifying the representation, we also use $(X - UV^T)_{Hx}$ to denote the objective function of Hx-NMF. Note that the new proposed Hx-NMF model is difficult to solve than the standard NMF. One main contribution of the paper is to build a connection between NMF and iterative reweighted algorithm, and derive an effective and elegant optimization method to solve Hx-NMF. To be specific, $Hx(x)$ function and $\|(X - UV^T)_i\|_2$ in Eq. (17) can be considered as $\Phi(\cdot)$ function and e_i in Eq. (1) respectively. According to the iterative reweighted algorithm in subsection 3.1, the proposed Hx-NMF (17) is first converted to the following iterative reweighted problem:

$$\begin{aligned} &\min_{U\geq 0, V\geq 0} \sum_i W_i * \|(X - UV^T)_i\|_2^2 \\ \Longleftrightarrow\ &\min_{U\geq 0, V\geq 0} Tr(X - UV^T)W(X - UV^T)^T, \end{aligned} \tag{18}$$

where $Tr(\cdot)$ stands for the trace of a matrix. W is a diagonal matrix whose diagonal element is the weight W_i of each sample. Next, we will discuss how to solve the proposed Hx-NMF. Denote A_{ik} and B_{jk} as the Lagrange multipliers for variables U_{ik} and V_{jk} respectively, and then we get the following Lagrangian function

$$\begin{aligned} \min_{U\geq 0, V\geq 0}\ & Tr(X - UV^T)W(X - UV^T)^T \\ & + Tr(AU^T) + Tr(BV^T). \end{aligned} \tag{19}$$

By making the partial derivatives with respect to U and V be 0, we obtain the following equations

$$U : 2UV^TWV - 2XWV + A = 0, \tag{20}$$

$$V : 2WVU^TU - 2WX^TU + B = 0. \tag{21}$$

According to KKT conditions $A_{ik}U_{ik} = 0$ and $B_{jk}V_{jk} = 0$, we have

$$U : (UV^TWV)_{ik}U_{ik} - (XWV)_{ik}U_{ik} = 0, \tag{22}$$

$$V : (WVU^TU)_{jk}V_{jk} - (WX^TU)_{jk}V_{jk} = 0. \tag{23}$$

The two equations suggest the following updating rules. Besides, combining the updating of the weight matrix W,

we obtain the final optimization algorithm to solve the proposed Hx-NMF,

$$U_{ik} = U_{ik}\frac{(XWV)_{ik}}{(UV^TWV)_{ik}}, \tag{24}$$

$$V_{jk} = V_{jk}\frac{(WX^TU)_{jk}}{(WVU^TU)_{jk}}, \tag{25}$$

$$W_i = \frac{1}{\|E_i\|_2 * (1 + \|E_i\|_2)}, \tag{26}$$

$$E_i = (X - UV^T)_i. \tag{27}$$

By observing the above updating rules, we find that the difference between the solving algorithms of the proposed Hx-NMF and $\ell_{2,1}$-NMF is only the updating of W. In fact, we can consider iterative reweighted algorithm as a unified framework to solve the standard NMF, $\ell_{2,1}$-NMF and the proposed Hx-NMF. For the standard NMF and $\ell_{2,1}$-NMF, the updating of W is $W_i = 1$ and $W_i = 1/\|E_i\|_2$ respectively according to Eq. (15) and Eq. (16).

4.3 Robust Bi-stochastic Graph Regularized NMF

In this subsection, we firstly analyse some problems of the existing graph regularized NMF. In order to handle the problems, we present to use bi-stochastic matrix to adaptively learn a better graph from an input low-quality graph.

Due to the non-negative constraints on U and V, NMF can acquire part-based representation for given data matrix X. Therefore, NMF gives a better physiological and psychological interpretation than other methods for non-negative data, such as documents and face images. However, as is seen from (7), NMF can only explore the global structure of data X. So some graph-regularized variants are presented. A typical variant is GNMF [19] which characterizes the local data relationships by k-nearest neighbor (k-nn) graph. GNMF is to solve the following problem

$$\min_{U\geq 0, V\geq 0} \quad \|X - UV^T\|_F^2 + \lambda Tr(V^T L_S V), \tag{28}$$

where L_S is graph Laplacian defined as $L_S = D_S - S$. S is similarity matrix, and the diagonal matrix D_S is called degree matrix with the elements $D_{ii} = \sum_j^n S_{ij}$. The second term encourages that samples with higher similarity have smaller distance, which improves the original NMF by exploiting local geometric features.

However, the graph constructed by k-nn is too simple and is not discriminative enough. It can not well represent the original data structure. Besides, the input graph is fixed in the process of matrix factorization. In this paper, we focus on the more high-quality graph construction to improve the performance of NMF, thus clustering result also would become better. Given an initial low-quality graph H, we want to learn a high-quality graph S based on H. Generally speaking, S should has the following properties. First, S depicts the probability of each two data samples to be neighbors. So the sum of every row of S is 1, i.e., $S\mathbf{1} = \mathbf{1}$. Then $L_S = D_S - S = I - S$, where I is identity matrix. Second, S represents the similarity of each pair of points. So it needs to be non-negative and symmetric, i.e., $S = S^T$, and $S \geq 0$. With the above constraints (i.e., $S \geq 0$, $S = S^T$, and $S\mathbf{1} = \mathbf{1}$) on similarity matrix S, it is also called bi-stochastic matrix [34] [35] [36]. Besides, according to the graph theory [5], each vertex is not allowed to be connected by itself. So we add the constraint $diag(S) = \mathbf{0}$. Finally, we can learn a bi-stochastic matrix from the input graph H as the new graph by solving

$$\begin{aligned} \min_{S} \quad & \|S - H\|_F^2 \\ s.t. \quad & S \geq 0, S = S^T, S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0}. \end{aligned} \tag{29}$$

Combining the robust Hx-NMF (17), we propose the following robust bi-stochastic graph regularized matrix factorization (RBSMF) framework

$$\begin{aligned} \min_{U,V,S} \quad & (X - UV^T)_{Hx} + \alpha Tr(V^T L_S V) + \beta\|S - H\|_F^2 \\ s.t. \quad & V \geq 0, V^TV = I, S \geq 0, S = S^T, \\ & S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0}. \end{aligned} \tag{30}$$

where α and β are two regularization parameters, which control the weights of manifold learning term and graph learning term respectively. $(X - UV^T)_{Hx}$ is the simplification of $\sum_{i=1}^n log(1 + \|(X - UV^T)_i\|_2)$. Furthermore, the orthogonal constraint $V^TV = I$ has the following two advantages: 1) It guarantees the solution of RBSMF is unique. Assume that U^* and V^* are the solutions of the proposed RBSMF (30), and then for any given positive diagonal matrix D, UD and VD^{-1} would have the same value in $(X - UV^T)_{Hx}$ and a lower value in $Tr(V^T L_S V)$. To eliminate the uncertainty, the constraint $V^TV = I$ is added into the proposed RBSMF. 2) The final clustering results can be obtained directly from clustering indicator matrix V without employing other post-processing algorithms (e.g., k-means) like NMF and GNMF. Note that the non-negative constraint on U is dropped, suggested by [20], to make the proposed RBSMF applicable for mixed data.

In a word, the proposed RBSMF can improve the robustness and learn a high-quality graph adaptively to obtain more important and accurate local structure information, which would definitely give a better performance of the clustering task. Moreover, compared with NMFAN which is also an adaptive graph regularized NMF method, RBSMF does not have many hyperparameters to be tuned and is more suitable for practical applications.

5 Optimization

Since we learn the bi-stochastic graph S and clustering indicator matrix V simultaneously, and the original F-norm in NMF is replaced with $Hx(x)$ function, the conventional iterative updating algorithm in [8] is no longer suitable for our proposed RBSMF. In this section, we develop a new iterative algorithm, which is based on Augmented Lagrange Multiplier (ALM) [37], to solve the proposed objective function (30). In order to make the problem (30) easily solvable, we firstly add two auxiliary variables Z and E. Then the equivalent problem of (30) can be obtained as follows

$$\begin{aligned} \min_{U,V,Z,E,S} \quad & (E)_{Hx} + \alpha Tr(V^T L_S Z) + \beta\|S - H\|_F^2 \\ s.t. \quad & Z \geq 0, Z = V, V^TV = I, S \geq 0, S = S^T, \\ & S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0}, E = X - UV^T. \end{aligned} \tag{31}$$

Based on problem (31), the augmented Lagrangian function is written as

$$\begin{aligned}&\min_{U,V,Z,E,S} \quad (E)_{Hx} + \alpha Tr(V^T L_S Z) + \beta\|S-H\|_F^2 \\ &+ \frac{\mu}{2}\|Z - V + \frac{Y_1}{\mu}\|_F^2 + \frac{\mu}{2}\|X - UV^T - E + \frac{Y_2}{\mu}\|_F^2 \\ &s.t. \quad Z \geq 0, V^T V = I, S \geq 0, S = S^T, \\ &\quad S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0},\end{aligned} \tag{32}$$

where μ is the penalty coefficient to control the unequal level of three equation constrains, and Y_1, Y_2 are Lagrange multipliers. Then it is easy to solve problem (32) by alternative optimization strategy, i.e., when updating a variable, all the other variables are fixed and viewed as constants. The detailed updating steps can be seen as follows

Update E: fix U, V, Z, S, and then E can be solved as follows

$$\min_{E} \quad (E)_{Hx} + \frac{\mu}{2}\|X - UV^T - E + \frac{Y_2}{\mu}\|_F^2. \tag{33}$$

Another contribution of this paper is that we derive an effective optimization algorithm to solve a general problem like (33). We summarize the solving algorithm as the following Theorem 1 and also give the detailed proof.

Theorem 1 *Given a positive constant λ and a matrix $W = [W_1, W_2, ..., W_n] \in \mathbb{R}^{m\times n}$, let X^* be the optimal solution of the following general problem*

$$\min_{X} \quad \lambda(X)_{Hx} + \frac{1}{2}\|X - W\|_F^2, \tag{34}$$

and then the i-th column of X^* is

$$X^*(:,i) = \begin{cases} (1 - \frac{\lambda}{\lambda+a+a^2})W_i, & if\ \lambda < \|W_i\|_2 \\ 0, & otherwise, \end{cases} \tag{35}$$

where $a = (\|W_i\|_2 - 1 + \sqrt{(\|W_i\|_2+1)^2 - 4\lambda})/2$.

Proof: Problem (34) is firstly expanded as the following problem

$$\min_{X} \quad \lambda\sum_{i=1}^{n} log(1 + \|X_i\|_2) + \frac{1}{2}\|X - W\|_F^2. \tag{36}$$

Note that problem (36) is independent for each X_i, so it can be converted to the equivalent problem

$$\min_{X_i} \quad \lambda * log(1 + \|X_i\|_2) + \frac{1}{2}\|X_i - W_i\|_2^2. \tag{37}$$

As we know, the derivative of $\|X_i\|_2$ with respect to X_i is

$$\frac{d\|X_i\|_2}{dX_i} = \begin{cases} \mathbf{r}, & X_i = \mathbf{0} \\ \frac{X_i}{\|X_i\|_2}, & otherwise, \end{cases} \tag{38}$$

where $\mathbf{r}$ is called subgradient, and $\|\mathbf{r}\|_2 \leq 1$.

We can set the derivative of (37) with respect to X_i as 0 to solve problem (37). The detailed solving process is divided into the following two cases.

1) For the case $X_i = \mathbf{0}$, we have

$$\lambda\mathbf{r} - W_i = 0, \tag{39}$$

which indicates $\lambda \geq \|W_i\|$.

2) For the case $X_i \neq \mathbf{0}$ (i.e., $\lambda < \|W_i\|_2$), we have

$$\frac{\lambda X_i}{(1 + \|X_i\|_2)\|X_i\|_2} + X_i - W_i = 0. \tag{40}$$

Let $a = \|X_i\|_2$, and then we have

$$X_i = \frac{a + a^2}{\lambda + a + a^2} W_i. \tag{41}$$

By using the operator $\|\cdot\|_2$ on both sides of Eq. (41), we have

$$a = \frac{a + a^2}{\lambda + a + a^2} b, \tag{42}$$

where $b = \|W_i\|_2$. After solving Eq. (42), we have

$$a = \frac{b - 1 + \sqrt{(b+1)^2 - 4\lambda}}{2}. \tag{43}$$

Substituting Eq. (43) into Eq. (41) and combining the case $X_i = \mathbf{0}$, we obtain the final optimal solution Eq. (35) of problem (34). □

Let $Q = X - UV^T + Y_2/\mu$, and then we obtain the following optimal solution of problem (33) according to Theorem 1,

$$E(:,i) = \begin{cases} (1 - \frac{1/\mu}{1/\mu+a+a^2})Q_i, & if\ \frac{1}{\mu} < \|Q_i\|_2 \\ 0, & otherwise, \end{cases} \tag{44}$$

where $a = (\|Q_i\|_2 - 1 + \sqrt{(\|Q_i\|_2+1)^2 - 4/\mu})/2$.

Update U: fix V, Z, E, S, and then U can be solved as follows

$$\min_{U} \quad \frac{\mu}{2}\|X - UV^T - E + \frac{Y_2}{\mu}\|_F^2. \tag{45}$$

By making the derivative of problem (45) be 0, we obtain the following optimal U,

$$U = (X - E + \frac{Y_2}{\mu})V. \tag{46}$$

Update V: fix U, Z, E, S, and then V can be solved as follows

$$\begin{aligned}&\min_{V^TV=I} \quad \alpha Tr(V^T L_S Z) + \frac{\mu}{2}\|Z - V + \frac{Y_1}{\mu}\|_F^2 \\ &\qquad + \frac{\mu}{2}\|X - UV^T - E + \frac{Y_2}{\mu}\|_F^2 \\ &\Longleftrightarrow \min_{V^TV=I} \quad \|V - P\|_F^2,\end{aligned} \tag{47}$$

where $P = (Z + \frac{Y_1}{\mu}) + (X^T - E^T + \frac{Y_2^T}{\mu})U - \frac{\alpha}{\mu}L_S Z$. According to [20], the solution is $V = FG^T$, where F and G are left and right singular vectors of SVD decomposition of P.

Update Z: fix U, V, E, S, and then Z can be solved as follows

$$\begin{aligned}&\min_{Z\geq 0} \quad \alpha Tr(V^T L_S Z) + \frac{\mu}{2}\|Z - V + \frac{Y_1}{\mu}\|_F^2 \\ &\Longleftrightarrow \min_{Z\geq 0} \quad \|Z - R\|_F^2,\end{aligned} \tag{48}$$

where $R = V - \frac{Y_1}{\mu} - \frac{\alpha}{\mu}L_S^T V$. The solution is obtained by

$$Z = max(R, 0). \tag{49}$$

Update S: fix U, V, Z, E, and then S can be solved as follows

$$\begin{aligned} \min_{S} \quad & \alpha Tr(V^T L_S Z) + \beta \|S - H\|_F^2 \\ s.t. \quad & S \geq 0, S = S^T, S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0}. \end{aligned} \tag{50}$$

Let $M = H + \frac{\alpha}{2\beta} V Z^T$, and then problem (50) can be rewritten as

$$\begin{aligned} \min_{S} \quad & \|S - M\|_F^2 \\ s.t. \quad & S \geq 0, S = S^T, S\mathbf{1} = \mathbf{1}, diag(S) = \mathbf{0}. \end{aligned} \tag{51}$$

In order to solve problem (51), it is firstly divided into the following two subproblems

$$\min_{S} \quad \|S - M\|_F^2, \quad s.t. \quad S = S^T, S\mathbf{1} = \mathbf{1}. \tag{52}$$

and

$$\min_{S} \quad \|S - M\|_F^2, \quad s.t. \quad S \geq 0, diag(S) = \mathbf{0}. \tag{53}$$

Then we solve the above two subproblems alternately, and project their solutions mutually. Specifically, we iteratively carry out the following two steps until convergence: 1) Obtain the optimal solution S_1 of subproblem (52), and view S_1 as M of subproblem (53); 2) Obtain the optimal solution S_2 of subproblem (53), and view S_2 as M of subproblem (52).

The convergence of the above solving strategy is guaranteed by Von Neumann successive projection lemma [38]. The lemma proves theoretically that the solution of mutual projection strategy finally converges to the global optimal solution of the original problem (51).

According to [36], the optimal solution of subproblem (52) is

$$S_1 = T + \frac{n + \mathbf{1}^T T \mathbf{1}}{n^2} \mathbf{1}\mathbf{1}^T - \frac{1}{n} T \mathbf{1}\mathbf{1}^T - \frac{1}{n} \mathbf{1}\mathbf{1}^T T, \tag{54}$$

where $T = \frac{M + M^T}{2}$. $\mathbf{1}$ is a vector with all elements 1, while $\mathbf{11}$ is a square matrix with all elements 1.

Subproblem (53) is easily solved by

$$S_2 = max(M, 0), \quad diag(S_2) = \mathbf{0}. \tag{55}$$

Update ALM parameters: some parameters with respect to ALM algorithm need to be updated as follows

$$\begin{aligned} Y_1 &= Y_1 + \mu(Z - V), \\ Y_2 &= Y_2 + \mu(X - UV^T - E), \\ \mu &= \rho\mu. \end{aligned} \tag{56}$$

Then, we evaluate the computational complexity of the proposed RBSMF. Because multiplication operations dominate all the computational complexity, we use the number of multiplication operations as the computational complexity of the proposed RBSMF. The computation of RBSMF can be divided into five parts as follows. Note that m, n and k represent the number of samples, features, and clusters respectively. 1) When updating variable E, the complexity is $O(mnk)$; 2) When updating variable U, the complexity is $O(mnk)$; 3) When updating variable V, the complexity is $O(n^2k)$; 4) When updating variable Z, the complexity is $O(n^2k)$; 5) When updating variable S, the complexity is $O(n^2k)$. In most cases, $m < n$. So computational complexity of the proposed RBSMF is $O(n^2k)$ each iteration.

At last, we summarize the whole optimization algorithm in Algorithm 2 for solving the proposed RBSMF (30). Similar to the arguments in [20] and [39], the convergence of Algorithm 2 relies on the convergence of ALM framework. The convergence of ALM framework has been discussed and proved in previous literatures [37] [40] [41].

Algorithm 2 ALM for solving RBSMF

Input: Data matrix X, cluster number k, the initial graph H, parameter α, β.

Output: Cluster centroid matrix U, cluster indicator matrix V, the learned bi-stochastic graph S.

Initialize: $\varepsilon = 10^{-2}$, $h = 1$, $maxiter = 500$.

While not converged **do**

1) Update E according to (44).
2) Update U according to (46).
3) Update V by solving (47).
4) Update Z according to (49).
5) Update S according to (54) and (55).
6) Update ALM parameters according to (56).
7) Check the convergence conditions $h > maxiter$, or $\|Z - V\|_\infty < \varepsilon$, $\|X - UV^T - E\|_\infty < \varepsilon$.
8) $h \leftarrow h + 1$.

End while

6 Experiments

In this section, we firstly perform some experiments on synthetic and real-world data to verify the robustness of the proposed Hx-NMF. Then extensive experiments on eight popular datasets are conducted to validate the clustering performance of the proposed RBSMF framework.

6.1 Robustness of Hx-NMF

6.1.1 Experiments on Synthetic Data

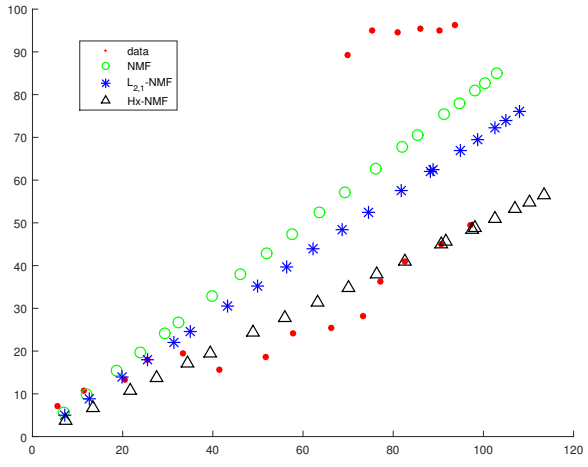


Fig. 2. Comparison on synthetic data of three NMF methods.

Before demonstrating the superior clustering performance of the proposed RBSMF, we firstly validate the robustness of the proposed Hx-NMF for noises and outliers.

The comparison algorithms are as follows: 1) the standard NMF [8]: using Frobenius norm as objective function. The corresponding weight function is constant 1; 2) $\ell_{2,1}$-NMF [18]: using $\ell_{2,1}$ norm as objective function. The corresponding weight function is $\frac{1}{x}$; 3) the proposed Hx-NMF: using our proposed robust $Hx(x)$ (14) as objective function. The corresponding weight function is $\frac{1}{x(1+x)}$.

We firstly utilize 2-dimensional synthetic data to verify the robustness of the proposed Hx-NMF. As is seen from Fig. 2, we generate 20 data points, which are drawn in red. Among 20 original samples, the top 6 points in Fig. 2 are outliers, which can be easily observed that they deviate from the normal data structure. After running the standard NMF, $\ell_{2,1}$-NMF, and the proposed Hx-NMF, we project all data points into 1-dimensional subspace and draw the reconstructed samples with different colors and shapes in Fig. 2. It can be clearly seen that the performance is "Hx-NMF $>$ $\ell_{2,1}$-NMF $>$ the standard NMF" according to robustness. To be specific, the standard NMF is strongly influenced by 6 outliers because the square errors dominate the objective function. $\ell_{2,1}$-NMF improves the robustness of NMF by replacing Frobenius norm with $\ell_{2,1}$ norm, where the square operation of errors is removed. Although $\ell_{2,1}$-NMF can decrease the sensitivity of noises and outliers, it is still affected by 6 outliers according to Fig. 2. The proposed Hx-NMF is the most robust method among all comparison algorithms. The fact is also verified from Table 2, which shows the weights of 6 outliers for all methods. As is shown in Table 2, these weights for the proposed RBSMF are the smallest (close to 0). So the proposed Hx-NMF can avoid the negative influence of six outliers and achieve the best robustness.

TABLE 2
Weights of 6 outliers for all methods

Methods	1	2	3	4	5	6
NMF	1	1	1	1	1	1
$\ell_{2,1}$-NMF	0.0305	0.0292	0.0351	0.0326	0.0402	0.0390
Hx-NMF	4.1e-4	3.7e-4	4.4e-4	4.1e-4	4.9e-4	4.9e-4

6.1.2 Experiments on Real-world Data

In order to further prove the robustness of the proposed Hx-NMF, we conduct a series of experiments on a real-world dataset. The ORL face database (its detailed descriptions are in section 6.2.2) is selected to perform the experiments. All the images are resized to 23×28 here. We randomly add some square block noises with different sizes to four face images of each person. Some noisy images of ORL datasets are shown in Fig. 3.

We set the sizes of block as 2×2, 4×4, 6×6, and 8×8. The comparison methods are k-means, the standard NMF, $\ell_{2,1}$-NMF, and the proposed Hx-NMF. We report the clustering results of all algorithms in Table 3. ACC, NMI and PUR (their detailed descriptions are in section 6.2.1) are considered as the evaluation indexes. Following [18], we initialize U and V of NMF, $\ell_{2,1}$-NMF and the proposed Hx-NMF using k-means. It is easily seen from Table 3 that the proposed Hx-NMF can achieve the best clustering performance no matter how large the noises are. Therefore, the experiments on real-world data also demonstrate that the proposed Hx-NMF is more robust than the standard NMF and $\ell_{2,1}$-NMF.

Futuremore, we conduct the experiment using ORL dataset with 8×8 block noises to verify the convergence of the proposed Hx-NMF. As is shown in Fig. 4, Hx-NMF also has a good convergence like NMF and $\ell_{2,1}$-NMF.

Fig. 3. Some noisy images of ORL dataset. The upper images have 4×4 block noises, and the lower images have 8×8 block noises.

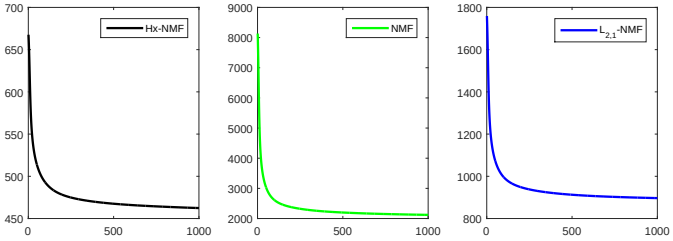


Fig. 4. Convergence of the proposed Hx-NMF, NMF and $\ell_{2,1}$-NMF.

TABLE 3
Experimental results of ORL dataset with some noises

block	Metric	k-means	NMF	$\ell_{2,1}$-NMF	Hx-NMF
2×2	ACC	0.6525	0.6727	0.6730	**0.6815**
	NMI	0.8261	0.8266	0.8309	**0.8318**
	PUR	0.6973	0.7195	0.7165	**0.7220**
4×4	ACC	0.6205	0.6520	0.6560	**0.6662**
	NMI	0.8000	0.8089	0.8080	**0.8183**
	PUR	0.6543	0.6923	0.6943	**0.7121**
6×6	ACC	0.5350	0.5495	0.5617	**0.5855**
	NMI	0.7219	0.7386	0.7479	**0.7571**
	PUR	0.5615	0.5855	0.6007	**0.6275**
8×8	ACC	0.4137	0.3888	0.3915	**0.4292**
	NMI	0.6299	0.6130	0.6101	**0.6324**
	PUR	0.4355	0.4122	0.4200	**0.4525**

6.2 Clustering Performance of RBSMF

6.2.1 Evaluation Indexes

In the subsection, we will introduce three widely used quantitative metrics to evaluate the clustering performance. They can be seen in detail as follows

Clustering Accuracy (ACC): It can not only discover the one-to-one relationship between clustering results and true

classes, but also acquire the data points that each cluster contains from the corresponding class. Specifically, it is defined as below

$$ACC = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \quad (57)$$

where r_i is the clustering label of sample x_i, and l_i denotes the true label of x_i. n is the number of all data samples. $map(r_i)$ is the optimal matching function which can permutate all clustering results to best map clustering labels to the true labels. $\delta(a, b)$ is the indicator function which equals 1 if $a = b$, and equals 0 otherwise.

Normalized Mutual Information (NMI): It is another commonly used metric to measure the clustering quality. The detailed definition of NMI is given as follows

$$NMI = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{i,j} log\frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i log\frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j log\frac{\hat{n}_j}{n})}}, \quad (58)$$

where c denotes the number of classes, and n_i is the number of samples belonging to the cluster $C_i (1 \leq i \leq c)$. $\hat{n}_j$ denotes the number of samples contained in the class $L_j (1 \leq j \leq c)$, and $n_{i,j}$ represents the number of overlapped samples between cluster C_i and class L_j.

Purity (PUR): It aims to measure the extent to which every cluster covers data samples from primarily one class. The purity of clustering results is computed by the weighted sum of all individual cluster purity values. The value of purity is obtained by

$$PUR = \sum_{i=1}^{c} \frac{n_i}{n} P(S_i), \quad P(S_i) = \frac{1}{n_i} \max_j n_i^j, \quad (59)$$

where S_i represents the particular cluster with size n_i, and n_i^j is the number of samples of the i-th class which are assigned to the j-th cluster.

More detailed descriptions about these evaluation indexes can be found in [18] [20].

TABLE 4
Descriptions of eight datasets

Datasets	#Samples (n)	#Dimensions (m)	#Class (c)
ORL	400	168	40
Yale	165	256	15
USPS	300	256	10
UMIST	575	644	20
JAFFE	213	256	10
Seeds	210	7	3
Ecoli	336	7	8
Vote	435	16	2

6.2.2 Dataset Descriptions

In this subsection, we introduce eight publicly available datasets, which are used to evaluate the clustering performance of the proposed RBSMF and other comparison methods. The eight datasets include four face datasets, one hand-written digit dataset, and three other datasets from UCI Machine Learning Repository (UCI for short) [42]. Table 4 summarizes the statistics of the eight datasets.

1) **ORL**: The ORL[1] face recognition database contains 400 grayscale images in total. It consists of 40 different persons and each subject has 10 images, which are taken under various conditions. All images have a homogeneous dark background with each person in a frontal, upright position. Besides, we resize all images to 12×14 in the comparison experiments.
2) **Yale**: The Yale[1] face recognition database has 15 different individuals, each of which is taken 11 images. All images are acquired from different conditions, such as normal/sleepy/sad, center/left/right-lighting, with/without glasses, and so on. Each image is scaled to 16×16 in our experiments.
3) **USPS**: The USPS[2] database is the US Postal handwritten digit dataset. It consists of 8-bit grayscale images from 0 to 9. The whole dataset has 9298 images in total. In our experiments we randomly select 30 images per class and keep the original pixel size of 16×16.
4) **UMIST**: The UMIST [43] database is another face recognition dataset. It includes 20 persons with 19-36 grayscale images with each individual. The 575 images are obtained from various views. In our experiments all images are resized to 23×28.
5) **JAFFE**: The JAFFE[3] database is a female expression dataset. It includes 213 images of 7 different emotional faces from 10 Japanese females. The 7 facial expressions contain 1 natural and 6 basic expressions. All images are resized to 16×16 in our experiments.
6) **Seeds**: The Seeds[4] database is from UCI [42]. It covers 210 different instances of 7 attributes. The number of classes is 3.
7) **Ecoli**: The Ecoli[5] database is also from UCI [42]. It consists of 336 objects, and each of them has 7 features. There are 8 different categories in Ecoli.
8) **Vote**: The Vote[6] database is another dataset from UCI [42]. It constains 435 different samples, each of which has 16 attributes. The number of classes is 2.

6.2.3 Comparison Methods

To verify the effectiveness and superiority of the proposed RBSMF for clustering, we consider other eight clustering algorithms as the comparison methods. They are listed in detail as follows.

1) **k-means** [44]: the most commonly used clustering method.
2) **Rcut** [45]: a spectral clustering using ratio cut.
3) **Ncut** [46]: another spectral clustering using normalized cut.
4) **NMF** [8]: the original NMF using Frobenius norm.
5) $\ell_{2,1}$**-NMF** [18]: a robust NMF using $\ell_{2,1}$ norm.

1. http://www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html
2. http://www.cad.zju.edu.cn/home/dengcai/Data/MLData.html
3. http://www.kasrl.org/jaffe.html
4. https://archive.ics.uci.edu/ml/datasets/seeds
5. https://archive.ics.uci.edu/ml/datasets/Ecoli
6. https://archive.ics.uci.edu/ml/datasets/congressional+voting+records

TABLE 5
ACC of ten clustering algorithms on eight datasets

Datasets	k-means	RCut	NCut	NMF	$\ell_{2,1}$-NMF	Hx-NMF	GNMF	RMNMF	NMFAN	RBSMF
ORL	0.6645	0.6743	0.7125	0.6863	0.6877	0.6945	0.7177	0.7395	0.7421	**0.7625**
Yale	0.5148	0.5524	0.5585	0.5279	0.5381	0.5533	0.5315	0.5437	0.5476	**0.5636**
USPS	0.6785	0.6907	0.6867	0.6313	0.6487	0.6547	0.6563	0.6585	0.6626	**0.6967**
UMIST	0.4357	0.5979	0.5704	0.4120	0.4228	0.4275	0.6052	0.5814	0.6135	**0.6221**
JAFFE	0.8333	0.9226	0.9231	0.8704	0.8948	0.9173	0.9343	0.9156	0.9203	**0.9577**
Seeds	0.8852	0.8429	0.8429	0.8385	0.8425	0.8571	0.8614	0.8625	0.8678	**0.8900**
Ecoli	0.5747	0.5655	0.5512	0.5658	0.5458	0.5539	0.5907	0.6125	0.5945	**0.6332**
Vote	0.8387	0.8493	0.8467	0.8018	0.8046	0.8082	0.8255	0.8432	0.8348	**0.8621**

TABLE 6
NMI of ten clustering algorithms on eight datasets

Datasets	k-means	RCut	NCut	NMF	$\ell_{2,1}$-NMF	Hx-NMF	GNMF	RMNMF	NMFAN	RBSMF
ORL	0.8328	0.8308	0.8553	0.8347	0.8413	0.8422	0.8628	0.8632	0.8693	**0.8711**
Yale	0.5789	0.5827	0.5744	0.5541	0.5673	0.5926	0.5642	0.5702	0.5802	**0.5967**
USPS	0.6176	0.6362	0.6349	0.6349	0.5936	0.5978	0.6071	0.6087	0.6115	**0.6372**
UMIST	0.6521	0.7768	0.7589	0.6003	0.6163	0.6179	0.7944	0.7756	0.8125	**0.8153**
JAFFE	0.8632	0.9265	0.9261	0.8795	0.8872	0.9050	0.9356	0.9145	0.9234	**0.9358**
Seeds	0.6634	0.6305	0.6305	0.5617	0.5948	0.6146	0.6134	0.6146	0.6175	**0.6720**
Ecoli	0.5402	0.5371	0.5294	0.5048	0.4567	0.4752	0.5485	0.5634	0.5512	**0.5821**
Vote	0.3738	0.3895	0.3847	0.3065	0.3064	0.3148	0.3585	0.3785	0.3645	**0.4128**

TABLE 7
PUR of ten clustering algorithms on eight datasets

Datasets	k-means	RCut	NCut	NMF	$\ell_{2,1}$-NMF	Hx-NMF	GNMF	RMNMF	NMFAN	RBSMF
ORL	0.7028	0.7133	0.7483	0.7220	0.7296	0.7308	0.7510	0.7785	0.7854	**0.7900**
Yale	0.5355	0.5717	0.5766	0.5382	0.5515	0.5673	0.5394	0.5524	0.5632	**0.5818**
USPS	0.6922	0.7113	0.7127	0.6433	0.6533	0.6617	0.6670	0.6696	0.6732	**0.7133**
UMIST	0.5178	0.6868	0.6697	0.4824	0.5007	0.5066	0.7061	0.6825	0.7047	**0.7121**
JAFFE	0.8451	0.9277	0.9282	0.8803	0.9001	0.9174	0.9366	0.9162	0.9225	**0.9577**
Seeds	0.8852	0.8429	0.8429	0.8386	0.8425	0.8571	0.8614	0.8571	0.8596	**0.8900**
Ecoli	0.8243	0.8280	0.8030	0.7991	0.7875	0.7830	0.8314	0.8452	0.8275	**0.8521**
Vote	0.8387	0.8375	0.8426	0.8018	0.8051	0.8085	0.8268	0.8432	0.8352	**0.8621**

6) **Hx-NMF**: a more robust NMF using the proposed $Hx(x)$ function.
7) **GNMF** [8]: the graph regularized NMF incorporating the low dimensional manifold into NMF.
8) **RMNMF** [20]: a more effective NMF which uses $\ell_{2,1}$ norm and simultaneously combines the local data structures.
9) **NMFAN** [32]: an improved NMF by utilizing adaptive neighbors to better incorporate geometric information .
10) **RBSMF**: the proposed method in our paper.

6.2.4 Experimental Settings

Before conducting the experiments, every value of all datasets is normalized to [0, 1]. For RCut, NCut, GNMF and RMNMF, we construct the k-nn graphs to exploit the latent data structure and the corresponding parameter k is set by searching from {3, 5, 7, 9, 11 ,13}. For GNMF, RMNMF and NMFAN, the regularization parameters are determined by searching the grid {0.001, 0.01, 0.1, 1, 10, 100}. Because the number of clusters is given, no other parameters need to be tuned for k-means, NMF, $\ell_{2,1}$-NMF and the proposed Hx-NMF.

For the proposed RBSMF, we set the regularization parameters α and β by searching the grid {0.001, 0.01, 0.1, 1, 10, 100}. Because our RBSMF can learn adaptively the more suitable graph, we just construct a simple 5-nn graph as the input graph. The experimental machine is a PC with 1) CPU: Intel Core i7-3770 3.40 GHz; 2) Memory: 32-GB RAM; 3) Sofeware: MATLAB R2015a.

Because some comparison methods need to tune one or more parameters, to make experiments more fair, we run these algorithms with different parameters and choose the best clustering results. The ground truth label number is set to the number of clusters for all methods. Besides, k-means is used to initialize the values of U and V for NMF-based algorithms. Note that NMF, $\ell_{2,1}$-NMF, Hx-NMF, GNMF and NMFAN need to employ k-means as post-processing method to get the final clustering results. RMNMF and the proposed RBSMF can directly obtain the clustering results by the largest element in every row of V, i.e., $class(X_i) = \arg\max_{k=1,\ldots,c} V_{ik}$. In addition, we repeat inde-

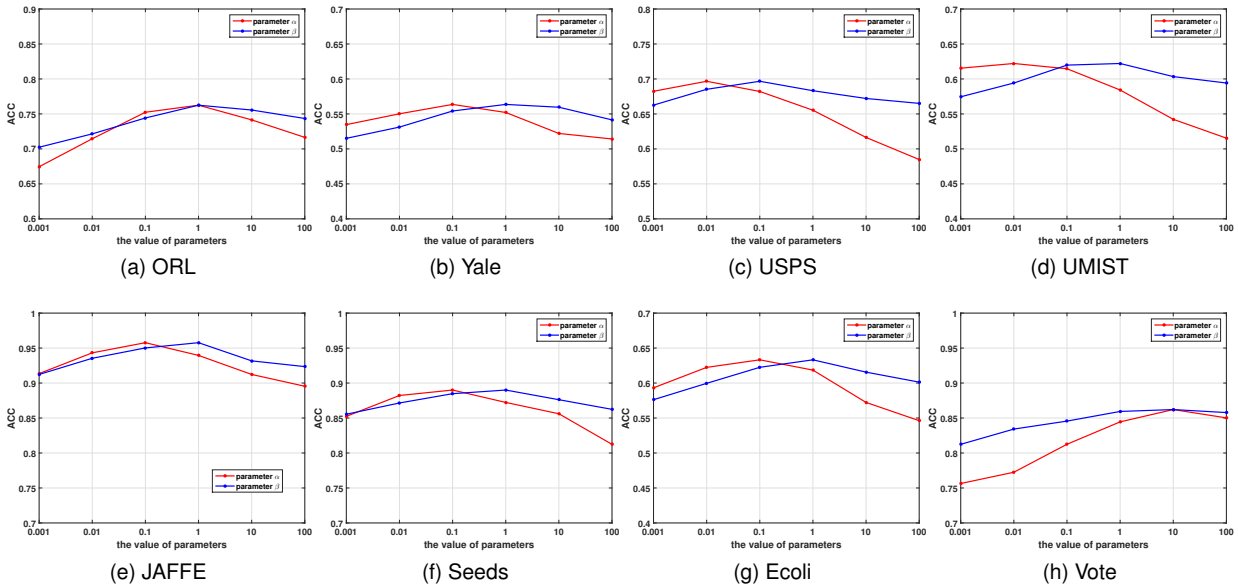


Fig. 5. Clustering ACC of the proposed RBSMF with respect to the parameters α and β in eight datasets.

pendently each method ten times and report the average clustering results under the best parameter settings.

6.2.5 Experimental Results and Analyses

After running every method under the best paremeter settings, all the clustering results of ten different algorithms are demonstrated in Table 5, 6, and 7. They are reported by the aforementioned evaluation indexes, i.e., ACC, NMI and PUR. As is seen from the experimental results, we can obtain the follows interesting points and some detailed analyses.

- In most datasets, the clustering performance of RCut and NCut is better than k-means. Meanwhile, GNMF, RMNMF, NMFAN, and the proposed RBSMF (all of them incorporate local geometrical structures) outperforms the standard NMF. The fact indicates that the local manifold information between neighboring data can bring great help for clustering task.
- $\ell_{2,1}$-NMF, Hx-NMF, RMNMF, and the proposed RBSMF can achieve better clustering performance than the original NMF. Among them, $\ell_{2,1}$-NMF and RMNMF replace Frobenius norm with $\ell_{2,1}$ norm as the objective function. Besides, Hx-NMF and RBSMF propose a more robust loss function $Hx(x)$ to measure the reconstructed errors. It reveals that robust objective function, which is insensitive to noises and outliers, can improve the clustering results of NMF.
- It can be noticed that both NMFAN and RBSMF yield better clustering performance than GNMF and RMNMF in most datasets. This is because NMFAN and RBSMF can learn an adaptive similarity graph. The dynamic learning process makes the obtained graph more suitable for clustering. Although GNMF and RMNMF, which also combine the local data structures, can enhance the clustering performance of NMF, the input data graph is fixed in the whole matrix factorization process and may not the optimal graph.
- It can be clearly observed that the proposed RBSMF consistently outperforms other nine comparison algorithms and achieves the best clustering results. On one hand, the proposed RBSMF presents a more robust loss function than the widely-used $\ell_{2,1}$ norm. This decreases the negative influence of noises and outliers. On the other hand, the proposed RBSMF proposes an adaptive updating scheme to learn dynamically an optimal graph instead of keeping the initial graph fixed.

6.2.6 Parameter Sensitivity

Similar to most other algorithms, the proposed RBSMF also needs to tune some hyperparameters, i.e., the two important regularization parameters α and β. In particular, α balances the manifold learning term, and β controls the weight of the graph learning term. In this subsection, we conduct a great number of experiments to investigate the sensitivity of α and β. The parameter of β is set as the optimal value when studying α. The same setting is also applied when studying β. Fig. 5 exhibits the clustering ACC curves of eight datasets when α and β change from the range of $\{0.001, 0.01, 0.1, 1, 10, 100\}$. It can be seen that both two parameters are not very sensitive. The clustering results of the proposed RBSMF are good generally when α and β are within the range of $\{0.1, 1, 10\}$.

7 Conclusion and Future Work

In this paper, a robust bi-stochastic graph regularized NMF (RBSMF) framework is proposed for data clustering. In order to address the existing problems of two kinds of widely concerned NMF, i.e., robust NMF and graph-based NMF, the proposed RBSMF develops a general robust loss function $Hx(x)$ and an adaptive graph learning strategy. Specifically, $Hx(x)$ function decreases the adverse impact of large noises and it is more robust than L_2 and L_1 functions. Besides,

the proposed loss function $Hx(x)$ can be extensively employed in other fields. For the graph learning strategy, it can learn a better graph using bi-stochastic matrix, which is proved more suitable than heat kernel matrix to capture the local data structures. Furthermore, the proposed RBSMF can simultaneously perform matrix factorization and graph learning. This makes the learned graph is more helpful to indicate the true cluster structures. Experimental results on eight benchmark datasets also confirm the effectiveness and superiority of our proposed RBSMF.

In the futher work, we will investigate other graph learning methods. For our proposed RBSMF, an initial graph is required to get a new graph. So it is desirable to study other ways, which aim to learn directly a better adaptive graph from the original data. Moreover, lots of NMF-based methods are used to cluster the given data. So we want to integrate some priori information and labeled data into NMF, and design a novel semisupervised NMF.

REFERENCES

[1] F. Nie, X. Wang, and H. Huang, "Clustering and projected clustering with adaptive neighbors," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 977–986.

[2] F. Nie, X. Wang, M. I. Jordan, and H. Huang, "The constrained laplacian rank algorithm for graph-based clustering." in *AAAI*, 2016, pp. 1969–1976.

[3] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.

[4] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[5] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[6] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.

[7] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

[8] ——, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[9] C. Ding, X. He, and H. D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," in *Proceedings of the 2005 SIAM International Conference on Data Mining*. SIAM, 2005, pp. 606–610.

[10] X. Luo, M. Zhou, Y. Xia, and Q. Zhu, "An efficient non-negative matrix-factorization-based approach to collaborative-filtering for recommender systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273–1284, 2014.

[11] X. Luo, M. Zhou, S. Li, Z. You, Y. Xia, and Q. Zhu, "A non-negative latent factor model for large-scale sparse matrices in recommender systems via alternating direction method," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 524–537, 2016.

[12] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari, and A. Alabdulwahab, "Generating highly accurate predictions for missing qos-data via aggregating non-negative latent factor models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 579–592, 2016.

[13] S. Z. Li, X. W. Hou, H. J. Zhang, and Q. S. Cheng, "Learning spatially localized, parts-based representation," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[14] Z. Xiong, Y. Zang, X. Jiang, and X. Hu, "Document clustering with an augmented nonnegative matrix factorization model," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2014, pp. 348–359.

[15] D. Tao, D. Tao, X. Li, and X. Gao, "Large sparse cone non-negative matrix factorization for image annotation," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 8, no. 3, p. 37, 2017.

[16] M. Chen, Q. Wang, and X. Li, "Anchor-based group detection in crowd scenes," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1378–1382.

[17] F. Shang, L. Jiao, and F. Wang, "Graph dual regularization non-negative matrix factorization for co-clustering," *Pattern Recognition*, vol. 45, no. 6, pp. 2237–2250, 2012.

[18] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l21-norm," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 673–682.

[19] D. Cai, X. He, J. Han, and T. S. Huang, "Graph regularized nonnegative matrix factorization for data representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1548–1560, 2011.

[20] J. Huang, F. Nie, H. Huang, and C. Ding, "Robust manifold nonnegative matrix factorization," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 8, no. 3, p. 11, 2014.

[21] X. Li, G. Cui, and Y. Dong, "Graph regularized non-negative low-rank matrix factorization for image clustering," *IEEE Trans. Cybern*, vol. 47, no. 99, pp. 1–14, 2016.

[22] L. Zong, X. Zhang, L. Zhao, H. Yu, and Q. Zhao, "Multi-view clustering via multi-manifold regularized non-negative matrix factorization," *Neural Networks*, vol. 88, pp. 74–89, 2017.

[23] Z. Zhang and K. Zhao, "Low-rank matrix approximation with manifold regularization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 7, pp. 1717–1729, 2013.

[24] M. Belkin and P. Niyogi, "Laplacian eigenmaps and spectral techniques for embedding and clustering," in *Advances in neural information processing systems*, 2002, pp. 585–591.

[25] X. He and P. Niyogi, "Locality preserving projections," in *Advances in neural information processing systems*, 2004, pp. 153–160.

[26] A. B. Hamza and D. J. Brady, "Reconstruction of reflectance spectra using robust nonnegative matrix factorization," *IEEE Transactions on Signal Processing*, vol. 54, no. 9, pp. 3637–3642, 2006.

[27] L. Xiong, X. Chen, and J. Schneider, "Direct robust matrix factorizatoin for anomaly detection," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*. IEEE, 2011, pp. 844–853.

[28] L. Zhang, Z. Chen, M. Zheng, and X. He, "Robust non-negative matrix factorization," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, no. 2, pp. 192–200, 2011.

[29] B. Shen, B.-D. Liu, Q. Wang, and R. Ji, "Robust nonnegative matrix factorization via l 1 norm regularization by multiplicative updating rules," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 5282–5286.

[30] S. Peng, W. Ser, Z. Lin, and B. Chen, "Robust sparse nonnegative matrix factorization based on maximum correntropy criterion," in *Circuits and Systems (ISCAS), 2018 IEEE International Symposium on*. IEEE, 2018, pp. 1–5.

[31] X. Li, G. Cui, and Y. Dong, "Graph regularized non-negative low-rank matrix factorization for image clustering," *IEEE transactions on cybernetics*, vol. 47, no. 11, pp. 3840–3853, 2017.

[32] S. Huang, Z. Xu, and F. Wang, "Nonnegative matrix factorization with adaptive neighbors," in *Neural Networks (IJCNN), 2017 International Joint Conference on*. IEEE, 2017, pp. 486–493.

[33] Z. Zhang, "Parameter estimation techniques: A tutorial with application to conic fitting," *Image and vision Computing*, vol. 15, no. 1, pp. 59–76, 1997.

[34] X. Wang, F. Nie, and H. Huang, "Structured doubly stochastic matrix for graph based clustering: Structured doubly stochastic matrix," in *Proceedings of the 22nd ACM SIGKDD International conference on Knowledge discovery and data mining*. ACM, 2016, pp. 1245–1254.

[35] F. Wang, P. Li, A. C. König, and M. Wan, "Improving clustering by learning a bi-stochastic data similarity matrix," *Knowledge and information systems*, vol. 32, no. 2, pp. 351–382, 2012.

[36] R. Zass and A. Shashua, "Doubly stochastic normalization for spectral clustering," in *Advances in Neural Information Processing Systems*, 2007, pp. 1569–1576.

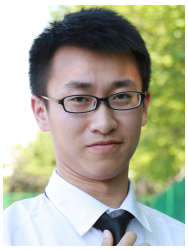
[37] Z. Lin, M. Chen, and Y. Ma, "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices," *arXiv preprint arXiv:1009.5055*, 2010.

[38] J. V. NEUMANN, *Functional Operators (AM-22), Volume 2: The Geometry of Orthogonal Spaces. (AM-22)*. Princeton University Press, 1950. [Online]. Available: http://www.jstor.org/stable/j.ctt1bc543b

[39] C. Peng, Z. Kang, Y. Hu, J. Cheng, and Q. Cheng, "Robust graph regularized nonnegative matrix factorization for clustering," *ACM*

Transactions on Knowledge Discovery from Data (TKDD), vol. 11, no. 3, p. 33, 2017.
[40] M. R. Hestenes, "Multiplier and gradient methods," *Journal of Optimization Theory and Applications*, vol. 4, no. 5, pp. 303–320, 1969.
[41] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Computer Science and Applied Mathematics, Boston: Academic Press, 1982.
[42] D. Dheeru and E. Karra Taniskidou, "UCI machine learning repository," 2017. [Online]. Available: http://archive.ics.uci.edu/ml
[43] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 3, pp. 328–340, 2005.
[44] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 7, pp. 881–892, 2002.
[45] P. K. Chan, M. D. Schlag, and J. Y. Zien, "Spectral k-way ratio-cut partitioning and clustering," *IEEE Transactions on computer-aided design of integrated circuits and systems*, vol. 13, no. 9, pp. 1088–1096, 1994.
[46] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

Qi Wang (M'15-SM'15) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Computer Science, and with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision and pattern recognition.



Xiang He received the B.E. degree in automation from Northwestern Polytechnical University, Xi'an, China, in 2017. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include machine learning and computer vision.

Xu Jiang received the B.E. degree in electronics and information engineering from Northwestern Polytechnical University, Xi'an, China, in 2019. He is currently working toward the M.S. degree in computer science in the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an, China. His research interests include machine learning and data mining.

Xuelong Li (M'02-SM'07-F'12) is a full professor with the School of Computer Science and the Center for OPTical IMagery Analysis and Learning (OPTIMAL), Northwestern Polytechnical University, Xi'an 710072, Shaanxi, China.