Multimed Tools Appl (2018) 77:3029–3048
DOI 10.1007/s11042-017-4970-9

Nonnegative matrix factorization by joint locality-constrained and $\ell_{2,1}$-norm regularization

Ling Xing[1] · Hao Dong[2] · Wei Jiang[2] · Kewei Tang[2]

Received: 21 April 2017 / Revised: 15 June 2017 / Accepted: 20 June 2017 /
Published online: 7 July 2017

Abstract Nonnegative matrix factorization has been widely applied recently. The nonnegativity constraints result in parts-based, sparse representations which can be more robust than global, non-sparse features. However, existing techniques could not accurately dominate the sparseness. To address this issue, we present a unified criterion, called Nonnegative Matrix Factorization by Joint Locality-constrained and $\ell_{2,1}$-norm Regularization(NMF2L), which is designed to simultaneously perform nonnegative matrix factorization and locality constraint as well as to obtain the row sparsity. We reformulate the nonnegative local coordinate factorization problem and use $\ell_{2,1}$-norm on the coefficient matrix to obtain row sparsity, which results in selecting relevant features. An efficient updating rule is proposed, and its convergence is theoretically guaranteed. Experiments on benchmark face datasets demonstrate the effectiveness of our presented method in comparison to the state-of-the-art methods.

Keywords Nonnegative matrix factorization · Local constraint · Clustering

✉ Wei Jiang
swxxjw@aliyun.com

Ling Xing
xinglinghaust@163.com

Hao Dong
haodong@163.com

Kewei Tang
tkwliaoning@gmail.com

[1] School of Information Engineering, Henan University of Science and Technology, Luoyang 471023, China

[2] School of Mathematics, Liaoning Normal University, Dalian 116029, China

1 Introduction

Data representation [15, 16, 27] is a fundamental issue in the machine learning, computer vision etc. In many applications, the data are usually high dimension. Traditional methods that perform well in low-dimensional space can become entirely impractical in high-dimensional feature space. Therefore, dimensionality reduction has become increasingly important since it can alleviate the curse of dimensionality, accelerate learning process, and even provide significant insights into the nature of the problem. Generally speaking, dimensionality reduction techniques [1, 18, 19, 25, 26, 29, 33] can be divided into two categories, that is, feature extraction [1, 18, 19, 33] and feature selection [6, 8, 17]. Feature extraction combines all original features to form new representations while feature selection tries to select a subset of most discriminative features. Compared with feature selection which does not change the original representations of data, feature extraction can create new features.

The most popular feature extraction approaches include Principal Component Analysis (PCA) [1, 22], Nonnegative Matrix Factorization (NMF) [7, 18, 19, 23, 31], Singular Value Decomposition (SVD), and Concept Factorization (CF) [33]. Although these methods have different motivations, they can all be interpreted in matrix decomposition, which usually finds two or more lower dimensional matrices to approximate the original one. The factorization leads to a reduced representation of the initial data, and thus belongs to the technologies for dimension reduction.

Unlike PCA [1] and SVD, NMF [18, 19] factorizes the original data matrix as a multiplication of two ones which are constrained by having nonnegative elements. One matrix consists of basis vectors which reveals the latent semantic structure, and the other matrix can be considered as the coefficients where each sample point is a linear combination of the bases. NMF can be recognized as a part-based representation of the data because only additive, not subtractive, combinations are applied. Such a representation encodes the data applying few components, which makes the encoding easy to interpret. Due to the capability of being able to extract the most discriminative features and feasibility in computation, NMF and its extension versions [5, 11, 21] have been widely applied in computer vision, especially for face recognition task.

Data locality has been widespread applied in many machine learning problems such as dimension reduction [16, 28], clustering [10, 20], and classification [9, 30, 32, 36, 37]. NMF yield sparse codings such that each data point is a linear representation of few basis vectors. However, the sparsity achieved by NMF does not always satisfy data locality properties. As suggested by Yu [20], locality must result in sparsity but not necessary vice versa. It has been stated in [36] that applying the locality constraint would means the sparsity for the encoding matrix, since only the basis vectors close to the original input data would be chosen for data representation. In NMF approach, a sample point might be reconstructed by basis vectors, which are far from the sample point and thus result in unsatisfying classification results. The standard NMF does not preserve the locality during its decomposition process, while local line coding(LLC) [20] can preserve such properties.

Given a loss function during optimization, sparsity regularization has been widely investigated. Bradley et al. [3] proposed ℓ_1-SVM to perform feature selection adopting the ℓ_1-norm which lead to sparse solution. Hoyer [14] extended NMF to sparse constraint explicitly by a ℓ_1-norm minimization on the coefficient and basis matrices, which makes us to discover sparse representations better than those given by standard NMF. Cai et al. [4] proposed a unified sparse subspace learning (SSL) approach based on ℓ_1-norm regularization. The shortcoming of the ℓ_1-norm regularization could not ensure that all the data vectors are sparse in the same features, so it is not feasible to conduct feature selection. To address such an issue, Nie et al. [24] proposed a robust feature selection approach by imposing $\ell_{2,1}$-norm on both loss function and regularization term. Gu et al. [12], Hou et al. [13], and Yang et al. [35] used the $\ell_{2,1}$-norm in subspace learning, sparse regression, and discriminative feature selection respectively. The $\ell_{2,1}$-norm regularization term results in the row sparsity as well as the correlations of all the features.

In this paper we present a novel matrix factorization approach, called Non-negative Matrix Factorization by Joint Locality-constrained and $\ell_{2,1}$-norm Regularization(NMF2L), which is designed to include row sparsity and locality constraints at the same time. The contribution of this paper is summarized as follows.

1. By making the basis vectors to be as close to the original input data points as possible, we incorporate local coordinate coding [36] into non-negative matrix factorization objective function. By adding the $\ell_{2,1}$-norm regularization, we can achieve a row sparse coefficient matrix.
2. The proposed NMF2L performs feature selection on the coefficient matrix, rather than performs feature selection on the original matrix used in traditional feature selection methods.
3. We provide an efficient and effective multiplicative updating procedure for NMF2L approach, and rigorous convergence analysis of our approach is given.

The rest part of this paper is organized as follows. Firstly, it reviews the NMF and Non-negative Sparse Coding(NNSC) methods, and then introduces our NMF2L method and the optimization scheme, convergence study is provided. Furthermore, it describes and analyzes the experimental results. Finally, it concludes and discusses future work.

2 Related works

In this section, we review briefly NMF and NNSC.

2.1 NMF

NMF is a decomposition approach of data matrices whose elements are nonnegative. Assume a nonnegative matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_N] \in \mathbb{R}_+^{M\times N}$. Each column of $\mathbf{X}$ is a data point. NMF aims to find two non-negative matrices $\mathbf{U} = [u_{ik}] \in \mathbb{R}_+^{M\times K}$ and $\mathbf{V} = [v_{jk}] \in \mathbb{R}_+^{K\times N}$ which solve the following optimization problem:

$$\begin{aligned} \mathcal{J}_{NMF} &= \|\mathbf{X} - \mathbf{UV}\|_F^2, \\ s.t.\ \ & \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{1}$$

where $\|\cdot\|_F$ is Frobenius norm. Although the loss function $\mathcal{J}_{NMF}$ is convex in $\mathbf{U}$ only or $\mathbf{V}$ only, they are nonconvex in both matrices together. Therefore, it is impractical to use an algorithm to find the global optimum solution of $\mathcal{J}_{NMF}$. To solve the optimization problem, Lee et al. [18] proposed an iterative updating rule as follows:

$$\begin{aligned} u_{jk} &\leftarrow u_{jk}\frac{(\mathbf{XV}^T)_{jk}}{(\mathbf{UVV}^T)_{jk}}, \\ v_{ki} &\leftarrow v_{ki}\frac{(\mathbf{U}^T\mathbf{X})_{ki}}{(\mathbf{U}^T\mathbf{UV})_{ki}}. \end{aligned}$$

2.2 Non-negative sparse coding (NNSC)

NMF produces sparse representations, which can encode the data with only a few basis vectors. This property further promotes the interpretability of practical problem. However, the sparseness introduced by nonnegativity may not be enough and is difficult to control. To address the difficulty, the Nonnegative Sparse Coding(NNSC) [14] method decomposes multivariate data into a set of positive sparse components by using ℓ_1-norm penalty function to measure the sparseness. Combining reconstruction loss with a sparseness constraint, the following optimization problem is solved:

$$\begin{aligned} \mathcal{J}_{NNSC} &= \|\mathbf{X} - \mathbf{UV}\|_F^2 + \lambda \|\mathbf{V}\|_1, \\ &s.t.\ \ \mathbf{U} \geq 0, \mathbf{V} \geq 0, \end{aligned} \tag{2}$$

where $\|\cdot\|_1$ is the ℓ_1-norm. λ is the tradeoff parameter. The loss function can be solved under the following update rules:

$$\begin{aligned} \mathbf{U} &\leftarrow \mathbf{U} - \mu(\mathbf{UV} - \mathbf{X})\mathbf{V}^T, \\ v_{ki} &\leftarrow v_{ki} \frac{(\mathbf{U}^T\mathbf{X})_{ki}}{((\mathbf{U}^T\mathbf{UV})_{ki} + \lambda)}, \end{aligned}$$

where μ denotes the step-size.

3 Non-negative matrix factorization by joint locality-constrained and $\ell_{2,1}$-norm regularization

In this section, we present a novel and effective approach for data representation. To the end, we consider two regularization terms, i.e., preserving the locality and generating row sparsity of coefficient matrix. We would like to introduce them in sequence.

3.1 The objective function

The first regularization term is motivated from the concept of LLC [36]. we give the concept of coordinate coding in the following.

Definition A coordinate coding is a pair (γ, C),where $C \subset \mathbb{R}^d$ is a set of anchor points, and γ is a map of $\mathbf{x} \in \mathbb{R}^d$ to $[\gamma_v(\mathbf{x})]_{v \in C} \in R^{|C|}$. It induces the following physical approximation of $\mathbf{x}$ in $\mathbb{R}^d$: $\gamma(\mathbf{x}) = \sum_{v \in C} \gamma_v(\mathbf{x})v$.

On the basis of this definition, the NMF can be considered as a coordinate coding where the columns of the basis matrix $\mathbf{U}$ can be considered as a set of anchor points, and each column of $\mathbf{V}$ is the coordinates of each data point in connection with the anchor points. In order to preserve the local structure of the data, only a few anchor points close to the original data would be chosen for data representation. The local coordinate constraint can be formulated as the following problem:

$$\mathcal{Q} = \sum_{k=1}^{K} |v_{ki}| \|\mathbf{u}_k - \mathbf{x}_i\|^2. \tag{3}$$

The above constraint leads to a heavy penalty if $\mathbf{x}_i$ is far away from the basis vector $\mathbf{u}_k$ while its coordinate v_{ki} with respect to $\mathbf{u}_k$ is large.

We choose the second regularization term to distinguish the importance of different features. It is desirable to let the significant features be represented by non-zero values and the nonsignificant features by zeros after the iterative update. Since each row of the coefficient matrix $\mathbf{V}$ is in correspondence to a feature in the original space. This motivates us to add $\ell_{2,1}$-norm regularization on the coefficient matrix $\mathbf{V}$, which impels many rows in $\mathbf{V}$ decline to zero. Then we choose the important feature (i.e. the feature with non-zero values) and discard the unimportant features. $\ell_{2,1}$-norm, as the second regularization term, is given as follows:

$$\|\mathbf{V}\|_{2,1} = \sum_{j=1}^{K} \|\mathbf{V}^{(j)}\|_2, \tag{4}$$

where $\mathbf{V}^{(j)}$ is the jth row of matrix $\mathbf{V}$, which reveals the important degree of the jth feature to all the data points.

By integrating (3) and (4) into the traditional NMF, the overall loss function of NMF2L is defined as:

$$\begin{aligned} \mathcal{O} &= \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \mu \sum_{i=1}^{N} \sum_{k=1}^{K} |v_{ki}| \|\mathbf{u}_k - \mathbf{x}_i\|^2 + \lambda \|\mathbf{V}\|_{2,1}, \\ s.t.\ \mathbf{U} &= [\mathbf{u}_1, \cdots, \mathbf{u}_K] \in \mathbb{R}^{M \times K} > 0, \\ \mathbf{V} &= [\mathbf{v}_1, \cdots, \mathbf{v}_K] \in \mathbb{R}^{K \times N} > 0, \end{aligned} \tag{5}$$

where μ and λ are positive regularization parameters. We call (5) Nonnegative Matrix Factorization by Joint Locality-constrained and $\ell_{2,1}$-norm Regularization(NMF2L). Let $\mu = 0$ and $\lambda = 0$, (5) degenerates to the original NMF.

3.2 The update rules

The loss function $\mathcal{O}$ of NMF2L in (5) is nonconvex in both $\mathbf{U}$ and $\mathbf{V}$ together. Therefore, it is unrealistic to find an algorithm to achieve the global optimal solution. We give an iterative algorithm which can achieve local optimal solution in the following. Following some algebraic steps, the objective function can be rewritten as follows:

$$\begin{aligned} \mathcal{O} = \|\mathbf{X} - \mathbf{U}\mathbf{V}\|_F^2 + \mu \sum_{i=1}^{N} \|(\mathbf{x}_i 1^T - \mathbf{U})\mathbf{\Lambda}_i^{1/2}\|^2 + \lambda \|\mathbf{V}\|_{2,1}, \\ s.t.\ \ \mathbf{U} \geq 0,\ \ \mathbf{V} \geq 0, \end{aligned} \tag{6}$$

where $\mathbf{\Lambda}_i = diag(|\mathbf{V}_i|) \in \mathbb{R}^{K \times K}$. According to the matrix property $\mathrm{Tr}(\mathbf{AB}) = \mathrm{Tr}(\mathbf{BA})$, $\|\mathbf{A}\|_F^2 = \mathrm{Tr}(\mathbf{A}^{\mathrm{T}}\mathbf{A})$ and $\mathrm{Tr}(\mathbf{A}) = \mathrm{Tr}(\mathbf{A}^{\mathbf{T}})$, we have

$$\begin{aligned} \mathcal{O} = \mathrm{Tr}\Bigg(\mathbf{X}\mathbf{X}^T + \mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T - 2\mathbf{X}\mathbf{V}^T\mathbf{U}^T + \mu \sum_{i=1}^{N} (\mathbf{x}_i \mathbf{1}^T \mathbf{\Lambda}_i \mathbf{1} \mathbf{x}_i^T \\ -2\mathbf{x}_i \mathbf{1}^T \mathbf{\Lambda}_i \mathbf{U}^T + \mathbf{U}\mathbf{\Lambda}_i \mathbf{U}^T)\Bigg) + \lambda \|\mathbf{V}\|_{2,1}. \end{aligned} \tag{7}$$

Due to $\mathbf{U} \geq 0$, $\mathbf{V} \geq 0$, we introduce the Lagrangian multiplier $\mathbf{\Psi} = [\psi_{jk}]$ and $\mathbf{\Phi} = [\phi_{ki}]$. Therefore, the objective function could be rewritten as Lagrangian multiplier

$$\mathcal{L} = \mathrm{Tr}\Bigg(\mathbf{X}\mathbf{X}^T + \mathbf{U}\mathbf{V}\mathbf{V}^T\mathbf{U}^T - 2\mathbf{X}\mathbf{V}^T\mathbf{U}^T + \mu\sum_{i=1}^{N}\Big(\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i\mathbf{1}\mathbf{x}_i^T - 2\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i\mathbf{U}^T + \mathbf{U}\mathbf{\Lambda}_i\mathbf{U}^T\Big)\Bigg) + \lambda\|\mathbf{V}\|_{2,1} - \mathrm{Tr}(\mathbf{\Psi}\mathbf{U}^T) - \mathrm{Tr}(\mathbf{\Phi}\mathbf{V}^T). \quad (8)$$

Setting $\frac{\partial\mathcal{L}}{\partial\mathbf{U}} = 0$ and $\frac{\partial\mathcal{L}}{\partial\mathbf{V}} = 0$, we obtain

$$\mathbf{\Psi} = 2\mathbf{U}\mathbf{V}\mathbf{V}^T - 2\mathbf{X}\mathbf{V}^T + \mu\sum_{i=1}^{N}(-2\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i + 2\mathbf{U}\mathbf{\Lambda}_i), \quad (9)$$

$$\mathbf{\Phi} = 2\mathbf{U}^T\mathbf{U}\mathbf{V} - 2\mathbf{U}^T\mathbf{X} + \mu(\mathbf{C} - 2\mathbf{U}^T\mathbf{X} + \mathbf{D}) + \lambda\mathbf{G}\mathbf{V}, \quad (10)$$

where $\mathbf{G}$ is a diagonal matrix with the i-th diagonal element as $\mathbf{G}_{ii} = \frac{1}{2\|\mathbf{V}^{(i)}\|}$. Define column vector $\mathbf{c} = diag(\mathbf{X}^T\mathbf{X}) \in \mathbb{R}^N$. Let $\mathbf{C} = (\mathbf{c}, \cdots, \mathbf{c})^T$ be a $K \times N$ matrix whose rows are $\mathbf{c}^T$. Define column vector $\mathbf{d} = diag(\mathbf{U}^T\mathbf{U}) \in \mathbb{R}^K$. Let $\mathbf{D} = (\mathbf{d}, \cdots, \mathbf{d})$ be a $K \times N$ matrix whose columns are $\mathbf{d}$.

Applying the Karush-Kuhn-Tucker conditions [2] $\psi_{jk}u_{jk} = 0$ and $\phi_{ki}v_{ki} = 0$, we get the following equations:

$$(\mathbf{U}\mathbf{V}\mathbf{V}^T)_{jk}u_{jk} - (\mathbf{X}\mathbf{V}^T)_{jk}u_{jk} + \mu\left(\sum_{i=1}^{N}\mathbf{U}\mathbf{\Lambda}_i\right)_{jk}u_{jk} - \mu\left(\sum_{i=1}^{N}(\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i)_{jk}\ u_{jk} = 0, \quad (11)$$

$$2(\mathbf{U}^T\mathbf{U}\mathbf{V})_{ki}v_{ki} - 2(\mathbf{U}^T\mathbf{X})_{ki}v_{ki} + \lambda(\mathbf{G}\mathbf{V})_{ki}v_{ki} + \mu(\mathbf{C} - 2\mathbf{U}^T\mathbf{X} + \mathbf{D})_{ki}v_{ki} = 0. \quad (12)$$

We can achieve the following update rules:

$$u_{jk} \leftarrow u_{jk}\frac{\left(\mathbf{X}\mathbf{V}^T + \mu\sum_{i=1}^{N}\left(\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i\right)_{jk}\right.}{\left(\mathbf{U}\mathbf{V}\mathbf{V}^T + \mu\sum_{i=1}^{N}\mathbf{U}\mathbf{\Lambda}_i\right)_{jk}}, \quad (13)$$

$$v_{ki} \leftarrow v_{ki}\frac{2(\mu+1)(\mathbf{U}^T\mathbf{X})_{ki}}{(2\mathbf{U}^T\mathbf{U}\mathbf{V} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{G}\mathbf{V})_{ki}}. \quad (14)$$

3.3 Convergence analysis

In this section, we apply the auxiliary function method [19] to prove the convergence. We first introduce the definition of auxiliary function in the following.

Definition 1 $Z(h, h')$ is an auxiliary function of $F(h)$ if the conditions

$$Z(h, h') \geq F(h), \quad Z(h, h) = F(h)$$

are satisfied.

Lemma 1 *If Z is an auxiliary function for F, then F is non-increasing under the update*

$$h^{(t+1)} = \arg\min_h Z(h, h^{(t)}).$$

Proof.

$$F(h^{(t+1)}) \leq Z(h^{(t+1)}, h^{(t)})) \leq Z(h^{(t)}, h^{(t)}) = F(h^{(t)}).$$

The convergence of the algorithm are addressed in the following:

For any element v_{ab} in $\mathbf{V}$, we apply $F_{v_{ab}}$ to denote the part of $\mathcal{O}$ which is only related to v_{ab}. It is easy to check that

$$F'_{v_{ab}} = (2\mathbf{U}^T\mathbf{U}\mathbf{V} - 2\mathbf{U}^T\mathbf{X} + \mu(\mathbf{C} - 2\mathbf{U}^T\mathbf{X} + \mathbf{D}) + \lambda\mathbf{G}\mathbf{V})_{ab},$$

$$F''_{v_{ab}} = 2(\mathbf{U}^T\mathbf{U})_{aa} + \lambda(\mathbf{G} - \mathbf{G}^3(\mathbf{V} \odot \mathbf{V}))_{aa},$$

where $\odot$ denotes the element-wise multiplication.

Theorem 1 *The function*

$$\begin{aligned} Z\left(v, v_{ab}^{(t)}\right) = {} & F_{v_{ab}}\left(v_{ab}^{(t)}\right) + F'_{v_{ab}}\left(v_{ab}^{(t)}\right)\left(v - v_{ab}^{(t)}\right) \\ & + \frac{(2\mathbf{U}^T\mathbf{U}\mathbf{V} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{G}\mathbf{V})_{ab}}{v_{ab}^{(t)}}\left(v - v_{ab}^{(t)}\right)^2 \end{aligned} \tag{15}$$

is an auxiliary function for $F_{v_{ab}}$.

Proof Since $Z(v, v) = F_{v_{ab}}(v)$ is obvious, we need only indicate that $Z(v, v_{ab}^{(t)}) \geq F_{v_{ab}}(v)$. To do this, we compare the Taylor series expansion of $F_{v_{ab}}(v)$

$$\begin{aligned} F_{v_{ab}}(v) = {} & F_{v_{ab}}\left(v_{ab}^{(t)}\right) + F'_{v_{ab}}\left(v_{ab}^{(t)}\right)\left(v - v_{ab}^{(t)}\right) + \Bigg((\mathbf{U}^T\mathbf{U})_{aa} \\ & + \lambda\frac{\left(\mathbf{G} - \mathbf{G}^3(\mathbf{V} \odot \mathbf{V})\right)_{aa}}{2}\Bigg)\left(v - v_{ab}^{(t)}\right)^2 \end{aligned} \tag{16}$$

with (15) to obtain that $Z(v, v_{ab}^{(t)}) \geq F_{v_{ab}}(v)$ is equal to

$$\begin{aligned} \frac{(2\mathbf{U}^T\mathbf{U}\mathbf{V} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{G}\mathbf{V})_{ab}}{v_{ab}^{(t)}} \geq {} & (\mathbf{U}^T\mathbf{U})_{aa} \\ & + \frac{\lambda(\mathbf{G} - \mathbf{G}^3(\mathbf{V} \odot \mathbf{V}))_{aa}}{2}. \end{aligned} \tag{17}$$

We have

$$(2\mathbf{U}^T\mathbf{U}\mathbf{V})_{ab} = 2\sum_{l=1}^{K}(\mathbf{U}^T\mathbf{U})_{al}v_{lb}^{(t)} \geq (\mathbf{U}^T\mathbf{U})_{aa}v_{ab}^{(t)}$$

and

$$\begin{aligned}\lambda(\mathbf{GV})_{ab} &= \lambda\sum_{l=1}^{K}\mathbf{G}_{al}v_{lb}^{(t)} \geq \lambda\mathbf{G}_{aa}v_{ab}^{(t)}\\ &\geq \lambda(\mathbf{G}_{aa} - (\mathbf{G}^3(\mathbf{V}\odot\mathbf{V}))_{aa})v_{ab}^{(t)}\\ &= \lambda(\mathbf{G} - \mathbf{G}^3(\mathbf{V}\odot\mathbf{V}))_{aa}v_{ab}^{(t)}\\ &\geq \frac{\lambda(\mathbf{G} - \mathbf{G}^3(\mathbf{V}\odot\mathbf{V}))_{aa}}{2}v_{ab}^{(t)}.\end{aligned} \tag{18}$$

Thus, (17) holds and $Z(v, v_{ab}^{(t)}) \geq F_{ab}(v)$. □

Theorem 2 *Equation* (14) *could be obtained by minimizing function* $Z(v, v_{ab}^{(t)})$, *where* $v_{ab}^{(t)}$ *is the iterative solution at the* t*-th step.*

Proof To obtain the minimum, we only need set the derivative $\frac{\partial Z(v, v_{ab}^{(t)})}{\partial v_{ab}} = 0$, and have

$$\frac{\partial Z\left(v_{ab}, v_{ab}^{(t)}\right)}{\partial v_{ab}} = F'_{v_{ab}}\left(v_{ab}^{(t)}\right) + \frac{2(2\mathbf{U}^T\mathbf{UV} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{GV})_{ab}}{v_{ab}^{(t)}}\left(v_{ab} - v_{ab}^{(t)}\right) = 0. \tag{19}$$

Thus, by simple algebra formulation, we can obtain (14).

From Theorem 1, we know that $Z(v, v_{ab}^{(t)})$ is an auxiliary function for $F_{v_{ab}}$. According to Lemma 1 and Theorem 2, updating v_{ab} using (14) will decrease monotonically the objective function in (7), therefore it converge local optimal.

The converge proof that updating u_{ab} using (13) is similar to the above. □

3.4 Connection to gradient descent method

The objective function of NMF2L can be minimized by gradient descent algorithm. Using gradient descent method results in the update rules as follows:

$$u_{jk} \leftarrow u_{jk} + \eta_{jk}\frac{\partial \mathcal{O}}{\partial v_{jk}}, \tag{20}$$

$$v_{ki} \leftarrow v_{ki} + \delta_{ki}\frac{\partial \mathcal{O}}{\partial v_{ki}}, \tag{21}$$

where the δ_{jk} and η_{ki} are the parameters of the step size.

It is difficult to set these size parameters, and maintain the non-negativity of u_{jk} and v_{ki}. we set

$$\eta_{jk} = -\frac{u_{jk}}{2(\mathbf{UVV}^T + \mu\sum_{i=1}^{N}\mathbf{U}\mathbf{\Lambda}_i)_{jk}}, \tag{22}$$

$$\delta_{ki} = -\frac{v_{ki}}{(2\mathbf{U}^T\mathbf{UV} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{GV})_{ki}}, \tag{23}$$

then we can obtain

$$u_{jk} + \eta_{jk}\frac{\partial \mathcal{O}}{\partial u_{jk}} = u_{jk}\frac{\left(\mathbf{X}\mathbf{V}^T + \mu\sum_{i=1}^{N}(\mathbf{x}_i\mathbf{1}^T\mathbf{\Lambda}_i)_{jk}\right.}{\left(\mathbf{U}\mathbf{V}\mathbf{V}^T + \mu\sum_{i=1}^{N}\mathbf{U}\mathbf{\Lambda}_i\right)_{jk}}, \tag{24}$$

$$v_{ki} + \delta_{ki}\frac{\partial \mathcal{O}}{\partial v_{ki}} = v_{ki}\frac{2(\mu+1)(\mathbf{U}^T\mathbf{X})_{ki}}{(2\mathbf{U}^T\mathbf{U}\mathbf{V} + \mu\mathbf{C} + \mu\mathbf{D} + \lambda\mathbf{G}\mathbf{V})_{ki}}, \tag{25}$$

which are the update rules in (13) and (14). It is clear that the (13) and (14) are special cases of gradient descent.

4 Experiments

In this section, we systematically investigated the NMF2L for clustering task. Some experiments were performed to indicate the effectiveness of our algorithm.

4.1 Data preparation

Three different publicly available database are widespread adopted as benchmark datasets. These datasets are described as follows:

The ORL face dataset consists of 10 different face images for 40 distinct persons. All the 400 have been captured against a dark homogeneous background with the subjects in an upright, frontal position with tolerance for some side movement. For some persons, the faces were taken at different times, varying the lighting, facial expressions (open/closed eyes, smiling/not smiling) and facial details (glasses/no glasses).

The Yale face dataset consists of 165 gray scale face images of 15 persons. All faces show variations in lighting condition (left-light, center-light, right-light), facial expression(normal, happy, sad, sleepy, surprised and wink), and with/without glasses.

The CMU PIE face dataset consists of 68 persons with 41368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying poses, illuminations and expressions. In our experiments, one near frontal pose (C27) is chosen under different illuminations, lightings and expressions which leaves us about 49 near frontal face images for each individual.

In the experiments, face images were preprocessed so that faces were located. Firstly, original face images were normalized in scale and orientation such that the two eyes were aligned at the same position. Then the facial areas were cropped into the final images for clustering. We resized them to 32 × 32 pixels with 256 gray levels per pixel for computational convenience.

4.2 Experimental design

This section contains the description of evaluation metrics, compared methods, and parameter settings.

4.2.1 Evaluation metrics

In our experiments, we set the number of clusters equal to the number of classes for all algorithms. We evaluated the clustering performance by comparing the cluster results obtained by algorithms with its true classes. The Accuracy (Acc) and the Normalized Mutual Information metric (NMI) were adopted to measure the clustering results [34].

The Accuracy(Acc) is defined as follows:

$$\mathrm{Acc} = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{26}$$

where r_i is the cluster label of x_i, and l_i is the true class label, n denotes the total number of samples, $\delta(x, y)$ represents the delta function that equals one if $x = y$ and equals zero otherwise, and $map(r_i)$ is the permutation mapping function that maps the obtained label r_i to the equivalent label from the data set.

Normalized Mutual Information(NMI) is defined as follows:

$$\mathrm{NMI} = \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{\left(\sum_{i=1}^{c} n_i \log \frac{n_i}{n}\right) \left(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n}\right)}}, \tag{27}$$

where n_i is the number of samples in the i-th cluster $\mathcal{C}_i$ according to clustering results and $\hat{n}_j$ is the number of samples in the j-th ground truth class $\mathcal{C}'_j$. $n_{i,j}$ denotes the number of samples that are in the intersection between $\mathcal{C}_i$ and $\mathcal{C}'_j$.

4.2.2 Compared methods

We showed the data clustering performance of the NML2L algorithm, and compared the result with the state-of-the-art algorithms using the same dataset. The algorithms that we evaluated are listed below:

- Traditional k-means clustering algorithm (KM).
- Principal Component Analysis(PCA) [1].
- Nonnegative Matrix Factorization (NMF) [18].
- Nonnegative Matrix Factorization with Sparseness Constraints(NMFSC) [14].
- Graph regularized Non-negative Matrix Factorization (GNMF) [5].
- Nonnegative Local Coordinate Factorization (NLCF) [10] for feature extraction.
- Our proposed Nonnegative Matrix Factorization by Joint Locality-constrained and $\ell_{2,1}$-norm Regularization(NMF2L).

4.2.3 Parameter settings

We evaluated the clustering performance on three face datasets. For each dataset, the evaluations were performed under different numbers of clusters. The dimensionality of the new space was set to the number of clusters. For NMF, once the number of clusters is given, there is no parameter selection. The regularization parameters were set by the grid $\{0.001, 0.01, 0.1, 1, 10, 100, 500, 1000\}$. The neighborhood size in GNMF was tuned by searching the grid $\{2, 3, \cdots, 10\}$.

For a fixed cluster number k, we randomly chose k classes from the dataset, and used different algorithms to achieve new data representations $\mathbf{V}$. K-means was performed based on the new data representation $\mathbf{V}$. Because of depending on initialization, k-means was repeated 20 times with random initializations and the average result was reported. We compared the obtained clusters with the original image class to compute the Acc and NMI.

4.3 Clustering results

Tables 1, 2 and 3 displayed the values of Acc and NMI when using different the number of clusters based on various algorithms. The last row of table gived the average clustering results over k. The mean values and standard deviations were available.

On ORL data set, our NMF2L algorithm achieves performance gain of 9.08% in Acc and 9.96% in NMI over the best performance by the other four algorithms. On Yale data set, our proposed algorithm achieves performance gain of 5.31% in Acc and 6.97% in NMI over the best performance by the other algorithms. On PIE data set, our algorithm outperform significantly other six algorithms in terms of both Acc and NMI.

4.4 Parameter sensitivity

Parameter selection for unsupervised algorirthm is one of challenges to machine learning tasks. In this part, we studied the clustering results of NMF2L in regard to the variations of different parameters' settings. Our algorithm have two regularization parameters, which are denoted as μ and λ in (5). We ploted the Acc and NMI of K-means with different μ and λ in a searching grid $\{0.0001, 0.01, 0.1, 1, 10, 100, 1000\}$. In the experiment, we randomly chose half of samples per class for parameter selection. We averaged the clustering results of 20 times with random initializations. The average result were recorded, and the 3D plots were shown in Figs. 1, 2, and 3. The horizontal axes denote the various value of the parameter μ and λ, while the vertical axis is the evaluation metric. In the 3D plots, the square/circle marker in the direction of X/Y-axis indicated the best μ/λ for varying μ/λ. Beside the intersect point, there is a digit number showing the value of Acc or NMI.

4.5 Convergence analysis

In this subsection, we performed an experiment to validate convergence of our algorithm. Following the above experiments, we randomly chose half of samples per class for convergence analysis. The two parameters α and β were both fixed at 10. We chose NMF and NMF2L for comparison to study the convergence speed. We drew the convergence curves on all the three datasets in Fig. 4. The horizontal axis represents the number of iterations and the vertical axis denotes the value of objective function. We can see in Fig. 4 that the objective function value becomes stable after about 100 iterations on all datasets. The convergence experiment shows the efficiency of our algorithm.

4.6 Overall observations and discussion

In the face recognition experiments above, we consider several groups of experiments based on different databases. From the above experiment results, we can obtain several attractive insights as follows:

1) The performance of K-means are the worst for all algorithms. This shows feature extraction is necessary to enhance the clustering performance. The performance of PCA is inferior to other NMF-based algorithms, which demonstrates the superiority of parts-based representation idea. In most cases, GNMF performs better than NMF, which verifies the major role of the geometric structure in matrix decomposition algorithm. But we can see from Table 2 that it is not right on the Yale dataset. A possible explanation is that GNMF cannot guarantee that nearby points have the same class labels,

Table 1 Clustering performance on the ORL dataset

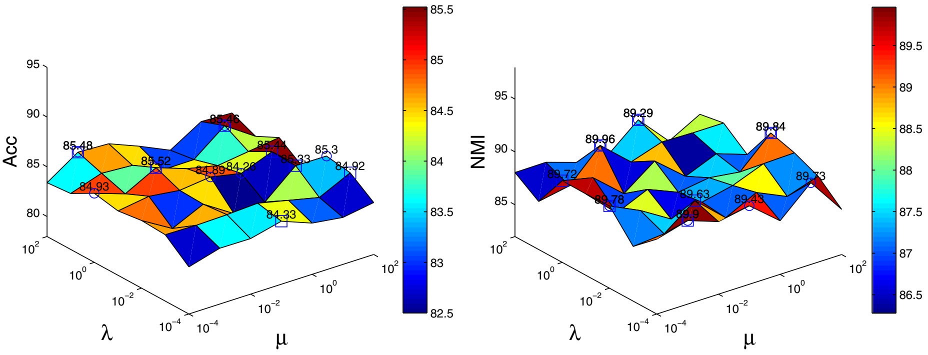
k	KM	PCA	NMF	NMFSC	GNMF	NLCF	NMF2L
	Acc(%)						
2	73.51 ± 17.61	72.17 ± 16.52	94.43 ± 13.23	97.20 ± 12.01	90.02 ± 7.83	98.54 ± 3.26	98.22 ± 8.92
4	52.23 ± 11.46	55.21 ± 13.76	79.52 ± 15.01	81.03 ± 13.43	82.33 ± 6.76	82.03 ± 12.74	90.44 ± 7.83
8	49.67 ± 5.81	51.09 ± 3.94	62.45 ± 5.55	72.54 ± 8.74	75.23 ± 5.66	76.64 ± 13.24	87.76 ± 8.54
12	46.92 ± 7.71	48.44 ± 5.91	56.76 ± 3.46	63.03 ± 6.64	74.39 ± 5.78	70.83 ± 7.33	85.67 ± 6.49
16	44.20 ± 4.35	47.83 ± 4.86	53.24 ± 3.47	58.33 ± 3.75	72.58 ± 2.43	67.42 ± 4.87	81.99 ± 4.38
20	41.75 ± 3.54	45.95 ± 4.72	48.11 ± 4.53	50.82 ± 3.23	70.34 ± 2.76	74.72 ± 7.23	80.89 ± 3.67
25	40.05 ± 4.10	43.07 ± 3.54	46.05 ± 3.26	50.28 ± 3.82	68.16 ± 2.14	73.52 ± 2.13	79.79 ± 2.56
30	39.21 ± 5.38	40.29 ± 3.41	41.97 ± 3.17	45.75 ± 2.14	62.35 ± 1.82	70.46 ± 3.14	80.53 ± 3.15
40	37.99 ± 3.91	38.99 ± 4.36	39.53 ± 3.56	41.02 ± 3.57	63.09 ± 2.98	71.83 ± 3.57	82.42 ± 3.87
Avg	47.28	49.26	58.04	62.22	73.17	76.22	85.30
	NMI(%)						
2	76.61 ± 21.68	78.54 ± 19.37	92.91 ± 18.13	95.23 ± 19.73	90.33 ± 19.88	93.71 ± 17.15	94.56 ± 15.54
4	54.78 ± 15.71	75.86 ± 12.57	70.63 ± 17.92	75.22 ± 17.23	86.63 ± 16.74	75.53 ± 15.94	92.57 ± 17.84
8	57.67 ± 7.86	63.48 ± 9.78	64.42 ± 5.21	73.14 ± 6.82	78.66 ± 5.75	76.82 ± 12.03	91.94 ± 8.74
12	55.34 ± 6.67	58.58 ± 8.03	64.14 ± 2.96	70.01 ± 4.87	72.68 ± 7.57	77.34 ± 5.02	85.62 ± 6.95
16	50.42 ± 4.93	55.78 ± 6.57	65.06 ± 3.01	68.15 ± 2.54	74.53 ± 5.81	77.16 ± 2.92	84.94 ± 8.45
20	48.98 ± 5.74	52.37 ± 5.81	62.41 ± 3.05	64.36 ± 1.72	72.71 ± 3.82	75.47 ± 3.77	86.45 ± 4.36
25	45.67 ± 4.48	50.59 ± 4.39	62.99 ± 2.11	65.38 ± 2.65	71.24 ± 2.93	74.55 ± 1.76	82.87 ± 2.64
30	43.38 ± 6.38	56.87 ± 5.28	61.23 ± 2.17	63.44 ± 1.64	73.16 ± 2.64	75.46 ± 2.44	85.59 ± 3.75
40	44.57 ± 3.06	58.56 ± 3.43	61.69 ± 2.89	61.49 ± 1.97	75.63 ± 2.79	76.59 ± 2.81	87.75 ± 2.98
Avg	53.05	61.18	67.28	70.71	77.29	78.07	88.03

Table 2 Clustering performance on the yale dataset

k	KM	PCA	NMF	NMFSC	GNMF	NLCF	NMF2L
	Acc(%)						
2	78.18 ± 16.86	75.92 ± 18.74	73.18 ± 18.00	79.09 ± 17.51	80.45 ± 15.82	86.00 ± 16.01	92.45 ± 16.78
3	59.09 ± 12.14	61.59 ± 9.38	62.73 ± 11.42	63.33 ± 9.91	63.64 ± 10.90	76.36 ± 15.25	78.52 ± 12.22
4	51.36 ± 7.20	55.48 ± 7.11	56.14 ± 5.75	56.82 ± 7.87	52.27 ± 8.76	59.63 ± 9.52	68.54 ± 11.43
5	50.36 ± 3.36	51.66 ± 6.13	52.55 ± 5.30	54.18 ± 6.94	46.73 ± 7.12	58.18 ± 7.62	61.68 ± 8.53
6	46.52 ± 6.85	48.38 ± 5.41	49.55 ± 8.91	51.06 ± 8.63	39.09 ± 5.76	52.42 ± 9.72	56.90 ± 4.87
7	43.25 ± 6.42	45.73 ± 6.23	46.49 ± 5.97	47.66 ± 5.29	41.17 ± 4.56	50.38 ± 4.72	57.83 ± 5.69
8	44.20 ± 3.78	45.46 ± 7.32	45.00 ± 3.56	47.05 ± 2.84	34.43 ± 7.74	50.34 ± 4.39	59.65 ± 5.42
9	41.74 ± 7.03	42.76 ± 4.34	43.23 ± 4.33	42.32 ± 4.50	35.45 ± 3.45	48.78 ± 4.61	53.56 ± 6.39
10	39.45 ± 4.49	40.14 ± 3.26	41.36 ± 3.98	41.00 ± 4.20	36.36 ± 4.45	50.00 ± 3.88	51.39 ± 5.42
Avg.	50.46	51.90	52.25	53.61	47.73	59.19	64.50
	NMI(%)						
2	35.32 ± 19.62	38.12 ± 19.67	32.53 ± 20.35	42.85 ± 20.21	41.27 ± 19.23	56.53 ± 21.23	68.17 ± 24.11
3	33.49 ± 18.12	34.16 ± 12.77	34.28 ± 17.48	32.70 ± 13.32	35.81 ± 12.12	35.81 ± 14.23	57.35 ± 16.33
4	29.16 ± 6.33	30.28 ± 8.32	32.78 ± 16.27	35.88 ± 8.19	34.85 ± 8.95	34.85 ± 10.23	48.54 ± 10.21
5	37.79 ± 7.43	38.99 ± 9.49	40.11 ± 12.75	40.21 ± 8.35	31.97 ± 8.65	31.97 ± 9.76	50.66 ± 9.54
6	29.46 ± 5.28	37.24 ± 8.05	37.74 ± 6.21	39.83 ± 6.78	28.24 ± 6.54	28.24 ± 5.47	49.32 ± 8.32
7	34.45 ± 5.19	36.81 ± 7.64	36.89 ± 6.53	41.27 ± 6.57	34.53 ± 5.45	34.53 ± 4.71	47.23 ± 6.43
8	38.02 ± 4.92	39.37 ± 4.63	41.07 ± 4.54	41.34 ± 5.16	28.18 ± 4.32	28.18 ± 5.31	47.44 ± 5.32
9	37.36 ± 3.03	38.63 ± 4.59	40.94 ± 5.32	42.25 ± 5.43	32.63 ± 4.12	32.63 ± 6.23	49.51 ± 4.45
10	36.58 ± 3.65	37.45 ± 3.31	39.83 ± 5.49	42.30 ± 4.01	35.34 ± 3.21	35.34 ± 4.45	49.85 ± 5.46
Avg.	34.65	36.78	37.35	39.85	33.65	45.04	52.01

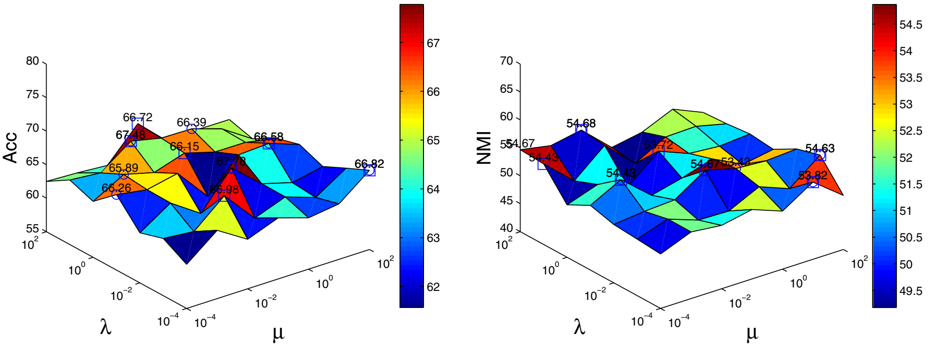
Table 3 Clustering performance on the PIE dataset

k	KM	PCA	NMF	NMFSL	GNMF	NLCF	NMF2L
	Acc(%)						
10	66.68 ± 11.32	68.83 ± 9.79	66.84 ± 12.30	78.26 ± 10.34	86.35 ± 11.71	89.34 ± 11.81	93.67 ± 12.57
20	64.33 ± 8.28	66.63 ± 8.54	77.06 ± 8.57	75.95 ± 8.94	88.45 ± 9.23	87.63 ± 10.53	90.43 ± 6.67
30	60.36 ± 3.37	63.44 ± 3.23	80.32 ± 3.75	73.44 ± 5.6	89.91 ± 4.55	90.64 ± 9.27	92.73 ± 5.74
40	58.34 ± 2.45	62.19 ± 2.87	82.73 ± 2.44	79.83 ± 6.87	88.71 ± 3.43	91.68 ± 6.29	93.55 ± 3.34
50	57.29 ± 1.29	60.06 ± 1.89	83.24 ± 2.97	77.45 ± 3.98	89.47 ± 3.04	91.83 ± 3.35	92.78 ± 3.52
60	56.03 ± 1.32	59.62 ± 1.56	84.18 ± 1.49	76.65 ± 4.53	86.21 ± 2.74	90.56 ± 2.54	91.74 ± 2.71
68	54.27 ± 2.27	58.79 ± 1.92	85.29 ± 2.32	77.87 ± 4.28	86.71 ± 3.45	75.14 ± 3.45	90.31 ± 3.27
Avg.	59.61	62.79	79.95	77.06	87.97	89.97	92.17
	NMI(%)						
10	69.37 ± 11.86	72.89 ± 13.29	72.23 ± 10.03	74.93 ± 13.74	81.41 ± 9.75	85.28 ± 10.32	90.22 ± 11.51
20	67.22 ± 12.63	69.77 ± 12.22	73.22 ± 9.73	73.33 ± 9.64	83.50 ± 7.68	84.51 ± 8.51	89.34 ± 9.53
30	66.13 ± 2.41	68.41 ± 2.63	75.44 ± 2.41	75.23 ± 2.34	84.71 ± 2.46	83.92 ± 3.48	85.56 ± 3.81
40	65.14 ± 2.28	67.37 ± 3.55	77.06 ± 2.61	72.15 ± 1.13	81.86 ± 3.72	84.73 ± 2.61	88.75 ± 3.62
50	65.08 ± 1.52	66.61 ± 2.28	78.64 ± 2.91	76.96 ± 2.85	82.48 ± 2.31	85.07 ± 2.50	87.52 ± 2.53
60	64.91 ± 1.55	65.59 ± 2.11	79.13 ± 2.85	75.35 ± 2.56	83.77 ± 1.69	86.85 ± 1.78	87.53 ± 1.45
68	63.54 ± 1.17	65.25 ± 2.12	79.18 ± 2.01	72.97 ± 3.24	82.61 ± 2.98	85.44 ± 2.11	87.22 ± 1.98
Avg.	65.91	67.98	76.41	74.42	82.91	85.11	88.02



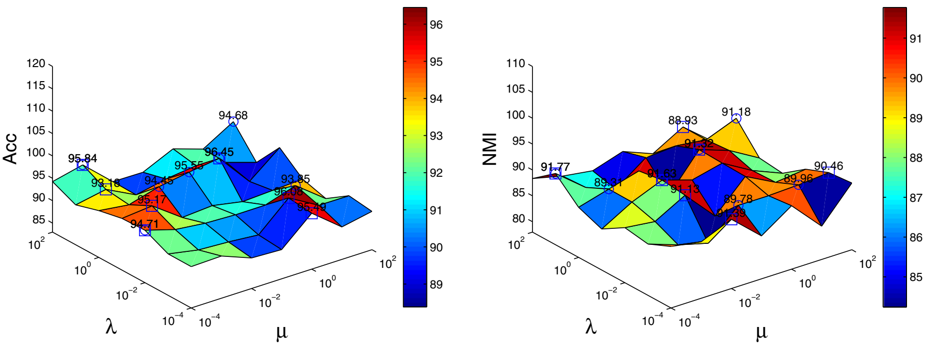
(a) The Acc results (b) The NMI results

Fig. 1 Acc and NMI of Kmeans on ORL data set



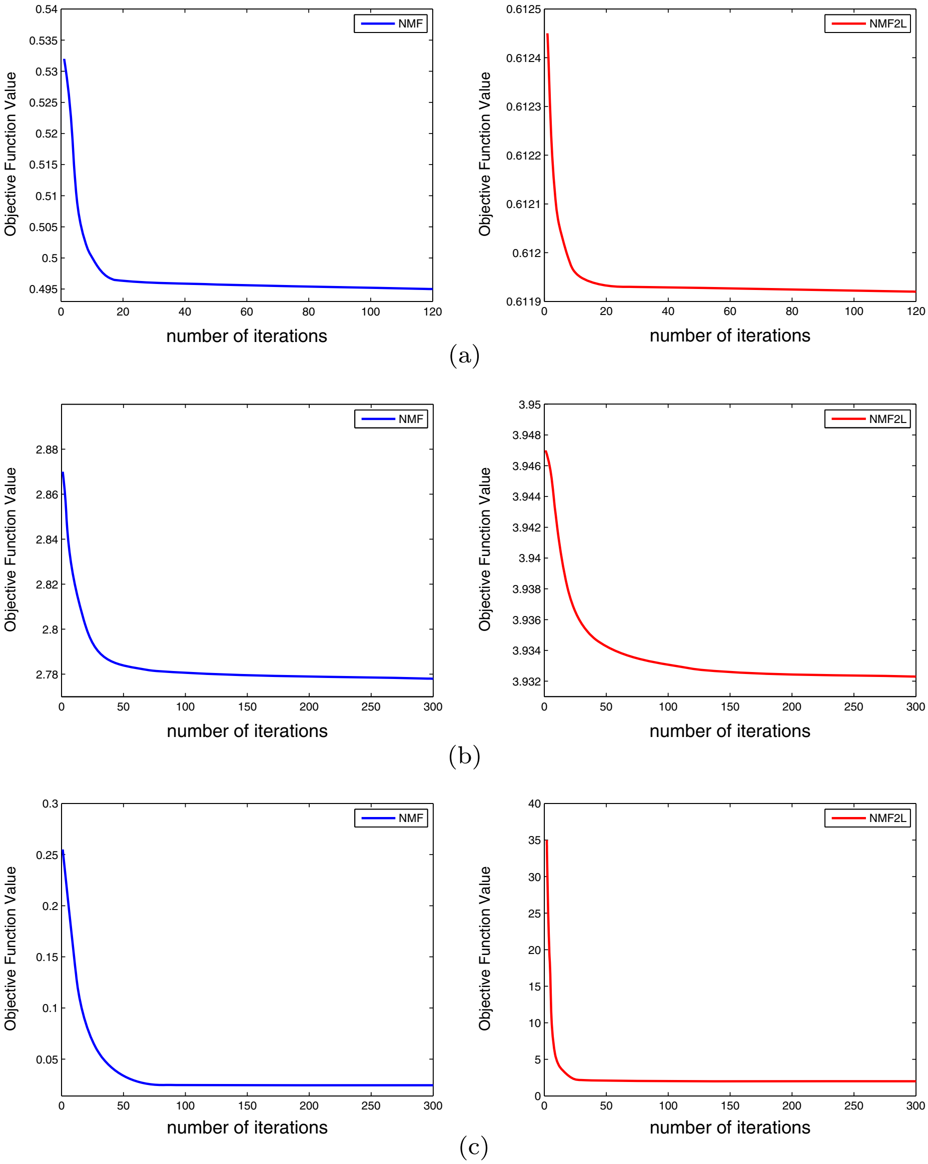
(a) The Acc results (b) The NMI results

Fig. 2 Acc and NMI of Kmeans on Yale data set



(a) The Acc results (b) The NMI results

Fig. 3 Acc and NMI of Kmeans on PIE data set



(a)

(b)

(c)

Fig. 4 Convergence curve of NMF and NMF2L. **a** ORL, **b** Yale, and **c** PIE

and therefore, NMF based on graph regularization may even bring negative effective effects.

2) For all datasets, our NMF2L algorithm is superior to all other algorithms. The reason lies in the fact that NMF2L is designed to use the locality constraint to preserve the geometrical structure, and employ the $\ell_{2,1}$-norm to generate the row sparsity.

3) We can notice from Figs. 1–3 that the clustering performance varies different combinations of μ and λ. The impact of different values of the regularization parameters is

involved in the trait of the data set. We can see from Fig. 4 that the objective function value rapidly converges.

5 Conclusion

In this correspondence, we have presented a novel matrix factorization algorithm, called Non-negative Matrix Factorization by Joint Locality-constrained and $\ell_{2,1}$-norm Regularization(NMF2L) for feature extraction, in which feature selection and non-negative local coordinate factorization problem are simultaneously solved by optimizing a single objective function. The experimental results on three datasets have demonstrated the effectiveness of our approach over other matrix factorization algorithms. Further research on this topic includes: 1) how to apply it to some large-scale and real-life applications; 2) how to extend the current framework for tensor-based nonnegative data decomposition.

Acknowledgments We would like to thank all anonymous reviewers for their helpful comments. This work is supported by the Natural Science Foundation of Liaoning No. 2015020070, and the Natural Science Foundation of China No.61171109 and 61175048.

References

1. Belhumeur PN, Hespanha JP, Kriegman D et al (1997) Eigenfaces vs. fisherfaces: recognition using class specific linear projection. IEEE Trans Pattern Anal Mach Intell 19(7):711–720
2. Boyd S, Vandenberghe L (2004) Convex optimization. Cambridge University Press
3. Bradley PS, Mangasarian OL (1998) Feature selection via concave minimization and support vector machines. In: International conference on machine learning, pp 82–90
4. Cai D, He X, Han J et al (2007) Spectral regression: a unified approach for sparse subspace learning. In: International conference on data mining, pp 73–82
5. Cai D, He X, Han J et al (2011) Graph regularized nonnegative matrix factorization for data representation. IEEE Trans Pattern Anal Mach Intell 33(8):1548–1560
6. Chang X, Nie F, Yang Y et al (2014) A convex formulation for semi-supervised multi-label feature selection. In: Twenty-eighth AAAI conference on artificial intelligence. AAAI Press, pp 1171–1177
7. Chang X, Nie F, Yang Y et al (2016) Convex sparse PCA for unsupervised feature learning. ACM Trans Knowl Discov Data (TKDD) 11(1):3:1–3:16
8. Chang X, Yang Y (2016) Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Transactions on Neural Networks and Learning Systems. doi:10.1109/TNNLS.2016.2582746
9. Chao Y, Yeh Y, Chen Y et al (2011) Locality-constrained group sparse representation for robust face recognition. In: International conference on image processing, pp 761–764
10. Chen Y, Zhang J, Cai D et al (2013) Nonnegative local coordinate factorization for image representation. IEEE Trans Image Process 22(3):969–979
11. Geng B, Tao D, Xu C et al (2012) Ensemble manifold regularization. IEEE Trans Pattern Anal Mach Intell 34(6):1227–1233
12. Gu Q, Li Z, Han J et al (2011) Joint feature selection and subspace learning. In: International joint conference on artificial intelligence, pp 1294–1299
13. Hou C, Nie F, Yi D et al (2011) Feature selection via joint embedding learning and sparse regression. In: International joint conference on artificial intelligence, pp 1324–1329
14. Hoyer PO (2002) Non-negative sparse coding. In: Proceedings of IEEE workshop on neural networks for signal processing, pp 557–565
15. Jiang W, Li M, Zhang Y (2014) Neighborhood preserving convex nonnegative matrix factorization. Math Probl Eng 2014(2):1–8
16. Jiang W, Liu J, Qi H et al (2016) Robust subspace segmentation via nonconvex low rank representation. Inf Sci 340:144–158
17. Kotsiantis SB (2014) RETRACTED ARTICLE: feature selection for machine learning classification problems: a recent overview[J]. Artif Intell Rev 42(1):157–157

18. Lee D, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788–791
19. Liu H, Wu Z, Li X et al (2012) Constrained nonnegative matrix factorization for image representation. IEEE Trans Pattern Anal Mach Intell 34(7):1299–1311
20. Liu H, Yang Z, Wu Z et al (2011) Locality-constrained concept factorization. In: International joint conference on artificial intelligence, pp 1378–1383
21. Liu H, Yang Z, Wu Z et al (2012) A-optimal non-negative projection for image representation. Comput Vision Pattern Recogn, 1592–1599
22. Luo M, Nie F, Chang X et al (2016) Avoiding optimal mean robust PCA/2DPCA with non-greedy l1-norm maximization. In: International joint conference on artificial intelligence
23. Luo M, Nie F, Chang X et al (2017) Probabilistic non-negative matrix factorization and its robust extensions for topic modeling. In: Thirty-first AAAI conference on artificial intelligence
24. Nie F, Huang H, Cai X et al (2010) Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. Neural Inform Process Syst, 1813–1821
25. Nie L, Song X, Chua TS (2016) Learning from multiple social networks. Synthesis Lect Inform Concepts Retriev Serv 8(2):118–129
26. Nie L, Zhang L, Wang M et al (2017) Learning user attributes via mobile social multimedia analytics. ACM Trans Intell Syst Technol (TIST) 8(3):36–47
27. Qi H, Li K, Shen Y et al (2012) Object-based image retrieval with kernel on adjacency matrix and local combined features. ACM Trans Multimed Comput Commun Appl 8(4):1–18
28. Roweis S, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
29. Song X, Nie L, Zhang L et al (2015) Interest inference via structure-constrained multi-source multi-task learning. In: International conference on artificial intelligence. AAAI Press, pp 2371–2377
30. Wang J, Yang J, Yu K et al (2010) Locality-constrained linear coding for image classification. Comput Vision Pattern Recogn, 3360–3367
31. Wang R, Nie F, Yang X et al (2015) Robust 2DPCA with non-greedy, ℓ_1-norm maximization for image analysis[J]. IEEE Trans Cybern 45(5):1108–1112
32. Wei C, Chao Y, Yeh Y et al (2013) Locality-sensitive dictionary learning for sparse representation based classification. Pattern Recogn 46(5):1277–1287
33. Xu W, Gong Y (2004) Document clustering by concept factorization. In: Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval. ACM, pp 202–209
34. Xu W, Liu X, Gong Y et al (2003) Document clustering based on non-negative matrix factorization. In: International ACM SIGIR conference on research and development in information retrieval, pp 267–273
35. Yang Y, Shen HT, Ma Z et al (2011) $\ell_{2,1}$-norm regularized discriminative feature selection for unsupervised learning[c]. In: International joint conference on artificial intelligence, pp 1589–1594
36. Yu K, Zhang T, Gong Y et al (2009) Nonlinear learning using local coordinate coding. Neural information processing systems, 2223–2231
37. Zheng M, Bu J, Chen C et al (2011) Graph regularized sparse coding for image representation. IEEE Trans Image Process 20(5):1327–1336

Ling Xing is a Professor in School of Information Engineering, Henan University of Science and Technology, China. She received the Ph.D. degree in Communication and Information System from Beijing Institute of Technology in 2008. Her research interests include information intelligent management, multimedia semantic concept discovery, and big data mining.

Hao Dong received the B.S. degree from Shenyang Normal University, Shenyang, China. she is currently a Ph.D. candidate in the School of Mathematics, Liaoning Normal University, Dalian, China. Her research interests include computer vision and machine learning.

Wei Jiang received the B.S and Ph.D. degrees from the School of Computer and Communication Engineering, University of Science and Technology Beijing, China, in 2004 and 2011, respectively. He became a faculty member in July 2004 in the School of Mathematics, Liaoning Normal University, Dalian, China, where he is currently an professor. His research interests include computer vision and machine learning.

Kewei Tang received the B.Sc. degree in Mathematics from Liaoning Normal University, the M.Sc. degree and the Ph.D. degree in Mathematics from the Dalian University of Technology, Dalian, China, in 2008, 2011, and 2015, respectively. He was a visiting scholar with the Department of Statistical Science, Duke University from 2013-2014. He is currently a lecturer in the School of Mathematics, Liaoning Normal University. His research interests include computer vision, machine learning, etc.