Signal Processing 124 (2016) 103–114

Contents lists available at ScienceDirect

Signal Processing

journal homepage: www.elsevier.com/locate/sigpro

Local coordinate based graph-regularized NMF for image representation

Qing Liao, Qian Zhang*

Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong

ARTICLE INFO

Article history:
Received 1 May 2015
Received in revised form
6 September 2015
Accepted 30 September 2015
Available online 6 November 2015

Keywords:
Local coordinate
Graph regularization
Non-negative matrix factorization (NMF)

ABSTRACT

Non-negative matrix factorization (NMF) has been a powerful data representation tool which has been widely applied in pattern recognition and computer vision due to its simplicity and effectiveness. However, existing NMF methods suffer from one or both of the following deficiencies: (1) they cannot theoretically guarantee the decomposition results to be sparse, and (2) they completely neglect geometric structure of data, especially when some examples are heavily corrupted. In this paper, we propose a local coordinate based graph regularized NMF method (LCGNMF) to simultaneously overcome both deficiencies. In particular, LCGNMF enforces the learned coefficients to be sparse by incorporating the local coordinate constraint over both factors meanwhile preserving the geometric structure of the data by incorporating graph regularization. To enhance the robustness of NMF, LCGNMF removes the effect of the outliers via the maximum correntropy criterion (MCC). LCGNMF is difficult because the MCC induced objective function is neither quadratic nor convex. We therefore developed a multiplicative update rule to solve LCGNMF and theoretically proved its convergence. Experiments of image clustering on several popular image datasets verify the effectiveness of LCGNMF compared to the representative methods in quantities.

1. Introduction

Data representation learns the intrinsic structure of the data and reduces data redundancy to facilitate subsequent data analysis. It has played an important role in pattern recognition [1,2], computer vision [3,4], and biological tasks [5–7] due to its efficacy and efficiency. Recently, non-negative matrix factorization (NMF [8,9]) has been proven to be a powerful data representation method. It represents data matrix as the product of two lower-dimensional factors, i.e., the bases and coefficients of examples on these bases. Since NMF learns sparse representation, it has been successfully applied in computer vision.

Since NMF preserves the non-negativity property of practical data and learns intuitive interpretation consistent with human brain [5–7], it has been widely used in computer vision [10] and data mining[11]. However, NMF neglects the geometric structure proven beneficial for various vision tasks. Cai et al. [12] proposed graph regularized NMF (GNMF) which can preserve the geometric structure of dataset in the lower-dimensional space. Guan et al. [13] proposed a manifold regularized discriminative NMF (MD-NMF) by preserving both the neighborhood relationships and marginal maximization among examples. Shen and Si [14] developed the multiple manifold NMF method (MM-NMF) to model the intrinsic geometrical structure of data on multiple manifolds. Guan et al. [15] proposed a non-negative patch alignment framework (NPAF) to unify NMF, GNMF, MD-NMF, MM-NMF, and other related methods. However, both NMF and NPAF

* Corresponding author.
E-mail addresses: qnature@ust.hk (Q. Liao), qzhang@ust.hk (Q. Zhang).

http://dx.doi.org/10.1016/j.sigpro.2015.09.038
0165-1684/© 2015 Elsevier B.V. All rights reserved.

cannot guarantee any of decomposition results to be sparse in theory.

Many NMF methods have been developed by imposing sparseness constraints over the factors to overcome this deficiency. For instance, Hoyer et al. [16] proposed the sparse NMF method (SNMF) which explicitly incorporates sparseness constraints over both factors via the L_1-norm regularization. Li et al. also proposed the local NMF (LNMF [1]) which imposes localization constraint over basis to learn spatially localized, parts-based representation of visual patterns. Besides, Yuan et al. [17] developed the projective NMF (PNMF) to learn sparse representation by implicitly enforcing orthogonal constraint over the basis. However, PNMF cannot effectively learn the basis, and thus fails to reveal the grouping memberships of examples. Chen et al. [18] proposed the non-negative local coordinate factorization (NLCF) to induce sparse coefficients via the local coordinate constraint. But it often easily induces trivial basis. To overcome this deficiency, Liu et al. [19] developed the local coordinate concept factorization method (LCF) to learn sparse coefficients. Since LCF implicitly requires that the learned basis vector be close to several original data points, each data point can be approximated by a linear combination of as few basis vectors as possible. However, since these methods assume that data noises follow either Gaussian or Poisson distribution, they often fail in situation where some examples are heavily corrupted.

To address this issue, Zhang et al. [20] and Shen et al. [21] proposed the sparse robust NMF (SR-NMF) method to decompose the original matrix into sparse and low-rank components. The sparse component captures the outliers, and meanwhile the low-rank component models the intrinsic structure of the data. Kong et al. [22,23] proposed the $L_{2,1}$-NMF which penalizes the reconstruction with the $L_{2,1}$-norm. It has been claimed to be robust to the outliers. Besides, Du et al. [24] proposed the CIM-NMF method which employs the correntropy induced metric (CIM) to remove the effect of the outliers. It possesses the ability to handle the non-Gaussian noises. However, both traditional NMF and robust NMF methods cannot guarantee the decomposition results of NMF to be sparse in theory and do not consider the geometric structure of the datasets. However, they still obtain unsatisfactory results in clustering tasks because they neither guarantee the learned factors to be sparse, nor preserve the geometric structure, both of which have been proven beneficial for clustering tasks.

In this paper, we propose a local coordinate based NMF method with the signed graph regularization (LCGNMF) to overcome the above deficiencies. Particularly, LCGNMF enforces the learned coefficients to be sparse by using local coordinate constraint over both factors meanwhile preserving the geometric structure of the data by incorporating graph regularization. To further boost the robustness of NMF, LCGNMF removes the effect of the outliers via the maximum correntropy criterion (MCC). It is well-known that the MCC induced loss function is non-quadratic and NMF's objective function is non-convex, and thus it is difficult to optimize LCGNMF. In this paper, we developed a multiplicative update rule to optimize LCGNMF, and theoretically proved its convergence. Experiments of image clustering on several popular image datasets including Yale [25], Extended Yale B [26], UMIST [27] and ORL [28] datasets verify the effectiveness of LCGNMF compared to the representative methods in terms of average accuracy and normalized mutual information.

The rest of this paper is organized as follows: Section 2 briefly reviews related works on NMF and its variants. Section 3 proposes LCGNMF and optimizes it via the multiplicative update rule (MUR). Section 4 conducts experiments to evaluate the effectiveness of LCGNMF, and Section 5 concludes this paper.

2. Related works

This section briefly reviews most related works with LCGNMF, including non-negative matrix factorization (NMF [8,9]), its robust variants [22–24] and non-negative local coordinate factorization (NLCF [18]) .

2.1. Non-negative matrix factorization

Given any non-negative matrix $X \in R_{+}^{m\times n}$, whose rows correspond to examples and columns to features, NMF decomposes X into the product of two lower-dimensional non-negatives $U \in R_{+}^{m\times r}$ and $V \in R_{+}^{r\times n}$ by minimizing the distance between X and UV, i.e.,

$$\min_{U \in R_{+}^{m\times r}, V \in R_{+}^{r\times n}} \| X - UV \|_F^2, \tag{1}$$

where U and V denote two factor matrices. $\| \cdot \|_F$ denotes the Frobenius norm, and r denotes the reduced dimensionality which satisfies $r \lll \min\{m, n\}$. Frobenius norm can also be replaced by Kullback–Leibler divergence.

Although the objective (1) is not convex with both factors, it is lucky to yield the local minimum by the multiplicative update rules

$$U_{ij} = U_{ij}\frac{(XV^T)_{ij}}{(UVV^T)_{ij}}, \tag{2}$$

$$V_{ij} = V_{ij}\frac{(U^TX)_{ij}}{(U^TUV)_{ij}}. \tag{3}$$

However, NMF still obtains unsatisfactory results in clustering tasks because it cannot always guarantee the learned factors to be sparse in theory. To induce sparse coefficients, Chen et al. [18] proposed the non-negative local coordinate factorization method (NLCF) to learn effective data representation by imposing the local coordinate constraint over both factors as follows:

$$\min_{U \in R_{+}^{m\times r}, V \in R_{+}^{r\times n}} \|X - UV\|_F^2 + \lambda \sum_{i=1}^{n} \sum_{k=1}^{r} |v_{ki}| \|u_k - x_i\|^2, \tag{4}$$

where λ signifies the predefined constant, and r the reduced dimensionality. The local coordinate constraint penalizes the distance of any pair-wise far-away examples meanwhile encourages the similarity among close examples. However, since they assume that data noises follow the Gaussian or Poisson distribution, they often do not work well in this case where some examples are heavily corrupted.

2.2. Robust NMF

To address the above issue, Kong et al. [22,23] proposed $L_{2,1}$-NMF which replaces the loss function with the $L_{2,1}$ norm, i.e.,

$$\min_{U \in R_+^{m\times r}, V \in R_+^{r\times n}} \|X - UV\|_{2,1}^2, \tag{5}$$

where norm $\|H\|_{2,1} = \sum_{i=1}^{n} \sqrt{\sum_{j=1}^{m} H_{ji}^2}$. $L_{2,1}$-norm reduces the effect of large outliers such that the corrupted examples never dominate the objective. Thus, $L_{2,1}$-NMF is more robust than NMF. Besides, Du et al. [24] proposed CIM-NMF which can process non-Gaussian and impulsive noises due to the correntropy induced metric (CIM). The loss function of CIM-NMF can be written as

$$\min_{U,V \geq 0} \sum_{a=1}^{n} \sum_{b=1}^{m} \left(1 - g\left(X_{ab} - \sum_{k=1}^{r} U_{ak} V_{kb}, \sigma\right)\right), \tag{6}$$

where $g(e,\sigma)$ is the Gaussian kernel $g(e,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} \exp(-e^2/2\sigma^2)$. Since CIM-NMF decreases the weight of large outliers, it is robust to the outliers. But it still has enough room to boost NMF for clustering tasks due to neglecting both geometric structure and sparse constraints.

3. Local coordinate based graph-regularized NMF

To address the above issues, this section introduces a local coordinate based graph-regularized NMF method (LCGNMF) via maximum correntropy criterion. The maximum correntropy criterion is robust to the non-Gaussian large outliers because it is dominated by slightly corrupted examples. Particularly, LCGNMF enforces the learned coefficients to be sparse by using local coordinate constraint over both factors meanwhile preserving the geometric structure of the data by incorporating graph regularization.

3.1. The proposed model

3.1.1. Unsupervised signed graph

Manifold learning [29–32] is to uncover the underlying geometric structure within the dataset. It assumes that high-dimensional data lie on a lower-dimensional manifold. A variety of methods have been developed to address this issue by preserving the neighborhood relationships among examples. Their success implies the importance of preserving the geometric structure in visual applications. However, traditional graph models only consider the similarity among nearby examples but often neglects the dissimilarity among far-away examples if none of labels are available. To overcome this deficiency, this section deploys an unsupervised signed graph Laplacian to take both similarity and dissimilarity information of examples into consideration.

For clarity, we first briefly review traditional unsigned graph whose weights are always non-negative. Assume that a graph $G = (V, E)$ represents neighborhood relationships among all examples in $X \in R^{m\times n}$, where V and E denote vertexes and edges, respectively. Each vertex stands for one example while the edges represent the link relationships among vertexes, i.e., the edge weighting matrix. For a k-NN graph, the edge weighting matrix W results from $W_{ij} = \exp\|x_i - x_j\|^2/\sigma^2$ (σ is the kernel width), if vertex x_i is one of the k nearest neighbors of vertex x_j, otherwise $W_{ij} = 0$. Therefore, manifold learning methods [29,30] restrict that the lower-dimensional representations of examples share the identical neighborhood relationships among examples, i.e., the edge weighting matrix, in original high-dimensional data space, i.e.,

$$\sum_{i,j=1}^{n} W_{ij} \|v_i - v_j\|_2^2, \tag{7}$$

where v_i denotes the lower-dimensional representation of the ith example x_i. By simple algebra, (7) can be rewritten as the so-called the Laplacian regularization term:

$$\mathrm{tr}(VLV^T), \tag{8}$$

where $L = D - W$, where $D_{ii} = \sum_j W_{ij}$, for $i = 1, 2, \ldots n$.

Interestingly, recent works [33,34] show that the edge weights allowing negative values, i.e., the signed graph can better incorporate both similarity and dissimilarity information to enhance the discrimination performance. Different from previous works [33,34], we focus on deploying an unsupervised signed graph, especially when none of labels can be available. We encode the edge weights of vertexes by using the sparse coding technique [35–37]. Sparse coding requires representing any vector as the linear combination of as few atoms as possible, and has been widely applied in computer vision applications due to its efficacy. Given the ith example x_i, we take all examples as the atom matrix X, and calculate the corresponding coding z_i by solving

$$\min_{z_i} \|x_i - Xz_i\|_F^2 + \gamma\|z_i\|_1,$$
$$\text{s.t., } e_i^T z_i = 0 \tag{9}$$

where $\|\cdot\|_1$ denotes the l_1-norm, and γ a predefined regularization parameter, and e_i denotes a unit vector whose ith entry is 1. The above problem (9) is equivalent to the following form:

$$\min_{c} \|x_i - X_{-i}c\|_F^2 + \gamma\|c\|_1, \tag{10}$$

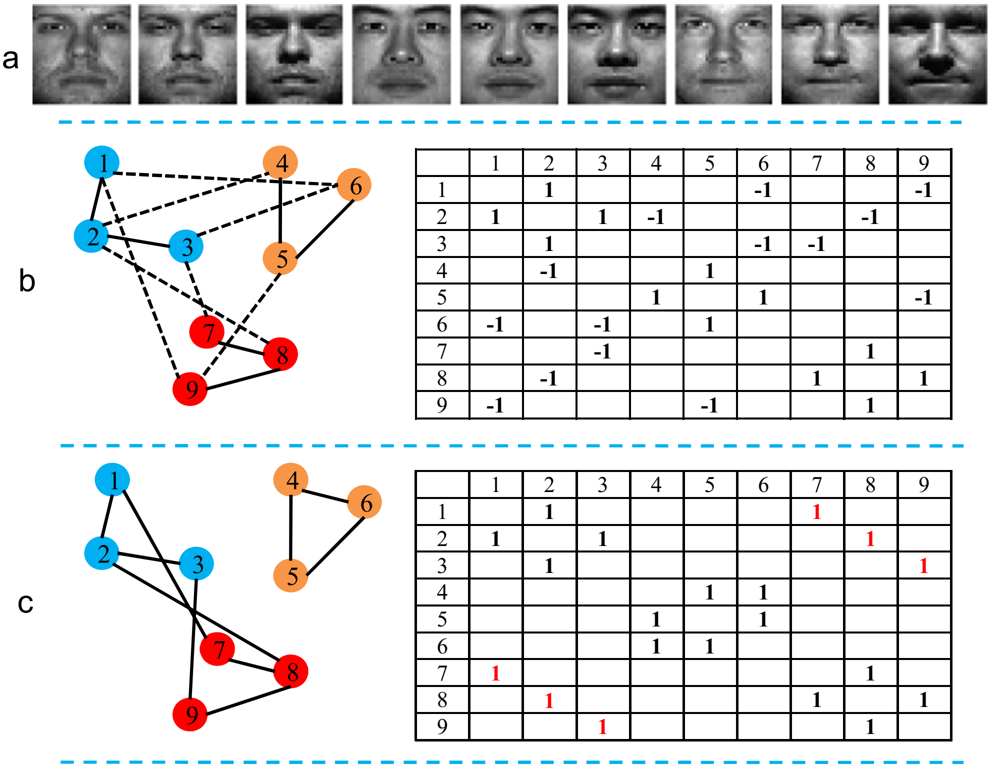
where X_{-i} denotes that X excludes the ith example x_i. For the above problem (10), many optimization algorithms such as LARS [38] and LeastR [39] have been developed to efficiently optimize it. By solving (10), we can obtain $z_i = [c_1, \ldots, c_{i-1}, 0, c_i, \ldots, c_{n-1}]$.

For all examples, we concatenate their coding vectors into a matrix S, and its entries have both positive and negative values. When all entries are positive values, it will degenerate to an unsigned graph. Therefore, the proposed signed graph can accommodate all situations in an adaptive manner. That is because coding vectors are driven by the dataset without manual settings. In addition, it can better reflect both similarity and dissimilarity relationships among examples. Fig. 1 verifies this point clearly. Fig. 1(a) shows nine images from three individuals which includes three images per individual. Fig. 1(b) and (c) demonstrates both signed and unsigned graphs, respectively, and their corresponding edge weighting matrices. When the edge weight is 1, it denotes the neighborhood relationships between two nodes. But when the edge weight is -1, it implies that two vertices cannot be neighborhood. Since the signed graph allows the edge weights to be negative, the constructed graph distinguishes the similar and dissimilar nodes very clearly. Fig. 1(b) shows that the proposed signed graph not only adaptively selects the neighborhood relationships among examples but also contains dissimilarity of different individuals. In Fig. 1(c), the unsigned graph is built based on k-NN with k=2. The unsigned k-NN graph neglects the dissimilarity information of examples and induces incorrect neighborhood relationships among examples from different individuals.

Thus, the signed graph is an effective way to encode the geometric structure of the dataset. Intrinsically, sparse coding vectors implicitly imply the neighborhood relationships among all examples. For the coding vector of a specific example, the greater values the positive entries have, the closer the examples corresponding to the position of the positive entries are to this example; meanwhile, the larger the negative entries are, the bigger their discrepancies are. Thus, the intuition induces the following definition of our signed edge weighting matrix:

$$\tilde{W}_{ij}=\begin{cases}1, & S_{ij}\ \ or\ \ S_{ji}\in \textit{maximal positive value}\\ 0, & \textit{otherwise}\\ -1, & S_{ij}\ \ or\ \ S_{ji}\in \textit{negative value}\end{cases} \tag{11}$$

where S_{ij} signifies the i-row j-column entry of S. Note that the maximal/minimal value indicates the maximum/minimum each column of S. After that, it is easy to yield the corresponding Laplacian matrix $\tilde{L}=\tilde{D}-\tilde{W}$, where $\tilde{D}_{ii}=\sum_j \tilde{W}_{ij}$. Thus, we can preserve the geometric structure of the dataset by incorporating the signed graph based



	1	2	3	4	5	6	7	8	9
1		1				-1			-1
2	1		1	-1				-1	
3		1				-1	-1		
4		-1			1				
5				1		1			-1
6	-1		-1		1				
7			-1					1	
8		-1					1		1
9	-1				-1			1	

	1	2	3	4	5	6	7	8	9
1		1					1		
2	1		1					1	
3		1							1
4					1	1			
5				1		1			
6				1	1				
7	1							1	
8		1					1		1
9			1					1	

Fig. 1. Examples of signed and unsigned graph, respectively. (a) Displays nine images of three individuals, i.e., three images each individual, and their sequence number corresponds to the nodes of (b) and (c). (b) and (c) show the signed and unsigned graphs (left) and the corresponding edge weight matrices (right), respectively. The negative edge weights in the graph are shown in dash line while the positive are in solid line. By comparing (b) and (c), some examples of different individuals in the signed graph take on the negative values without positive ones, while the unsigned k-NN graph contains the positive edges between examples of different individuals. Thus, signed graph can reflect similarity and dissimilarity information within the dataset.

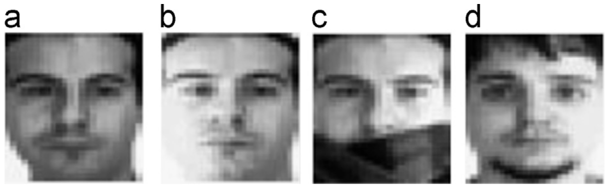


Fig. 2. Examples motivating the use of the CIM distance measure, including (a) the original image, (b) a second image of the same individual, (c) an occluded image of the identical individual, and (d) an image of another individual.

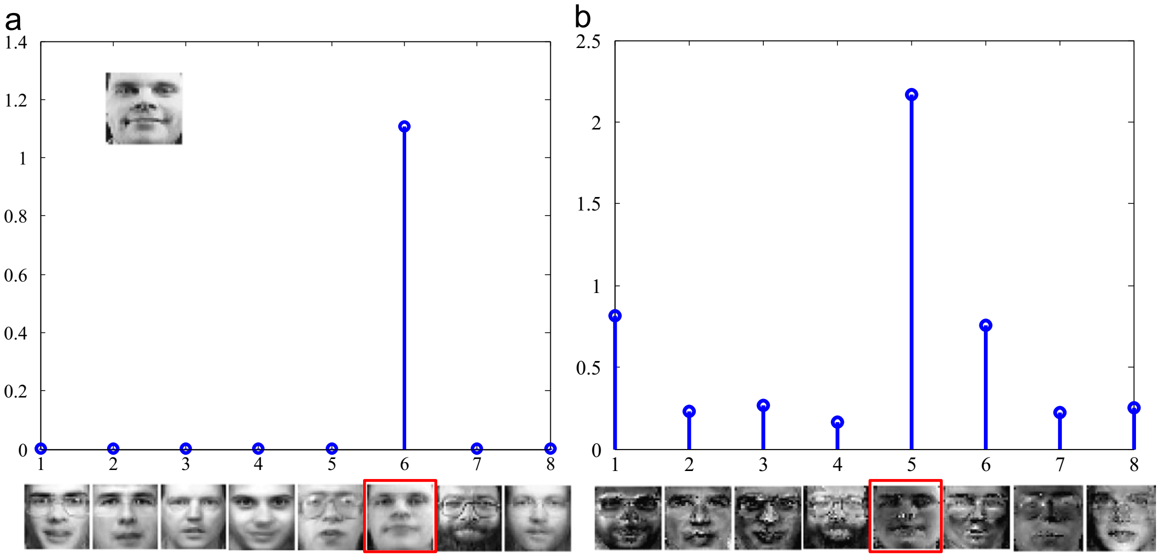


Fig. 3. Coefficients of an image learned by (a) LCGNMF and (b) CIM-NMF. (a) Displays an illustrative image and its learned coefficient and basis by LCGNMF, and meanwhile the learned coefficient and basis by CIM-NMF are shown in (b). By contrast, LCGNMF easily induces sparse coefficient and this property is beneficial for identifying cluster identity in terms of the bounding box.

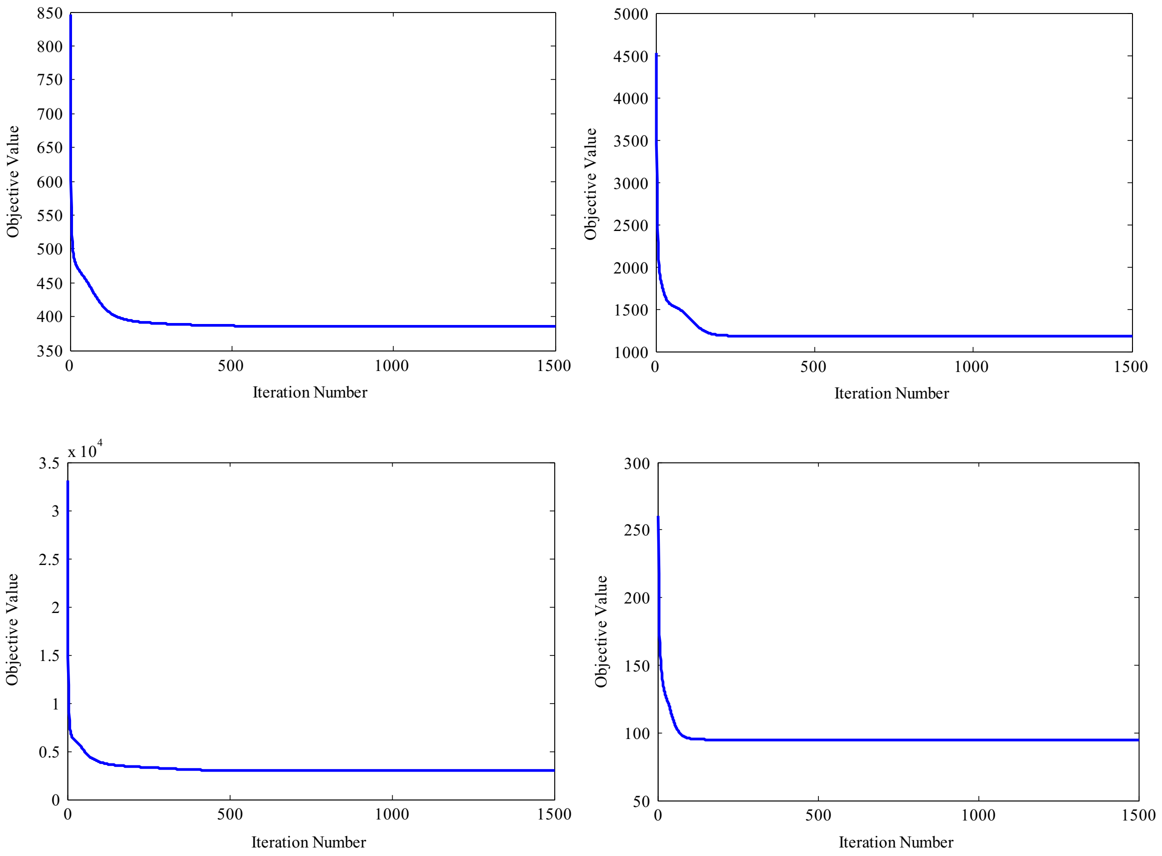


Fig. 4. Convergent curves of LCGNMF on (a) Yale, (b) YaleB, (c) UMIST and (d) ORL datasets.

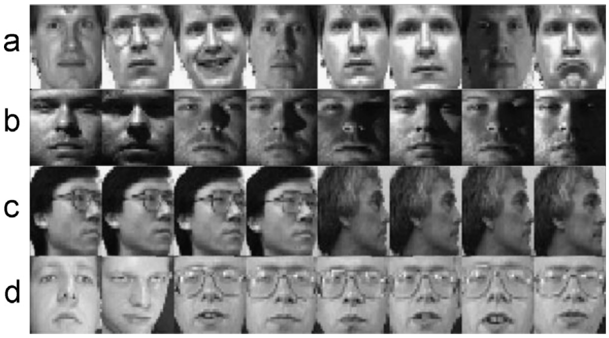


Fig. 5. Image instances of (a) Yale, (b) YaleB, (c) UMIST, and (d) ORL datasets.

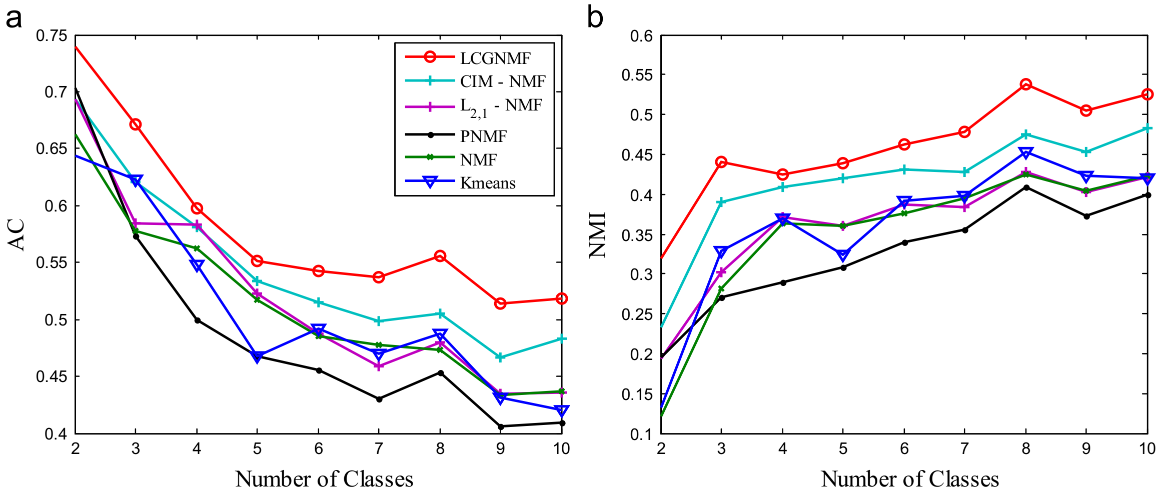


Fig. 6. Average accuracy (AC) and normalized mutual information (NMI) versus different numbers of classes on the Yale datasets.

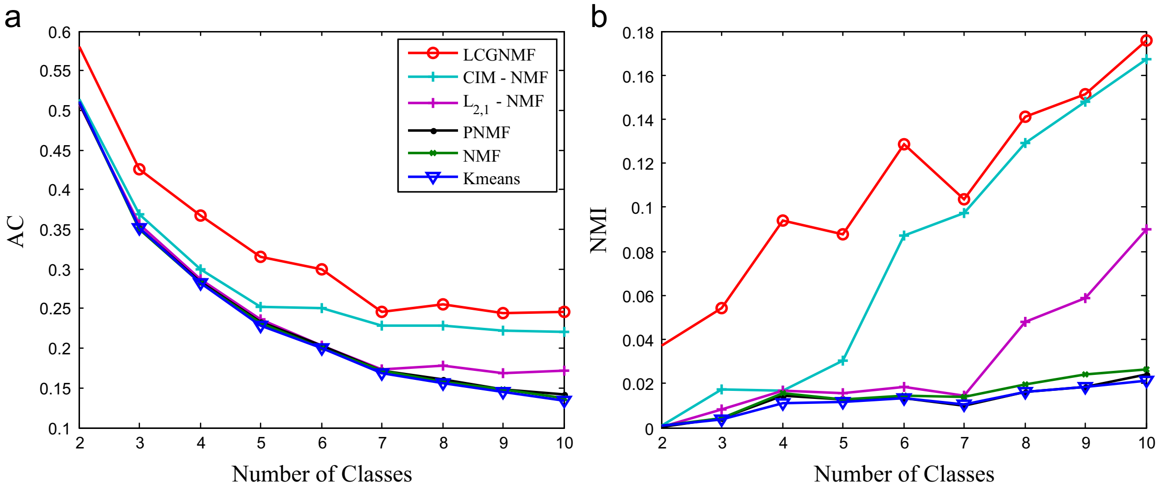


Fig. 7. Average accuracy (AC) and normalized mutual information (NMI) versus different numbers of classes on the Extended Yale B dataset.

manifold regularization term as follows:

$$\min_{U,V \geq 0} \|X - UV\|_F^2 + \alpha\, tr(V\tilde{L}V^T), \tag{12}$$

where α denotes the regularization parameter.

3.1.2. CIM-based local coordinate constraint

Local coordinate technique [18,19,40,1] is investigated to induce sparse coding. In practice, since the basis also contains the data noises, the sparsity may be easily ruined. Recent theoretical works [24] claim that CIM can remove

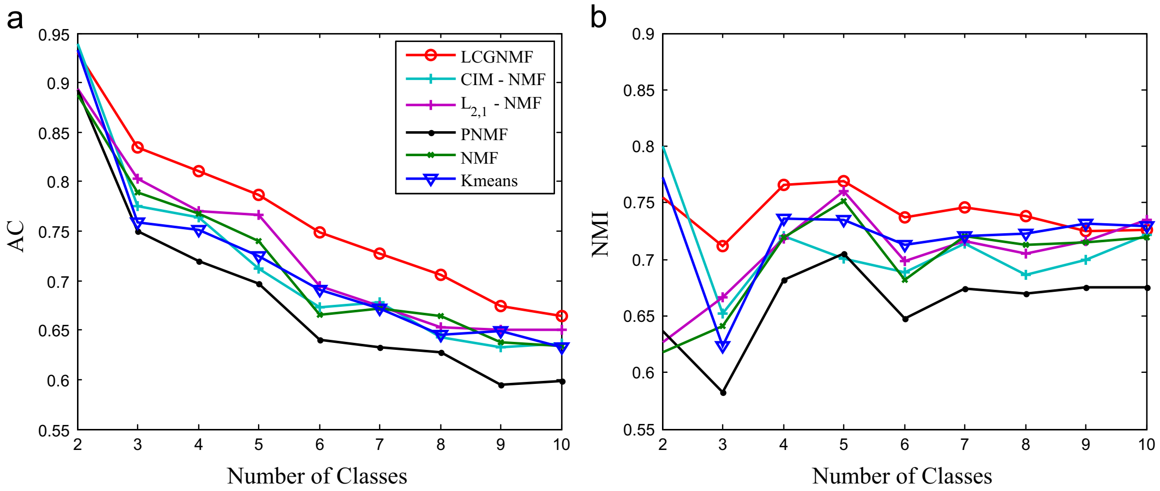


Fig. 8. Average accuracy (AC) and normalized mutual information (NMI) versus different numbers of classes on the UMIST datasets.

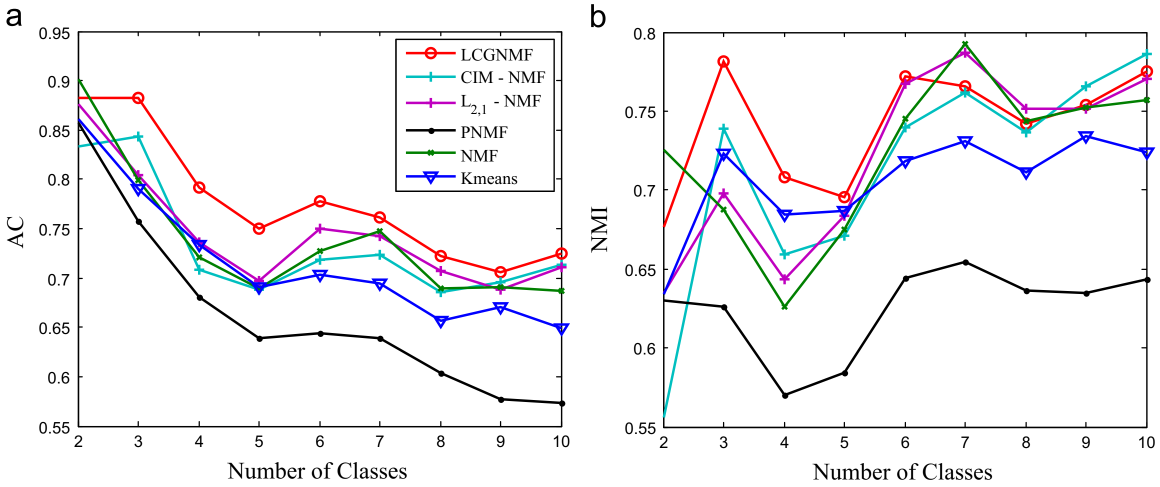


Fig. 9. Average accuracy (AC) and normalized mutual information (NMI) versus different numbers of classes on the ORL datasets.

the effect of large outliers. Thus, we propose a CIM-based local coordinate constraint to address the above issue. Both Fig. 2 and Table A1 show the examples' distances measured by l_2-norm and CIM metric, respectively, on real dataset. Obviously, CIM can mitigate the effect of noises over coefficients and reduce the risk of incorrectly distinguishing different subjects.

Motivated by the above observation, we can rewrite the traditional local coordinate as the following form:

$$\sum_{a=1}^{n}\sum_{b=1}^{r} V_{ba}(1-g(U_{\cdot b}-X_{\cdot a},\sigma_1)), \tag{13}$$

where $\sigma_1 = \frac{1}{nr}\sum_{a=1}^{n}\sum_{b=1}^{r}\left\|U_{\cdot b}-X_{\cdot a}\right\|_2^2$ which signifies the bandwidth of the Gaussian kernel. By substituting (13) into (12), we can obtain

$$\min_{U,V\geq 0}\|X-UV\|_F^2+\alpha\ \mathrm{tr}(V\tilde{L}V^T)+\beta\sum_{a=1}^{n}\sum_{b=1}^{r} V_{ba}(1-g(U_{\cdot b}-X_{\cdot a},\sigma_1)), \tag{14}$$

where β stands for the predefined regularization parameter. Actually, the CIM-based coordinate constraint still induces the sparse coefficients. This point can be verified in Fig. 3.

Although objective (14) can maintain the data locality and relieve the effect of data noises to coefficients, it often fails in practical situations where some examples are heavily corrupted. This is because data noises do not follow the Gaussian or Poisson distribution. To address this issue, we also employ the correntropy induced metric (CIM [24]) to measure the reconstruction error. Thus, the final objective of LCGNMF becomes

$$\begin{aligned}\min_{U,V\geq 0} J(U,V) = &\sum_{a=1}^{m}\sum_{b=1}^{n}(1-g(X_{ab}\\ &-\sum_{k=1}^{r}U_{ak}V_{kb},\sigma))+\alpha\ \mathrm{tr}(V\tilde{L}V^T)\\ &+\beta\sum_{a=1}^{n}\sum_{b=1}^{r}V_{ba}(1-g(U_{\cdot b}-X_{\cdot a},\sigma_1)).\end{aligned} \tag{15}$$

Table A1
Comparison of normalized distance measures based on L_2-norm and CIM.

Normalized distance measurement	L_2-norm	CIM
Same individual	0.0632	0.3722
Same individual with occlusions	0.0747	0.3435
Different individuals	0.0353	0.4010

According to [24], we empirically set $\sigma = \frac{1}{2mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(X_{ij}-\sum_{k=1}^{r}U_{ik}V_{jk})^2$ in our experiments. The proposed LCGNMF greatly differs from both $L_{2,1}$-NMF and CIM-NMF and has the following distinct aspects: (1) LCGNMF takes the geometric structure of the data to boost performance of image clustering, and (2) based on the robustness of CIM, LCGNMF adopts a novel CIM-based local coordinate constraint to reduce the effect of noises to coefficients. Both of them can significantly boost NMF in image clustering. This point has been verified in the Experiments section.

3.2. Optimization algorithm

The function (15) is not jointly convex over U and V, and thus it is impossible to obtain the global solution. Fortunately, it is convex with respect to U with V fixed, and vice versa. Thus we developed a multiplicative update rule (MUR) to optimize LCGNMF by alternatively updating both factors.

Similar to [8,9], the optimization problem of LCGNMF equals to minimizing the following augmented objective function in an enlarged parameter space

$$\begin{aligned}\min_{U,V\ge 0} J(U,V) = &\sum_{a=1}^{m}\sum_{b=1}^{n} W_{ab}\left(X_{ab}-\sum_{k}U_{ak}V_{kb}\right)^2\\ &+\phi(W_{ab})+\beta\sum_{a=1}^{n}\sum_{b=1}^{r}V_{ba}\left\|(A^{ab})^{1/2}\otimes(U_{\cdot b}-X_{\cdot a})\right\|_2^2\\ &+\varphi(A^{ab})+\alpha\,\mathrm{tr}(V\tilde{L}V^T),\end{aligned} \tag{16}$$

where $A^{ab}\in R^{m\times 1}$ is a vector calculated by $X_{\cdot a}$ and $U_{\cdot b}$. By simple algebra, we can obtain the following two forms:

$$\begin{aligned}\min_{U,V\ge 0} J(U,V) = &\sum_{a=1}^{m}\sum_{b=1}^{n} W_{ab}\left(X_{ab}-\sum_{k=1}^{r}U_{ak}V_{kb}\right)^2\\ &+\phi(W_{ab})+\beta\sum_{a=1}^{n}\left\|(L^a\otimes(X_{\cdot a}\mathbf{1}^T-U))\Lambda_a^{1/2}\right\|_2^2\\ &+\varphi(L^a)+\alpha\,\mathrm{tr}(V\tilde{L}V^T),\end{aligned} \tag{17}$$

and

$$\begin{aligned}\min_{U,V\ge 0} J(U,V) = &\sum_{a=1}^{m}\sum_{b=1}^{n} W_{ab}\left(X_{ab}-\sum_{k=1}^{r}U_{ak}V_{kb}\right)^2\\ &+\phi(W_{ab})+\beta\sum_{b=1}^{r}\left\|(P^b\otimes(X-U_{\cdot b}\mathbf{1}^T))\Pi_b^{1/2}\right\|_2^2\\ &+\varphi(P^b)+\alpha\mathrm{tr}(V\tilde{L}V^T),\end{aligned} \tag{18}$$

where $\Lambda_a\in R^{r\times r}$ and $\Pi_b\in R^{n\times n}$ are the diagonal matrix of $V_{\cdot a}$ and $V_{b\cdot}$, respectively, and $L^a=[(A^{a1})^{1/2},\dots,(A^{ar})^{1/2}]\in R^{m\times r}$ and $P^b=[(A^{1b})^{1/2},\dots,(A^{nb})^{1/2}]\in R^{m\times n}$. Thus, both L^a and P^b can easily result from A^{ab}. In addition, $\phi(W_{ab})$ and $\phi(A^{ab})$ denote the conjugate function of $\sum_{a=1}^{n}\sum_{b=1}^{m}(1-g(X_{ab}-\sum_{k=1}^{r}U_{ak}V_{kb},\sigma))$ and $\sum_{a=1}^{n}\sum_{b=1}^{r}V_{ba}(1-g(U_{\cdot b}-X_{\cdot a},\sigma_1))$, respectively. Both W_{ab} and A^{ab} indicate the corresponding auxiliary variables. The reason why we deduce the above two forms, i.e., ((17) and (18)), is to reduce computation overhead of optimization algorithm. We optimize (16) with respect to one variable with the other fixed as follows: *Computation of W*: When U and V are fixed, the optimization problem with respect to W_{ab} is

$$W_{ab}=\exp\left(-\frac{(X_{ab}-\sum_{k=1}^{r}U_{ak}V_{kb})^2}{2\sigma^2}\right). \tag{19}$$

Computation of A: When U are fixed, the optimization value for A^{ab} is

$$A_i^{ab}=\exp\left(-\frac{(X_{ia}-U_{ib})^2}{2\sigma_1^2}\right). \tag{20}$$

Computation of U: Given V, we can yield the derivative of (17) with respect to U

$$\Delta U=-(W\otimes X)V^T+(W\otimes(UV))V^T+\alpha\sum_{a=1}^{n}((L^a\otimes(U-X_{\cdot a}\mathbf{1}^T))\Lambda_a)\otimes L^a. \tag{21}$$

By setting the derivative of U to zero, we obtain

$$U=U\otimes\frac{(W\otimes X)V^T+\alpha\sum_{a=1}^{n}((L^a\otimes X_{\cdot a}\mathbf{1}^T)\Lambda_a)\otimes L^a}{(W\otimes(UV))V^T+\alpha\sum_{a=1}^{n}((L^a\otimes U)\Lambda_a)\otimes L^a}, \tag{22}$$

where $\otimes$ is the element-wise multiplication. *Computation of V*: Given U, we can yield the derivative of (18) with respect to V

$$\begin{aligned}\Delta V_{j\cdot}=&(-2U^T(W\otimes X)+2U^T(W\otimes(UV)))_{j\cdot}\\ &+2\beta(V(\tilde{D}-\tilde{W}))_{j\cdot}+\alpha(Diag(((X-U_{\cdot j}\mathbf{1}^T)\otimes P^{jT})(P^j\otimes(X-U_{\cdot j}\mathbf{1}^T))),\end{aligned} \tag{23}$$

where the operator $Diag(\cdot)$ selects the diagonal elements of specific square matrix as a row vector.

By setting the derivative of V to zero, we obtain:

$$V_{j\cdot}=V_{j\cdot}\otimes\frac{(2U^T(W\otimes X)_{j\cdot}+2\alpha Diag(C_j^TD_j)+2\beta(V\tilde{W})_{j\cdot}}{(2U^T(W\otimes(UV))_{j\cdot}+\alpha(Diag(C_j^TC_j)+Diag(D_j^TD_j))+2\beta(V\tilde{D})_{j\cdot}}, \tag{24}$$

where $C_j=P^j\otimes X$, $D_j=P^j\otimes U_{j\cdot}\mathbf{1}^T$, and $\mathbf{1}^T\in R^{1\times n}$. For clarity, we summarize the optimization procedure of LCGNMF into Algorithm 1. In addition, we can obtain the following theorem:

Theorem 1. *The objective function in (16) is non-increasing under the update rules in (22) and (24). The objective function is invariant under these updates if and only if U and V are at a stationary point.*

We leave the proof of Theorem 1 in Appendix A. As shown in Fig. 4, the iteration curves of LCGNMF on Yale [25], Extended Yale B [26], UMIST [27], and ORL [28] datasets remain convergent after the 1000th iteration round. To check the stationary of the solution, we utilize the stopping criterion as follows:

$$\frac{\left|J(U^t,V^t)-J(U^{t-1},V^{t-1})\right|}{\left|J(U^t,V^t)-J(U^0,V^0)\right|}\le\varepsilon, \tag{25}$$

where t denotes the iteration round, and the tolerance ε is set to 10^{-5} empirically in our experiments.

Algorithm 1. MUR for LCGNMF.

```
Input:   Example $X \in R^{m \times n}$, number of cluster $r$, parameters $\alpha$, $\beta$.
Output:  $U$ and $V$.
1:       Initialize: $U, V$ with random initialization.
2:       Construct signed Laplacian matrix $\tilde{W}$ via (11)
3:       repeat
4:         Update $W$ via (19).
5:         Update $U$ via (22).
6:         Update $A$ via (20).
7:         Update $V$ via (24).
8:       until {stopping criterion (25) is satisfied.}
```

The major computation cost of Algorithm 1 lies in Steps 4–7. The time complexity of Step 4 is $O(mn)$ and Step 5 takes $O(mnr+mn+mr+mr^2)$ in time. Both Steps 6 and 7 take time of $O(m^2n)$ and $O(mnr+rn^2)$, respectively. Therefore, the total time complexity of Algorithm 1 is $O(mn+mnr+mr\ +mr^2 +rn^2)$.

4. Experiments

This section verifies the effectiveness by comparing the clustering performance of LCGNMF with the representative methods including NMF [8,9], $L_{2,1}$-NMF [22], PNMF [17], CIM-NMF [24] and Kmeans on four popular datasets including Yale [25], Extended Yale B [26], UMIST [27], and ORL [28]. The image instances of these datasets are shown in Fig. 5. In clustering tasks, we adopt the raw pixels to learn the coefficients of examples and we choose K-means to cluster the coefficients. The number of clusters equals to the number of selected subjects. Besides, we do not distinguish the training set and testing set yet we randomly collect the images from specific numbers $(2 \sim 10)$ of individuals as the training set meanwhile validate the clustering performance on this dataset. For fair comparison, each experiment was independently conducted 10 times and then we validate the clustering performance in terms of average accuracy (AC) and normalized mutual information (NMI).

4.1. Yale dataset

The Yale face image dataset [25] contains 165 frontal view images from 15 individuals. And each individual has 11 different images under various facial expressions and lighting conditions. All images are cropped to 32 × 32-pixel grayscale images and we reshape them into a 1024-dimensional vector. We set the parameters $\alpha = 0.1$ and $\beta = 0.1$ for LCGNMF on this dataset. The compared methods involve no parameters.

Fig. 6 shows that LCGNMF consistently outperforms the compared methods in terms of clustering accuracy and normalized mutual information. This is because LCGNMF can learn effective cluster centroids and the learned coefficients can indicate the true cluster identity to some extent. Besides, the signed graph regularization contains discrimination information and thus can significantly boost the clustering performance.

4.2. Extended Yale B dataset

The Extended Yale B face image dataset [26] is an extension of the Yale dataset. By contrast with the Yale dataset, it is tougher to perform clustering tasks on Extended Yale B dataset. There are totally 2424 face images of 38 individuals. Each individual has 59 images at least and 64 images at most under different illumination conditions. All images are still cropped to 32 × 32-pixel gray scale images and we reshape them into a 1024-dimensional vector. We set the parameters $\alpha = 0.1$ and $\beta = 1$ for LCGNMF on this dataset. The compared methods involve none of parameters.

Fig. 7 reports the clustering accuracy and normalized mutual information of the compared methods on the Extended Yale B dataset. This also implies that LCGNMF is superior to the representative methods in quantities. However, since this dataset involves drastic illumination variations, the learned bases by both the compared methods and LCGNMF still contain the noises. Thus, it is inevitable to obtain the relative lower performance compared with the other datasets.

4.3. UMIST dataset

The UMIST face image dataset [27] contains 575 frontal view images from 20 individuals. And each individual has no less than 41 images which vary in poses. All images are cropped to 40 × 40 pixel gray scale images, and we reshape them into 1600-dimensional vector. We set the parameters $\alpha = 0.1$ and $\beta = 0.1$ for LCGNMF on this dataset. The compared methods involve none of parameters.

Fig. 8 shows the clustering accuracy and normalized mutual information of all the methods on the UMIST dataset. The results imply that LCGNMF outperforms other methods under different class numbers. The improvement in clustering accuracy can be attributed to the signed graph regularization and local coordinate constraint. The former not only preserves the geometric structure of the dataset but also considers discrimination information into the learned basis, while the latter induces sparse coefficients to be as sparse as the cluster indicator vectors.

4.4. ORL dataset

The ORL face image dataset [28] contains 400 frontal view images from 40 individuals. Each individual has 10 images which vary in lighting, poses and facial expressions. All images are cropped to 32 × 32 pixel gray scale images and we reshape them into 1024-dimensional vector. We set the parameters $\alpha = 0.5$ and $\beta = 0.1$ for LCGNMF on this dataset. The compared methods involve none of parameters.

Fig. 9 shows the clustering accuracy and normalized mutual information of all the methods on the ORL dataset. In terms of the normalized mutual information, LCGNMF is comparable to both $L_{2,1}$-NMF and CIM-NMF. However, the results imply that LCGNMF outperforms the other methods in terms of clustering accuracy.

Considering the high effectiveness of LCGNMF, we are expected to extend this model to incorporate more information such as advanced Fisher's discriminant analysis techniques including geometric mean [41] and tensor discriminant analysis [42] in our future works.

5. Conclusion

This paper proposes a local coordinate based graph-regularized NMF (LCGNMF) to induce the sparse coefficients and consider the geometric structure of data space under the real noise datasets. Benefiting from these effective strategies, LCGNMF enhances the representation ability of NMF. Besides, LCGNMF utilizes the correntropy induced metric as the loss function to remove the effect of the outliers. To optimize LCGNMF, we developed a multiplicative update rule and proved its convergence. Experimental results of image clustering on four popular face datasets verify the effectiveness of LCGNMF in quantities.

Acknowledgment

The research was supported in part by grants from 973 project 2013C B329006, China NSFC under Grant 61173156, RGC under the contracts CERG 622613, 16212714, HKUST6/CRF/12R, and M-HKUST609/13, as well as the grant from Huawei-HKUST joint lab.

Appendix A. Convergence analysis

This section utilizes the auxiliary function to prove the convergence of Algorithm 1. We first give the following definition:

Definition 1. $G(x,x')$ is an auxiliary function of $F(x)$ if the following conditions:

$$G(x,x') \geq F(x), \quad G(x,x') = F(x) \tag{A.1}$$

are satisfied.

Based on Definition 1 and (16), we have the following observations.

Lemma 1. *If G is an auxiliary function of F, then F is nonincreasing under the update*:

$$x^{t+1} = \arg\min_x G(x,x'). \tag{A.2}$$

Proof. $F(u^{(t+1)}) \leq G(u^{(t+1)},u^{(t)}) \leq G(u^{(t)},u^{(t)}) = F(u^{(t)})$. Thus, this completes the proof.□

Let F_{ab} denotes the objective with respect to U with V fixed

$$F'_{ab} = \frac{\partial F_{ab}}{\partial u_{ab}} = -2\left((W \otimes X)V^T\right)_{ab} + 2\left(W \otimes (UV)V^T\right)_{ab} + 2\alpha \sum_{i=1}^{n} (((L^i \otimes (U - X_{.i}\mathbf{1}^T))\Lambda_i) \otimes L^i)_{ab} \tag{A.3}$$

$$F''_{ab} = 2\left(W(V \otimes V)^T\right)_{ab} + 2\alpha \sum_{i=1}^{n} (L^i \otimes L^i)_{ab}(\Lambda_i)_{bb} \tag{A.4}$$

Lemma 2. *The function*:

$$G\left(u,u_{ab}^{(t)}\right) = F_{ab}\left(u_{ab}^{(t)}\right) + F'_{ab}\left(u_{ab}^{(t)}\right)\left(u - u_{ab}^{(t)}\right) + \frac{(W \otimes (UV)V^T)_{ab} + \alpha \sum_{i=1}^{n} (((L^i \otimes U)\Lambda_i) \otimes L^i)_{ab}}{u_{ab}^{(t)}}(u - u_{ab}^{(t)})^2 \tag{A.5}$$

is an auxiliary function for F_{ab}.

Proof. Since $G(u,u) = F_{ab}(u)$ is obvious, we only need to show that $G(u,u_{ab}^{(t)}) \geq F_{ab}(u)$. To do this, we compare the Taylor series expansion of $F_{ab}(u)$:

$$F_{ab}(u) = F_{ab}(u_{ab}^t) + F'_{ab}(u - u_{ab}^t) + ((W(V \otimes V)^T)_{ab} + \alpha \sum_{i=1}^{n} (L^i \otimes L^i)_{ab}(\Lambda_i)_{bb})(u - u_{ab}^t)^2. \tag{A.6}$$

According to (A.5), we find that the inequality $G(u,u_{ab}^t) \geq F_{ab}(u)$ is equivalent to:

$$(W \otimes (UV(V^T)_{ab} + \alpha \sum_{i=1}^{n} (((L^i \otimes U)\Lambda_i) \otimes L^i)_{ab} \geq ((W(V \otimes V)^T)_{ab} + \alpha \sum_{i=1}^{n} (L^i \otimes L^i)_{ab}(\Lambda_i)_{bb})u_{ab}^t. \tag{A.7}$$

By simple algebra, we obtain

$$(W \otimes (UV)V^T)_{ab} = \sum_k w_{ak}(UV)_{ak}v_{bk} \geq \sum_k w_{ak}v_{bk}v_{bk}u_{ab}^t \geq (W(V \otimes V)^T)_{ab}u_{ab}^t, \tag{A.8}$$

and

$$\sum_{i=1}^{n} (((L^i \otimes U)\Lambda_i) \otimes L^i)_{ab} = \sum_{i=1}^{n} (L^i \otimes L^i)_{ab}(\Lambda_i)_{bb}u_{ab}^t. \tag{A.9}$$

Thus, the inequality (A.8) holds and thus $G(u,u_{ab}^t) \geq F_{ab}(u)$. Let H_{ab} denote the objective with respect to V with U fixed,

$$H'_{ab} = \frac{\partial H_{ab}}{\partial v_{ab}} = -2(U^T(W \otimes X))_{ab} + 2(U^T(W \otimes (UV)))_{ab} + \alpha(Diag((X - U_{.a}\mathbf{1}^T) \otimes P^{a^T}P^a \otimes (X - U_{.a}\mathbf{1}^T)))_b + 2\beta(V(\tilde{D} - \tilde{W}))_{ab} \tag{A.10}$$

and

$$H''_{ab} = 2((U \otimes U)^T W)_{aa} + 2\beta\tilde{L}_{bb}.\square \tag{A.11}$$

Lemma 3. *The function:*

$$G\left(v,v_{ab}^{(t)}\right) = H_{ab}\left(v_{ab}^{(t)}\right) + H'_{ab}\left(v_{ab}^{(t)}\right)\left(v - v_{ab}^{(t)}\right) + \frac{\left(U^T(W \otimes (UV))\right)_{ab} + \alpha Diag((X \otimes P^a)^T(P^a \otimes X))_b + 2\beta(V\tilde{D})_{ab}}{v_{ab}^{(t)}}(v - v_{ab}^{(t)})^2 \tag{A.12}$$

is an auxiliary function for H_{ab}.

Proof. Since $G(v,v) = H_{ab}(v)$ is obvious, we only need to show that $G(v,v_{ab}^t) \geq H_{ab}(v)$. To do this, we calculate the

Taylor series expansion of $H_{ab}(v)$:

$$H_{ab}(v) = H_{ab}(v^t_{ab}) + H'_{ab}(v - v^t_{ab}) + (((U \otimes U)^T W)_{aa} + 2\beta \tilde{L}_{bb})(v - v^t_{ab})^2 \tag{A.13}$$

According to (A.12), we find that the inequality $G(v, v^t_{ab}) \geq H_{ab}(v)$ equals to

$$(U^T(W \otimes (UV)))_{ab} + \alpha Diag((X \otimes P^a)^T (P^a \otimes X))_b + 2\beta(V\tilde{D})_{ab} \geq (((U \otimes U)^T W)_{aa} + 2\beta \tilde{L}_{bb}) v^t_{ab}. \tag{A.14}$$

By simple algebra

$$\begin{aligned}(U^T(W \otimes (UV)))_{ab} &= \sum_k u_{ka} w_{kb} (UV)_{kb} \\ &\geq \sum_k u_{ka} u_{ka} w_{ak} v^t_{ab} \\ &\geq ((U \otimes U)^T W + 2\lambda_2 L)_{ab} v^t_{ab}\end{aligned} \tag{A.15}$$

and

$$(V\tilde{D})_{ab} = \sum_k v^t_{ak} \tilde{D}_{kb} \geq v^t_{ak} \tilde{D}_{bb} \geq (\tilde{D} - \tilde{W})_{bb} v^t_{ak} = L_{bb} v^t_{ak}. \tag{A.16}$$

Thus, the objective (A.7) holds and $G(v, v^t_{ab}) \geq H_{ab}(v)$. This completes the proof.□

According to the above lemmas, we prove the convergence of Theorem 1. We will show that the objective function (16) of LCGNMF is bounded from below and non-increasing under the update steps in (22) and (24). Since the objective function (16) is greater than zero, we only need to verify that the objective function (16) is non-increasing under the updates steps in (22) and (24). Our proof will make use of an auxiliary function similar to that used in the Expectation-Maximization algorithm [43].

Proof. We substitute $G(u, u^t_{ab})$ of (A.2) into (A.5) and $G(v, v^t_{ab})$ of (A.2) into (A.12) to obtain the following update rules:

$$\begin{aligned}u^{t+1}_{ab} &= u^t_{ab} - u^t_{ab} \frac{F'_{ab}(u^t_{ab})}{(2W \otimes (UV)V^T + 2\alpha \sum_{a=1}^{n} ((L^a \otimes U)\Lambda_a) \otimes L^a)_{ab}} \\ &= u^t_{ab} \frac{(W \otimes XV^T + \alpha \sum_{a=1}^{n} ((L^a \otimes X_{\cdot a} \mathbf{1}^T)\Lambda_a) \otimes L^a)_{ab}}{(W \otimes (UV)V^T + \alpha \sum_{a=1}^{n} ((L^a \otimes U)\Lambda_a) \otimes L^a)_{ab}}\end{aligned} \tag{A.17}$$

and

$$\begin{aligned}v^t_{ab} &= v^t_{ab} - v^t_{ab} \frac{H'_{ab}(v^t_{ab})}{(2U^T(W \otimes (UV)))_{ab} + \alpha + \alpha(Diag(C_a^T C_a) + Diag(D_a^T D_a))_b + 2\beta(V\tilde{D})_{ab}} \\ &= \frac{(2U^T(W \otimes X))_{ab} + 2\alpha(Diag(C_a^T D_a)) + 2\beta(V\tilde{W})_{ab}}{(2U^T(W \otimes (UV)))_{ab} + \alpha(Diag(C_a^T C_a) + Diag(D_a^T D_a))_b + 2\beta(V\tilde{D})_{ab}}\end{aligned} \tag{A.18}$$

Since (A.5) and (A.12) are auxiliary function of (A.6) and (A.13), respectively, F_{ab} and H_{ab} are non-increasing using the update rules (22) and (24), respectively. This completes the proof.□

References

[1] S.Z. Li, X.W. Hou, H. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, IEEE, 2001, pp. I–207.

[2] A. Pascual-Montano, J.M. Carazo, K. Kochi, D. Lehmann, R.D. Pascual-Marqui, Nonsmooth nonnegative matrix factorization (nsnmf), IEEE Trans. Pattern Anal. Mach. Intell. 28 (3) (2006) 403–415.

[3] 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.

[4] D. Tao, X. Tang, X. Li, X. Wu, Asymmetric bagging and random subspace for support vector machines based relevance feedback in image retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 28 (7) (2006) 1088–1099.

[5] S.E. Palmer, Hierarchical structure in perceptual representation, Cogn. Psychol. 9 (4) (1977) 441–474.

[6] N.K. Logothetis, D.L. Sheinberg, Visual object recognition, Annu. Rev. Neurosci. 19 (1) (1996) 577–621.

[7] E. Wachsmuth, M. Oram, D. Perrett, Recognition of objects and their component parts: responses of single units in the temporal cortex of the macaque, Cerebral Cortex. 4 (5) (1994) 509–522.

[8] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, Adv. Neural Inf. Process. Syst. (2001) 556–562.

[9] D.D. Lee, H.S. Seung, Learning the parts of objects by non-negative matrix factorization, Nature 401 (6755) (1999) 788–791.

[10] N. Guan, D. Tao, Z. Luo, B. Yuan, Online non-negative matrix factorization with robust stochastic approximation, IEEE Trans. Neural Netw. Learn. Syst. 23 (7) (2012) 1087–1099.

[11] N. Guan, D. Tao, Z. Luo, B. Yuan, Nenmf: an optimal gradient method for nonnegative matrix factorization, IEEE Trans. Signal Process. 60 (6) (2012) 2882–2898.

[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 (8) (2011) 1548–1560.

[13] 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.

[14] B. Shen, L. Si, Non-negative matrix factorization clustering on multiple manifolds, in: Proceedings of the 24th AAAI Conference on Artificial Intelligence, 2010.

[15] N. Guan, D. Tao, Z. Luo, B. Yuan, Non-negative patch alignment framework, IEEE Trans. Neural Netw. 22 (8) (2011) 1218–1230.

[16] P.O. Hoyer, Non-negative matrix factorization with sparseness constraints, J. Mach. Learn. Res. 5 (2004) 1457–1469.

[17] Z. Yuan, E. Oja, Projective nonnegative matrix factorization for image compression and feature extraction, in: Proceedings of 14th Scandinavian Conference on Image Analysis (SCIA), Springer, 2005, pp. 333–342.

[18] Y. Chen, J. Zhang, D. Cai, W. Liu, X. He, Nonnegative local coordinate factorization for image representation, IEEE Trans. Image Process. 22 (3) (2013) 969–979.

[19] H. Liu, Z. Yang, Z. Wu, Locality-constrained concept factorization, in: IJCAI Proceedings-International Joint Conference on Artificial Intelligence, vol. 22, Citeseer, 2011, p. 1378.

[20] L. Zhang, Z. Chen, M. Zheng, X. He, Robust non-negative matrix factorization, Front. Electr. Electron. Eng. China 6 (2) (2011) 192–200.

[21] B. Shen, L. Si, R. Ji, B. Liu, Robust nonnegative matrix factorization via l1 norm regularization, arXiv:1204.2311.

[22] D. Kong, C. Ding, 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.

[23] C. Ding, D. Zhou, X. He, H. Zha, R1-pca: rotational invariant l1-norm principal component analysis for robust subspace factorization, in: Proceedings of the 23rd International Conference on Machine Learning, ACM, 2006, pp. 281–288.

[24] L. Du, X. Li, Y.-D. Shen, Robust nonnegative matrix factorization via half-quadratic minimization, 2012, pp. 201–210.

[25] P.N. Belhumeur, J.P. Hespanha, D. Kriegman, Eigenfaces vs. Fisherfaces: recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (7) (1997) 711–720.

[26] A.S. Georghiades, P.N. Belhumeur, D. Kriegman, From few to many: illumination cone models for face recognition under variable lighting and pose, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 643–660.

[27] D.B. Graham, N.M. Allinson, Characterising virtual eigensignature for general face recognition, in: Face Recognition, Springer, 1998, pp. 446–456.

[28] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, in: Proceedings of the Second IEEE Workshop on Applications of Computer Vision, IEEE, 1994, pp. 138–142.
[29] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: NIPS, vol. 14, 2001, pp. 585–591.
[30] X. He, P. Niyogi, Locality preserving projections, in: Neural Information Processing Systems, vol. 16, MIT, 2004, p. 153.
[31] L. Ding, P. Tang, H. Li, et al., Subspace feature analysis of local manifold learning for hyperspectral remote sensing images classification, Int. J. Appl. Math. Inf. Sci. 8 (4) (2014) 1987–1995.
[32] M. Belkin, P. Niyogi, V. Sindhwani, Manifold regularization: a geometric framework for learning from labeled and unlabeled examples, J. Mach. Learn. Res. 7 (2006) 2399–2434.
[33] J. Kunegis, S. Schmidt, A. Lommatzsch, J. Lerner, E.W. De Luca, S. Albayrak, Spectral analysis of signed graphs for clustering, prediction and visualization, in: SDM, vol. 10, SIAM, 2010, pp. 559–559.
[34] C. Gong, D. Tao, J. Yang, K. Fu, Signed Laplacian embedding for supervised dimension reduction, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.
[35] P.O. Hoyer, Non-negative sparse coding, in: Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing, IEEE, 2002, pp. 557–565.
[36] J. Wright, Y. Ma, J. Mairal, G. Sapiro, T.S. Huang, S. Yan, Sparse representation for computer vision and pattern recognition, Proc. IEEE 98 (6) (2010) 1031–1044.
[37] R. He, W.-S. Zheng, B.-G. Hu, X.-W. Kong, Nonnegative sparse coding for discriminative semi-supervised learning, in: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2011, pp. 2849–2856.
[38] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, et al., Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[39] M. Schmidt, Least Squares Optimization with l1-Norm Regularization, CS542B Project Report.
[40] H. Liu, Z. Yang, J. Yang, Z. Wu, X. Li, Local coordinate concept factorization for image representation, IEEE Trans. Neural Netw. Learn. Syst. 25 (6) (2014) 1071–1082.
[41] D. Tao, X. Li, X. Wu, S. Maybank, Geometric mean for subspace selection, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 260–274.
[42] D. Tao, X. Li, X. Wu, S. Maybank, General tensor discriminant analysis and Gabor features for gait recognition, IEEE Trans. Pattern Anal. Mach. Intell. 29 (10) (2007) 1700–1715.
[43] A.P. Dempster, N.M. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Stat. Soc. Ser. B (Methodol.) (1977) 1–38.