Signal Processing: Image Communication 62 (2018) 139–148

Contents lists available at ScienceDirect

Signal Processing: Image Communication

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

Label and orthogonality regularized non-negative matrix factorization for image classification

Wenjie Zhu *, Yunhui Yan

School of Mechanical Engineering and Automation, Northeastern University, Shenyang, 110819, China

ARTICLE INFO

Keywords:
Non-negative matrix factorization (NMF)
Orthogonal property
Label consistence
Image classification

ABSTRACT

As one of the most popular data-representation methods, non-negative matrix factorization (NMF) has been widely used in image processing and pattern recognition. Compared with other dimension reduction methods, we can interpret the data with psychological intuition using NMF since NMF can decompose the whole into visual parts by learning the non-negative basis. However, the original NMF lacks of extracting the discriminant information of the data for the image classification task. For enhancing the discriminant and parts-based interpretability, this work proposes a label and orthogonality regularized NMF (LONMF) algorithm based on the squared Euclidean distance. LONMF takes into account the label consistence with the low-dimensional projected data and orthogonal property of the non-negative basis. By integrating the non-negative constraint, label consistence, and orthogonal property into the objective function, the efficient updating procedure can obtain a discriminant basis matrix. Meanwhile, we design a linear classifier using the projected data to guide the label for efficient image classification task. Experiment results of the competitive NMF variants on the challenging digit and face databases demonstrate the effectiveness of the proposed LONMF algorithm.

1. Introduction

Many applications of image processing aim at searching for the representation by using the prior knowledge. A suitable representation plays a fundamental role in the further processing. There have been many works on image representation for the classification task, such as sparse coding [1,2], dictionary learning [3–5], and dimension reduction [6–8]. All of the mentioned methods learn the coefficient or the basis (projection) matrix for minimizing the reconstruction error while boosting the discriminative ability. However, the methods ignore the non-negative property of the related variables. Both of the coefficient and basis matrices derived from the mentioned methods have negative elements. Intuitively, the image handled is non-negative and should be represented by the combination of the non-negative sub images which are expected to indicate the visual parts of the original image. For instance, the face image can be recognized by observing the discriminant parts, such as nose, eyes, and mouth. From this perspective, non-negative property is more consistent with the psychological intuition.

Linear discriminant analysis (LDA) and principal component analysis (PCA) are the famous approaches to dimension reduction [9]. Later, matrix factorization has become popular for data representation. In the real applications of pattern recognition, computer vision, and information retrieval, the original input data is of high dimension which increases the pressure of data processing. Accordingly, the matrix factorization technique aims to decompose the high-dimensional input data matrix into some low-dimensional matrices. Singular value decomposition (SVD) [10], vector quantization (VQ) [11], and non-negative matrix factorization (NMF) [12] are some of representative matrix factorization techniques. SVD decomposes the input data matrix into the left singular vector, the right singular vector, and a diagonal matrix whose diagonal entries indicate the singular values of the input matrix. Representing the input matrix in a low-rank approximation, SVD has been applied to face recognition successfully. VQ maps the input data matrix into binary vectors and always is used for information retrieval task, while NMF aims to search for two matrices whose entries are non-negative and product is approximate to constructing the input data matrix.

There have been many works indicating that the non-negativity constraint leads to a parts-based representation of the data. Some studies have shown that there are psychological and physiological evidences for parts-based representation in the human brain [13,14]. Diverse from the PCA, LDA, VQ, and SVD, NMF only allows additive, not subtractive combination of the input data, and thus it is naturally favor to sparse, parts-based representation which is more robust than non-sparse, global representations.

* Corresponding author.
E-mail addresses: zwenjie@stumail.neu.edu.cn (W. Zhu), yanyh@mail.neu.edu.cn (Y. Yan).

https://doi.org/10.1016/j.image.2018.01.001
Received 30 May 2017; Received in revised form 4 January 2018; Accepted 4 January 2018
Available online 10 January 2018
0923-5965/

Normally, the NMF variants can be categorized into four classes [15], including sparse NMF, orthogonal NMF, manifold NMF and discriminant NMF. Sparse NMF focuses the sparseness property of the matrices. The standard NMF enforces the matrices to be non-negative. Meanwhile, NMF may extract the sparse information. However, NMF is different from sparse coding. Sparse coding learns a full rank representation basically, whilst NMF pursues for a low-rank representation. Furthermore, NMF is not dictionary learning. Dictionary learning aims to optimize an over-complete basis matrix, however, basis matrix generated by NMF is under-complete. Orthogonal NMF can obtain the parts-based representation to boost the psychological and physiological representation ability. Manifold NMF expects to extend the NMF on the manifold structure. This work focuses on the discriminant NMF to enhance the discriminant ability of NMF for image classification tasks.

In the past decade, a number of works related to NMF have been proposed. Li et al. imposed extra constraints to solve the localized and part-based decomposition by extending the standard NMF and this work is called LNMF [16]. To obtain sparse encoding vectors, Hoyer incorporated the sparseness constraint with standard NMF and proposed the non-negative sparse coding (NSC) [17]. In addition, the non-negative property of the sparse constraint has been used in the image classification task [18]. The Fisher-NMF (FNMF) was proposed to encode discriminant information into NMF [19]. The work proposed in [20] extended FNMF by adding an extra term of scatter difference to the objective function of NMF to obtain the discriminant subspace. Employing the data geometric structure, Cai et al. proposed a graph-regularized NMF (GNMF). The geometric structure encoded by k-nearest-neighbor (KNN) measurement is usually used in dimension reduction methods [21]. Li et al. proposed an approach named discriminative orthogonal non-negative matrix factorization (DON) [22], which preserves both the local manifold structure and the global discriminant information simultaneously through manifold discriminant learning. The work first learned a weight matrix in the same manner with GNMF to measure the relationship among the input data, and then computed the central matrix which indicated the discriminant information. With the successful development of deep learning technique, the combination of non-negative constraint and deep learning are proposed recently. Zhang et al. [23] propose a nonlinear NMF learning method which formulates the original data with deep learning technique and then conduct the procedure of NMF. Trigeorgis et al. [24] present a semi-supervised deep matrix factorization with non-negative constraint which can learn a low-dimensional representation of a dataset which is suited for clustering as well as classification. Moreover, Zurada et al. [25] adopts non-negative constraint to regularize the sparse autoencoders in the framework of deep learning. In [25], the recognition result on the famous MNIST digit dataset can achieve 97%+.

The latent discriminant information plays a vital role for the image classification task. However, the unsupervised NMF methods lack of extracting enough discriminant information. The supervised NMF variants only take into account the inner-class and intra-class constraints for the coefficient while ignoring the discriminant constraint for the basis matrix. To this end, this paper proposes a label and orthogonal regularized NMF (LONMF) based on the squared Euclidean distance. LONMF integrates the orthogonal constraint and supervised label information for the basis matrix into the objective function. By doing so, the parts-based representation can extract the discriminant information which is consistent with the label. According to the designed classifier, LONMF fulfills a novel discriminant NMF method in the image classification task.

Most of the contributions of this paper are as follows:

(i) Orthogonal constraint involved into the proposed LONMF guarantees the parts-based representation and indicates the discriminant localization;

(ii) LONMF enforces the label to be consistent with the projected representation which is robust to the discriminant feature extraction and guides the latent discriminant information of data to the right label;

(iii) The designed linear classifier increases the effectiveness of the image classification task. Comparison experiments on the challenging databases well validate the performance of LONMF.

The remainder of this paper is organized as follows: Section 2 briefly introduces the NMF variants; Section 3 describes the proposed algorithm in detail; Section 4 presents the experimental results and Section 5 concludes this paper.

2. Brief review of non-negative matrix factorization variants

For ease of presentation, we introduce the notations used in the whole work. The bold uppercase letter denotes the matrix and the lowercase letter denotes the vector. Given a real $m \times n$ matrix $\boldsymbol{A} = (\boldsymbol{A}_{ij})_{m\times n}$, $a^i \in \mathbb{R}^n (i = 1, 2, \ldots, m)$ and $a_j \in \mathbb{R}^m (j = 1, 2, \ldots, n)$ are respectively the ith row and jth column vectors of $\boldsymbol{A}$. The Frobenius norm of the matrix is denoted as $\|\cdot\|_F$, and $\|\cdot\|_q$ denotes the ℓ_q norm of a vector.

2.1. Standard NMF

Given an input data matrix $\boldsymbol{X} = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^{m\times n}$, NMF aims to search for two non-negative matrices $\boldsymbol{B} = (b_1, b_2, \ldots, b_r) \in \mathbb{R}^{m\times r}$ and $\boldsymbol{C} = (c_1, c_2, \ldots, c_n) \in \mathbb{R}^{r\times n}$ whose product can approach to $\boldsymbol{X}$.

Commonly, there are two common-used criteria measuring the cost function. The first one is the square of Euclidean distance,

$$f_1 = \|\boldsymbol{X} - \boldsymbol{BC}\|_F^2 = \sum_{i,j}(x_{ij} - \sum_{k=1}^{r} b_{ik}c_{kj})^2, \tag{1}$$

and another one is the Kullback–Leibler (KL) divergence [26] between $\boldsymbol{X}$ and $\boldsymbol{BC}$,

$$f_2 = D(\boldsymbol{X} \parallel \boldsymbol{BC}) = \sum_{i,j}(x_{ij} \log \frac{x_{ij}}{(\boldsymbol{BC})_{ij}} - x_{ij} + (\boldsymbol{BC})_{ij}). \tag{2}$$

This paper focuses on NMF based on the square of Euclidean distance. It can be found that NMF searches for a linear combination of $r (r < n)$ non-negative components and the representation coefficients are non-negative too. Lee et al. [12,26] have given the iterative update rules:

$$b_{jk}^{(t+1)} = b_{jk}^{(t)} \frac{(\boldsymbol{XC}^T)_{jk}}{(\boldsymbol{BCC}^T)_{jk}}, c_{jk}^{(t+1)} = c_{jk}^{(t)} \frac{(\boldsymbol{B}^T\boldsymbol{X})_{kj}}{(\boldsymbol{B}^T\boldsymbol{BC})_{kj}}. \tag{3}$$

2.2. Local NMF

To learn the spatially localized and parts-based representation of the data, a local NMF (LNMF) algorithm was proposed in [16]. The objective function of LNMF based on KL divergence is defined as follows:

$$f_{LNMF} = f_2 + \lambda_1 \sum_{i,j}(\boldsymbol{B}^T\boldsymbol{B})_{ij} - \lambda_2 \sum_{i}(\boldsymbol{CC}^T)_i, \tag{4}$$

where λ_1 and λ_2 are the positive parameters. LNMF minimizes the number of components of decomposition while retaining the components giving most important information.

2.3. Discriminant NMF

Different from LNMF, DNMF [20] introduced discrimination criterion into the objective function,

$$f_{DNMF} = f_2 + \lambda_1 tr(S_w) - \lambda_2 tr(S_b), \tag{5}$$

where $tr(\cdot)$ indicates the trace of the matrix, and S_w and S_b are within-class scatter matrix and between-class scatter matrix of $\boldsymbol{C}$, respectively.

$$\begin{aligned} S_w &= \sum_{r=1}^{c}\sum_{j=1}^{n_r}(x_{r,j} - \mu^r)(x_{r,j} - \mu^r)^T, \\ S_b &= \sum_{r=1}^{c} n_r(\mu^r - \mu)(\mu^r - \mu)^T, \end{aligned} \tag{6}$$

where c is the number of classes, n_r is the number of samples belonging to the rth class, $x_{r,j}$ is the jth sample belonging to the rth class, and μ^r and μ denote the mean vector of the rth class and the whole class, respectively. By involving the maximum margin criterion (MMC) acting on the coefficient, the basis matrix of DNMF is same with that of the standard NMF.

3. Label and orthogonality regularized non-negative matrix factorization

In this section, we introduce the proposed label and orthogonality regularized non-negative matrix factorization (LONMF). Based on the analysis of the related works to NMF, we first state the basic idea of LONMF. Then, we give the update rules derived from the standard NMF framework with two optimization methods. At last, we analyze the convergence of the proposed algorithm.

3.1. Basic idea of LONMF

In the general processing of NMF for data representation, the basis matrix $\boldsymbol{B}$ is used to map the original data $y \in \mathbb{R}^m$ into a low-dimensional representation $y^* = \boldsymbol{B}^T y$. It is vital to learn a discriminant basis matrix while employing NMF technique to accomplish the image classification task. The related NMF variants ignore the discriminant regularized to the basis matrix. LNMF only requires the basis matrix to generate the localized and parts-based representation. DNMF takes into account the discriminant criterion of the coefficient matrix. Similarly, DONMF designs a complex regularization of the coefficient matrix in local and global manner. The basis matrix takes a construction role in the related NMFs.

Considering the problem mentioned, this paper aims to adopt the label and orthogonal regularization acting on the basis matrix directly to define the objective function:

$$f_{LONMF} = f_1 + \alpha f_L + \beta f_O, \tag{7}$$

where α and β are the positive constants. We introduce the label information into the second term of the objective function. As analyzed before, the basis matrix is directly responsible for acting on the original data. Hence, we have $f_L = \|\boldsymbol{B}^T\boldsymbol{X} - \boldsymbol{H}\|_F^2$ to enforce the projected low-dimensional representation to be consistent with the corresponding label involved in $\boldsymbol{H}$. At last, we add the orthogonal regularized term $f_O = \|\boldsymbol{B}^T\boldsymbol{B} - \boldsymbol{I}\|_F^2$ ($\boldsymbol{I}$ is the identity matrix with proper size).

Combined with the constraints of NMF, the proposed LONMF aims to optimize the following objective function,

$$\begin{aligned} &\min_{\boldsymbol{B},\boldsymbol{C}} \|\boldsymbol{X} - \boldsymbol{B}\boldsymbol{C}\|_F^2 + \alpha\|\boldsymbol{B}^T\boldsymbol{X} - \boldsymbol{H}\|_F^2 + \beta\|\boldsymbol{B}^T\boldsymbol{B} - \boldsymbol{I}\|_F^2 \\ &s.t.\ \boldsymbol{B} \geqslant \boldsymbol{0}, \boldsymbol{C} \geqslant \boldsymbol{0}. \end{aligned} \tag{8}$$

In (8), the constraint condition indicates all the elements of $\boldsymbol{B}$ and $\boldsymbol{C}$ are non-negative, and $\boldsymbol{H}$ is the label matrix. $\boldsymbol{H} = (h_1, h_2, \ldots, h_n) \in \mathbb{R}^{r\times n}$ is a sufficiently sparse matrix. The sparse vector $h_k (k = 1, 2, \ldots, n)$ indicates the desired form of the low-dimensional representation. Normally, the element of h_k is 1 that indicates the category of the data, otherwise 0. However, it has a latent condition that the dimension of the basis matrix should equal the number of classes, $r = c$, where c denotes the number of classes. To make the choice of r flexible for the real scenario, we set $r = Kc$, where K is a positive integer. Hence, $\boldsymbol{H}$ is the block-label matrix in this work since it has a block distribution and also indicates the label information. By doing so, the vector h_k is divided into c parts. For each part, the K elements are set to 1 to indicate the class label of the data. The label matrix can be generated by employing the Kronecker product of the original label matrix when $r = c$ and an all ones vector with the size of K. The label consistency term fully exploits the discrimination embedded in data and enforces the input samples belong to the same class to have similar low-dimensional representations.

3.2. Iterative updating algorithm

The objective function of LONMF is not convex with regard to the variables $\boldsymbol{B}$ and $\boldsymbol{C}$. The common approach to the problem consists of two categories: multiplicative approach and the gradient descend (GD) method [27]. The two approaches are specified in Appendix A.

From the two methods, we can find that multiplicative approach is a special case of the additive one by setting the update step size. According to [26], gradient descent is perhaps the simplest technique to implement, but convergence can be slow. The convergence of gradient based methods also has the disadvantage of being very sensitive to the choice of step size, which can be very inconvenient for large applications. In this paper, we adopt the multiplicative update rules for optimizing LONMF.

3.3. LONMF for image classification

Given a testing sample $y \in \mathbb{R}^m$, the basis matrix derived from LONMF maps it into a low-dimensional representation $y^* = \boldsymbol{B}^T y$. Before using the training low-dimensional samples to categorize y^*, we design a simple linear classifier using the training low-dimensional samples $\boldsymbol{X}^* = \boldsymbol{B}^T\boldsymbol{X}$. The linear classifier $\boldsymbol{W}$ is obtained via minimizing the cost function as follows,

$$\min_{\boldsymbol{W}} \|\boldsymbol{W}\boldsymbol{X}^* - \boldsymbol{Q}\|_F^2 + \gamma\|\boldsymbol{W}\|_F^2, \tag{9}$$

where γ is a small positive number, and the second term is added into the cost function to prevent from arbitrary solution. $\boldsymbol{Q} \in \mathbb{R}^{c\times n}$ is the label matrix which is also used in LDA as the spectral matrix. The problem (9) is convex optimization w.r.t. the variable, and has a analyzed solution,

$$\begin{aligned} \boldsymbol{W} &= \boldsymbol{Q}\boldsymbol{X}^{*T}(\boldsymbol{X}^*\boldsymbol{X}^{*T} + \gamma\boldsymbol{I})^{-1} \\ &= \boldsymbol{Q}\boldsymbol{X}^T\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B} + \gamma\boldsymbol{I})^{-1}. \end{aligned} \tag{10}$$

The class label of the testing sample y is determined by searching for the index of the maximum value:

$$l(y) = \max_k (\boldsymbol{W}\boldsymbol{B}^T y)_k. \tag{11}$$

The procedure of the proposed LONMF for image classification is outlined in Algorithm 1. Furthermore, the convergence analysis of LONMF algorithm is conducted in Appendix B.

4. Experimental results and analysis

In this section, we will evaluate the proposed LONMF algorithm for image classification on the challenging MNIST digit database [28] and face databases: ORL [29], YALE [30], FERET [31], and CMU PIE [32].

4.1. Baseline and setting

The compared methods used in this experiment including famous subspace learning algorithm LDA, and the five representative NMF variants: NMF, LNMF, DNMF_KL, PGDNMF_KL, and MD-NMF_KL. Moreover, we also compare the NMF variants with the discriminative dictionary learning method named LC-KSVD which involves sparse coding and K-SVD learning. It is worth noting that the discriminative algorithms are formulated with KL divergence. Both LC-KSVD1 and LC-KSVD2 proposed in [4] are adopted. We set the size of dictionary with different numbers according to the character of data set and report the best results. The number of feature dimension of LDA is set as the same one reported in [29]. As for the parameters of DNMF, we choose the them in the range of $[0.1, 0.5]$ as reported in [20]. The selection of the parameters of PGDNMF [33] is all the same with DNMF. Since there are three parameters involved in MD-NMF [34], we set them as the selection reported in that $\alpha = 10^{-2}$, $\beta = 10^{-1}$, and $\gamma = 10^2$. The parameter γ used in LNOMF is 10^{-2}. The other parameters α, β, K are tuned by cross validation from $\{10^{-2}, 5\times10^{-2}, \ldots, 1\}$, $\{10^{-2}, 5\times10^{-2}, \ldots, 1\}$, and

Algorithm 1. Procedure of LONMF for image classification.

Input: training image sets $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ and the corresponding label matrix $\boldsymbol{Q} \in \mathbb{R}^{c\times n}$ ($\boldsymbol{H} \in \mathbb{R}^{r\times n}$), parameters α, β, K, and the testing image $y \in \mathbb{R}^{m}$;

Initialize the non-negative matrices $\boldsymbol{B}^0 \in \mathbb{R}^{m\times r}$, $\boldsymbol{C}^0 \in \mathbb{R}^{r\times n}$ and $t = 0$;

Repeat for fixed number of iterations or until convergence:

Loop

Update $\boldsymbol{B}^{(t+1)}$: $b_{ij}^{(t+1)} \leftarrow b_{ij}^{(t)} \cdot (\boldsymbol{X}\boldsymbol{C}^{(t)T} + \alpha\boldsymbol{X}\boldsymbol{H}^T + 2\beta\boldsymbol{B}^{(t)})_{ij}/(\boldsymbol{B}^{(t)}\boldsymbol{C}^{(t)}\boldsymbol{C}^{(t)T} + \alpha\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B}^{(t)} + 2\beta\boldsymbol{B}^{(t)}\boldsymbol{B}^{(t)T}\boldsymbol{B}^{(t)})_{ij}$;

Update $\boldsymbol{C}^{(t+1)}$: $c_{ij}^{(t+1)} \leftarrow c_{ij}^{(t)} \cdot (\boldsymbol{B}^{(t)T}\boldsymbol{X})_{ij}/(\boldsymbol{B}^{(t)T}\boldsymbol{B}^{(t)}\boldsymbol{C}^{(t)})_{ij}$;

End loop

Compute the linear classifier: $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{X}^T\boldsymbol{B}(\boldsymbol{B}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{B} + \gamma\boldsymbol{I})^{-1}$;

Output:the label of the testing image: $l(y) = \max_k(\boldsymbol{W}\boldsymbol{B}^T y)_k$.

Table 1
Parameters of LONMF used in different databases for image classification.

Parameters	α	β	K
MNIST	0.5	1.0	5
YALE	0.5	1.0	3
ORL	0.5	1.0	2
FERET	0.01	0.05	1
CMU PIE	0.5	0.5	5

$\{1, 2, \ldots, 10\}$. The best parameters used for the corresponding database are shown in Table 1.

The max iteration of all the compared methods except LDA is 200. For the NMF variants, the standard NMF and LNMF is unsupervised and others employ the label information to conduct the supervised NMF for the image classification task. For DNMF, PGDNMF, and MD-NMF, the with-in/between class scatter matrix are computed employing the label information. While LONMF generates the (block-)label matrix employing the label information. The algorithms are implemented in Matlab R2014a and run on desktop PC with 3.5 GHz Intel CPU and 8 GB memory.

We first conduct the experiment on the MNIST digit database and analyze the convergence, parameter selection, and basis matrix at same time. Then, we evaluate the algorithms on the four face databases and discuss the results of the competitive algorithms. Meanwhile, the basis matrix and ROC analysis are also operated by taking the ORL face dataset as an example.

4.2. MNIST digit database

In this subsection, we conduct experiments on the MNIST database to verify the performance of the proposed LONMF algorithm for hand-written digit recognition task. Employing this database, we analyze the convergence, parameters selection and basis matrix.

The MNIST database contains 60,000 training images and 10,000 test images, both drawn from the same distribution. Some example images are shown in Fig. 1. From Fig. 1, we can see that images of ten classes with the number (0–9) written by different persons are variable. Diverse from the face images, the digit images used are sparse. Most of the entries of the digit images are zero. Hence, the competitive algorithms attempt to learn the patterns using the limited information actually. All these images are size normalized to 28 × 28 pixels. We randomly select 3000 and 4000 images from 60,000 training samples

Fig. 1. Example images collected from MNIST digit database.

to construct the training set (*ntr* = 3000, 4000, where *ntr* denotes the number of the training samples for each class)and randomly select 5000 test samples to construct the test set, respectively.

As proved in the previous sections, the proposed LONMF convergences using the update rules. Here, we experimentally show the convergent speed of LONMF on the MNIST database. Fig. 2 shows the decrease cost function versus the number of iterations. In Fig. 2, the number of iterations is shown on the x-axis and the value of cost function is shown on the y-axis. We can see that the value of the cost function of LONMF algorithm drops very fast in the early iterations and begins to get stable at the 40th iteration.

Table 2 illustrates the classification accuracy of the MNIST database. From Table 2, we can see that the proposed methods has the best recognition rate among all the compared dimension reduction methods, and LC-KSVD1 obtains the best classification result which is marked in red. The result marked in blue denotes the second best one. However, LC-KSVD takes a longer time than the proposed LONMF in training.

Considering the visual character of the digit pattern, we employ this database to present the basis matrix derived from NMF, LNMF, and the proposed LONMF. Before this, we first show the classification accuracy versus the parameters. The parameters are selected with others fixed as setting in Table 1. Fig. 3 shows the digit recognition accuracy versus the parameters α, β, and K. From Fig. 3, we can see that the classification accuracy varies between 72.5% and less than 73.5%, 72% and less than 73.2% versus α and β, respectively. Besides, the classification accuracy

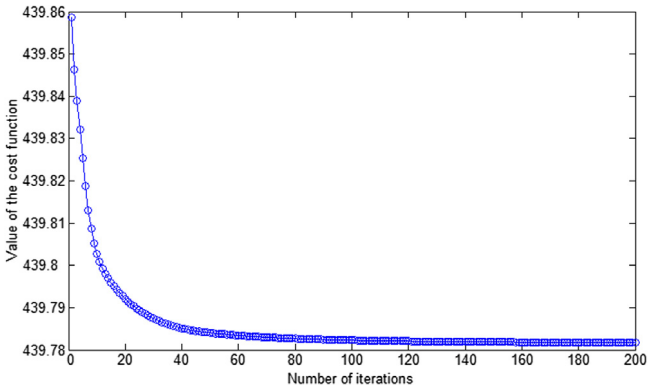


Fig. 2. The value of the cost function versus the number of iterations on the MNIST digit database.

improves little when $K \geqslant 5$. Hence, it is apt to select the applicable parameters.

The visual basis matrices derived from NMF, LNMF, and the proposed LONMF are presented in Fig. 4. We select the first 50 vectors of the basis matrix and transform them into matrix with the same size of the digit image. From the basis matrices shown in Fig. 4, we can find that NMF keeps the most information of the input image while LNMF generates a sparse basis matrix. We can hardly recognize any information consistent with physiology. The basis obtained by the proposed LONMF is sparser than that of NMF while denser than that of LNMF. Obviously, LONMF keeps some parts-based representations and enhances the discriminant information of the input image.

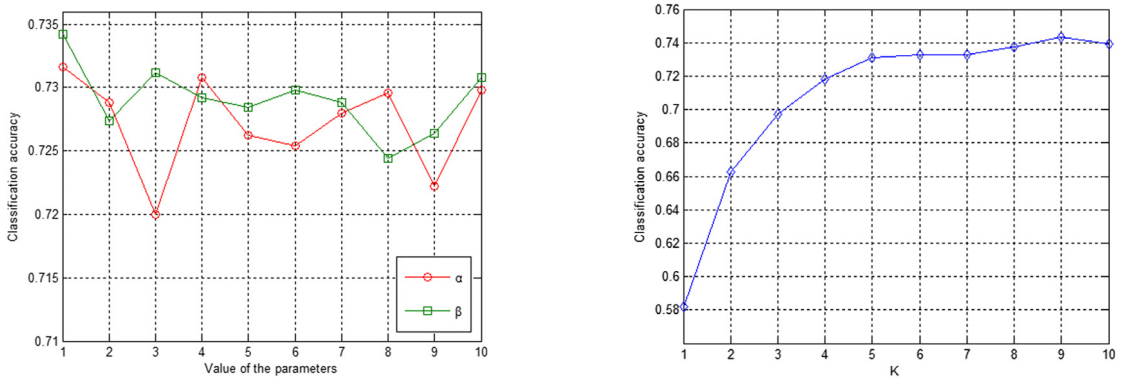
Table 2
MNIST digit recognition results (average recognition rate standard deviation)% of the competitive algorithms.

Methods	$ntr = 3000$	$ntr = 4000$
LDA	74.22 ± 2.05	77.34 ± 1.69
NMF	70.28 ± 3.11	72.06 ± 4.51
LNMF	71.24 ± 4.15	74.24 ± 3.52
DNMF_KL	73.90 ± 1.62	76.18 ± 3.68
PGDNMF_KL	72.10 ± 4.56	74.20 ± 2.24
MD-NMF_KL	75.32 ± 5.12	78.06 ± 4.16
LONMF	82.26 ± 0.61	84.20 ± 0.39
LC-KSVD1	82.39 ± 0.99	83.45 ± 0.75
LC-KSVD2	81.51 ± 1.15	82.59 ± 0.60

4.3. ORL face database

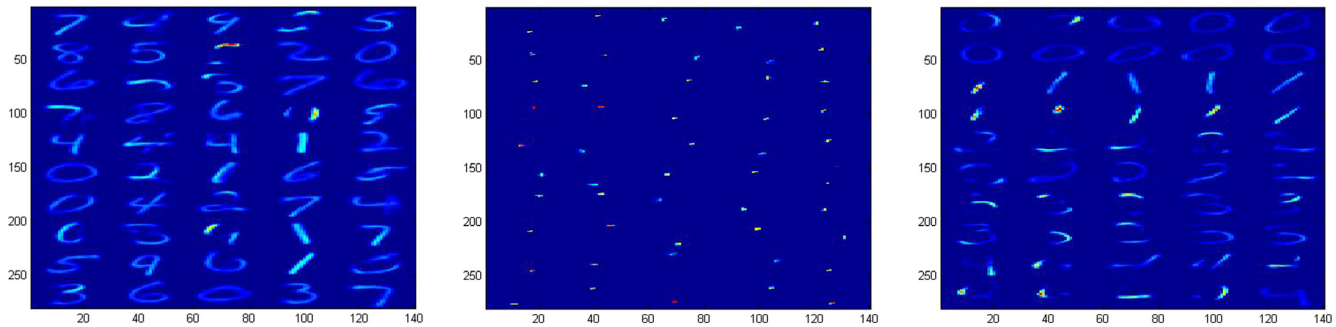
There are 400 frontal (with tolerance for some side movement) face images of 40 individuals for the ORL face database. For each individual, a total number of 10 face images are taken in the same dark background. The original face images are normalized to 40×40 pixels and some of the example images used in the experiment are shown in Fig. 5.

We randomly select $ntr(4, 5)$ images for each individual as training samples. The rest of the images are used for testing. Table 3 shows the average recognition accuracy and the standard deviations of the competing algorithms. It can be seen that both LONMF and LC-KSVD1 (LC-KSVD2) achieves much better results than other competitive methods. Both LC-KSVD1 and LC-KSVD2 obtains top 2 best result when $nr = 4$, and the proposed LONMF gets the second best result when $nr = 5$. Generally speaking, LONMF takes less training time than LC-KSVDs with similar results. The visualized basis matrices learned by NMF, LNMF, and the proposed LONMF are expressed in Fig. 6. Similar observation



(a) α and β. (b) K.

Fig. 3. Parameters (a) α and β and (b) K selected and the corresponding classification accuracy on the MNIST database with $ntr = 4000$.



(a) (b) (c)

Fig. 4. Visual basis matrices derived from (a) NMF, (b) LNMF, and (c) LONMF on the MNIST database with $ntr = 4000$.

Fig. 5. Example images from the ORL face database.

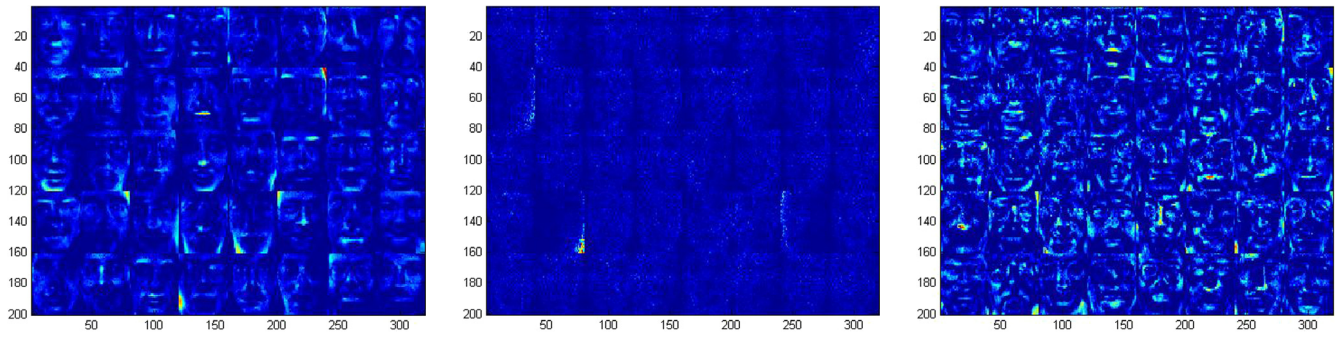


Fig. 6. Visual basis matrices derived from (a) NMF, (b) LNMF, and (c) LONMF on the ORL database with $ntr = 5$.

Table 3
ORL face recognition results (average recognition rate standard deviation)% of the competing algorithms.

Methods	$ntr = 4$	$ntr = 5$
LDA	87.08 ± 3.74	88.15 ± 3.22
NMF	83.75 ± 2.38	84.90 ± 1.87
LNMF	84.16 ± 4.16	85.80 ± 2.66
DNMF_KL	85.66 ± 2.63	86.77 ± 5.11
PGDNMF_KL	84.64 ± 3.36	87.64 ± 4.61
MD-NMF_KL	87.18 ± 3.12	90.41 ± 2.88
LONMF	90.00 ± 1.34	92.05 ± 1.51
LC-KSVD1	90.12 ± 1.66	91.52 ± 1.36
LC-KSVD2	90.71 ± 1.48	92.75 ± 1.60

Table 4
YALE face recognition results (average recognition rate standard deviation)% of the competing algorithms.

Methods	$ntr = 4$	$ntr = 5$
LDA	76.22 ± 6.68	80.22 ± 3.18
NMF	71.33 ± 3.72	77.33 ± 4.12
LNMF	74.25 ± 3.13	78.25 ± 2.13
DNMF_KL	76.21 ± 2.15	80.21 ± 3.25
PGDNMF_KL	75.66 ± 4.23	79.66 ± 1.63
MD-NMF_KL	80.18 ± 2.65	82.18 ± 5.25
LONMF	82.86 ± 2.92	86.11 ± 4.04
LC-KSVD1	83.75 ± 4.89	86.00 ± 4.61
LC-KSVD2	83.76 ± 4.95	86.11 ± 4.48

with MNIST, the proposed LONMF achieves more sparse and localized bases.

4.4. YALE face database

There are 165 frontal view face images of 15 individuals for the YALE face database. For each individual, a total number of 11 face images with different facial expression or configuration: center-light, with/without glasses, happy, left-light, normal, right-light, sad, sleepy, surprised, and wink. The original face images are cropped and normalized to 40×40 pixels. Some of the example images used in the experiment are shown in Fig. 7.

In the same way, $ntr(4, 5)$ images are selected for each individual as training samples. The rest of the images are used for testing. Table 4 shows the average recognition accuracy and the standard deviations of the competing algorithms. LDA, a discriminant analysis tool, performs better than most of the NMF algorithms. However, the manifold-based NMF and the proposed LONMF perform better than other discriminative NMF algorithms. Similar with the results on the ORL database, LC-KSVD2 gets the best result when $nr = 4$ and $nr = 5$.

4.5. FERET face database

The facial recognition technology (FERET) database consists of 1400 images from 200 individuals. For each subject, there are 7 images and $ntr(4, 5)$ of them are selected randomly for training and the rest for testing. All images are normalized to 40×40 pixels. Fig. 8 shows the examples of images from FERET database.

Table 5
FERET face recognition results (average recognition rate standard deviation)% of the competing algorithms.

Methods	$ntr = 4$	$ntr = 5$
LDA	64.17 ± 2.05	61.75 ± 3.33
NMF	65.10 ± 3.43	60.45 ± 2.44
LNMF	66.47 ± 3.99	63.15 ± 2.65
DNMF_KL	67.89 ± 3.62	66.20 ± 3.89
PGDNMF_KL	67.01 ± 4.57	67.10 ± 4.36
MD-NMF_KL	68.84 ± 2.48	68.70 ± 2.95
LONMF	71.12 ± 6.18	73.30 ± 7.27
LC-KSVD1	60.75 ± 6.89	68.25 ± 8.19
LC-KSVD2	60.56 ± 6.95	67.00 ± 7.23

Table 5 shows the average recognition accuracy and the standard deviations of the competing algorithms. It can be seen obviously that the proposed LONMF outperforms the other algorithms. It is worth noting that LC-KSVDs performs worse than others on FERET face database. It is possible that there is much more classes (200) than other database, while much less images (only 7) per class. Thus, it is difficult to capture the discriminative information among the sub dictionaries.

Fig. 7. Example images from the YALE face database.

Fig. 8. Example images from the FERET face database.

Fig. 9. Example images from the CMU PIE face database.

4.6. CMU PIE face database

The CMU PIE database contains 41,368 face images collected from 68 subjects. Each subject has 13 images of different poses, 43 different illumination conditions, and with 4different expressions. In this experiment, we select a subset of 5 near frontal poses (C05, C07, C09, C27, and C29) and illuminations indexed as 08 and 11. Therefore, each subject has ten images. All images were normalized to 32 × 32 pixel array and reshaped to a vector. Fig. 9 shows the face examples of the CMU PIE face database.

Table 6 lists the performance of different methods on the CMU PIE database. In the experiments, 4 and 5 images of each individual were randomly selected and used as training set, and the rest of images were used as test set. As can be seen from Table 6, LONMF obtains the best recognition rates in all the cases when there are variations in pose and illumination among the dimension reduction methods. In addition, the discriminative dictionary learning algorithm LC-KSVD2 performs better than LONMF. However, the training time of LC-KSVDs is much more than the dimension reduction methods.

From the experiment results on the face databases, we can see that LDA, as the discriminant analysis technique, performs better than NMF and LNMF. In the family of NMF, DNMF and PGDNMF perform better than NMF and LNMF due to that DNMF and PGDNMF combine discriminant information in nonnegative factorization. Although taking into account the discriminant constraint, LDA underperforms DNMF, PGNMF, and MD-NMF. DNMF, PGDNMF and MD-NMF all encode discriminant information for classification. However, MD-NMF performs better than DNMF and PGDNMF in our experiments, due to the reason that MD-NMF not only introduces marginal information to NMF, but also introduces manifold structure of the data in the learning steps. LONMF regularizes the basis matrix directly and involves the label information into the optimization. The label regularized term enforces the basis matrix to extract the latent discriminant information for image classification. Furthermore, the discriminative dictionary learning methods obtains a relatively similar results with the proposed LONMF. However, dictionary learning has a complexity computation time on training.

Table 6
CMU PIE face recognition results (average recognition rate standard deviation)% of the competing algorithms.

Methods	$ntr = 4$	$ntr = 5$
LDA	84.51 ± 2.87	91.33 ± 2.33
NMF	78.18 ± 3.41	82.16 ± 3.31
LNMF	80.11 ± 6.25	84.81 ± 2.80
DNMF_KL	85.90 ± 5.89	91.38 ± 4.67
PGDNMF_KL	86.00 ± 5.21	92.61 ± 5.52
MD-NMF_KL	86.33 ± 6.67	94.26 ± 5.81
LONMF	87.72 ± 3.00	95.68 ± 3.84
LC-KSVD1	87.62 ± 3.09	95.41 ± 3.77
LC-KSVD2	88.24 ± 3.19	95.76 ± 3.94

It is difficult to reconstruct the model and tune the parameters of deep learning works [23–25] on the same data set with same settings, however, it should be pointed that deep learning with non-negative constraint methods on data representation for clustering and classification have achieved the state-of-the-art performance.

5. Conclusion

We propose a discriminant NMF algorithm with label and orthogonality regularization for image classification in this paper. We directly formulate the objective function by regularizing the basis matrix with the label consistent and orthogonal constraints. By using the update

rules, our LONMF convergences. The objective function of LONMF enforces the basis matrix to extract the latent information which has similar representation with the data from same class. Furthermore, we design a linear classifier using the LONMF-generated low-dimensional representation for the image classification task. Experimental results on the digit database and the challenging face databases demonstrate that the proposed LONMF algorithm is overall superior to the famous competing algorithms. In the next work, we will study for the deep learning involved NMF for image representation.

Acknowledgments

This work is supported by the National Key Research and Development Program of China (2017YFB0304200), the National Natural Science Foundation of China (51374063), and the Fundamental Research Funds for the Central Universities (N141008001, N150308001).

Appendix A. Update rules using multiplicative approach and GD method

This section describes the update rules using multiplicative approach and GD method.

(i) Multiplicative approaches to LONMF

The objective function in (8) can be rewritten as follows,

$$\begin{aligned} f_{LONMF} &= tr((\boldsymbol{X}-\boldsymbol{BC})(\boldsymbol{X}-\boldsymbol{BC})^T) \\ &+ \alpha tr((\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})(\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})^T) \\ &+ \beta tr(\boldsymbol{B}^T\boldsymbol{B}-\boldsymbol{I})^2. \end{aligned} \tag{12}$$

Then, we transform the constrained optimization into unconstrained one by introducing the Lagrangian multipliers $(\boldsymbol{\Phi}_B)_{ij}$ and $(\boldsymbol{\Phi}_C)_{ij}$ for $b_{ij} \geqslant 0$ and $c_{ij} \geqslant 0$, respectively. The Lagrange function is expressed as follows

$$\begin{aligned} &L(\boldsymbol{B},\boldsymbol{C},\boldsymbol{\Phi}_B,\boldsymbol{\Phi}_C) = \\ &tr((\boldsymbol{X}-\boldsymbol{BC})(\boldsymbol{X}-\boldsymbol{BC})^T) + \alpha tr((\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})(\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})^T) \\ &+ \alpha tr((\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})(\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})^T) + \beta tr(\boldsymbol{B}^T\boldsymbol{B}-\boldsymbol{I})^2 \\ &+ tr(\boldsymbol{\Phi}_B\boldsymbol{B}^T) + tr(\boldsymbol{\Phi}_C\boldsymbol{C}^T). \end{aligned} \tag{13}$$

The partial derivatives of the Lagrange function w.r.t. $\boldsymbol{B}$ and $\boldsymbol{C}$ are,

$$\begin{aligned} \frac{\partial L}{\partial \boldsymbol{B}} &= 2\boldsymbol{BCC}^T - 2\boldsymbol{XC}^T + 2\alpha\boldsymbol{XX}^T\boldsymbol{B} - 2\alpha\boldsymbol{XH}^T \\ &+ 4\beta\boldsymbol{BB}^T\boldsymbol{B} - 4\beta\boldsymbol{B} + \boldsymbol{\Phi}_B, \\ \frac{\partial L}{\partial \boldsymbol{C}} &= 2\boldsymbol{B}^T\boldsymbol{BC} - 2\boldsymbol{B}^T\boldsymbol{X} + \boldsymbol{\Phi}_C. \end{aligned} \tag{14}$$

Enforcing the partial derivatives to be zero and employing the Karush–Kuhn–Tucker condition $\boldsymbol{\Phi}_B \odot \boldsymbol{B} = \boldsymbol{0}$ and $\boldsymbol{\Phi}_C \odot \boldsymbol{C} = \boldsymbol{0}$. We have

$$\begin{aligned} &(\boldsymbol{BCC}^T)_{ij}b_{ij} + \alpha(\boldsymbol{XX}^T\boldsymbol{B})_{ij}b_{ij} + 2\beta(\boldsymbol{BB}^T\boldsymbol{B})_{ij}b_{ij} \\ &= (\boldsymbol{XC}^T)_{ij}b_{ij} + \alpha(\boldsymbol{XH}^T)_{ij}b_{ij} + 2\beta\boldsymbol{B}_{ij}b_{ij} \end{aligned} \tag{15}$$

and

$$(\boldsymbol{B}^T\boldsymbol{BC})_{ij}c_{ij} - (\boldsymbol{B}^T\boldsymbol{X})_{ij}c_{ij} = 0. \tag{16}$$

Combined (15) and (16), we obtain the following update rules:

$$b_{ij} \leftarrow b_{ij}\frac{(\boldsymbol{XC}^T + \alpha\boldsymbol{XH}^T + 2\beta\boldsymbol{B})_{ij}}{(\boldsymbol{BCC}^T + \alpha\boldsymbol{XX}^T\boldsymbol{B} + 2\beta\boldsymbol{BB}^T\boldsymbol{B})_{ij}}, \tag{17}$$

$$c_{ij} \leftarrow c_{ij}\frac{(\boldsymbol{B}^T\boldsymbol{X})_{ij}}{(\boldsymbol{B}^T\boldsymbol{BC})_{ij}}. \tag{18}$$

(ii) GD method approaches to LONMF

In the related work to DNMF, the GD algorithm was used to optimize the problem (8). According to GD method, the additive rules are

$$\begin{aligned} b_{ij} &\leftarrow b_{ij} + \mu_{ij}\frac{\partial L}{\partial b_{ij}}, \\ c_{ij} &\leftarrow c_{ij} + \eta_{ij}\frac{\partial L}{\partial c_{ij}}. \end{aligned} \tag{19}$$

where μ_{ij} and η_{ij} control the step size of GD. Now we set the parameters as following,

$$\begin{aligned} \mu_{ij} &= \frac{-b_{ij}}{(2\boldsymbol{BCC}^T + 2\alpha\boldsymbol{XX}^T\boldsymbol{B} + 4\beta\boldsymbol{BB}^T\boldsymbol{B})_{ij}}, \\ \eta_{ij} &= \frac{-c_{ij}}{(2\boldsymbol{B}^T\boldsymbol{BC})_{ij}}. \end{aligned} \tag{20}$$

Combined (19) and (20), we have

$$\begin{aligned} &b_{ij} + \mu_{ij}\frac{\partial L}{\partial b_{ij}} = \\ &b_{ij} - \frac{b_{ij}}{(2\boldsymbol{BCC}^T + 2\alpha\boldsymbol{XX}^T\boldsymbol{B} + 4\beta\boldsymbol{BB}^T\boldsymbol{B})_{ij}}\frac{\partial L}{\partial b_{ij}} \\ &= b_{ij} - \frac{b_{ij}(\boldsymbol{M})_{ij}}{(2\boldsymbol{BCC}^T + 2\alpha\boldsymbol{XX}^T\boldsymbol{B} + 4\beta\boldsymbol{BB}^T\boldsymbol{B})_{ij}} \\ &= b_{ij}\frac{(\boldsymbol{XC}^T + \alpha\boldsymbol{XH}^T + 2\beta\boldsymbol{B})_{ij}}{(\boldsymbol{BCC}^T + \alpha\boldsymbol{XX}^T\boldsymbol{B} + 2\beta\boldsymbol{BB}^T\boldsymbol{B})_{ij}}, \end{aligned} \tag{21}$$

where $\boldsymbol{M} = 2\boldsymbol{BCC}^T - 2\boldsymbol{XC}^T + 2\alpha\boldsymbol{XX}^T\boldsymbol{B} - 2\alpha\boldsymbol{XH}^T + 4\beta\boldsymbol{BB}^T\boldsymbol{B} - 4\beta\boldsymbol{B} + \boldsymbol{\Phi}_B$, and

$$\begin{aligned} &c_{ij} + \eta_{ij}\frac{\partial L}{\partial c_{ij}} \\ &= c_{ij} - \frac{c_{ij}}{(2\boldsymbol{B}^T\boldsymbol{BC})_{ij}}\frac{\partial L}{\partial C_{ij}} \\ &= c_{ij} - \frac{c_{ij}(2\boldsymbol{B}^T\boldsymbol{BC} - 2\boldsymbol{B}^T\boldsymbol{X} + \boldsymbol{\Phi}_C)_{ij}}{(2\boldsymbol{B}^T\boldsymbol{BC})_{ij}} \\ &= c_{ij}\frac{(\boldsymbol{B}^T\boldsymbol{X})_{ij}}{(\boldsymbol{B}^T\boldsymbol{BC})_{ij}}. \end{aligned} \tag{22}$$

Appendix B. Convergence proof of LONMF algorithm

The update rules given by (17) and (18) make the iterative process of LONMF converge under the condition that f_{LONMF} in (12) is nonincreasing using the rules. Define the part of f_{LONMF} w.r.t. $\boldsymbol{B}$ and $\boldsymbol{C}$. We have

$$\begin{aligned} f_{LONMF}(\boldsymbol{B}) &= tr((\boldsymbol{X}-\boldsymbol{BC})(\boldsymbol{X}-\boldsymbol{BC})^T) \\ &+ \alpha tr((\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})(\boldsymbol{B}^T\boldsymbol{X}-\boldsymbol{H})^T) + \beta tr(\boldsymbol{B}^T\boldsymbol{B}-\boldsymbol{I})^2 \\ &= tr(\boldsymbol{XX}^T) + tr(\boldsymbol{BCC}^T\boldsymbol{B}^T) - 2tr(\boldsymbol{XC}^T\boldsymbol{B}^T) + \alpha tr(\boldsymbol{HH}^T) \\ &+ \alpha tr(\boldsymbol{B}^T\boldsymbol{XX}^T\boldsymbol{B}) - 2\alpha tr(\boldsymbol{B}^T\boldsymbol{XH}^T) + \beta tr(\boldsymbol{B}^T\boldsymbol{B}-\boldsymbol{I})^2 \\ &= tr(\boldsymbol{BCC}^T\boldsymbol{B}^T) - 2tr(\boldsymbol{XC}^T\boldsymbol{B}^T) + \alpha(\boldsymbol{B}^T\boldsymbol{XX}^T\boldsymbol{B}) \\ &- 2\alpha tr(\boldsymbol{B}^T\boldsymbol{XH}^T) + \beta tr((\boldsymbol{B}^T\boldsymbol{B})^2) - 2\beta\boldsymbol{B}^T\boldsymbol{B} + constant\,B \end{aligned} \tag{23}$$

and

$$\begin{aligned} f_{LONMF}(\boldsymbol{C}) &= tr((\boldsymbol{X}-\boldsymbol{BC})(\boldsymbol{X}-\boldsymbol{BC})^T) \\ &= tr(\boldsymbol{XX}^T) + tr(\boldsymbol{BCC}^T\boldsymbol{B}^T) - 2tr(\boldsymbol{BCX}^T) \\ &= tr(\boldsymbol{BCC}^T\boldsymbol{B}^T) - 2tr(\boldsymbol{BCX}^T) + constant C, \end{aligned} \tag{24}$$

where $constant\,B$ and $constant C$ are the constant variables w.r.t. the part function of $\boldsymbol{B}$ and $\boldsymbol{C}$.

Then, we should prove that $f_{LONMF}(\boldsymbol{B})$ and $f_{LONMF}(\boldsymbol{C})$ are non-increasing under the update rules. According to **Lemma** 1 [35,36], we should generate the auxiliary function and prove that the update rules are satisfied to (25).

Lemma 1. *If there is an auxiliary function $z(s, s^{(t)})$ for $f(s)$ under the conditions $f(s) \leqslant z(s, s^{(t)})$, $f(s) = z(s, s)$, then f is nonincreasing using the update*

$$s^{(t+1)} = arg \min_{s} z(s, s^{(t)}). \tag{25}$$

Proof. We give priority to $f(s) = z(s, s)$, and the corresponding auxiliary functions (we denote f instead of f_{LONMF} for formulation) are as follows:

$$\begin{aligned} & z(b_{ij}, b_{ij}^{(t)}) \\ & = f(b_{ij}^{(t)}) + f'(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)}) \\ & + \frac{1}{2} f''(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)})^2 \\ & + \frac{1}{3!} f'''(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)})^3 + \frac{\beta b_{ij}}{b_{ij}^{(t)}}(b_{ij} - b_{ij}^{(t)})^4, \end{aligned} \tag{26}$$

and

$$\begin{aligned} z(c_{ij}, c_{ij}^{(t)}) = f(c_{ij}^{(t)}) + f'(c_{ij}^{(t)})(c_{ij} - c_{ij}^{(t)}) \\ + \frac{(\boldsymbol{B}^T \boldsymbol{B}\boldsymbol{C})_{ij}}{c_{ij}^{(t)}}(c_{ij} - c_{ij}^{(t)})^2. \end{aligned} \tag{27}$$

To prove that $z(b_{ij}, b_{ij}^{(t)}) \geqslant f(b_{ij})$ and $z(c_{ij}, c_{ij}^{(t)}) \geqslant f(c_{ij})$, we first show the Taylor series expansion of $f(b_{ij})$ and $f(c_{ij})$:

$$\begin{aligned} f(b_{ij}) = f(b_{ij}^{(t)}) + f'(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)}) \\ + \frac{1}{2} f''(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)})^2 \\ + \frac{1}{3!} f'''(b_{ij}^{(t)})(b_{ij} - b_{ij}^{(t)})^3 + \frac{1}{4!}(b_{ij} - b_{ij}^{(t)})^4 \end{aligned} \tag{28}$$

and

$$\begin{aligned} f(c_{ij}) = f(c_{ij}^{(t)}) + f'(c_{ij}^{(t)})(c_{ij} - c_{ij}^{(t)}) \\ + \frac{1}{2} f''(c_{ij}^{(t)})(c_{ij} - c_{ij}^{(t)})^2. \end{aligned} \tag{29}$$

Compare (26) and (28), (27) and (29), respectively, we should prove the following inequalities:

$$\frac{(\beta \boldsymbol{B})_{ij}}{b_{ij}^{(t)}} \geqslant \frac{1}{4!} f^{(4)}(b_{ij}^{(t)}) = \frac{1}{4!} \times 24\beta = \beta \tag{30}$$

and

$$\frac{(\boldsymbol{B}^T \boldsymbol{B}\boldsymbol{C})_{ij}}{c_{ij}^{(t)}} \geqslant \frac{f''(c_{ij}^{(t)})}{2} = \frac{1}{2} \times 2(\boldsymbol{B}^T \boldsymbol{B})_{ij} = (\boldsymbol{B}^T \boldsymbol{B})_{ij}. \tag{31}$$

From (30) and (31), it is obvious that $(\beta \boldsymbol{B})_{ij} \geqslant \beta b_{ij}^{(t)}$ and $(\boldsymbol{B}^T \boldsymbol{B}\boldsymbol{C})_{ij} \geqslant (\boldsymbol{B}^{(T)} \boldsymbol{B})_{ij} c_{ij}^{(t)}$. Hence, $z(b_{ij}, b_{ij}^{(t)}) \geqslant f(b_{ij})$ and $z(c_{ij}, c_{ij}^{(t)}) \geqslant f(c_{ij})$ hold. We present the first, second, third and fourth order derivative w.r.t. $\boldsymbol{B}$:

$$\begin{aligned} f'(b_{ij}) &= 2(\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^T - \boldsymbol{X}\boldsymbol{C}^T + \alpha \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{B} - \alpha \boldsymbol{X}\boldsymbol{H}^T \\ &\quad + 2\beta \boldsymbol{B}\boldsymbol{B}^T \boldsymbol{B} - 2\beta \boldsymbol{B})_{ij}, \\ f''(b_{ij}) &= (2\boldsymbol{C}\boldsymbol{C}^T + 2\alpha \boldsymbol{X}\boldsymbol{X}^T + 12\beta \boldsymbol{B}^T \boldsymbol{B} - 4\beta \boldsymbol{I})_{ij}, \\ f'''(b_{ij}) &= (24\beta \boldsymbol{B})_{ij}, \\ f^{(4)}(b_{ij}) &= 24\beta. \end{aligned} \tag{32}$$

The first, and second order derivative w.r.t. $\boldsymbol{C}$:

$$\begin{aligned} f'(c_{ij}) &= 2(\boldsymbol{B}^T \boldsymbol{B}\boldsymbol{C} - \boldsymbol{B}^T \boldsymbol{X})_{ij}, \\ f''(b_{ij}) &= 2(\boldsymbol{B}^T \boldsymbol{B})_{ij}. \end{aligned} \tag{33}$$

Then, we put (32) and (33) into (25) and obtain:

$$\begin{aligned} b_{ij}^{(t+1)} &= arg \min_{b_{ij}} z(b_{ij}, b_{ij}^{(t)}) \\ &= b_{ij}^{(t)} \frac{(\boldsymbol{X}\boldsymbol{C}^T + \alpha \boldsymbol{X}\boldsymbol{H}^T + 2\beta \boldsymbol{B})_{ij}}{(\boldsymbol{B}\boldsymbol{C}\boldsymbol{C}^T + \alpha \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{B} + 2\beta \boldsymbol{B}\boldsymbol{B}^T \boldsymbol{B})_{ij}}, \end{aligned} \tag{34}$$

and

$$\begin{aligned} c_{ij}^{(t+1)} &= arg \min_{c_{ij}} z(c_{ij}, c_{ij}^{(t)}) \\ &= c_{ij}^{(t)} \frac{(\boldsymbol{B}^T \boldsymbol{X})_{ij}}{(\boldsymbol{B}^T \boldsymbol{B}\boldsymbol{C})_{ij}}. \end{aligned} \tag{35}$$

From the solutions obtained, we can see that (34) and (35) are consistent with the update rules (17) and (18). By employing the auxiliary functions (26) and (27), $f_{LONMF}(\boldsymbol{B})$ and $f_{LONMF}(\boldsymbol{C})$ are nonincreasing under the update rules. Hence, the update rules guarantee the convergence of the proposed algorithm.

References

[1] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1794–1801.

[2] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2009) 210–227.

[3] M. Yang, L. Zhang, X. Feng, D. Zhang, Sparse representation based fisher discrimination dictionary learning for image classification, Int. J. Comput. Vis. 109 (2014) 209–232.

[4] Z. Jiang, Z. Lin, L.S. Davis, Label consistent k-svd: learning a discriminative dictionary for recognition, IEEE Trans. Pattern Anal. Mach. Intell. 35 (2013) 2651–2664.

[5] W. Zhu, Y. Yan, Y. Peng, Dictionary learning based on discriminative energy contribution for image classification, Knowl.-Based Syst. 113 (2016) 116–124.

[6] Y. Zhou, J. Peng, C.L.P. Chen, Dimension reduction using spatial and spectral regularized local discriminant embedding for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 53 (2015) 1082–1095.

[7] L. Qiao, S. Chen, X. Tan, Sparsity preserving projections with applications to face recognition, Pattern Recognit. 43 (2010) 331–341.

[8] W. Zhu, Y. Yan, Y. Peng, Pair of projections based on sparse consistence with applications to efficient face recognition, Signal Process., Image Commun. 55 (2017) 32–40.

[9] A.M. Martí nez, A.C. Kak, Pca versus lda, IEEE Trans. Pattern Anal. Mach. Intell. 23 (2001) 228–233.

[10] R.O. Duda, P.E. Hart, D.G. Stork, Pattern Classification, Wiley, 2001.

[11] A. Gersho, R.M. Gray, Vector Quantization and Signal Compression, Vol. 159, Springer International, 1992, pp. 407–485.

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

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

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

[15] Y.X. Wang, Y.J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Trans. Knowl. Data Eng. 25 (2013) 1336–1353.

[16] S.Z. Li, X.W. Hou, H.J. Zhang, Q. Cheng, Learning spatially localized, parts-based representation, in: Computer Vision and Pattern Recognition, pp. 207–212.

[17] P.O. Hoyer, Non-negative sparse coding, in: Neural Networks for Signal Processing, 2002. Proceedings of the 2002 IEEE Workshop on, pp. 557–565.

[18] Y. Shi, Y. Wan, K. Wu, X. Chen, Non-negativity and locality constrained laplacian sparse coding for image classification, Expert Syst. Appl. 72 (2017) 121–129.

[19] Y. Wang, Y. Jia, C. Hu, M. Turk, Fisher non-negative matrix factorization for learning local features, in: Asian Conference on Computer Vision, pp. 27–30.

[20] S. Zafeiriou, A. Tefas, I. Buciu, I. Pitas, Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification, IEEE Trans. Neural Netw. 17 (2006) 683–695.

[21] D. Cai, X. He, X. Wu, J. Han, Non-negative matrix factorization on manifold, in: IEEE International Conference on Data Mining, pp. 63–72.

[22] P. Li, J. Bu, Y. Yang, R. Ji, C. Chen, D. Cai, Discriminative orthogonal nonnegative matrix factorization with flexibility for data representation, Expert Syst. Appl. 41 (2014) 1283–1293.

[23] H. Zhang, H. Liu, R. Song, F. Sun, Nonlinear non-negative matrix factorization using deep learning, in: International Joint Conference on Neural Networks, pp. 477–482.

[24] G. Trigeorgis, K. Bousmalis, S. Zafeiriou, B.W. Schuller, A deep matrix factorization method for learning attribute representations, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2015) 417–429.

[25] E. Hosseinisl, J.M. Zurada, O. Nasraoui, Deep learning of part-based representation of data using sparse autoencoders with nonnegativity constraints, IEEE Trans. Neural Netw. Learn. Syst. 27 (2016) 2486–2498.

[26] D.D. Lee, H.S. Seung, Algorithms for non-negative matrix factorization, in: Annual Conference on Neural Information Processing Systems, pp. 556–562.

[27] J. Kivinen, M.K. Warmuth, Additive versus exponentiated gradient updates for linear prediction, in: Twenty-Seventh ACM Symposium on Theory of Computing, pp. 209–218.

[28] Y. Lé, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (1998) 2278–2324.
[29] P.N. Belhumeur, J. Hespanha, D.J. Kriegman, Eigenfaces vs fisherfaces: Recognition using class specific linear projection, IEEE Trans. Pattern Anal. Mach. Intell. 19 (1997) 711–720.
[30] F.S. Samaria, A.C. Harter, Parameterisation of a stochastic model for human face identification, in: Applications of Computer Vision, 1994. Proceedings of the Second IEEE Workshop on, pp. 138–142.
[31] P.J. Phillips, H. Moon, S.A. Rizvi, P.J. Rauss, The feret evaluation methodology for face-recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (2000) 1090–1104.
[32] T. Sim, S. Baker, M. Bsat, The cmu pose, illumination, and expression database, IEEE Trans. Pattern Anal. Mach. Intell. 25 (2003) 1615–1618.
[33] I. Kotsia, S. Zafeiriou, I. Pitas, A novel discriminant nonnegative matrix factorization algorithm with applications to facial image characterization problems, IEEE Trans. Inform. Forensics Secur. 2 (2007) 588–595.
[34] N. Guan, D. Tao, Z. Luo, B. Yuan, Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent, IEEE Trans. Image Process. 20 (2011) 20–30.
[35] X. Liu, S. Yan, H. Jin, Projective nonnegative graph embedding, IEEE Trans. Image Process. 19 (2010) 1126–1137.
[36] H. Liu, Z. Wu, D. Cai, T.S. Huang, Constrained nonnegative matrix factorization for image representation, IEEE Trans. Softw. Eng. 34 (2012) 1299–1311.