Robust Semi-supervised Nonnegative Matrix Factorization

Jing Wang, Feng Tian
Faculty of Science and Technology
Bournemouth University, UK
Email: {jwang, ftian}@bournemouth.ac.uk

Chang Hong Liu
Department of Psychology
Bournemouth University, UK
Email: liuc@bournemouth.ac.uk

Xiao Wang
School of Computer Science and Technology
Tianjin University, China
Email: wangxiao_cv@tju.edu.cn

***Abstract*—Nonnegative matrix factorization (NMF), which aims at finding parts-based representations of nonnegative data, has been widely applied to a wide range of applications such as data clustering, pattern recognition and computer vision. Real-world data are often sparse and noisy which may reduce the accuracy of representations. And a small part of data may have prior label information, which, if utilized, may improve the discriminability of representations. In this paper, we propose a robust semi-supervised nonnegative matrix factorization (RSSNMF) approach which takes all factors into consideration. RSSNMF incorporates the label information as an additional constraint to guarantee that the data with the same label have the same representation. It addresses the sparsity of data and accommodates noises and outliers consistently via $L_{2,1}$-norm. An iterative updating optimization scheme is derived to solve RSSNMF's objective function. We have proven the convergence of this optimization scheme by utilizing auxiliary function method and the correctness based on the Karush-Kohn-Tucker condition of optimization theory. Experiments carried on well-known data sets demonstrate the effectiveness of RSSNMF in comparison to other existing state-of-the-art approaches in terms of accuracy and normalized mutual information.**

I. Introduction

Finding an optimal data representation is a fundamental problem in many data analysis tasks. A good data representation [1], [2], [3], [4], [5] can typically reveal the latent structure of data and facilitate further data processing. Especially, matrix factorization techniques [6], [7], have been demonstrated to produce superior data representation. Central to the matrix factorization is to find two or more matrix factors whose product is a good approximation to the original matrix. Among these methods, nonnegative matrix factorization (NMF) [8], with the nonnegative constraint, is widely investigated and applied to analyze real-world data, such as images and texts, because it possesses parts-of-whole interpretations and better practical performance.

However, data often contain noises and outliers. The standard NMF uses the least square error function which is unstable with respect to noises and outliers [9], because a few noisy features or outliers with large errors will dominate objective function. Thus, a more robust NMF is needed to tackle the issue of noises or outliers [10]. Also, the sparsity is one of important characters of data, which can be considered as lack of useful labels or sufficient high quality data in the data set. That is, not all the features or data show the positive effect on the final results and only a few provide meaningful and useful information. Usually, adding sparsity regularization to select the most useful features can improve the generalization of a method, and thus avoiding the over-fitting problem [11]. Meanwhile, sparsity regularization can discover the most relevant features [12]. Therefore, it is also important to consider the sparsity of data. Moreover, some literatures have shown that utilizing a small amount of labeled data can produce considerable improvements in learning accuracy [13], [14]. The cost associated with the labeling process may render a fully labeled training set infeasible, whereas acquisition of a small set of labeled data is relatively inexpensive. In such situations, semi-supervised NMFs [15], [16] are proposed with great practical value and great benefits compared with unsupervised NMF.

To our best knowledge, there is no such a NMF which takes all the factors mentioned above into consideration. In this paper, we propose a robust semi-supervised nonnegative matrix factorization (RSSNMF), which can not only employ the label information with a constraint matrix, but also address the noisy and sparse data simultaneously. Specifically, we first utilize a constraint matrix, which guarantees that data with the same label have the same representation, as a hard constraint to integrate the label information. Thus, the learned new representation has a better discriminative power. Then, we adopt the $L_{2,1}$-norm as our loss function, so that RSSNMF can accommodate the outliers and noises in a better way than the standard NMF which uses F-norm as loss function. Besides, a sparse regularization term is added to RSSNMF to avoid the over-fitting problem and select the most relevant features. Furthermore, we derive an efficient and elegant iterative updating rule with convergence and correctness analysis. Our experimental results demonstrate the superiority of our proposed RSSNMF.

The rest of this paper is organized as follows. The Section 2 discusses the previous related work and put forward motivation of our approach. In Section 3, we present our RSSNMF framework and the corresponding solutions. The experimental results on image data sets are discussed in Section 4. Finally, we draw a conclusion and discuss future work.

II. Related Work

Nonnegative Matrix Factorization (NMF) has gained great success in many applications such as image processing, face recognition [17], [18], document clustering [19], [20]. However, as an unsupervised learning method, NMF does not use any prior knowledge of data to guide the learning process. Nevertheless, there is certain amount of prior knowledge in the real world applications, and it is natural to use accessorial

information such as class labels to improve the performance. Therefore, it would be beneficial to extend the usage of NMF to a semi-supervised manner. Cai *et al*. [21] proposed a graph regularized nonnegative matrix factorization (GNMF) model to preserve geometrical information by constructing the nearest neighbor graph. When label information is available, it can be naturally incorporated into the graph structure. Liu *et al*. [15] proposed constrained nonnegative matrix factorization (CNMF), which uses label information as additional hard constraints. CNMF forces samples with identical class label to have consistent coordinates in the reduced dimensional space, and thus the samples show more discriminative. Semi-NMF [22] as another variation of CNMF, not only utilizes the local structure of the data characterized by the graph Laplacian, but also incorporates the label information as the fitting constraints to learn. These semi-supervised NMF models incorporate prior information only, but do not take sparsity into consideration and are not robust to noises.

To alleviate the issue caused by noises and outliers, some robust methods are proposed. For example, Kong *et al*. [10] proposed a robust formulation of NMF (RNMF) using $L_{2,1}$-norm loss function, with which the error for each data point is not squared, and thus the large errors due to outliers do not dominate the objective function. Zhang *et al*. [23] proposed a robust non-negative matrix factorization algorithm (Robust NMF), which decomposed the data matrix as the summation of one sparse error matrix and the two nonnegative matrices, and the L_1-norm regularization term is added to the error matrix to get a sparse solution. These robust NMF models can tackle the issue of noises or outliers effectively but fail to make full use of available label information.

Recently, the sparse regularization technique is also investigated due to some important practical benefits. Usually, the sparse regularization term can avoid over-fitting problems, and discover the most relevant samples or features [11]. For example, [24] proposed a sparse NMF method with L_1-norm regularization. [25] proposed a strategy to compute an approximate NMF by using L_0-norm as sparse constraints. However, they are both unsupervised methods and cannot integrate the label information, and also they do not take the accommodation of outliers and noises into consideration.

In this paper, we take all above factors into consideration. Our proposed RSSNMF model not only integrates the data label information as an additional constraint to improve learning accuracy, but also can address the sparse and noisy data simultaneously.

III. Robust Semi-supervised Nonnegative Matrix Factorization (RSSNMF)

A. RSSNMF model

Suppose we have N data points $\{v_i\}_{i=1}^N$, each data point $v_i \in \mathbb{R}^M$ is M-dimensional and is represented by a vector. The vectors are placed in the columns and the whole data set is represented by a matrix $\mathbf{V} = [v_1, v_2, ..., v_N] \in \mathbb{R}^{M\times N}$. NMF aims to find two nonnegative matrix factors $\mathbf{W} \in \mathbb{R}^{M\times K}$ and $\mathbf{H} \in \mathbb{R}^{K\times N}$ where the product of the two factors can well approximate the original matrix, represented as $\mathbf{V} \approx \mathbf{WH}$. In particular, the $\mathbf{H}$ can be considered as the new representations of data in terms of the basis $\mathbf{W}$, where the ith column, h_i, is the new representation of the ith data. The approximation is quantified by a cost function which can be constructed by distance measures [15]. The standard NMF measures dissimilarity between $\mathbf{V}$ and $\mathbf{WH}$ by using the least squares error divergence. The error function of the standard NMF is

$$\|\mathbf{V} - \mathbf{WH}\|_F^2 = \sum_{i=1}^{N} \|v_i - \mathbf{W}h_i\|^2, \tag{1}$$

so the objective function is defined as

$$\min_{\mathbf{W}\geq 0, \mathbf{H}\geq 0} \|\mathbf{V} - \mathbf{WH}\|_F^2. \tag{2}$$

In many real world applications, a small amount of labeled data could be used to aid and bias the learning of unlabeled data. Motivated by CNMF [15], we suppose the first l data points are labeled with c classes. We first build an indicator matrix $\mathbf{C}$, where $c_{i,j} = 1$ if v_i is labeled with jth class; $c_{i,j} = 0$ otherwise. Then, with the indicator matrix C, we build the label constraint matrix $\mathbf{A}$ as follows,

$$\mathbf{A} = \begin{pmatrix} \mathbf{C}_{l\times c} & 0 \\ 0 & \mathbf{I}_{N-l} \end{pmatrix}, \tag{3}$$

where $\mathbf{I}_{N-l}$ is a $(N-l)\times(N-l)$ identity matrix. Recall that NMF maps each data point v_i to h_i from M-dimensional space to K-dimensional space. To incorporate label information, we introduce an auxiliary matrix $\mathbf{Z}$, and we have $\mathbf{H} = \mathbf{ZA}^T$. Thus, with the constraint matrix $\mathbf{A}$, we can see that if v_i and v_j have the same label, then the ith row and jth row of $\mathbf{A}$ must be the same, and so h_i=h_j, which guarantees that data sharing the same label have the same new representation. Therefore, (2) could be rewritten as follows,

$$\min_{\mathbf{W}\geq 0, \mathbf{Z}\geq 0} \|\mathbf{V} - \mathbf{WZA}^T\|_F^2. \tag{4}$$

However, using F-norm is unstable and sensitive to outliers, because errors are squared so can easily dominate the objective function. To overcome this limitation, we employ $L_{2,1}$-norm loss function to weaken the impact of noises and outliers effectively. $L_{2,1}$-norm was first proposed in [26], and defined as

$$\|\mathbf{G}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} \mathbf{G}_{ji}^2} = \sum_{i=1}^{N} \|g_i\|, \tag{5}$$

where g_i is the ith column of $\mathbf{G}$. Thus, the robust formulation of the error function can be written as

$$\|\mathbf{V} - \mathbf{WH}\|_{2,1} = \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{M} (\mathbf{V} - \mathbf{WH})_{ji}^2} = \sum_{i=1}^{N} \|v_i - \mathbf{W}h_i\|. \tag{6}$$

In this robust formulation that compared with (1), we can see that the error for each data point is $\|v_i - \mathbf{W}h_i\|$, which is not squared, and thus the large errors due to outliers do

not dominate the objective function. Therefore, our objective function can be reformulated as

$$\min_{\mathbf{W}\geq 0,\mathbf{Z}\geq 0} \|\mathbf{V}-\mathbf{W}\mathbf{Z}\mathbf{A}^T\|_{2,1}. \tag{7}$$

Besides, the real data may be sparse, i.e., not all the features are important to learning procedure. With regard to this, the $L_{2,1}$-norm regularization term is designed to generate column sparsity of representation of data to select correlated samples or features. Usually, we can get the group sparsity of the representation matrix $\mathbf{H}^T$ as follows,

$$\min_{\mathbf{H}\geq 0} \|\mathbf{H}^T\|_{2,1}. \tag{8}$$

As we let $\mathbf{H}=\mathbf{Z}\mathbf{A}^T$, here it is equal to minimize the matrix $\mathbf{A}\mathbf{Z}^T$. Because the matrix $\mathbf{A}$ is known, we just need to facilitate the unknown matrix $\mathbf{Z}^T$.

To incorporate the label information as a hard constraint, and deal with noisy and sparse data effectively, we propose the final formulation as follows,

$$\min_{\mathbf{W}\geq 0,\mathbf{Z}\geq 0} \|\mathbf{V}-\mathbf{W}\mathbf{Z}\mathbf{A}^T\|_{2,1}+\alpha\|\mathbf{Z}^T\|_{2,1}, \tag{9}$$

where $\mathbf{V}\in\mathbb{R}^{M\times N}$, $\mathbf{W}\in\mathbb{R}^{M\times K}$, $\mathbf{Z}\in\mathbb{R}^{K\times(N-l+c)}$ and $\mathbf{A}\in\mathbb{R}^{N\times(N-l+c)}$. Besides, the coefficient of sparse item α is the only parameter in this function, which is a nonnegative real value used to adjust the weight of sparse regularization.

B. Algorithm of RSSNMF model

The solution for RSSNMF model via iterative updating algorithm is presented as follows,

$$\mathbf{Z}_{ki}\leftarrow\mathbf{Z}_{ki}\frac{(\mathbf{W}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z})_{ki}}, \tag{10}$$

$$\mathbf{W}_{jk}\leftarrow\mathbf{W}_{jk}\frac{(\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T)_{jk}}{(\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A}\mathbf{Z}^T)_{jk}}, \tag{11}$$

where $\mathbf{D}_1$ and $\mathbf{D}_2$ are diagonal matrices with the diagonal elements given by

$$(\mathbf{D_1})_{ii}=\frac{1}{\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}\mathbf{A}^T)_i\|}, i=1,2...,N. \tag{12}$$

$$(\mathbf{D_2})_{ii}=\frac{1}{\|(\mathbf{Z}^T)_i\|}, i=1,2...,K. \tag{13}$$

C. Convergence of RSSNMF model

In this section, we prove the convergence of the algorithm described in the following Theorem 1 and Theorem 2.

Theorem 1. Updating $\mathbf{Z}$ using the rule of (10) while fixing $\mathbf{W}$, the objective function of (9) decreases monotonically, that is,

$$\begin{aligned}&\|\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T\|_{2,1}+\alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1}\\&-\|\mathbf{V}-\mathbf{W}\mathbf{Z}^t\mathbf{A}^T\|_{2,1}-\alpha\|(\mathbf{Z}^t)^T\|_{2,1}\leq 0,\end{aligned} \tag{14}$$

where t is the number of iteration times.

Theorem 2. Updating $\mathbf{W}$ using the rule of (11) while fixing $\mathbf{Z}$, the objective function of (9) decreases monotonically, that is,

$$\|\mathbf{V}-\mathbf{W}^{t+1}\mathbf{Z}\mathbf{A}^T\|_{2,1}-\|\mathbf{V}-\mathbf{W}^t\mathbf{Z}\mathbf{A}^T\|_{2,1}\leq 0, \tag{15}$$

where t is the number of iteration times.

To prove the Theorem 1, we first have the following Lemma 1.

Lemma 1. Under the updating rule of (10), the following inequation holds

$$\begin{aligned}&Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)^T)\\&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\&\leq Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^t\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^t\mathbf{A}^T)^T)\\&+\alpha Tr((\mathbf{Z}^t)^T\mathbf{D}_2\mathbf{Z}^t).\end{aligned} \tag{16}$$

Proof. We prove Lemma 1 by using the auxiliary function approach [27]. First of all, we define

$$\begin{aligned}J(\mathbf{Z})=&Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}\mathbf{A}^T)^T)\\&+\alpha Tr(\mathbf{Z}^T\mathbf{D}_2\mathbf{Z}).\end{aligned} \tag{17}$$

Then we can reformulate (16) as

$$J(\mathbf{Z}^{t+1})\leq J(\mathbf{Z}^t). \tag{18}$$

According to (17), we can get

$$\begin{aligned}J(\mathbf{Z})=&Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T)\\&+Tr(\mathbf{W}\mathbf{Z}\mathbf{A}^T\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T)+\alpha tr(\mathbf{Z}^T\mathbf{D}_2\mathbf{Z})\\&\leq Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T)\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{S}_1\mathbf{H}'\mathbf{B}_1)_{ki}(\mathbf{H}^2)_{ki}}{\mathbf{H}'_{ki}}\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{S}_2\mathbf{H}'\mathbf{B}_2)_{ki}(\mathbf{H}^2)_{ki}}{\mathbf{H}'_{ki}}\quad\text{(by Lemma 2)}\\&=Tr(\mathbf{V}\mathbf{D}_1\mathbf{V}^T-2\mathbf{V}\mathbf{D}_1\mathbf{A}\mathbf{Z}^T\mathbf{W}^T)\\&+\sum_{k=1}^{K}\sum_{i=1}^{(N-l+c)}\frac{(\mathbf{W}^T\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z}')_{ki}(\mathbf{Z^2})_{ki}}{\mathbf{Z}'_{ki}}\\&=F(\mathbf{Z},\mathbf{Z}'),\end{aligned} \tag{19}$$

where $\mathbf{S}_1=\mathbf{W^TW}$, $\mathbf{B}_1=\mathbf{A}^T\mathbf{D}_1\mathbf{A}$, $\mathbf{H}=\mathbf{Z},\mathbf{H}'=\mathbf{Z}'$, $\mathbf{S}_2=\alpha\mathbf{D}_2$, and $\mathbf{B}_2=\mathbf{I}$. The equality holds when $\mathbf{Z}=\mathbf{Z}'$. So $F(\mathbf{Z},\mathbf{Z}')$ is an auxiliary function of $J(\mathbf{Z})$.

Let

$$\mathbf{Z}^{t+1}=\arg\min_{\mathbf{Z}} F(\mathbf{Z},\mathbf{Z}^t), \tag{20}$$

and we can get

$$J(\mathbf{Z}^{t+1})=F(\mathbf{Z}^{t+1},\mathbf{Z}^{t+1})\leq F(\mathbf{Z}^{t+1},\mathbf{Z}^t)\leq J(\mathbf{Z}^t), \tag{21}$$

this proves that $J(\mathbf{Z}^t)$ decreases monotonically.

Then we let $f(\mathbf{Z})=F(\mathbf{Z},\mathbf{Z}')$, the gradient of $f(\mathbf{Z})$ is

$$\begin{aligned}\frac{\partial f(\mathbf{Z})}{\partial\mathbf{Z}_{ki}}=&-2(\mathbf{W}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}\\&+2\frac{(\mathbf{W}^T\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z}')_{ki}(\mathbf{Z})_{ki}}{\mathbf{Z}'_{ki}}.\end{aligned} \tag{22}$$

The second-order derivatives (Hessian matrix) is

$$\frac{\partial^2 f(\mathbf{Z})}{(\partial\mathbf{Z}_{ki})(\partial\mathbf{Z}_{lj})}=2\frac{(\mathbf{W}^T\mathbf{W}\mathbf{Z}'\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z}')_{ki}}{\mathbf{Z}'_{ki}}\delta_{ij}\delta_{kl}. \tag{23}$$

Because the Hessian matrix is semi-positive definite, so $f(\mathbf{Z})$ is a convex function and there is an unique global minima for $f(\mathbf{Z})$. By forcing (22) to zero, we can get the solution of $\mathbf{Z}$ as follows,

$$\mathbf{Z}_{ki} = \mathbf{Z'}_{ki}\frac{(\mathbf{W}^T\mathbf{V}\mathbf{D}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{W}\mathbf{Z'}\mathbf{A}^T\mathbf{D}_1\mathbf{A}+\alpha\mathbf{D}_2\mathbf{Z'})_{ki}}. \tag{24}$$

When we set $\mathbf{Z}^{t+1} \leftarrow \mathbf{Z}$, $\mathbf{Z}^t \leftarrow \mathbf{Z'}$, (24) can derive updating rule of (10). Under this updating rule, objective function of (17) decreases monotonically.

Until now, we have completed the proof of Lemma 1.

Lemma 2. To prove, we apply the matrix inequality [28]. If matrices $\mathbf{S}$, $\mathbf{B}$, $\mathbf{H}$ are nonnegative matrices with appropriate sizes and $\mathbf{S} = \mathbf{S}^T$, $\mathbf{B} = \mathbf{B}^T$, then we have the following matrix inequality

$$Tr(\mathbf{H}^T\mathbf{S}\mathbf{H}\mathbf{B}) \leq \sum_{ik}(\mathbf{S}\mathbf{H'}\mathbf{B})\frac{\mathbf{H}^2{}_{ik}}{\mathbf{H'}_{ik}}. \tag{25}$$

Lemma 3. Under the updating rule of (10), the following inequation holds

$$\begin{aligned}
&\|\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T\|_{2,1}+\alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1}\\
&-\|\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T\|_{2,1}-\alpha\|(\mathbf{Z}^{t})^T\|_{2,1}\\
&\leq\frac{1}{2}[Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\
&-Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T)^T)\\
&-\alpha Tr((\mathbf{Z}^{t})^T\mathbf{D}_2\mathbf{Z}^{t})].
\end{aligned} \tag{26}$$

Proof. We borrow the idea of [10] to prove Lemma 3. Then, (26) can be reformulated as follows.

According to the definition of $\mathbf{D}_1$ and $\mathbf{D}_2$, we can get

$$\begin{aligned}
&Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t+1})^T\mathbf{D}_2\mathbf{Z}^{t+1})\\
&=\sum_{i=1}^{N}\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&+\alpha\sum_{i=1}^{K}\|(\mathbf{Z}_i^{t+1})^T\|^2(\mathbf{D}_2)_{ii},
\end{aligned} \tag{27}$$

$$\begin{aligned}
&Tr((\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T)\mathbf{D}_1(\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T)^T)\\
&+\alpha Tr((\mathbf{Z}^{t})^T\mathbf{D}_2\mathbf{Z}^{t})\\
&=\sum_{i=1}^{N}\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&+\alpha\sum_{i=1}^{K}\|(\mathbf{Z}^{t})_i^T\|^2(\mathbf{D}_2)_{ii}.
\end{aligned} \tag{28}$$

Then the right-hand side (RHS) of (26) is

$$\begin{aligned}
RHS=&\frac{1}{2}\sum_{i=1}^{N}(\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}\\
&-\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii})\\
&+\frac{1}{2}\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\|^2(\mathbf{D}_2)_{ii}-\|(\mathbf{Z}^{t})_i^T\|^2(\mathbf{D}_2)_{ii}).
\end{aligned} \tag{29}$$

According to the (12) and (13), we can get

$$\begin{aligned}
RHS=&\frac{1}{2}\sum_{i=1}^{N}(\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|^2(\mathbf{D}_1)_{ii}-\frac{1}{(\mathbf{D}_1)_{ii}})\\
&+\frac{1}{2}\alpha\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\|^2(\mathbf{D}_2)_{ii}-\frac{1}{(\mathbf{D}_2)_{ii}}).
\end{aligned} \tag{30}$$

The left-hand side (LHS) of (26) is

$$\begin{aligned}
LHS=&\|\mathbf{V}-\mathbf{W}\mathbf{Z}^{t+1}\mathbf{A}^T\|_{2,1}+\alpha\|(\mathbf{Z}^{t+1})^T\|_{2,1}\\
&-\|\mathbf{V}-\mathbf{W}\mathbf{Z}^{t}\mathbf{A}^T\|_{2,1}-\alpha\|(\mathbf{Z}^{t})^T\|_{2,1}\\
=&\sum_{i=1}^{N}(\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|-\frac{1}{(\mathbf{D}_1)_{ii}})\\
&+\alpha\sum_{i=1}^{K}(\|(\mathbf{Z}^{t+1})_i^T\|-\frac{1}{(\mathbf{D}_2)_{ii}}).
\end{aligned} \tag{31}$$

Finally, we get

$$\begin{aligned}
&LHS-RHS\\
&=\sum_{i=1}^{N}\frac{-(\mathbf{D}_1)_{ii}}{2}(\|\mathbf{V}_i-\mathbf{W}(\mathbf{Z}^{t+1}\mathbf{A}^T)_i\|-\frac{1}{(\mathbf{D}_1)_{ii}})^2\\
&+\sum_{i=1}^{K}\frac{-(\mathbf{D}_2)_{ii}}{2}(\|(\mathbf{Z}^{t+1})_i^T\|-\frac{1}{(\mathbf{D}_2)_{ii}})^2\leq 0.
\end{aligned} \tag{32}$$

Until now, we have completed the proof of Lemma 3.

By using Lemma 1, lemma 2 and Lemma 3, we can easily prove Theorem 1. That is to say, the objective function of (9) decreases monotonically under the updating rules of (10).

The proof of Theorem 2 is similar to Theorem 1, so the details will not be mentioned here.

D. Correctness of RSSNMF analysis

Here we prove correctness of our updating rules that the converged solution is the correct optimal solution, i.e., the converged solution satisfies the Karush-Kohn-Tucker(KKT) condition of the constrained optimization theory.

Theorem 3. At convergence, the converged solution $\mathbf{Z}$ of the updating rule of (10) satisfies the KKT condition of the optimization theory.

Proof. The KKT condition for $\mathbf{Z}$ with the constrains $(\mathbf{Z})_{ki} \geq 0$, $k = 1, 2, ..., K$; $i = 1, 2, ..., (N - l + c)$, is

$$\frac{\partial J(\mathbf{Z})}{\partial(\mathbf{Z})_{ki}}(\mathbf{Z})_{ki}=0, \forall k, i. \tag{33}$$

TABLE I: Description of dataset

Datasets	Size	Dimention	Class
Yale	165	1024	15
ORL	400	1024	40
COIL20	1440	1024	20

The derivative is

$$\frac{\partial J(\mathbf{Z})}{\partial(\mathbf{Z})_{ki}} = (\mathbf{W}^T(\mathbf{V} - \mathbf{WZA}^T)\mathbf{D}_1\mathbf{A})_{ki} + \alpha(\mathbf{D}_2\mathbf{Z})_{ki}. \quad (34)$$

Then, the KKT condition for $\mathbf{Z}$ is

$$\begin{aligned}&[-(\mathbf{W}^T\mathbf{VD}_1\mathbf{A})_{ki} + (\mathbf{W}^T\mathbf{WZA}^T\mathbf{D}_1\mathbf{A})_{ki} \\ &+ \alpha(\mathbf{D}_2\mathbf{Z})_{ki}](\mathbf{Z})_{ki} \\ &= 0, \forall k, i.\end{aligned} \quad (35)$$

If the $\mathbf{Z}$ converges according to the updating rule of (10), the converged solution $\mathbf{Z}^*$ satisfied

$$\mathbf{Z}^*_{ki} \leftarrow \mathbf{Z}^*_{ki}\frac{(\mathbf{W}^T\mathbf{VD}_1\mathbf{A})_{ki}}{(\mathbf{W}^T\mathbf{WZ}^*\mathbf{A}^T\mathbf{D}_1\mathbf{A} + \alpha\mathbf{D}_2\mathbf{Z}^*)_{ki}}, \quad (36)$$

which can be reformulated as

$$\begin{aligned}&[-(\mathbf{W}^T\mathbf{VD}_1\mathbf{A})_{ki} + (\mathbf{W}^T\mathbf{WZ}^*\mathbf{A}^T\mathbf{D}_1\mathbf{A})_{ki} \\ &+ \alpha(\mathbf{D}_2\mathbf{Z}^*)_{ki}](\mathbf{Z}^*)_{ki} \\ &= 0, \forall k, i.\end{aligned} \quad (37)$$

We can see that (37) is identical to (35). Then the converged solution $\mathbf{Z}^*$ satisfies the KKT condition. Until now, we have completed the proof of Theorem 3.

Theorem 4. At convergence, the converged solution $\mathbf{W}$ of the updating rule of (11) satisfies the KKT condition of the optimization theory.

The proof of Theorem 4 is similar to Theorem 3, so we omit the details here.

IV. Experiment Results

A. Database description

In this paper, we select three data sets to evaluate the effectiveness of our proposed RSSNMF method for data clustering. The details of three data sets are summarized in the Table I.

Yale Database[1]. The Yale database contains 11 facial images for each of 15 distinct subjects, thus 165 images in total. For each subject, the images are in great varieties such as different facial expressions or configurations. In the preprocessing step, we normalize the original images (in scale and direction) to keep the two eyes are aligned at the same position. Then, the facial areas are cropped into the final images for clustering. Each image is resized into 32×32 pixels with 256 gray levels per pixel.

ORL Database[2]. The ORL database consists of 400 facial images belonging to 40 different subjects. For each subject, the images are in great varieties because of different taking time with changing lighting variance, facial details and facial expressions. All the pictures are taken against dark homogeneous background with the subjects in an upright, frontal position. We do the same preprocessing for this data set as for the Yale data set.

COIL20 Database[3]. The COIL20 image library consists of 20 objects with 1440 images as a whole. The objects are placed on a motorized turntable against a black background. The turntable is rotated through 360deg and a fixed camera took images at a pose intervals of 5deg for each object. Thus, each object has 72 images in total. The size of each image is the same as Yale and ORL image, which is also represented by a 1024-dimensional feature vector in image space.

B. Evaluation metrics

Two metrics, the accuracy (AC) and the normalized mutual information metric (NMI) are used to measure the clustering performance [29], [2]. These measurements are widely used by comparing the obtained label of each sample with that provided by the data set in different clustering approaches.

Clustering accuracy (AC) is used to measure the percentage of correct labels obtained. Given a data set containing n images, let l_i and r_i be the the obtained cluster label and label provided from each sample image, respectively. The AC is defined as follows,

$$AC = \frac{\sum_{i=1}^{n}\delta(r_i, map(l_i))}{n} \quad (38)$$

where $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and $map(l_i)$ is the permutation mapping function that maps each cluster label l_i to the equivalent label from the data set. The best mapping can be found by using the Kuhn-Munkres algorithm [30].

Normalized mutual information (NMI) is used to measure the similarity between the cluster assignments and the pre-existing input labeling of the classes. Let C and C' denote the set of clusters obtained from the ground truth and obtained from our algorithm, respectively, their mutual information metric $MI(C, C')$ is defined as follows,

$$MI(C, C') = \sum_{c_i\in C, c_j{}'\in C'} p(c_i, c_j{}') \cdot log\frac{p(c_i, c_j{}')}{p(c_i)\cdot p(c_j{}')}, \quad (39)$$

where $p(c_i)$, $p(c_j{}')$ are the probabilities that an image randomly selected from the data set belongs to the clusters c_i and c_j, respectively, and $p(c_i, c_j{}')$ denotes the joint probability that this randomly selected image belongs to the cluster c_i as well as c_j at the same time. In our experiment, we used the normalized metric $NMI(C, C')$ as follows,

$$NMI(C, C') = \frac{MI(C, C')}{max(H(C), H(C'))}, \quad (40)$$

where $H(C)$ and $H(C')$ are the entropies of C and C', respectively. It is easy to check that $NMI(C, C')$ ranges from

[1]http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[2]http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[3]http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php

TABLE II: Clustering Results Comparison on the Yale Database

K	AC (%)				NMI (%)			
	NMF	RNMF	CNMF	RSSNMF	NMF	RNMF	CNMF	RSSNMF
2	78.64	77.73	81.64	**87.27**	40.76	35.89	48.99	**62.23**
3	66.36	66.97	68.48	**73.03**	37.69	40.32	42.90	**45.83**
4	63.18	68.18	64.25	**71.14**	43.50	50.58	47.80	**52.91**
5	58.91	68.00	65.82	**73.09**	42.39	52.89	53.99	**61.53**
6	49.09	62.88	60.12	**63.64**	36.07	50.44	50.63	**51.53**
7	50.52	60.78	59.25	**62.34**	44.25	52.07	53.03	**53.40**
8	45.45	54.66	54.40	**57.39**	40.59	47.29	**50.83**	50.80
9	46.87	52.83	54.85	**56.16**	43.34	49.44	53.05	**53.79**
10	48.36	51.91	52.68	**53.73**	46.22	48.73	**52.84**	51.37
Avg.	56.38	62.66	62.39	**66.42**	41.64	47.52	50.45	**53.71**

TABLE III: Clustering Results Comparison on the ORL Database

K	AC (%)				NMI (%)			
	NMF	RNMF	CNMF	RSSNMF	NMF	RNMF	CNMF	RSSNMF
2	89.00	96.00	**93.90**	89.55	67.00	**88.25**	78.24	62.45
3	78.67	84.33	84.33	**89.33**	67.54	73.05	76.61	**82.01**
4	81.75	83.57	84.05	**89.00**	75.13	73.81	78.69	**84.92**
5	78.40	87.25	78.38	**89.60**	74.73	79.34	76.69	**85.95**
6	71.17	86.60	76.73	**91.00**	68.75	82.98	76.52	**85.11**
7	78.29	**88.67**	79.04	86.29	79.82	**86.30**	82.45	85.65
8	75.25	83.87	77.03	**86.50**	77.02	84.54	80.57	**87.34**
9	80.11	82.44	82.23	**86.89**	83.79	84.66	86.03	**87.81**
10	74.00	81.40	76.88	**86.20**	78.96	83.36	81.77	**87.36**
Avg.	78.51	86.06	81.40	**88.26**	74.75	81.81	79.73	**83.18**

0 to 1. $NMI = 1$ when the two sets of image clusters are identical, and it becomes zero when the two sets are completely independent.

C. Clustering results

We present our clustering performance by making comparisons with other related NMF methods on three data sets. The algorithms that we choose to compare are listed below,

1. Standard Nonnegative Matrix Factorization algorithm (NMF) minimizing F-norm cost as suggested in [17].

2. Robust Nonnegative Matrix Factorization algorithm by using $L_{2,1}$-norm (RNMF) [10] is implemented in our experiment to compare the results.

3. Constrained Nonnegative Matrix Factorization algorithm (CNMF) minimizing the F-norm cost that was introduced in [15].

4. Our proposed robust semi-supervised NMF (RSSNMF) model.

For each data set, the evaluations are conducted with different numbers of clusters K varying from 2 to 10 categories. We randomly choose K categories from the data set, and mix the images of these K categories as the collection $\mathbf{V}$ for clustering. For the semi-supervised algorithms (RSSNMF, CNMF), we randomly pick up 10 percent of images from each category in $\mathbf{V}$ and use their category numbers as the available label information. However, there are only 10 images for each category in Yale and ORL, so 10 percent gives one image only. One label is meaningless for RSSNMF since this algorithm maps the images with the same label onto the same point. Thus, for Yale and ORL, we randomly choose two images from each category to provide the label information.

Then, we apply different matrix factorization algorithms as listed above to obtain new data representations. Because our method converges to a local optimum, we initialize $\mathbf{Z}$ and $\mathbf{W}$ randomly 10 times between 0 and 1. When the algorithm converges, there are 10 values of the loss function, respectively. We can then find the minimum value from these values, and consider the corresponding matrix $\mathbf{Z}$ and $\mathbf{W}$ as the final solution that will be not too far from the global optimum. Thereafter, K-means is applied to the new data representation for images clustering, which is repeated 10 times with different initial points and the average result in terms of the cost function of K-means is recorded.

Finally, we compare the obtained clusters with the original image category to compute the AC and NMI.

We run this experiment for t repetition until it converges. The convergence criterion we used is

$$|\frac{J_{t+1} - J_t}{J_t}| < 10^{-6} \tag{41}$$

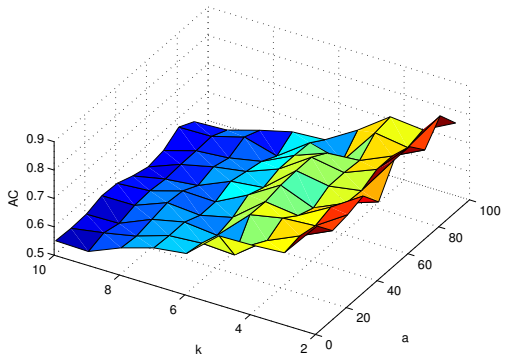
where J_t is the objective function value in the t-th iteration of each algorithm.

The detailed clustering results of Yale are shown in Table II. The last row shows the average AC and NMI over K. As we can see, RSSNMF outperforms others in all cases in terms of AC and achieves 7 best results in terms of NMI. And comparing to second best results, i.e., average results of RNMF in AC and average results of CNMF in NMI, our algorithm RSSNMF achieves 6 percent improvement and 6.46 percent improvement respectively.

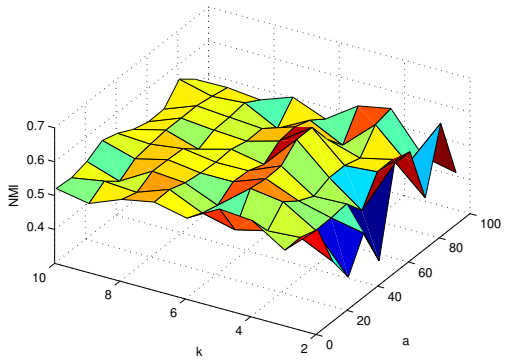
The clustering results of ORL are summarized in Table III. RSSNMF outperforms all other algorithms with 8 best results in AC and 7 best results in NMI. Moreover, for the average

TABLE IV: Clustering Results Comparison on the COIL20 Database

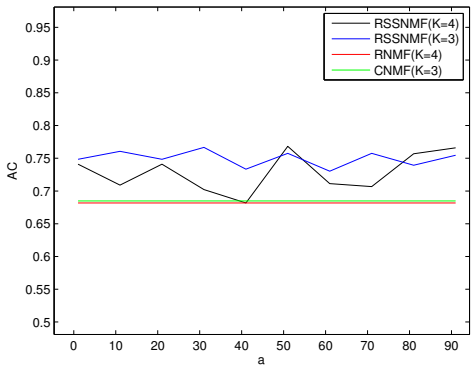
K	AC (%)				NMI (%)			
	NMF	RNMF	CNMF	RSSNMF	NMF	RNMF	CNMF	RSSNMF
2	89.86	89.58	94.72	**95.42**	76.89	75.79	78.55	**83.06**
3	87.22	90.83	91.48	**94.63**	75.53	80.02	84.07	**86.91**
4	82.64	83.82	83.96	**84.44**	74.32	75.39	71.40	**79.31**
5	**94.94**	93.22	74.83	87.72	**91.02**	89.68	66.74	80.40
6	80.69	79.07	77.50	**81.30**	78.31	**78.83**	76.37	77.49
7	81.83	80.79	78.53	**82.18**	81.45	80.86	77.82	**82.43**
8	76.77	74.58	72.36	**80.49**	77.18	75.86	72.63	**79.92**
9	73.95	72.01	**82.72**	81.08	74.46	72.93	**83.62**	80.36
10	76.14	75.17	73.44	**79.78**	79.36	77.40	76.02	**81.61**
Avg.	82.67	82.19	81.06	**85.23**	78.72	78.53	76.36	**81.28**



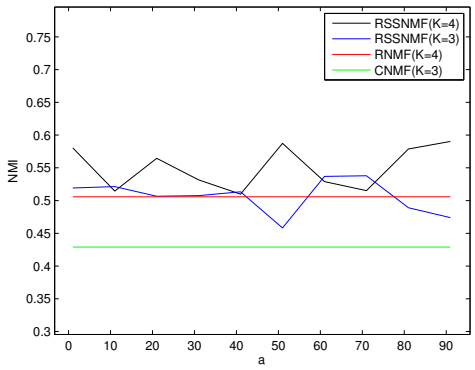
(a) AC of RSSNMF on Yale data set



(b) NMI of RSSNMF on Yale data set



(c) Comparison of AC on Yale data set



(d) Comparison of NMI on Yale data set

Fig. 1: The parameter sensitivity of RSSNMF

performances, RSSNMF achieves 2.56 percent improvement in AC and 1.67 percent improvement in NMI, compared to the second best algorithm, i.e., RNMF.

Table IV shows the detailed results of COIL20. We can see that NMF and RNMF have similar results, while RSSNMF still shows the best performance. RSSNMF gains 7 and 6 highest values in AC and NMI respectively. Besides, comparing to the best algorithm other than our proposed RSSNMF algorithms, i.e., NMF, RSSNMF achieves 3.10 percent and 3.52 percent improvement in average results with respective to AC and NMI.

D. Parameter analysis

In this section, we analyze the sensitivity of the parameter α of our proposed approach. In RSSNMF model, the second term is sparse regularization via $L_{2,1}$-norm minimization. The value of α affects the effectiveness of clustering. Taken the Yale data set as an example, Fig.1(a) and Fig.1(b) show that AC and NMI vary slightly with α. We can find that nearly for all values of α (from 1 to 90) with K varying from 2 to 10, the performance of RSSNMF is superior to the performances of other approaches shown in Table II. To clearly display this, in particular, we plot the performances of RSSNMF and second best methods when K=3 and K=4, respectively, shown in

Fig.1(c) and Fig.1(d). We can see that, when K=3 , the blue line that refers to the results of RSSNMF is above the green line that represents the results of CNMF. Similarly, the black line is above the red line when K=4. In our experiments, for the results given in Table II, Table III and Table IV, the values of α are 60, 65 and 50 respectively.

V. Conclusion

In this paper, we have presented a novel nonnegative matrix factorization method, called robust semi-supervised nonnegative matrix factorization (RSSNMF). Firstly, our proposed RSSNMF model imposes label information with a constraint matrix, so that RSSNMF guarantees that data with the same label have the same new representation. Thus, the new representations learned by RSSNMF have more discriminative power. Secondly, by utilizing $L_{2,1}$-norm, RSSNMF is more robust to the noises and outliers. Thirdly, incoporating the sparse regularization term, RSSNMF can address the sparsity of data more effectively compared to existing approaches. Efficient iterative update algorithms with rigorous convergence and correctness analysis are also given. Experiments on three well known data sets have demonstrated that our approach is superior to other algorithms nearly in all cases.

There are some potential improvements to our approach. Theoretical or mathematical analysis may be given on the influence of the parameter α to the RSSNMF's performance. Another route to further improve and extend RSSNMF is multi-view NMF. Often, data can be represented by different types of features, and different types of features may be complementary to each other. Instead of using only one type of feature, the new representation of data could be learnt by combining different types of features, so better clustering performance can be expected.

References

[1] Y.-H. Xiao, Z.-F. Zhu, Y. Zhao, and Y.-C. Wei, "Class-driven nonnegative matrix factorization for image representation," *Journal of Computer Science and Technology*, vol. 28, no. 5, pp. 751–761, 2013.

[2] D. Cai, X. He, X. Wu, and J. Han, "Non-negative matrix factorization on manifold," in *Data Mining, 2008. ICDM'08. Eighth IEEE International Conference on*. IEEE, 2008, pp. 63–72.

[3] Q.-L. Lin, B. Sheng, Y. Shen, Z.-F. Xie, Z.-H. Chen, and L.-Z. Ma, "Fast image correspondence with global structure projection," *Journal of Computer Science and Technology*, vol. 27, no. 6, pp. 1281–1288, 2012.

[4] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang, "Face recognition using laplacianfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 27, no. 3, pp. 328–340, 2005.

[5] X. Niyogi, "Locality preserving projections," in *Neural information processing systems*, vol. 16, 2004, p. 153.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*. John Wiley & Sons, 2012.

[7] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Springer, 1992.

[8] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in neural information processing systems*, 2001, pp. 556–562.

[9] W. Liu, N. Zheng, and Q. You, "Nonnegative matrix factorization and its applications in pattern recognition," *Chinese Science Bulletin*, vol. 51, no. 1, pp. 7–18, 2006.

[10] 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.

[11] S. Yang, C. Hou, C. Zhang, and Y. Wu, "Robust non-negative matrix factorization via joint sparse and graph regularization for transfer learning," *Neural Computing and Applications*, vol. 23, no. 2, pp. 541–559, 2013.

[12] P. O. Hoyer, "Non-negative matrix factorization with sparseness constraints," *The Journal of Machine Learning Research*, vol. 5, pp. 1457–1469, 2004.

[13] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *The Journal of Machine Learning Research*, vol. 7, pp. 2399–2434, 2006.

[14] X. Zhu, Z. Ghahramani, J. Lafferty *et al.*, "Semi-supervised learning using gaussian fields and harmonic functions," in *ICML*, vol. 3, 2003, pp. 912–919.

[15] H. Liu, Z. Wu, X. Li, D. Cai, and T. S. Huang, "Constrained nonnegative matrix factorization for image representation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1299–1311, 2012.

[16] L. Lan, N. Guan, X. Zhang, D. Tao, and Z. Luo, "Soft-constrained nonnegative matrix factorization via normalization," in *Neural Networks (IJCNN), 2014 International Joint Conference on*. IEEE, 2014, pp. 3025–3030.

[17] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[18] S. Z. Li, X. Hou, H. Zhang, and Q. 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–207.

[19] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, "Document clustering using nonnegative matrix factorization," *Information Processing & Management*, vol. 42, no. 2, pp. 373–386, 2006.

[20] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. ACM, 2003, pp. 267–273.

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

[22] Y. He, H. Lu, and S. Xie, "Semi-supervised non-negative matrix factorization for image clustering with graph laplacian," *Multimedia Tools and Applications*, pp. 1–23, 2013.

[23] 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.

[24] J. Kim and H. Park, "Sparse nonnegative matrix factorization for clustering," 2008.

[25] R. Peharz and F. Pernkopf, "Sparse nonnegative matrix factorization with l0-constraints," *Neurocomputing*, vol. 80, pp. 38–46, 2012.

[26] C. Ding, D. Zhou, X. He, and H. Zha, "R 1-pca: rotational invariant l 1-norm principal component analysis for robust subspace factorization," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 281–288.

[27] D. Seung and L. Lee, "Algorithms for non-negative matrix factorization," *Advances in neural information processing systems*, vol. 13, pp. 556–562, 2001.

[28] C. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, no. 1, pp. 45–55, 2010.

[29] D. Cai, X. He, and J. Han, "Document clustering using locality preserving indexing," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 12, pp. 1624–1637, 2005.

[30] M. D. Plummer and L. Lovász, *Matching theory*. Elsevier, 1986.