

Subspace clustering guided convex nonnegative matrix factorization

Guosheng Cui^{a,b}, Xuelong Li^a, Yongsheng Dong^{a,*}

^a The Center for OPTICAL IMagery Analysis and Learning (OPTIMAL), Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, Shaanxi 710119, PRChina

^b University of Chinese Academy of Sciences, 19A Yuquanlu, Beijing, 100049 P R China



ARTICLE INFO

Article history:

Received 13 December 2017

Revised 5 February 2018

Accepted 21 February 2018

Available online 28 February 2018

Communicated by Dr. Nianyin Zeng

Keywords:

Convex nonnegative matrix factorization

Subspace clustering

Multiple centroids

Geometry structure

Image clustering

ABSTRACT

As one of the most important information of the data, the geometry structure information is usually modeled by a similarity graph to enforce the effectiveness of nonnegative matrix factorization (NMF). However, pairwise distance based graph is sensitive to noise and can not capture the subspace structure of the data. Reconstruction coefficients based graph can capture the subspace structure of the data, but the procedure of building the representation based graph is usually independent to the framework of NMF. To address this issue, a novel subspace clustering guided convex nonnegative matrix factorization (SC-CNMF) is proposed. In this NMF framework, the nonnegative subspace clustering is incorporated to learning the representation based graph, and meanwhile, a convex nonnegative matrix factorization is also updated simultaneously. To tackle the noise influence of the dataset, only k largest entries of each representation are kept in the subspace clustering. To capture the complicated geometry structure of the data, multiple centroids are also introduced to describe each cluster. Additionally, a row constraint is used to remove the relevance among the rows of the encoding matrix, which can help to improve the clustering performance of the proposed model. For the proposed NMF framework, two different objective functions with different optimizing schemes are designed. Image clustering experiments are conducted to demonstrate the effectiveness of the proposed methods on several datasets and compared with some related works based on NMF together with k -means clustering method and PCA as baseline.

© 2018 Elsevier B.V. All rights reserved.

1. Introduction

As one feature extraction technique [1,2], nonnegative matrix factorization (NMF) has become more and more popular in the fields of the computer vision and pattern recognition thanks to the pioneering work of Lee and Seung [3]. In their works, they point out that the nonnegative constraints on the component matrices can automatically lead to the parts-based representation of the data which is closely related to the perception mechanism. Besides this finding, a simple yet effective algorithmic procedure is another contribution of their work. Due to these advantages of NMF, the research around the original NMF and its variants is becoming increasingly flourishing [4–9].

The original NMF method tries to factorize the original nonnegative data matrix into two nonnegative factorial matrices whose product can approximate the original data matrix. To improve the sparseness of NMF, Hoyer [10] has proposed nonnegative sparse coding (NSC) in which a ℓ_1 -norm sparse constraint is imposed on

the encoding matrix. Localized NMF (LNMF) [11] imposes some local constraints on the factor matrices to help learning a more localized features of the data. Both these two methods can ensure a parts-based representation. For unsupervised variants of NMF, geometry information is another important information that is frequently used to reinforce the effectiveness of NMF methods. Graph regularized NMF (GNMF) [12] first considers to improve the performance of NMF from the geometric perspective. An affinity graph is constructed to encode the local data distributing structure. With this information, the data local geometry structure in the original space can be retained in the learned low dimension space. Neighborhood preserving NMF (NPNMF) [13] tries to encode the local geometry structure with the coefficients of the k nearest neighbors. Although, NPNMF does not use the Heat Kernel to measure the similarity of the nearest neighbors. But the selection of the k nearest neighbors is still based on the Euclidean distance. In order to learn a discriminative and sparse representation, Ren et al. [14] add a rank constraint into the framework of NMF and proposed NMF with regularizations (RNMF). Nie et al. propose a semi-NMF named robust manifold NMF (RMNMF) [15] to improve the robustness of NMF. RMNMF uses $\ell_{2,1}$ -norm to measure the residual of the approximation instead F -norm which is sensitive to

* Corresponding author.

E-mail addresses: guosheng.cui.opt@gmail.com (G. Cui), xuelong_li@opt.ac.cn (X. Li), dongyongsheng98@163.com (Y. Dong).

outliers. Both RNMF and RMNMF have promoted the performance of NMF from different views, but they have one common regularizer in their object function, that is graph Laplacian regularizer. Actually, graph Laplacian regularizer is widely used in various NMF frameworks since its first utilization in GNMF.

For supervised NMF methods, label provided by the data is used to encode the discriminative structure of the data. Discriminative NMF (DNMF) [16] incorporates the Fisher's criterion, which is defined on the encoding matrix, into the framework of NMF directly. An et al. propose a method named manifold-respecting discriminant NMF (NMF-kNN) [17] in which two graph Laplacian regularizers are included. Penalty graph is designed to describe the distinctness among the different classes and intrinsic graph is used to encode the local data distributing structure within the same class. Manifold regularized discriminative NMF (MDNMF) [18] also uses these two graphs to capture the discriminative information of the data. The difference is that, NMF-kNN uses the difference of these two regularizers and MDNMF uses the ratio of these two regularizers. Constrained NMF (CNMF) [19] incorporates the label constraint matrix into the cost function of NMF directly. Graph regularized discriminant NMF (GDNMF) [20] approximates the label indicating matrix with the product of the encoding matrix and a random matrix. In addition, a Laplacian graph is also constructed using label information in this method.

Considering that a Laplacian graph is constructed by a similarity matrix, it can also be called a similarity graph, and used for capturing latent structure information of the data. Generally speaking, there are two ways to build a similarity graph: one way is based on pairwise distance (e.g. Euclidean distance), another way is based on reconstruction coefficients (e.g. sparse representation). Pairwise distance based graph is usually constructed using the Euclidean distance that fails to explore the multi-subspaces structure of the data. However, reconstruction coefficients based graph can be used to capture multi-subspaces structure of the data when building it by using subspace representation coefficients. In this paper, we propose a novel subspace clustering guided convex nonnegative matrix factorization (SC-CNMF). SC-CNMF uses nonnegative subspace clustering to guide convex nonnegative matrix factorization. In this framework, the learning of reconstruction coefficients based graph and the convex nonnegative matrix factorization can be implemented simultaneously. To our knowledge, this kind of unified framework has not been proposed before.

The contributions of the proposed model are listed as follow:

- A unified NMF framework named subspace clustering guided convex nonnegative matrix factorization (SC-CNMF) is proposed. In this framework, nonnegative subspace clustering term, which can capture the multi-subspaces structure of the data, is incorporated to guide the learning of convex NMF. The nonnegative constraint that is imposed on the subspace clustering facilitates the optimizing of the proposed unified framework.
- A local subspace constraint is imposed on nonnegative subspace clustering term to improve the robustness of the proposed model. This constraint is that only s largest values are kept in each column of the learned representation matrix that will be used to construct the similarity matrix in each iteration. The two different ways of using this constraint lead to two different implementations of the proposed model, SC-CNMF₁ and SC-CNMF₂. And two slightly different optimizing schemes are designed for these two methods.
- Image clustering experiments are conducted on six image datasets in which the proposed methods are compared with several related NMF methods together with k -means clustering method and PCA as baseline. The experimental results reveal the effectiveness of the proposed methods.

The rest of paper is organized as follow: in Section 2, some related NMF methods are briefly reviewed in Section 2.1, then subspace clustering is briefly introduced in the Section 2.2. In Section 3, the details of the proposed model and its two implementations together with their optimizing schemes are described. The experimental results and analysis are in Section 4. The conclusion of this paper is made in Section 5.

2. Reviews of some related works

2.1. NMF, GNMF and convex NMF

Nonnegative matrix factorization (NMF) [3] tries to find two nonnegative factorial matrices, $U \in \mathbb{R}_+^{m \times k}$ and $V \in \mathbb{R}_+^{k \times n}$, whose product can approximate the original data matrix $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}_+^{m \times n}$. The objective function of the standard NMF with F -norm is like follow:

$$D_{NMF} = \|X - UV\|_F^2. \quad (1)$$

In this equation, we can see that, each original data point can be represented with the linear combination of the column vectors in U weighted by the corresponding column vector in V . So matrix U can be regarded as a basis matrix and matrix V is an encoding matrix. Eq. (1) is non-convex for U and V . But when one of these two matrices is fixed, Eq. (1) is convex for the rest one matrix. Therefore this equation can be minimized with an iterative updating strategy. The nonnegative constraints imposed on U and V cause that the updating rules only allow additive operations. To this end, the following updating rules are obtained:

$$\begin{aligned} U &= U \odot \frac{XV^T}{UVV^T}, \\ V &= V \odot \frac{U^T X}{U^T U V}, \end{aligned} \quad (2)$$

where $\odot$ denotes the element-wise multiplication and the division used in above equations are also element-wise.

To make use of the latent embedding manifold structure of the data, Cai et al. [12] incorporate a Laplacian regularizer into the framework of the standard NMF method and propose graph regularized NMF (GNMF). To capture the manifold structure of the data in the original space, a Laplacian graph is constructed which can describe the local geometry structure of the data. The local near neighbor relationship in the original space, i.e. X , can be kept in the learned low-dimension space, i.e. V , by minimizing the graph Laplacian regularizer. With this regularizer, the objective function of GNMF is as follow:

$$D_{GNMF} = \|X - UV\|_F^2 + \alpha \text{tr}(V L V^T), \quad (3)$$

where α is a parameter to control the influence of the Laplacian regularizer.

Convex NMF [21] constraints each column of the basis matrix to be a convex combination of the data points for the interpretability reason. In Convex NMF, the basis matrix can be denoted as $U = XG$, then its objective function can be written as

$$D_{\text{convexNMF}} = \|X - XGV\|_F^2. \quad (4)$$

The advantage of this convex constraint is that each column of U can be interpreted as a weighted sum of certain data samples.

2.2. Subspace clustering

Subspace clustering can find out latent subspace structure of the data. Generally, subspace clustering can be divided into three categories: algebraic algorithms, statistical methods and spectral clustering based methods. Among all these three kind of subspace clustering methods, the performance of spectral clustering based

methods are quite well. Furthermore, spectral clustering based methods can be classified into local and global spectral clustering based methods. Local methods depend on the pairwise distance based similarity matrix. This causes that local methods can not tackle with the points around the intersections of different subspaces well and are sensitive to the near neighborhood size [22]. While, global methods try to find a better similarity that can capture multi-subspaces structure of the data. Sparse subspace clustering (SSC) [23] and low rank representation (LRR) [24] are two representative global methods. These two methods are based on the self-representation model that believes that each sample $x_i \in \mathbb{R}^{m \times 1}$ can be linearly represented by the whole dataset $X \in \mathbb{R}^{m \times n}$ itself. Specifically,

$$x_i = Xz_i, \text{ s.t. } z_{ii} = 0, \quad (5)$$

where x_i is the i th column of X and z_i is the reconstruction coefficient of x_i on the self-representation dictionary X . The constraint $z_{ii} = 0$ is imposed to avoid the trivial solution. Rewrite Eq. (5) for all samples in matrix format and obtain

$$X = XZ, \text{ s.t. } \text{diag}(Z) = 0, \quad (6)$$

where $Z \in \mathbb{R}^{n \times n}$ is the self-representation matrix and $\text{diag}(\cdot)$ denotes the operation that gets the diagonal entries of a matrix. With the obtained Z , a similarity matrix $W \in \mathbb{R}^{n \times n}$ can be calculated as $W = \frac{|Z| + |Z|^T}{2}$. Then spectral clustering can be conducted on this similarity matrix. SSC seeks a sparse representation and LRR try to find a low rank representation for the data points in X . Both these two methods can capture multi-subspaces structure of the data, but they can not tackle the influence of the inter-subspace points. The reconstruction coefficients from the inter-subspace points are always small and corresponding to errors. To tackle the noise influence in the projection space, Peng et al. [25] has proposed a thresholding operation to eliminate the affect of the errors. In this operation, only s largest values are kept in each column of Z . It has been proven that this thresholding operation can effectively improve the robustness of subspace clustering.

3. The proposed model

A similarity matrix is usually built to construct a Laplacian graph used for improving the performance of NMF methods. Pairwise distance based graph are widely used in these NMF methods. However, this kind of similarity graph can not dig out the latent subspace structure of the data. Reconstruction coefficients based graph, such as SSC and LRR, can explore the multi-subspaces structure of the data. Generally speaking, the latter kind of graph can also be used to improve the performance of the NMF methods. But subspace clustering, which can build the reconstruction coefficients based graph, is usually independent from graph based NMF methods. That is to say, subspace clustering is conducted firstly and then the similarity graph built in subspace clustering can be used in the framework of graph based NMF methods. To address this issue, a novel subspace clustering guided convex NMF (SC-CNMF) framework is proposed. To our knowledge, it is the first to simultaneously optimize subspace clustering and NMF in a unified learning framework.

3.1. Model

In the proposed model, subspace clustering is used to guide the learning of convex NMF, in which the Laplacian graph will be updated by the former. With this mechanism, the multi-subspaces structure information captured by subspace clustering will be utilized to improve the performance of convex NMF. The relation of subspace clustering and convex NMF is connected by a Laplacian

graph regularizer. For given data matrix $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}_+^{m \times n}$, the proposed model can be preliminarily written as

$$\min_{G, V \geq 0} \|X - XGV\|_F^2 + \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T), \quad (7)$$

where α is a parameter to control the weight of Laplacian regularizer. The former term is the objective function of convex NMF which has been introduced in the former section. $\|X - XZ\|_F^2$ is used to learn the self-representation coefficient matrix $Z \in \mathbb{R}^{n \times n}$ which can be used to calculate the similarity matrix as $W = \frac{|Z| + |Z|^T}{2}$. Then Laplacian matrix $L = D - W$, $D_{ii} = \sum_j W_{ij}$. Additionally, in order to facilitate the optimizing of the proposed model, a nonnegative constraint is imposed on the coefficient matrix in the subspace clustering. Then the model is

$$\min_{G, V, Z \geq 0} \|X - XGV\|_F^2 + \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T). \quad (8)$$

In this model, the similarity matrix is calculated as $W = \frac{|Z| + |Z|^T}{2}$. To alleviate the relevance among the rows of V , a constraint on V is added in the objective function with following format:

$$\text{tr}(V^T eV), \quad (9)$$

where $e = \bar{1} - I$, $\bar{1}$ is a $k \times k$ matrix with all ones and I is identity matrix. For now, the objective function of the model is

$$\min_{G, V, Z \geq 0} \|X - XGV\|_F^2 + \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T) + \beta \text{tr}(V^T eV), \quad (10)$$

where parameter β is set to control the weight of corresponding regularizing term. To avoid trivial solution of Z when optimizing above objective function, some constraint should be added on Z . There are two ways to restrict Z and these two different constraints lead to two implementation of the proposed model. One way is to put a square of F -norm term of Z into above objective function, another way is to restrict $\text{diag}(Z) = 0$. The implementations of the proposed model with former and latter constraint are respectively denoted as SC-CNMF₁ and SC-CNMF₂. Finally, the objective function of SC-CNMF₁ and SC-CNMF₂ can be written as

$$D_{\text{SC-CNMF}_1} = \|X - XGV\|_F^2 + \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T) + \beta \text{tr}(V^T eV) + \gamma \|Z\|_F^2, \text{ s.t. } G, V, Z \geq 0, \quad (11)$$

$$D_{\text{SC-CNMF}_2} = \|X - XGV\|_F^2 + \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T) + \beta \text{tr}(V^T eV), \text{ s.t. } G, V, Z \geq 0, \text{diag}(Z) = 0. \quad (12)$$

The parameter γ in Eq. (11) is set to balance the weight of $\|Z\|_F^2$. To improve the robustness of the proposed two methods, before using Z to construct the similarity matrix, only s largest values are kept in each column of Z . This operation can be implemented using thresholding operation [25] in SC-CNMF₁ and be realized as an optimizing constraint in the optimization of SC-CNMF₂ when solving Z . Specifically, the implementations of this operation in SC-CNMF₁ and SC-CNMF₂ will be described clearly in following subsections.

Additionally, to improve the representative ability of the proposed methods, the dimension of the learned low-dimension representation matrix $V \in \mathbb{R}_+^{k \times n}$ is set $k = n_s \times n_c$. Where n_c is the number of clusters of one dataset and n_s is the number of centroids of each cluster. For most traditional NMF methods, n_s is usually set 1 which means each cluster is represented by one single centroid. However, for complicated data distributing structure, it is not wise to represent one cluster with single centroid. So, in the proposed methods, for each cluster multiple centroids are adopted to capture the complicated manifold structure of the data.

3.2. Optimization of SC-CNMF₁

In this section, the details to optimizing the objective function of SC-CNMF₁ will be described. Three matrices, i.e. G , V and Z , need to be solved in Eq. (11). They could be solved in an iterative way. In following paragraphs, optimizing scheme will be clearly described.

When Z is fixed, Eq. (11) reduces to

$$\min_{G, V \geq 0} \|X - XGV\|_F^2 + \alpha \text{tr}(VLV^T) + \beta \text{tr}(V^T eV). \quad (13)$$

Above equation is only about G and V , and when one of G or V is fixed it is convex for the other one. So Eq. (13) can be solved iteratively using multiplicative updating rules as NMF [3] and GNMF [12]. Specifically, Eq. (13) can be rewritten as

$$\begin{aligned} O_1 &= \text{tr}((X - XGV)(X - XGV)^T) + \alpha \text{tr}(VLV^T) + \beta \text{tr}(V^T eV) \\ &= \text{tr}(XX^T) - 2\text{tr}(XGVX^T) + \text{tr}(XGVV^T G^T X^T) + \alpha \text{tr}(VLV^T) \\ &\quad + \beta \text{tr}(V^T eV). \end{aligned} \quad (14)$$

Then the Lagrangian function of Eq. (14) for the constraints $G \geq 0$ and $V \geq 0$ with the Lagrangian multipliers $\Psi \in \mathbb{R}^{n \times k}$ and $\Phi \in \mathbb{R}^{k \times n}$ is as follow

$$\begin{aligned} L(G, V) &= \text{tr}(XX^T) - 2\text{tr}(XGVX^T) + \text{tr}(XGVV^T G^T X^T) + \alpha \text{tr}(VLV^T) \\ &\quad + \beta \text{tr}(V^T eV) + \text{tr}(\Psi G^T) + \text{tr}(\Phi V^T). \end{aligned} \quad (15)$$

The partial derivatives of $L(G, V)$ with respect to G and V are

$$\begin{aligned} \frac{\partial L(G, V)}{\partial G} &= -2X^T X V^T + 2X^T X G V V^T + \Psi, \\ \frac{\partial L(G, V)}{\partial V} &= -2G^T X^T X + 2G^T X^T X G V + 2\alpha V L + 2\beta e V + \Phi. \end{aligned} \quad (16)$$

Using the KKT (Karush–Kuhn–Tucker) conditions [26] $\Psi_{it} G_{it} = 0$ and $\Phi_{tj} V_{tj} = 0$, we have

$$\begin{aligned} (-2X^T X V^T + 2X^T X G V V^T)_{it} G_{it} &= 0, \\ (-2G^T X^T X + 2G^T X^T X G V + 2\alpha V L + 2\beta e V)_{tj} V_{tj} &= 0. \end{aligned} \quad (17)$$

Then, we have following updating rules

$$\begin{aligned} G &= G \odot \frac{X^T X V^T}{X^T X G V V^T}, \\ V &= V \odot \frac{G^T X^T X + \alpha V W}{G X^T X G V + \alpha V D + \beta e V}, \end{aligned} \quad (18)$$

where $\odot$ denotes the element-wise multiplication and the division operation used in above rules is also element-wise.

When G and V are fixed, Eq. (11) reduces to

$$\min_{Z \geq 0} \|X - XZ\|_F^2 + \alpha \text{tr}(VLV^T) + \gamma \|Z\|_F^2. \quad (19)$$

According to [27], Eq. (19) can be reformulated as

$$\min_{Z \geq 0} \|X - XZ\|_F^2 + \alpha \text{tr}(Z^T P) + \gamma \|Z\|_F^2, \quad (20)$$

where $P_{ij} = \|v_i - v_j\|_2^2$, v_i denotes the i th column of V . Then Z could also be solved using multiplicative method [3,12] which can be deduced through a similar procedure as solving G and V . Then the updating rule for Z is

$$Z = Z \odot \frac{4X^T X}{4X^T X Z + \alpha P + 4\gamma Z}. \quad (21)$$

When Z is obtained, the similarity matrix can be calculated as $W = \frac{Z + Z^T}{2}$ and then the Laplacian graph is refined. However, in order to improve the robustness of the refined Laplacian graph, before calculating W , a thresholding operation is imposed on Z in which only s largest values in each column of Z are kept. In this way, the influence among inter-subspaces can be effectively reduced. The optimizing scheme of SC-CNMF₁ is summarized in Algorithm 1.

ALGORITHM 1: The optimizing scheme of SC-CNMF₁.

Require: $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}_+^{m \times n}$, $\alpha > 0$, $\beta > 0$, $\gamma > 0$, $s > 0$
Ensure: G, V, Z
 Initialize G, V with nonnegative random matrices, Z and L are initialized with Heat kernel metric, the near neighbor size is set 5 as GNMF [12]
repeat
 S1: Update G as
 $G = G \odot \frac{X^T X V^T}{X^T X G V V^T}$
 S2: Update V as
 $V = V \odot \frac{G^T X^T X + \alpha V W}{G X^T X G V + \alpha V D + \beta e V}$
 S3: Update Z as
 $Z = Z \odot \frac{4X^T X}{4X^T X Z + \alpha P + 4\gamma Z}$
 S4: Choose s largest values in each column of Z and obtain Z^*
 S5: Update Laplacian graph $L = D - W$, $W = \frac{Z^* + Z^{*T}}{2}$,
 $D_{ii} = \sum_j W_{ij}$
until Converges

3.3. Optimization of SC-CNMF₂

SC-CNMF₂ can also be optimized in an iterative manner. When Z is fixed, Eq. (12) also reduces to Eq. (13). This makes G and V can be solved by using multiplicative updating rules in Eq. (18) as well. But the different constraint imposed on Z to avoid trivial solution makes the optimizing manner of Z is different from SC-CNMF₁.

Because of the symmetry of P , $\text{tr}(Z^T P)$ in Eq. (20) can be rewritten as $\text{tr}(ZP)$. Therefore, when G and V are fixed, Eq. (12) reduces to

$$\min_{Z \geq 0, \text{diag}(Z)=0} \|X - XZ\|_F^2 + \alpha \text{tr}(ZP). \quad (22)$$

When optimizing above equation, the constraint that only s largest values are kept in each column of Z is also included. Then Eq. (22) with this constraint is

$$\begin{aligned} \min_Z & \|X - XZ\|_F^2 + \alpha \text{tr}(ZP), \\ \text{s.t. } & Z \geq 0, \text{diag}(Z) = 0, \|Z_{:,i}\|_0 \leq s, i = 1, 2, \dots, n. \end{aligned} \quad (23)$$

Above problem can be optimized iteratively by solving each column of Z . Eq. (23) can be reduced to following subproblem

$$\begin{aligned} \min_{y_i} & \|x_i - X_{-i} y_i\|_F^2 + \frac{\alpha}{2} q_i^T y_i, \\ \text{s.t. } & y_i \geq 0, \|y_i\|_0 \leq s, \end{aligned} \quad (24)$$

where $X_{-i} = \{x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n\} \in \mathbb{R}_+^{m \times (n-1)}$, i.e. the i th sample x_i is eliminated from X . $y_i \in \mathbb{R}_+^{(n-1) \times 1}$ is representation z_i without the i th entry and $q_i \in \mathbb{R}^{1 \times (n-1)}$ is the i th row of P without i -th entry. Subproblem Eq. (24) can be solved by projected gradient descent method in Algorithm 2. The obtained $\{y_1, \dots, y_i, \dots, y_n\}$ can be reorganized to obtain Z with zeros on the diagonal. Finally, the optimizing scheme of SC-CNMF₂ is summarized in Algorithm 3.

4. Experimental results

In this section, image clustering experiments will be conducted on six datasets comparing with several representative methods to demonstrate the effectiveness of our methods. The compared methods include k -means clustering method [28], principle component analysis (PCA) [29], NMF [3], GNMF [12], $\ell_{2,1}$ -norm NMF ($\ell_{2,1}$ -NMF) [30], robust structured NMF (RSNMF) [31], Capped-NMF [32], SC-CNMF₁ and SC-CNMF₂. $\ell_{2,1}$ -NMF employs $\ell_{2,1}$ -norm instead of the square of F -norm to measure the residual of

ALGORITHM 2: Projected gradient descent method.**Require:**
 $X_{-i} \in \mathbb{R}_+^{m \times (n-1)}, x_i \in \mathbb{R}_+^{m \times 1}, q_i \in \mathbb{R}_+^{1 \times (n-1)}, \alpha > 0, \eta > 0, s > 0$
Ensure: y_i **repeat**

S1: compute the gradient

$$\nabla_{y_i} = 2(\frac{\alpha}{4}q_i^T - X_{-i}^T x_i + X_{-i}^T X_{-i} y_i)$$

S2: projected gradient descent

$$y_i = \max(y_i - \eta \nabla_{y_i}, 0)$$

S3: solve ℓ_0 -norm constraintkeep the s largest values of y_i **until** Converges**ALGORITHM 3:** The optimizing scheme of SC-CNMF₂.**Require:** $X = \{x_1, x_2, \dots, x_n\} \in \mathbb{R}_+^{m \times n}, \alpha > 0, \beta > 0, s > 0$ **Ensure:** G, V, Z

Initialize G, V with nonnegative random matrices, Z and L are initialized with Heat kernel metric, the near neighbor size is set 5 as GNMF [12]

repeatS1: Update G as

$$G = G \odot \frac{X^T X V^T}{X^T X G V V^T}$$

S2: Update V as

$$V = V \odot \frac{G^T X^T X + \alpha V W}{G X^T X G V + \alpha V D + \beta e V}$$

S3: Get y_i ($i = 1, 2, \dots, n$) by solving Eq. (24) with Algorithm 2 and re-

organize $\{y_1, \dots, y_i, \dots, y_n\}$ to obtain Z .

S4: Update Laplacian graph $L = D - W$, $W = \frac{Z + Z^T}{2}$,
 $D_{ii} = \sum_j W_{ij}$

until Converges**Table 1**

Database description.

Databases	# Samples (n)	# Features (d)	# Classes (C)
Yale	165	1024	11
UMIST	575	644	20
ORL	400	1024	40
BinaryAlpha	1404	320	36
YaleB	2414	1024	38
Pointing04	2790	1120	15

reconstruction. $\ell_{2,1}$ -norm can restrain the influence of noises and outliers in the data. RSNMF uses label information to ensure the diagonal structure of the learned low dimension representation and measures the residual of reconstruction with $\ell_{2,p}$ -norm ($0 < p \leq 1$) which can inhibit noises and outliers to some extent. It has been reported that when $p = 0.5$ RSNMF has a better results. Without label information, RSNMF reduces to an standard NMF with $\ell_{2,p}$ -norm. In this paper, unsupervised version of RSNMF is used and p is set 0.5. For convenience, it is denoted as $\ell_{2,0.5}$ -NMF. CappedNMF uses a preset threshold to restrict the residual of each sample in order to inhibit the influence of noises and outliers.

4.1. Data sets

Six datasets are used in the experiments and some important statistical information of these datasets are listed in Table 1. Some sample images are shown in Fig. 1. The description of these datasets are summarized as follow:

1. Yale¹ has 165 grayscale images collected from 15 people. Each person has 11 images with different facial expressions or configurations and all these images are normalized to 32×32 pixel array.
2. UMIST² contains 575 grayscale images collected from 20 individuals in size of 28×23 pixels. The pose of each subject varies from frontal to side views uniformly.
3. ORL³ has 400 grayscale face images collected from 40 people and each subject has 10 samples with size of 32×32 . The images of each subject are captured under different configurations: different facial expressions and light conditions, with or without glasses.
4. BinaryAlpha⁴ consists of 1404 binary images of 10 digits from '0' to '9' and 26 capitals from 'A' to 'Z' with 20×16 . Each class has 39 samples. In the experiments, 20 samples of each class are selected for test.
5. YaleB⁵ has 2414 grayscale images collected from 38 people. Each person has around 64 samples taken under very different light and illumination conditions. In the experiments, 20 images are selected from each class for test.
6. Pointing04⁶ contains 2790 images collected from 15 individuals. Each individual has 186 images with head pose varies from up to down and left to right and with or without classes. In the experiments, 30 images of each class are selected for test.

Additionally, in the experiments, the samples in all datasets are normalized to unit vectors for all comparing methods.

4.2. Evaluation metrics

Two clustering metrics are used to evaluate the performance of the methods: clustering accuracy (AC) [33] and normalized mutual information (NMI) [34].

AC is defined as follow:

$$AC = \frac{\sum_{i=1}^n \delta(gnd_i, map(z_i))}{n}, \quad (25)$$

where n is the number of samples in each database. $\delta(a, b)$ is delta function. When $a = b$, $\delta(a, b) = 1$, otherwise, $\delta(a, b) = 0$. gnd_i is the ground truth label of x_i . z_i is the label obtained by clustering methods corresponding to the sample x_i . $map(\cdot)$ is a mapping function which can map the labels obtained by the clustering methods to the labels provided by the databases.

NMI is defined as

$$NMI(C, C') = \frac{MI(C, C')}{\sqrt{H(C)H(C')}}, \quad (26)$$

where C and C' separately denote the cluster set provided by the database and the cluster set obtained by the clustering methods. $MI(C, C')$ is the mutual information between C and C' . $H(C)$ and $H(C')$ denote the entropies of C and C' . $NMI(C, C')$ varies from 0 to 1. If C is independent from C' , $NMI(C, C') = 0$. If C is equivalent to C' , $NMI(C, C') = 1$.

4.3. Parameter selection

There are several parameters in the proposed methods, i.e. SC-CNMF₁ and SC-CNMF₂. Parameter analysis on the proposed methods will be conducted in this section. The Laplacian graph regularizer parameter α is set 100 as GNMF [12]. There are two

¹ <http://vision.ucsd.edu/content/yale-face-database>

² <http://researchweb.iit.ac.in/~chetan/face-retrieval-php/dataset.php>

³ <http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html>

⁴ <https://cs.nyu.edu/~roweis/data.html>

⁵ <http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html>

⁶ <http://www-prima.inrialpes.fr/Pointing04/data-face.html>

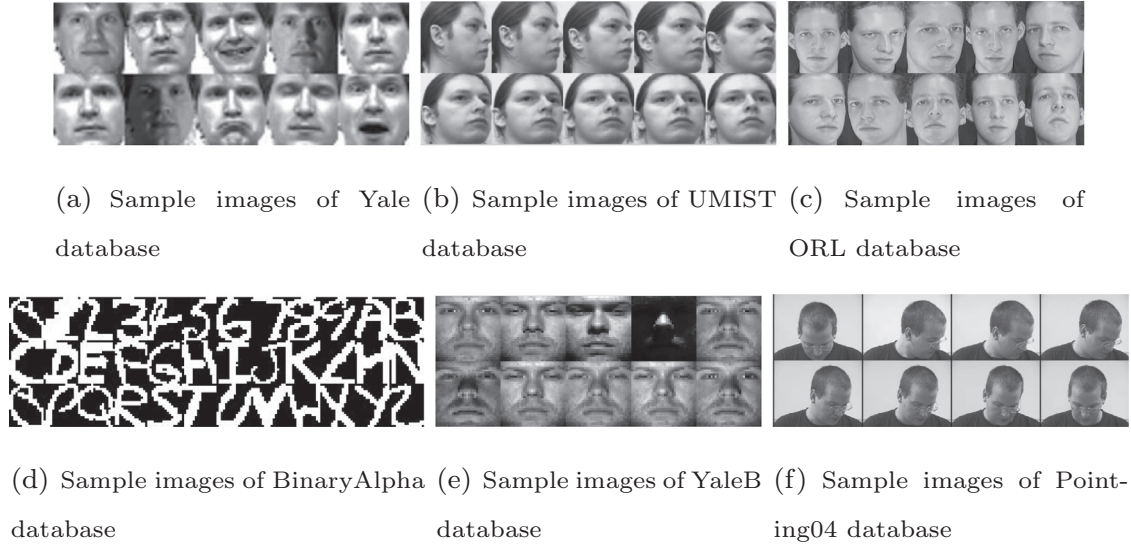
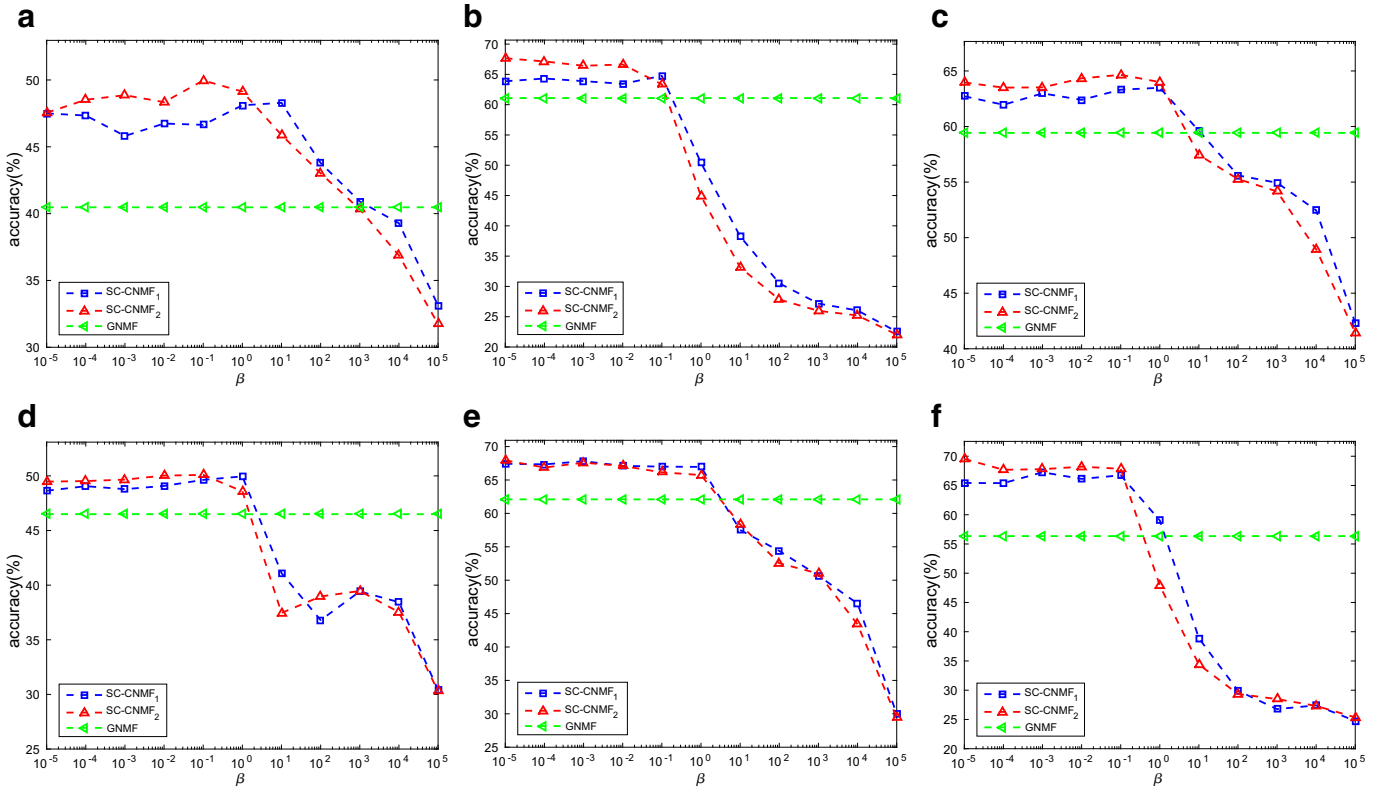


Fig. 1. Illustration of sample images from six databases.

Fig. 2. Performance of SC-CNMF₁ and SC-CNMF₂ versus parameter β on the datasets (a) Yale, (b) UMIST, (c) ORL, (d) BinaryAlpha, (e) YaleB, and (f) Pointing04.

common parameters in SC-CNMF₁ and SC-CNMF₂: β and s . The parameter γ only exists in SC-CNMF₁ and the parameter η is only used in SC-CNMF₂. Fig. 2–5 demonstrate the accuracy curves of the proposed methods with varying values of different parameters. In these figures, GNMf is plotted as baseline. Additionally, n_s is set 6 for BinaryAlpha dataset and $n_s = 10$ for other datasets. This is because that, if $n_s = 10$ on BinaryAlpha dataset, the dimension of V will be larger than the dimension of the original data.

Fig. 2 demonstrates the accuracy curves of SC-CNMF₁ and SC-CNMF₂ varying with different values of parameter β . The tendencies of these curves are very obvious on all datasets for both proposed methods. When parameter value is too large, the accuracy

goes down sharply. When value is small, the performance of both methods are relatively steady. The inflection points of the proposed methods on all datasets are round the parameter value 10^0 . Balancing the performance of the proposed methods on all datasets, we set $\beta = 10^{-2}$ for all datasets in the experiments.

Parameter γ only exists in SC-CNMF₁. Fig. 3 shows the varying accuracy curves of SC-CNMF₁ on six datasets with different values of γ . For UMIST, ORL, BinaryAlpha, YaleB, and Pointing04 datasets, a large value of γ is needed for SC-CNMF₁. On Yale and YaleB datasets, SC-CNMF₁ outperforms GNMf on the whole scale of the parameter values. The inflection points of SC-CNMF₁ on UMIST, ORL, BinaryAlpha and Pointing04 datasets are around the

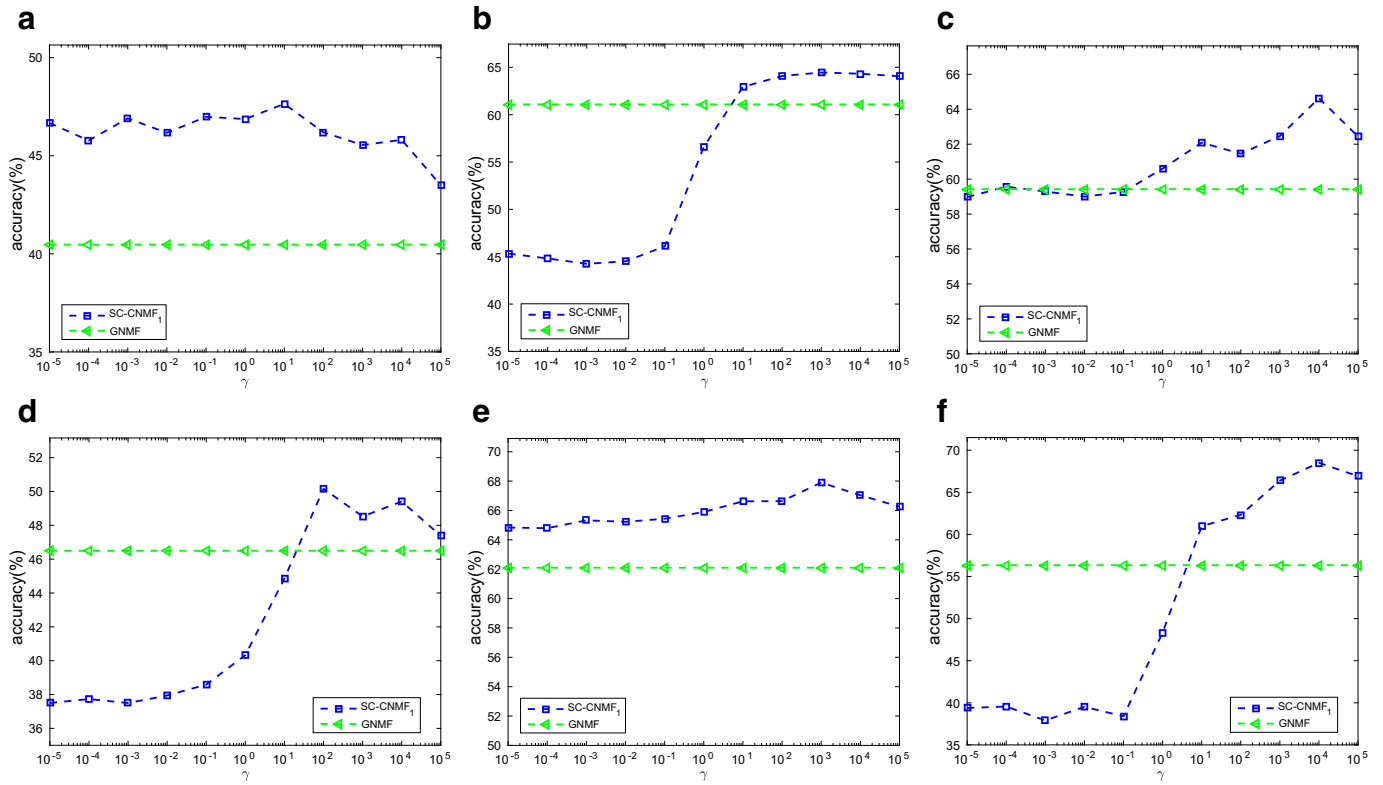


Fig. 3. Performance of SC-CNMF₁ versus parameter γ on the datasets (a) Yale, (b) UMIST, (c) ORL, (d) BinaryAlpha, (e) YaleB, and (f) Pointing04.

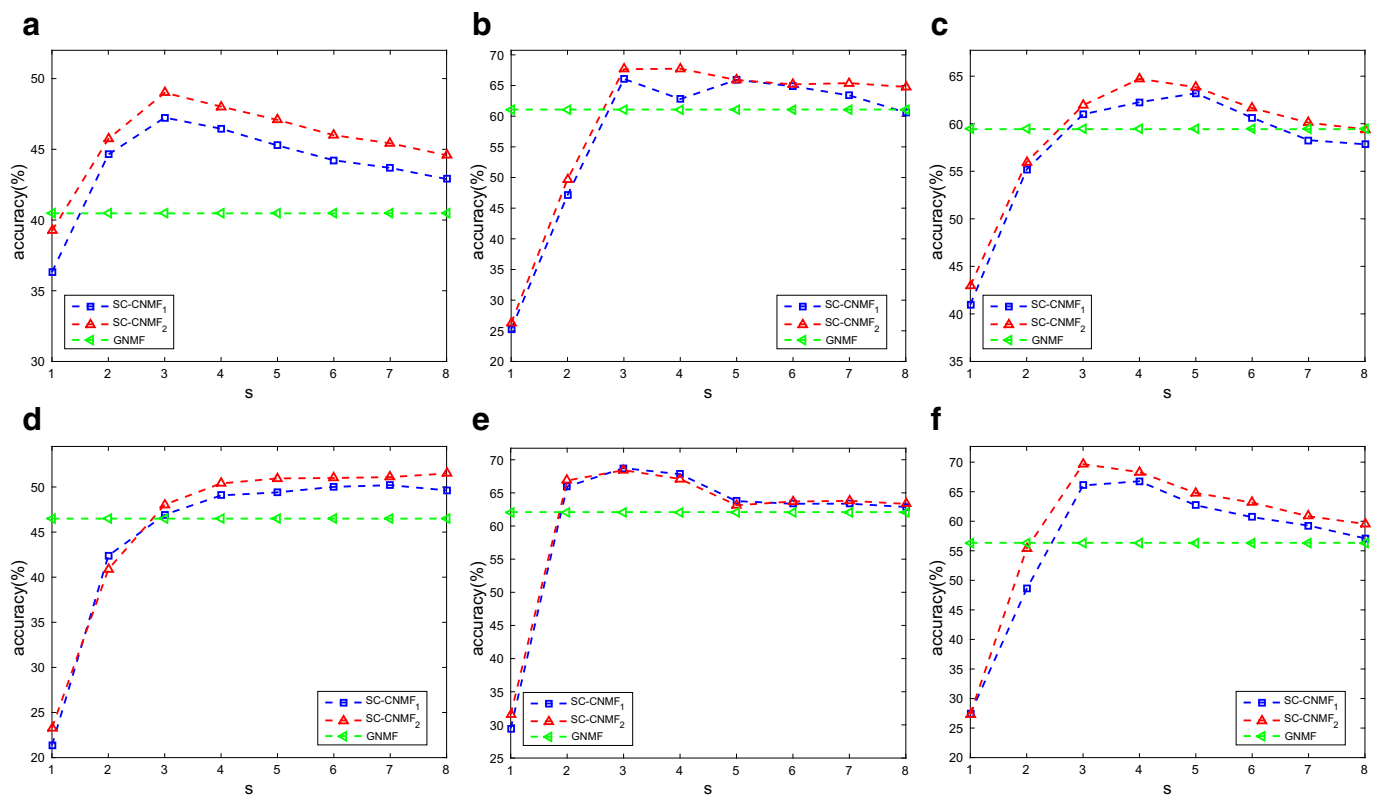


Fig. 4. Performance of SC-CNMF₁ and SC-CNMF₂ versus parameter s on the datasets (a) Yale, (b) UMIST, (c) ORL, (d) BinaryAlpha, (e) YaleB, and (f) Pointing04.

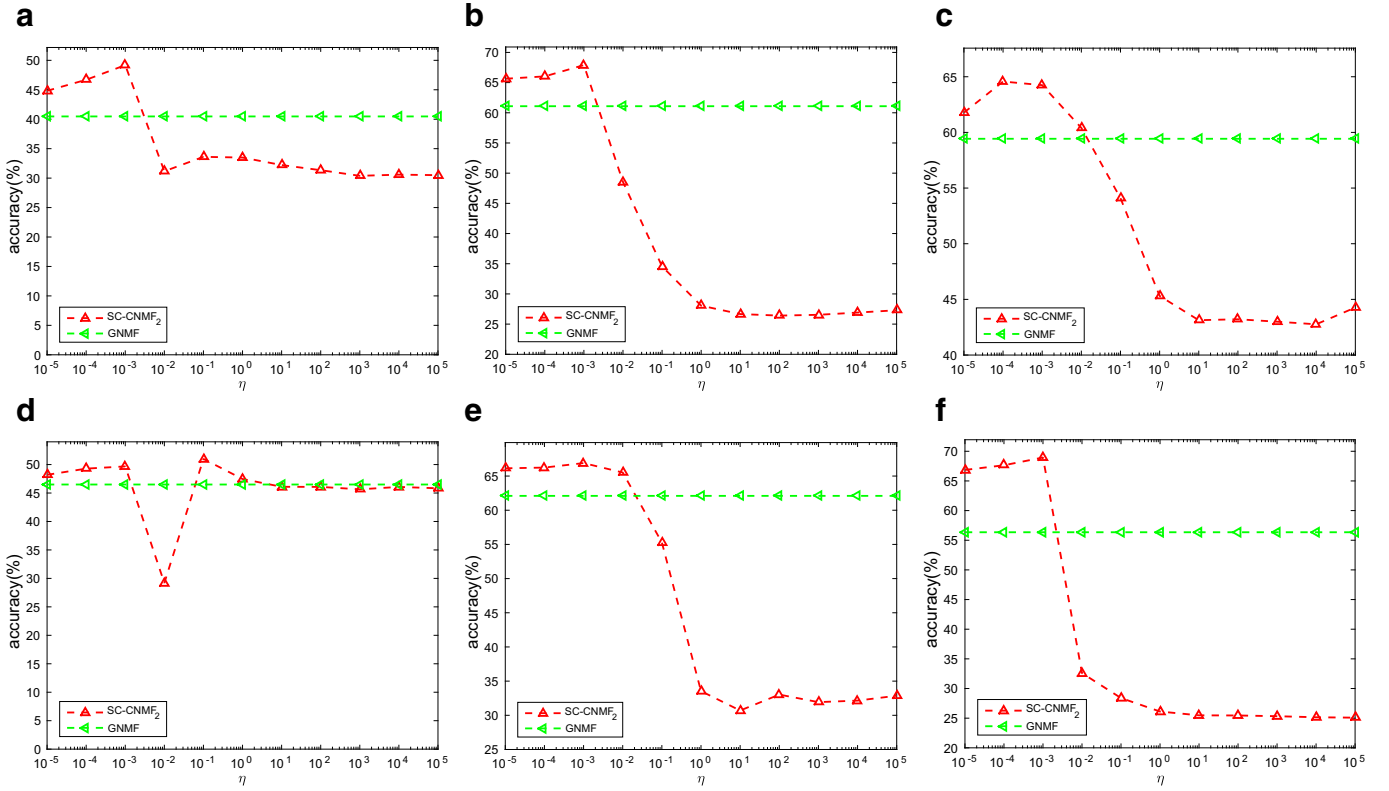


Fig. 5. Performance of SC-CNMF₂ versus parameter η on the datasets (a) Yale, (b) UMIST, (c) ORL, (d) BinaryAlpha, (e) YaleB, and (f) Pointing04.

parameter value 10^1 . Balancing the robustness and the performance of SC-CNMF₁ on all datasets, γ is set 10^3 for all datasets in the experiments.

Parameter s is used to eliminate the influence of inter-subspace coefficients. Intuitively speaking without experiments, s can not be set too large or too small. When s is set too large, the influence of inter-subspace can not be effectively inhibited. If s is set too small, the multi-subspaces structure of the data can not be explored well. Fig. 4 has verified this intuition. When s is set too large, the performance of these two methods go down gradually on most datasets except BinaryAlpha dataset, but it can be expected that if the value of s increases further, the curves will surely goes down. On the other hand, if s is set too small, the accuracy curves drop sharply. SC-CNMF₁ and SC-CNMF₂ have peak accuracies on Yale, UMIST, ORL, YaleB and Pointing04 datasets. The peak accuracies of the proposed methods on Yale, UMIST, YaleB and Pointing04 datasets are reached when $s = 3$. For ORL dataset, the best results of the proposed methods are obtained when $s = 4$. The proposed methods perform stably when $s \geq 4$ for the scale of $1 \sim 8$. Usually, it is proper to set s within $[3, 4, 5]$. In the experiments, s is set 4 for all datasets.

Parameter η is the step size used in projected gradient descent method when solving each column of self-representation matrix Z . Fig. 5 shows that the step size can not be too large, however, it also can be seen that if η is set too small, the performance of SC-CNMF₂ goes down slowly also. On Yale, UMIST, YaleB and Pointing04 datasets, SC-CNMF₂ has peak accuracies when $\eta = 10^{-3}$. For ORL and BinaryAlpha datasets, SC-CNMF₂ gets best results respectively when $\eta = 10^{-4}$ and $\eta = 10^{-1}$. Balancing the performance of the proposed method on all datasets, η is set 10^{-3} for all datasets in the experiments.

Additionally, from Figs. 2 and 4 it is can be seen that SC-CNMF₂ usually performs better than SC-CNMF₁ on most datasets. It indicates that it is better to optimize the proposed model with

thresholding constraint than implementing thresholding operation as a post-processing.

4.4. Comparative experiments

In this section, our proposed methods are compared with k -means clustering method, PCA and several representative NMF methods on six image datasets. For our proposed methods, the number of centroids n_s for each cluster is set 10 on Yale, UMIST, ORL, YaleB and Pointing04 datasets. For BinaryAlpha dataset, n_s is set 6. One single centroid is assigned to each cluster for other NMF methods. The comparing results are reported in Table 2. The results in Table 2 indicate the effectiveness of the proposed methods. And on the most of datasets, SC-CNMF₂ outperforms SC-CNMF₁. This phenomena shows that optimizing the proposed model with thresholding constraint is better than implementing thresholding operation as a post-processing. This finding is also verified in above Section 4.3. The price SC-CNMF₂ pays for a better performance is that it needs more computation than SC-CNMF₁ in optimizing Z . Because Z is iteratively solved in SC-CNMF₂ and is not solved as a whole matrix in SC-CNMF₁. In most cases, GNMF has better results than other basic NMF methods which have nonnegative constraint only. This indicates the geometry information can improve the performance of NMF methods. Both proposed methods outperform GNMF, this indicates that the multi-subspaces structure information can be used to improve the performance of NMF effectively. And the local subspace constraint imposed on the subspace clustering term ensures the robustness of the proposed methods just like above subsection mentioned. Because this constraint can eliminate the influence of inter-subspace coefficients.

To further figure out how n_s effects different NMF methods mentioned in the experiments, the performance of all NMF methods including the proposed SC-CNMF₁ and SC-CNMF₂ with

Table 2
Clustering performance on Yale, UMIST, ORL, BinaryAlpha, YaleB and Pointing04 datasets.

Datasets	Yale	UMIST	ORL	BinaryAlpha	YaleB	Pointing04
AC (%)						
k-means	37.15 ± 3.5	41.10 ± 1.6	52.93 ± 2.3	42.61 ± 2.5	26.42 ± 1.5	43.44 ± 2.2
PCA	41.09 ± 2.1	39.70 ± 2.6	53.97 ± 2.3	46.42 ± 1.4	24.91 ± 1.7	44.78 ± 2.3
NMF	43.29 ± 3.3	39.22 ± 1.3	57.64 ± 1.2	39.09 ± 1.5	44.85 ± 1.7	40.06 ± 1.5
$\ell_{2,1}$ -NMF	43.24 ± 2.3	39.78 ± 1.1	58.92 ± 1.0	37.29 ± 1.0	48.54 ± 2.0	40.60 ± 2.2
$\ell_{2,0.5}$ -NMF	42.22 ± 1.8	39.91 ± 1.9	59.04 ± 0.7	37.02 ± 1.2	49.69 ± 1.0	40.33 ± 2.7
CappedNMF	42.57 ± 2.7	40.63 ± 1.4	50.29 ± 1.7	42.96 ± 1.6	37.70 ± 1.0	43.68 ± 1.8
GNMF	40.42 ± 1.5	59.93 ± 2.0	59.64 ± 0.6	46.41 ± 0.8	62.50 ± 0.8	56.39 ± 0.6
SC-CNMF ₁	47.28 ± 1.3	64.42 ± 2.0	64.06 ± 1.5	48.46 ± 0.7	67.08 ± 0.5	69.24 ± 1.2
SC-CNMF ₂	49.36 ± 1.3	65.07 ± 2.1	65.23 ± 1.0	50.03 ± 0.5	66.90 ± 1.2	69.50 ± 1.8
NMI (%)						
k-means	42.34 ± 3.1	59.94 ± 1.5	71.78 ± 1.9	60.28 ± 1.1	42.75 ± 1.2	54.95 ± 1.5
PCA	47.48 ± 2.3	59.10 ± 1.3	72.70 ± 1.2	63.35 ± 1.0	42.23 ± 1.4	56.53 ± 1.9
NMF	46.69 ± 2.0	57.08 ± 1.6	75.42 ± 0.8	56.93 ± 1.1	61.75 ± 1.4	49.76 ± 2.0
$\ell_{2,1}$ -NMF	47.15 ± 2.3	57.44 ± 1.3	76.28 ± 0.7	55.46 ± 1.0	63.80 ± 1.4	49.90 ± 2.6
$\ell_{2,0.5}$ -NMF	46.27 ± 1.3	57.78 ± 1.9	76.41 ± 0.6	55.32 ± 1.4	64.44 ± 0.7	49.04 ± 3.5
CappedNMF	47.55 ± 1.5	58.75 ± 1.4	69.25 ± 1.0	59.22 ± 0.9	52.26 ± 0.8	55.52 ± 1.5
GNMF	45.88 ± 1.6	76.31 ± 1.3	74.80 ± 0.4	63.21 ± 0.6	73.85 ± 0.4	70.05 ± 0.6
SC-CNMF ₁	53.00 ± 1.0	80.62 ± 1.1	79.68 ± 0.7	64.18 ± 0.4	76.22 ± 0.3	79.33 ± 1.0
SC-CNMF ₂	54.84 ± 0.9	81.79 ± 1.0	80.18 ± 0.5	65.00 ± 0.4	76.24 ± 0.7	80.28 ± 0.8

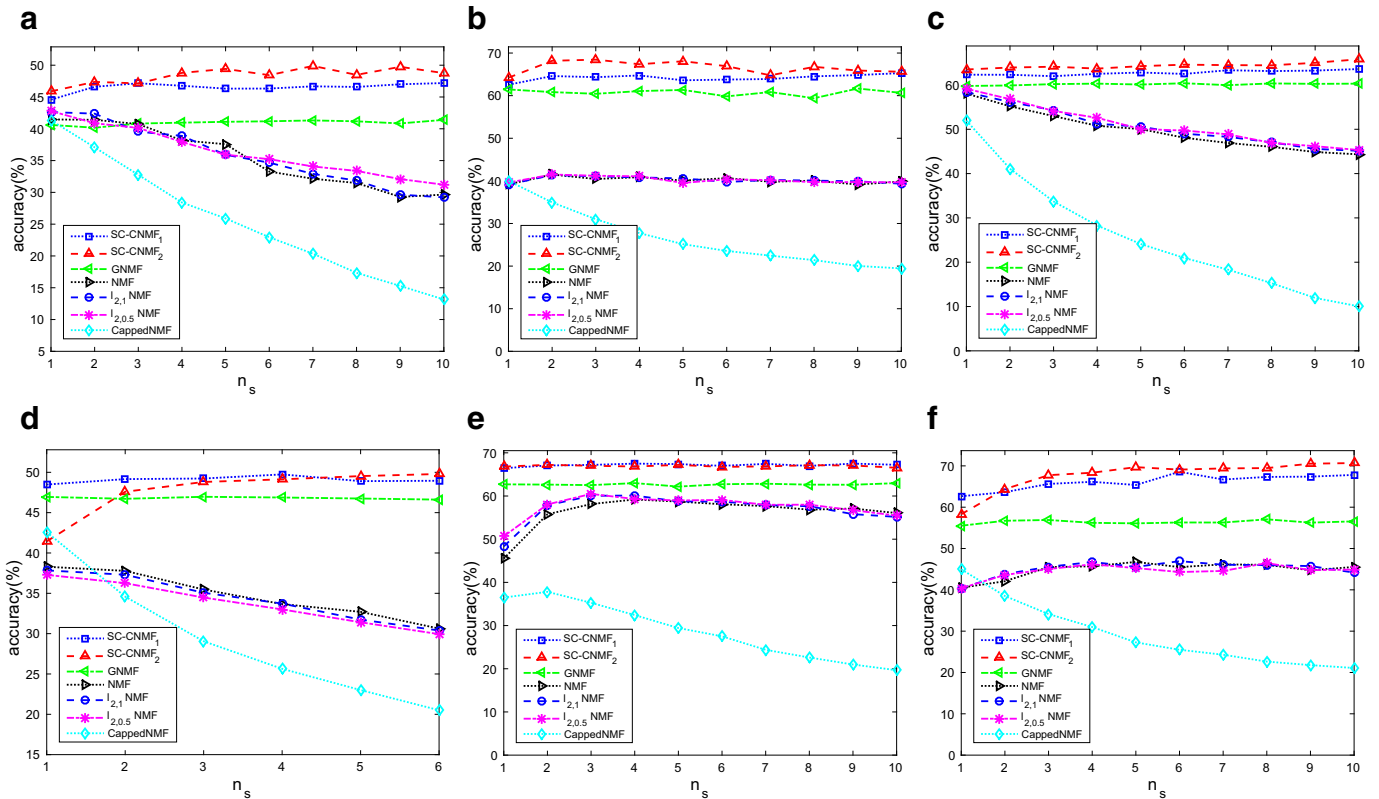


Fig. 6. Performance of all NMF methods mentioned in the experiments with different n_s on the datasets (a) Yale, (b) UMIST, (c) ORL, (d) BinaryAlpha, (e) YaleB, and (f) Pointing04.

different n_s , which varies from 1 to 10 (for BinaryAlpha dataset, the scale of n_s is 1 to 6), is reported in Fig. 6.

From Fig. 6, we can see that the performance of GNMF is stable on the whole scale of n_s . The performance of SC-CNMF₁ is steady on UMIST, ORL, BinaryAlpha and YaleB datasets. While SC-CNMF₂ is relatively stable on ORL and YaleB datasets. For other situations, the performance of SC-CNMF₁ and SC-CNMF₂ slightly grows with the values of n_s . On Yale, ORL and BinaryAlpha datasets, the accuracy curves of the basic NMF family methods drop with the increasing of n_s . On other datasets, the performance of the basic NMF family methods grows first and then slightly goes down with the

increasing of n_s . CappedNMF performs poorly with larger n_s . Now, we can find that multiple centroids not always can improve the performance of NMF methods. Basic NMF methods sometimes behave poorly with multiple centroids. However, the performance of the NMF methods with graph regularizer using multiple centroids are usually promoted. This is because that the geometry structure information discovered by the multiple centroids is imposed on the basis matrix and the same information discovered by graph Laplacian is emphasized on the encoding matrix. Therefore, for the basic NMF family methods, the geometry structure information discovered by the multiple centroids is only emphasized on the

basis matrix. While, for the image clustering task in this paper, the encoding matrix is directly used as a low dimension representation of the original data. So, the performance of the basic NMF family methods do not always behave well when n_s increases. And for the NMF methods with graph regularizer and multiple centroids, the manifold information is both considered in the basis matrix and the encoding matrix. Although multiple centroids can promote the performance of the proposed methods, the cost is that they need more computation because the dimensions of the learned feature are higher than other methods.

5. Conclusion

In this paper, a novel NMF framework called subspace clustering guided convex nonnegative matrix factorization (SC-CNMF) is proposed. In this framework, nonnegative subspace clustering term is used to digging out the multi-subspaces structure information of the data. To enforce the robustness of the proposed model, only s largest values are kept in each column of the self-representation matrix which will be used to constructed the similarity matrix in each iteration. This constraint can be applied in two different ways that lead to two implementations of the proposed model, i.e. SC-CNMF₁ and SC-CNMF₂, with two slightly different optimizing schemes. To improve the the clustering accuracy of the proposed model, one additional constraint is imposed on the encoding matrix to alleviate the relevance among the rows of the encoding matrix. At the same time, to enforce the representative ability in describing data with complicated geometry structure, multiple centroids are set for each cluster instead of single centroid. The two proposed methods are compared with several methods on six image datasets. And the experimental results reveal the effectiveness of the proposed methods.

Acknowledgments

This work was supported in part by the [National Natural Science Foundation of China](#) under Grants [61761130079](#) and [U1604153](#), in part by the Key Research Program of Frontier Sciences, CAS under Grant QYZDY-SSW-JSC044, in part by the Training Program for the Young-Backbone Teachers in Universities of Henan Province under Grant 2017GGJS065, and in part by the State Key Laboratory of Virtual Reality Technology and Systems under Grant BUAVR-16KF-04.

References

- [1] N. Zeng, Z. Wang, H. Zhang, W. Liu, F.E. Alsaadi, Deep belief networks for quantitative analysis of a gold immunochromatographic strip, *Cognit. Comput.* 8 (4) (2016) 684–692.
- [2] N. Zeng, H. Zhang, Y. Li, J. Liang, A.M. Dobaie, Denoising and deblurring gold immunochromatographic strip images via gradient projection algorithms, *Neurocomputing* 247 (2017) 165–172.
- [3] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, *Nature* 401 (6755) (1999) 788–791.
- [4] Y. Xiao, Z. Zhu, Y. Zhao, Y. Wei, S. Wei, X. Li, Topographic NMF for data representation, *IEEE Trans. Cybern.* 44 (10) (2014) 1762–1771.
- [5] Y. Dong, D. Tao, X. Li, Nonnegative multiresolution representation-based texture image classification, *ACM Trans. Intell. Syst. Technol.* 7 (1) (2015) 4:1–4:21.
- [6] X. Zhao, X. Li, Z. Zhang, C. Shen, Y. Zhuang, L. Gao, X. Li, Scalable linear visual feature learning via online parallel nonnegative matrix factorization, *IEEE Trans. Neural Netw. Learn. Syst.* 27 (12) (2016) 2628–2642.
- [7] X. Li, G. Cui, Y. Dong, Refined-graph regularization-based nonnegative matrix factorization, *ACM Trans. Intell. Syst. Technol.* 9 (1) (2017) 1:1–1:21.
- [8] Y. Lu, Z. Lai, Y. Xu, X. Li, D. Zhang, C. Yuan, Nonnegative discriminant matrix factorization, *IEEE Trans. Circuits Syst. Video Technol.* 27 (7) (2017) 1392–1405.
- [9] H. Gao, F. Nie, H. Huang, Local centroids structured non-negative matrix factorization, in: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, 2017, pp. 1905–1911.
- [10] P.O. Hoyer, Non-negative sparse coding, in: *Proceedings of the IEEE Workshop Neural Networks for Signal Processing*, IEEE, 2002, pp. 557–565.
- [11] S.Z. Li, X. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1, 2001, pp. 1–207.
- [12] D. Cai, X. He, J. Han, T.S. Huang, Graph regularized nonnegative matrix factorization for data representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (2011) 1548–1560.
- [13] Q. Gu, J. Zhou, Neighborhood preserving nonnegative matrix factorization, in: *Proceedings of the British Machine Vision Conference (BMVC)*, 2009, pp. 1–10.
- [14] W. Ren, G. Li, D. Tu, L. Jia, Nonnegative matrix factorization with regularizations, *IEEE J. Emerg. Sel. Topics Circuits Syst.* 4 (1) (2014) 153–164.
- [15] J. Huang, F. Nie, H. Huang, C. Ding, Robust manifold nonnegative matrix factorization, *ACM Trans. Knowl. Discov. Data* 8 (3) (2014) 11.
- [16] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, *IEEE Trans. Neural Netw.* 17 (3) (2006) 683–695.
- [17] S. An, J. Yoo, S. Choi, Manifold-respecting discriminant nonnegative matrix factorization, *Pattern Recognit. Lett.* 32 (6) (2011) 832–837.
- [18] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, *IEEE Trans. Image Process.* 20 (7) (2011) 2030–2048.
- [19] H. Liu, Z. Wu, X. Li, D. Cai, T.S. Huang, Constrained nonnegative matrix factorization for image representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 34 (7) (2012) 1299–1311.
- [20] X. Long, H. Lu, Y. Peng, W. Li, Graph regularized discriminative non-negative matrix factorization for face recognition, *Multimedia Tools Appl.* 72 (3) (2014) 2679–2699.
- [21] C.H.Q. Ding, T. Li, M.I. Jordan, Convex and semi-nonnegative matrix factorizations, *IEEE Trans. Pattern Anal. Mach. Intell.* 32 (1) (2010) 45–55.
- [22] T. Zhang, A. Szlam, Y. Wang, G. Lerman, Hybrid linear modeling via local best-fit flats, *Int. J. Comput. Vis.* 100 (3) (2012) 217–240.
- [23] E. Elhamifar, R. Vidal, Sparse subspace clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2009, pp. 2790–2797.
- [24] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: *Proceedings of the International Conference on Machine Learning (ICML)*, 2010, pp. 663–670.
- [25] X. Peng, Z. Yi, H. Tang, Robust subspace clustering via thresholding ridge regression, in: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2015, pp. 3827–3833.
- [26] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.
- [27] H. Gao, F. Nie, X. Li, H. Huang, Multi-view subspace clustering, in: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 4238–4246.
- [28] C. Elkan, Using the triangle inequality to accelerate k-means, in: *Proceedings of the 20th International Conference on Machine Learning (ICML)*, 3, Aug. 2003, pp. 147–153.
- [29] I. Jolliffe, *Principal Component Analysis*, Wiley Online Library, 2005.
- [30] D. Kong, C. Ding, H. Huang, Robust nonnegative matrix factorization using ℓ_{21} -norm, in: *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, ACM, 2011, pp. 673–682.
- [31] Z. Li, J. Tang, X. He, Robust structured nonnegative matrix factorization for image representation, *IEEE Trans. Neural Netw. Learn. Syst.* PP (99) (2017) 1–14.
- [32] H. Gao, F. Nie, W. Cai, H. Huang, Robust capped norm nonnegative matrix factorization: Capped norm nmf, in: *Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, ACM, 2015, pp. 871–880.
- [33] C.H. Papadimitriou, K. Steiglitz, *Combinatorial Optimization: Algorithms and Complexity*, Prentice-Hall, 1982.
- [34] A. Strehl, J. Ghosh, Cluster ensembles – a knowledge reuse framework for combining multiple partitions, *J. Mach. Learn. Res.* 3 (2002) 583–617.



Guosheng Cui is currently pursuing the Ph.D. degree with the Center for Optical Imagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China. His current research interests include artificial intelligence and computer vision.

Xuelong Li is a full professor with the Center for Optical Imagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.



Yongsheng Dong received his Ph.D. degree in applied mathematics from Peking University in 2012. He was a postdoctoral research fellow with the Center for Optical Imagery Analysis and Learning, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China from 2013 to 2016. From 2016 to 2017, he was a visiting research fellow at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently an associate professor with the School of Information Engineering, Henan University of Science and Technology. His current research interests include pattern recognition, machine learning, and computer vision.

He has authored and co-authored over 30 papers at famous journals and conferences, including IEEE TIP, IEEE TNNLS, IEEE TCYB, IEEE SPL and ACM TIST. He has served as a reviewer for over 30 international prestigious journals and conferences, such as IEEE TNNLS, IEEE TIP, IEEE TCYB, IEEE TIE, IEEE TSP, IEEE TKDE, IEEE TCDS and ACM TIST. He has also served as a Program Committee Member for more than 10 international conferences. He is a member of the IEEE, ACM and CCF.