www.ietdl.org

Published in IET Computer Vision
Received on 29th September 2012
Revised on 5th June 2013
Accepted on 6th June 2013
doi: 10.1049/iet-cvi.2013.0055

ISSN 1751-9632

Non-negative matrix factorisation based on fuzzy *K* nearest neighbour graph and its applications

Jun Ye[1,2], *Zhong Jin*[1]

[1]*School of Computer Science and Technology, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China*
[2]*School of Natural Sciences, Nanjing University of Posts and Telecommunications, Nanjing 210003, People's Republic of China*
E-mail: yj8422092@163.com

Abstract: Non-negative matrix factorisation (NMF) has been widely used in pattern recognition problems. For the tasks of classification, however, most of the existing variants of NMF ignore both the discriminative information and the local geometry of data into the factorisation. The actual conditions of the problems will be affected by the change of the environmental factors to affect the recognition accuracy. In order to overcome these drawbacks, the authors regularised NMF by intra-class and inter-class fuzzy *K* nearest neighbour graphs, leading to NMF-F*K*-NN in this study. By introducing two novel fuzzy *K* nearest neighbour graphs, NMF-F*K*-NN can contract the intra-class neighbourhoods and expand the inter-class neighbourhoods in the decomposition. This method not only exploits the discriminative information and uses the geometric structure in the data effectively, but also reduces the influence of the external factors to improve recognition effect. In the factorisation, the authors minimised the approximation error whilst contracting intra-class fuzzy neighbourhoods and expanding inter-class fuzzy neighbourhoods. The authors develop simple multiplicative updates for NMF-F*K*-NN and present monotonic convergence results. Experiments of the text clustering on the CLUTO toolkit and face recognition on ORL and YALE datasets show the effectiveness of our proposed method.

1 Introduction

Non-negative matrix factorisation (NMF) is a recent method for finding a non-negative decomposition of the original data matrix. Given an input data matrix $\boldsymbol{X}$, each column of which represents a sample, NMF produces two factor matrices $\boldsymbol{U}$ and $\boldsymbol{V}^{\mathrm{T}}$ using low-rank approximation such that $\boldsymbol{X} \simeq \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$. Each column of $\boldsymbol{U}$ represents a base vector, and each column of $\boldsymbol{V}^{\mathrm{T}}$ describes how these base vectors are combined fractionally to form the corresponding sample in $\boldsymbol{X}$. All entries in matrices are required to be non-negative. Compared to other methods, such as principal components analysis (PCA) [1], and independent component analysis [2], non-negativity enables a non-subtractive combination of parts to form a whole, and make the encoding of data easier to interpret [3]. Several varieties of NMF have been developed by introducing additional constraints to the original NMF. To incorporate the data geometric structure, Cai *et al.* [4] proposed graph-regularised NMF (GNMF). Since data are assumed to lie in a smooth sub-manifold embedded in high-dimensional space. The data geometric structure is encoded by a nearest neighbour (NN) graph, which plays an important role in popular dimension reduction algorithm, such as Laplacian eigenmap [5] and local preserving projections (LPP) [6, 7]. The localisation constraint in NMF leads to a part-based representation. Li *et al.* [8] presented the local NMF (LNMF) that learns spatially localised, part-based representation for images. By introducing the sparseness constraints, Hoyer [9] presented the sparse NMF (SNMF) that improves the ability of part-based representation. Most of the existing variants of NMF have good effect in data representation, but their performance was not yet satisfactory in the feature extraction since discriminative information was not used. By introducing the Fisher's discriminative information to NMF, Wang *et al.* [10] and Zafeiriou *et al.* [11] proposed fisher NMF (FNMF) and discriminant NMF, respectively. Without regard to the geometric structure in the data, they just only consider the label information. So their methods are all deficient.

In order to exploit discriminative information and consider geometric structure sufficiently in the data at the same time, An *et al.* [12] presented the manifold-respecting discriminant NMF *K* nearest neighbour (NMF-*K*-NN) which was based on the assumption that data points on the same structure were likely to have the same label. They regularised NMF by intra-class and inter-class *K*-NN graphs, each of which reflected intra-class neighbours and inter-class neighbours. The main purpose of their method was to seek a non-negative decomposition which minimised the approximation error whereas contracting intra-class neighbourhoods and expanding inter-class neighbourhoods in the decomposition and they had got good performance in the task of the pattern recognition. In fact, the pattern

recognition problems will be affected by the change of the environmental factors to affect the recognition accuracy, such as illumination, expression, viewing conditions. Using the samples which are significantly affected by numerous environmental conditions to construct the intra-class and inter-class K-NN graphs can lead to some uncertain impact. How to investigate these factors and quantify their impact on their 'internal' class assignment can determine the recognition performance stand or fall [13]? Interestingly, Keller *et al.* [14] proposed the algorithm of fuzzy K nearest neighbour (FK-NN), by using the fuzzy membership, the influence of the environmental conditions can be effectively reduced. Recently, Wan *et al.* [15] proposed the fuzzy local discriminant embedding (FLDE) algorithm, by implementing the FK-NN in LDE [16], it could reduce the environmental conditions effect to obtain the correct local distribution information. This method adjusts to the ultimate objective of using fuzzy sets to cope with the uncertainty factors which have been inherently appeared in the pattern recognition problems.

In this paper, considering the fact that the outlier samples in the patterns may have some adverse influences on the classification result, we develop a novel NMF algorithm regularised by intra-class and inter-class FK-NN graphs, leading to NMF-FK-NN. By introducing two novel FK-NN graphs, NMF-FK-NN can contract the intra-class neighbourhoods and expand the inter-class neighbourhoods in the decomposition. This method not only exploits the discriminative information, but also uses the geometric structure in the data effectively. Also we develop simple multiplicative updates for NMF-FK-NN and present monotonic convergence results. Finally, experiments of the text clustering on the CLUTO toolkit and face recognition on ORL and YALE datasets are presented to demonstrate the effectiveness of our proposed method.

The rest of the paper is organised as follows. Section 2 gives a brief introduction of NMF. Section 3 presents the proposed NMF-FK-NN and the multiplicative updates. Section 4 shows the experimental results on the task of text clustering and face recognition. We conclude this paper in Section 5. Detailed proofs of lemmas and theorems are given in Appendix.

2 Standard NMF

Consider a data matrix $\boldsymbol{X}=[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n] \in \boldsymbol{R}^{m \times n}$, each column of which consists of m features, and represents a sample such as a text document or a face image. NMF aims to decompose $\boldsymbol{X}$ into two low rank non-negative matrices, basis matrix $\boldsymbol{U}=[u_{ij}] \in \boldsymbol{R}^{m \times p}$ and feature matrix $\boldsymbol{V}=[v_{ij}] \in \boldsymbol{R}^{n \times p}$, such that $\boldsymbol{X} \simeq \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}$, where $p < \min\{m, n\}$. So we can view this approximation column by column as $\boldsymbol{x}_i \simeq \sum_{j=1}^{p} \boldsymbol{u}_j v_{ij}$, where $\boldsymbol{u}_j$ is the jth column vector of $\boldsymbol{U}$, and can be regarded as a basis vector. So each data vector $\boldsymbol{x}_i$ is approximated by a linear combination of the columns of $\boldsymbol{U}$, weighted by the components of $\boldsymbol{V}$. Therefore the objective optimisation problem of NMF can be concluded as follows

$$\min_{\boldsymbol{U},\boldsymbol{V}} : \boldsymbol{J}_{\mathrm{NMF}} = \left\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right\|_F^2 \text{s.t.} \quad \boldsymbol{U},\ \boldsymbol{V} \geq 0 \tag{1}$$

Several methods have been proposed to find a solution to this non-linear optimisation problem. The multiplicative updates rules were first investigated by Lee and Seung [17] as follows

$$\text{(a)} \quad \boldsymbol{U} \leftarrow \boldsymbol{U} \otimes \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}}; \quad \text{(b)} \quad \boldsymbol{V} \leftarrow \boldsymbol{V} \otimes \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}}{\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}} \tag{2}$$

where $\otimes$ denote elementwise multiplication.

3 NMF regularised by intra-class and inter-class NMF-FK-NN graphs

3.1 Intra-class and inter-class K-NN graphs

Suppose there are c known pattern classes $\omega_1, \ldots, \omega_c$ and a set of samples with m dimension $\boldsymbol{X}=[\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \in \boldsymbol{R}^{m \times n}$. Each sample in $\boldsymbol{X}$ belongs to one of the classes ω_i, that is, $\boldsymbol{x}_j \in \omega_i$, $i=1, \ldots, C$; $j=1, \ldots, n$. We denote l_i as the label of $\boldsymbol{x}_i$.

Recent studies on spectral graph theory and manifold learning theory have demonstrated that the local geometric structure can be effectively modelled through a NN graph on a scatter of data points, so in order to exploit both the geometrical structure of data and label information, An *et al.* [12] constructed the adjacency matrix for the intra-class K-NN graph $\boldsymbol{W}^W$ by using binary weights indicating neighbourhood relationships as

$$\boldsymbol{W}_{ij}^W = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N_K^W\left(\boldsymbol{x}_j\right) \text{ or } \boldsymbol{x}_j \in N_K^W\left(\boldsymbol{x}_i\right) \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

where $N_K^W\left(\boldsymbol{x}_i\right)$ denotes the set of K-NN of $\boldsymbol{x}_i$ with $l_j = l_i$. Meanwhile, the adjacency matrix for the inter-class K-NN graph $\boldsymbol{W}^B$ can be constructed by using binary weights indicating neighbourhood relationships as follows

$$\boldsymbol{W}_{ij}^B = \begin{cases} 1, & \text{if } \boldsymbol{x}_i \in N_K^B\left(\boldsymbol{x}_j\right) \text{ or } \boldsymbol{x}_j \in N_K^B\left(\boldsymbol{x}_i\right) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $N_K^B\left(\boldsymbol{x}_i\right)$ denotes the set of K-NN of $\boldsymbol{x}_i$ with $l_j \neq l_i$.

3.2 Intra-class and inter-class FK-NN graphs

As we all know that environmental conditions can influence the performance of the pattern recognition problem, in order to cope with the uncertainty factors, Keller *et al.* [14] proposed FK-NN algorithm, by using fuzzy membership, the influence of the environmental conditions can be effectively reduced. So in order to reduce the influence of outliers and exploit both geometrical structure of data and label information, we introduce the fuzzy membership to construct the intra-class compactness FK-NN graph and inter-class separability FK-NN graph. Moreover, we construct the adjacency matrix for the intra-class FK-NN graph $\boldsymbol{FW}^W$ as follows

$$\boldsymbol{FW}^W = \boldsymbol{\Delta}_{ij}^W .* \boldsymbol{W}_{\mathrm{ij}}^{\mathrm{W}} \tag{5}$$

where * denotes matrix elementwise multiplication. Moreover, the fuzzy membership matrix Δ can be constructed through the (FK-NN) algorithm [14] as follows

$$\boldsymbol{\Delta}_{ij}^W = \begin{cases} 0.51 + 0.49\left(n_{ij}/K\right), & \text{if } \boldsymbol{x}_i \in N_K^W\left(\boldsymbol{x}_j\right) \\ 0.49\left(n_{ij}/K\right), & \text{otherwise} \end{cases} \tag{6}$$

Moreover, the inter-class FK-NN graph $\boldsymbol{FW}^B$ can be constructed as follows

$$\boldsymbol{FW}^B = \boldsymbol{\Delta}_{ij}^B * \boldsymbol{W}_{\mathrm{ij}}^{\mathrm{B}} \tag{7}$$

where

$$\boldsymbol{\Delta}_{ij}^B = \begin{cases} 0.51 + 0.49\left(n_{ij}/K\right), & \text{if } \boldsymbol{x}_i \in N_K^B\left(\boldsymbol{x}_j\right) \\ 0.49\left(n_{ij}/K\right), & \text{otherwise} \end{cases}$$

From these adjacency matrices defined, we can compute the fuzzy Laplacian scatter matrix of the graph for the intra-class FK-NN graphs as follows

$$\boldsymbol{FL}^W = \boldsymbol{FD}^W - \boldsymbol{FW}^W \tag{8}$$

where $\boldsymbol{FD}^W = \sum_{j \neq i} \boldsymbol{\Delta}_{ij}^W \boldsymbol{W}_{ij}^W$, $\forall i$ is the fuzzy diagonal matrix whose entries are column sums of $\boldsymbol{FW}^W$. $\boldsymbol{FL}^W$ is called fuzzy graph Laplacian, which is a discrete approximation to the Laplace–Beltrami operator on the data manifold. Moreover, the fuzzy Laplacian scatter matrix of the graph for the inter-class FK-NN graphs can be constructed in a similar way

$$\boldsymbol{FL}^B = \boldsymbol{FD}^B - \boldsymbol{FW}^B \tag{9}$$

where $\boldsymbol{FD}^B = \sum_{j \neq i} \boldsymbol{\Delta}_{ij}^B \boldsymbol{W}_{ij}^B$, $\forall i$ is also the fuzzy diagonal matrix.

3.3 NMF-FK-NN

The main purpose of constructing the FK-NN graph is to exploit both the geometrical structure of data and label information. Moreover, we want to shrink the local regions of the intra-class neighbourhood and expanding the local regions of the inter-class neighbourhood. So, we define $\boldsymbol{J}_{FG}$ as the measure of the smoothness of mapping function along the geodesics in the intrinsic geometry of the data. By minimising $\boldsymbol{J}_{FG}$, we can obtain a mapping function which is sufficiently smooth on the data manifold. Let $f_k(\boldsymbol{x}_i) = v_{ik}$ be a function that produce the mapping of the original data point $\boldsymbol{x}_i$ onto the axis $\boldsymbol{u}_k$. A reasonable criterion for choosing a 'good' map is to minimise the following objective function $\boldsymbol{J}_{FG}$, and the intuitive explanation of minimising $\boldsymbol{J}_{FG}$ is that if two data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ with the same label are close, $f_k(\boldsymbol{x}_i)$ and $f_k(\boldsymbol{x}_j)$ are similar to each other. Otherwise, if two data points with different label are far, $f_k(\boldsymbol{x}_i)$ and $f_k(\boldsymbol{x}_j)$ are different to each other (see (10))

Our NMF-FK-NN incorporates the $\boldsymbol{J}_{FG}$ term and minimises the objective function

$$\begin{aligned} \boldsymbol{J}_{\mathrm{NMF-FK-NN}} &= \boldsymbol{J}_{\mathrm{NMF}} + \alpha \sum_{k=1}^{p} \boldsymbol{J}_{FG} \\ &= \left\| \boldsymbol{X} - \boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} \right\|_F^2 + \alpha \mathrm{tr}\left[\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{FL}^W - \boldsymbol{FL}^B\right)\boldsymbol{V}\right] \end{aligned} \tag{11}$$

With the constraint that $\boldsymbol{U}$ and $\boldsymbol{V}$ are non-negative. tr$(\cdot)$ denotes the trace of a matrix. The $\alpha \geq 0$ is the regularisation parameter.

3.4 Multiplicative update rules

The objective function $J_{\text{NMF-FK-NN}}$ of NMF-FK-NN in (11) is not convex in both $\boldsymbol{U}$ and $\boldsymbol{V}$ together. Therefore it is unrealistic to expect an algorithm to find the global minimum of $J_{\text{NMF-FK-NN}}$. In the following, we introduce an iterative algorithm which can achieve a local minimum.

The objective function $J_{\text{NMF-FK-NN}}$ can be rewritten as

$$\begin{aligned} \boldsymbol{J}_{\mathrm{NMF-FK-NN}} = {} & \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right) - 2\mathrm{tr}\left(\boldsymbol{X}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right) + \mathrm{tr}\left(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right) \\ & + \alpha \mathrm{tr}\left[\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{FL}^W - \boldsymbol{FL}^B\right)\boldsymbol{V}\right] \end{aligned} \tag{12}$$

Let $\Psi = [\psi_{ij}]$ and $\Phi = [\varphi_{ij}]$ be the Lagrange multiplier for constraint $\boldsymbol{U} \geq 0$ and $\boldsymbol{V} \geq 0$, respectively. So the Lagrange function L is

$$\begin{aligned} L = {} & \mathrm{tr}\left(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\right) - 2\mathrm{tr}\left(\boldsymbol{X}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right) + \mathrm{tr}\left(\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\right) \\ & + \alpha \mathrm{tr}\left[\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{FL}^W - \boldsymbol{FL}^B\right)\boldsymbol{V}\right] + \mathrm{tr}\left(\Psi \boldsymbol{U}^{\mathrm{T}}\right) + \mathrm{tr}\left(\Phi \boldsymbol{V}^{\mathrm{T}}\right) \end{aligned} \tag{13}$$

The partial derivatives of L with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$ are as follows

$$\frac{\partial L}{\partial \boldsymbol{U}} = -2\boldsymbol{X}\boldsymbol{V} + 2\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} + \Psi \tag{14}$$

$$\frac{\partial L}{\partial \boldsymbol{V}} = -2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + 2\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + 2\alpha\left(\boldsymbol{FL}^W - \boldsymbol{FL}^B\right)\boldsymbol{V} + \Phi \tag{15}$$

Using the Karush-Kuhn-Tucker (KKT) conditions, we can obtain the following update rules

$$\boldsymbol{U} \leftarrow \boldsymbol{U} \otimes \frac{\boldsymbol{X}\boldsymbol{V}}{\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}} \tag{16}$$

$$\boldsymbol{V} \leftarrow \boldsymbol{V} \otimes \frac{\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{FD}^B + \boldsymbol{FW}^W\right)\boldsymbol{V}}{\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{FD}^W + \boldsymbol{FW}^B\right)\boldsymbol{V}} \tag{17}$$

$$\begin{aligned} \boldsymbol{J}_{FG} &= \frac{1}{2}\sum_{i,j=1}^{n}\left(\left\|f_k(\boldsymbol{x}_i) - f_k\left(\boldsymbol{x}_j\right)\right\|^2 \boldsymbol{FW}_{ij}^W - \left\|f_k(\boldsymbol{x}_i) - f_k\left(\boldsymbol{x}_j\right)\right\|^2 \boldsymbol{FW}_{ij}^B\right) \\ &= \left(\sum_{i=1}^{n} v_{ik}^2 \boldsymbol{FD}_{ii}^W - \sum_{i,j=1}^{n} v_{ik}v_{jk}\boldsymbol{FW}_{ij}^W\right) - \left(\sum_{i=1}^{n} v_{ik}^2 \boldsymbol{FD}_{ii}^B - \sum_{i,j=1}^{n} v_{ik}v_{jk}\boldsymbol{FW}_{ij}^B\right) \\ &= \left(\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FD}^W\boldsymbol{v}_k - \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FW}^W\boldsymbol{v}_k\right) - \left(\boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FD}^B\boldsymbol{v}_k - \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FW}^B\boldsymbol{v}_k\right) \\ &= \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FL}^W\boldsymbol{v}_k - \boldsymbol{v}_k^{\mathrm{T}}\boldsymbol{FL}^B\boldsymbol{v}_k \end{aligned} \tag{10}$$

doi: 10.1049/iet-cvi.2013.0055

Table 1 Summary of datasets in the CLUTO toolkit

Dataset	Source	#Documents	#Terms	#Classes
reviews	San Jose Mercury (TREC)	4069	18 483	5
klb	WebACE	2340	21 839	6
sports	San Jose Mercury (TREC)	8580	14 870	7
tr12	TREC	313	5799	8

where the matrices $\boldsymbol{FD}^B$, $\boldsymbol{FW}^B$, $\boldsymbol{FD}^W$, $\boldsymbol{FW}^W$ are non-negative. When $\alpha = 0$, it is easy to check that the update rules in (16) and (17) reduce to the update rules of original NMF. When $\alpha > 0$, we have the following theorem:

Theorem 1: The objective function $J_{\text{NMF-F}K\text{-NN}}$ is non-increasing under the update rules. The objective function is invariant under these updates if and only if $\boldsymbol{U}$ and $\boldsymbol{V}$ are at a stationary point.

Theorem 1 guarantees that the update rules of $\boldsymbol{U}$ and $\boldsymbol{V}$ in (16) and (17) converge and the final solution will be a local optimum. Please see the Appendix for a detailed proof.

4 Numerical experiments

We evaluated the performance of our proposed algorithm in the task of text clustering and face recognition.

4.1 Text clustering

We applied the proposed algorithm to feature extraction for text clustering. NMF and NMF-*K*-NN algorithms were also tested on the same tasks to compare performance. Throughout this experiment, we empirically set the regularisation parameter α to 10, the number of NN K to 5 in the NMF-*K*-NN and NMF-F*K*-NN algorithms. Moreover, we selected four well-known preprocessed document databases from the CLUTO toolkit to evaluate our algorithm. Each dataset is represented by a term-by-document matrix of varying characteristics (see Table 1). Two popular metrics, the accuracy and the normalised mutual information (NMI) were used in these experiments. The accuracy is defined as

$$\text{AC} = \frac{\sum_{i=1}^{n} \delta\left(g_i,\ \text{map}\left(l_i\right)\right)}{n} \tag{18}$$

where n is the total number of documents, and g_i is the label given by the document corpus. $\delta(x, y)$ is the function that equals 1 when $x = y$ and is 0 otherwise. $\text{map}(l_i)$ is a mapping function which maps each cluster label to an equivalent given label. The Kuhn-Munkres algorithm [18] is used for the best mapping. The greater the accuracy, the better the clustering quality.

The second metric is the NMI metric. NMI is based on the mutual information (MI) between two sets of clusters, C and C' which correspond to the set of estimated clusters and the set of ground truth clusters, respectively. Denote by c_i and c'_j the set of documents grouped into cluster i and the set of documents in the ground truth cluster i, respectively. With these definitions, MI between C and C' is calculated as

$$\text{MI}\left(C,\ C'\right) = \sum_{c_i \in C,\ c'_j \in C'} p\left(c_i, c'_j\right) \log_2 \frac{p\left(c_i, c'_j\right)}{p\left(c_i\right) p\left(c'_j\right)} \tag{19}$$

MI is normalised by $\max(H(C), H(C'))$, which is defined by

$$\text{NMI}\left(C,\ C'\right) = \frac{\text{MI}\left(C,\ C'\right)}{\max\left(H(C),\ H(C')\right)} \tag{20}$$

where $H(C)$ and $H(C')$ denote the entropies of C and C', respectively. The normalised value varies between 0 and 1. The greater the NMI, the better the clustering quality.

Table 2 shows the evaluation results. Moreover, the results were aggregated by averaging over ten independent trials. In terms of the whole performance, our algorithm against other algorithms has more significant improvements on CLUTO toolkit. Therefore it can be concluded that our method can exploit the discriminative information better and use the geometric structure of the data effectively.

4.2 Face recognition

In this section, in order to compare the NMF-F*K*-NN that using the fuzzy K neighbour graph and NMF-*K*-NN that using K neighbour graph in dealing with the problems of the pattern recognition that affected by the external factors, we design the experiment of the face recognition. Here, we evaluate the performance of the proposed method comparing with four representative algorithms, which are NMF [3], LNMF [8], FNMF [10] and NMF-*K*-NN [12], on two popular face image databases including ORL and YALE to complete the face recognition tasks.

Figs. 1*a* and *b* show example images of the ORL and YALE databases, respectively. All face images of two databases were aligned according to the eye position. Each pixel of images was linearly rescaled to the grey level of 256, and each image was rearranged to a vector. Different INDICES (2, 3, 4, 5) and (3, 4, 5, 6), of the images were randomly selected from each individual to constitute the training set $\boldsymbol{X}_{\text{train}}$, and the rest images make up the test set $\boldsymbol{X}_{\text{test}}$ on the ORL and YALE databases, respectively. $\boldsymbol{X}_{\text{train}}$ was used to learn basis for the low-dimensional space. Moreover, $\boldsymbol{X}_{\text{test}}$ was used to report the accuracy of face recognition in the learned low-dimensional space. The

Table 2 Clustering performance comparison on the CLUTO toolkit

Dataset	Accuracy			Normalised mutual information		
	NMF	NMF-*K*-NN	NMF-F*K*-NN	NMF	NMF-*K*-NN	NMF-F*K*-NN
reviews	0.678	0.723	0.736	0.371	0.473	0.486
klb	0.593	0.609	0.612	0.498	0.464	0.485
sports	0.338	0.341	0.357	0.287	0.284	0.295
tr12	0.623	0.610	0.639	0.587	0.594	0.612

Fig. 1 *Face examples of one person*
a ORL
b YALE database

accuracy was calculated as the percentage of samples in the $\boldsymbol{X}_{\text{test}}$ that were correctly classified using the NN classifiers. In order to compare these algorithms fairly, we ran these algorithms under different parameter settings, and then the best average result was reported. We chose the same number of the neighbourhood K to construct the intra-class FK-NN graph and inter-class FK-NN graph. Moreover, the numerical value of K varied from 1 to 10. Also, because we want to compare the best average results between NMF-K-NN algorithm and our proposed algorithm, we chose the regularisation parameter α and the dimensionality of features with the same as the literature [12], that can be set by [0.01, 0.1, 1, 10, 100] and [200, 195, 190, … ,10, 5], respectively.

Since $\boldsymbol{X}$ can be approximated by columnwise $\boldsymbol{UV}^{\text{T}}$, we can naturally project a sample $\boldsymbol{x}_i$ from the original high-dimensional space to the low-dimensional space or equivalently $\boldsymbol{y} = \boldsymbol{U}^{\dagger}\boldsymbol{x}_i$, wherein the projection matrix $\boldsymbol{U}^{\dagger} = (\boldsymbol{U}^{\text{T}}\boldsymbol{U})^{-1}\boldsymbol{U}^{\text{T}}$ is the pseudo-inverse of $\boldsymbol{U}$. So, by computing the features of the $\boldsymbol{X}_{\text{test}}$ as $\boldsymbol{V}_{\text{test}} = (\boldsymbol{U}^{\dagger}\boldsymbol{X}_{\text{test}})^{\text{T}}$, we can obtain the face recognition results using the NN classifier.

4.2.1 ORL database: The ORL databases contain images from 40 individuals, each providing ten different images. All the images were taken with the people in the frontal position, but the times, lightings, facial expressions and facial details (glasses or no glasses) are different from image to image. The images were taken with a tolerance for some tilting and rotation of the face of up to 20°. Moreover, there is also some variation in the scale of up to about 10%. Each image in the databases was resized into 56×46 pixel array and reshaped to a vector. Table 3 shows the best recognition accuracy and corresponding dimension for all the algorithms, and these trials were independently conducted 50 times with different initialisations. From Table 3, we can see that all algorithms as the number of training samples increases, the recognition accuracy becomes better. As can be also seen, in all cases, our algorithm performs the best. When compared with the NMF-K-NN algorithm, ours provides better recognition accuracy.

4.2.2 YALE database: The YALE face database contain 165 images of 15 individuals (each person providing 11 different images) under various facial expressions (smile or sad) and lighting conditions. Each image in the databases was resized into 50×40 pixel array and reshaped to a vector for computational effectiveness. Table 4 shows the best recognition accuracy and corresponding dimension for all the algorithms, and these trials were also independently conducted 50 times with different initialisations. Also we can conclude that NMF-FK-NN outperforms NMF-K-NN.

From Tables 3 and 4, the method of NMF-K-NN by incorporating the label information and the geometric structure in the data can achieve good recognition, but face images are always affected by variations in illumination conditions and different facial expressions. These changes will reduce the recognition performance. Through the fuzzy K neighbour graphs, our proposed method has lower sensitivities to the sample variations caused by varying illumination, expression, viewing conditions and shapes. So the class of a new test point can be more reliably predicted by the NN criterion, owing to the locally discriminating nature. From Tables 3 and 4, we can also see that the number of extracted features showing the best performance in parentheses used by NMF-K-NN is always lower than that of our proposed method. The reason for this can explain that although our proposed method which using the FK-NN graph has lower sensitivities to the sample variations, our proposed method need more discriminative

Table 3 Maximal average recognition rates (percent) of five methods with different number of training samples on the ORL face database and the corresponding dimensions (show in parentheses)

Training	NMF	LNMF	FNMF	NMF-K-NN	NMF-FK-NN
2	59.67 (75)	63.43 (90)	67.48 (50)	69.68 (105)	70.42 (135)
3	62.28 (80)	71.33 (105)	74.14 (70)	76.28 (110)	77.83 (150)
4	70.41 (85)	81.67 (135)	84.16 (95)	85.41 (125)	86.87 (145)
5	81.12 (110)	85.62 (175)	87.51 (100)	89.37 (115)	91.83 (160)

Table 4 The maximal average recognition rates (percent) of five methods with different number of training samples on the YALE face database and the corresponding dimensions (show in parentheses)

Training	NMF	LNMF	FNMF	NMF-K-NN	NMF-FK-NN
3	48.33 (70)	58.21 (60)	57.46 (65)	59.17 (140)	63.76 (165)
4	56.67 (80)	63.76 (100)	62.52 (90)	64.82 (120)	67.13 (175)
5	60.98 (75)	66.64 (110)	64.21 (110)	68.38 (125)	74.62 (180)
6	68.89 (100)	74.67 (125)	67.37 (105)	77.62 (125)	80.98 (155)

doi: 10.1049/iet-cvi.2013.0055

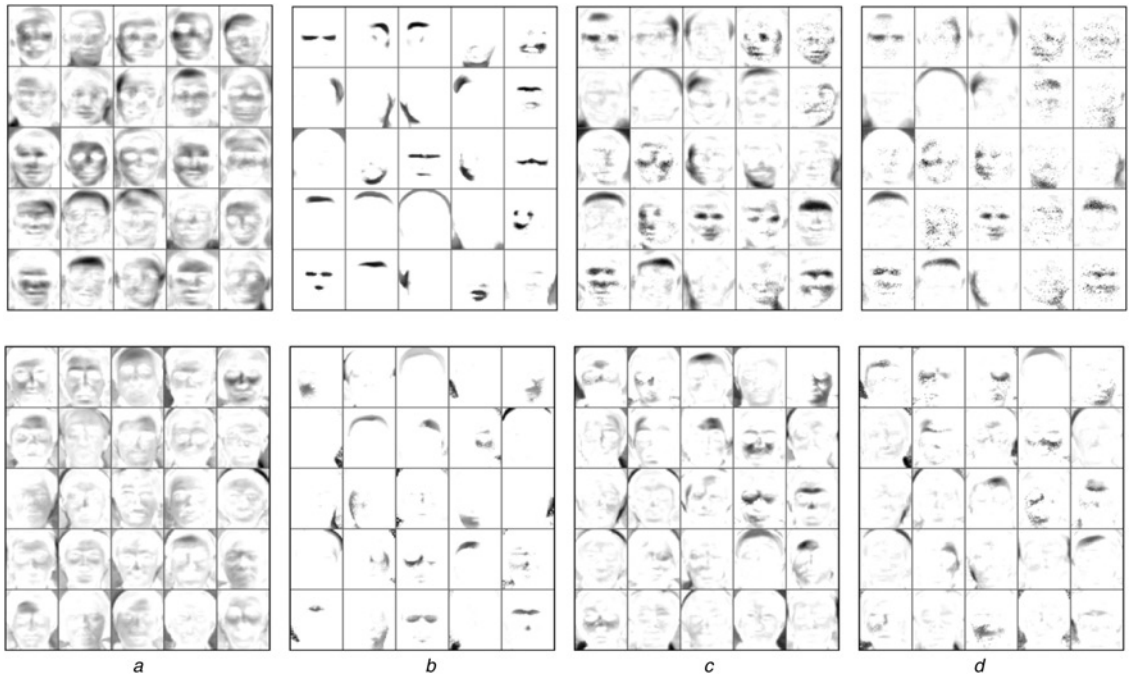

Fig. 2 *Learned bases of the (first row) ORL and (second row) YALE datasets*
a–d (column) sub-figures represent the basis vector of NMF, LNMF, NMF-*K*-NN and NMF-F*K*-NN, respectively

information to achieve better recognition performance. The advantage of the NMF-F*K*-NN is that the fuzzy neighbour membership degree can efficiently handle the vagueness and ambiguity of samples being degraded by poor illumination, shape and facial expression variations. In other words, the fuzzy neighbour membership degree helps to pull the near neighbour samples in same class nearer and nearer and repel the far neighbour samples of different classes farther and farther. So our proposed method based on the fuzzy neighbour membership degree can better characterise the compactness and separability.

4.3 Parts-based learning

In this section, we study the sparseness ability of parts-based representation of the proposed algorithm. We compare the base sparseness ability with those learned by NMF, LNMF, NMF-*K*-NN and NMF-F*K*-NN on ORL and YALE databases. Fig. 2 present these bases for sub-space of the same number of dimensionality 25. Moreover, Fig. 2*a* (first column) shows the results using NMF on two databases. Even if the factors' images give an intuitive notion of a parts-based representation, the factorisation is not really sparse enough to represent unique parts of an average face. In other words, the NMF allows some undesirable overlapping of parts, especially in those areas that are common to most of the faces in the input data. Fig. 2*b* (second column) shows the results using LNMF, and the local and independent features that is mouths, noses and other facial parts on faces can be extracted. However, it ignores both the discriminative information and the local geometry. Fig. 2*c* (third column) and 2*d* (fourth column) show the results using NMF-*K*-NN and NMF-F*K*-NN, respectively. Although more useful features of faces are retained compared to the LNMF, NMF-*K*-NN algorithm allows more undesirable overlapping of parts than the NMF-F*K*-NN. This allows us to tune the NMF-*K*-NN algorithm to more relevant parts to give more useful information about the data.

In order to quantify sparseness ability of the basis images, we also introduce the sparseness according to the Hoyer [9], and the sparseness was defined as

$$\text{sparseness}\ (\boldsymbol{X}) = \frac{\sqrt{n} - \left(\sum_i |x_i|\right)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1} \tag{21}$$

where n is the dimensionality of vector $\boldsymbol{X}$. Table 5 shows the average sparseness of the columns in the learned basis by NMF, LNMF, NMF-*K*-NN and NMF-F*K*-NN. It can be seen from the results that both NMF-*K*-NN and NMF-F*K*-NN bases are sparser than NMFs. LNMF bases are sparser than our methods. Moreover, the NMF-F*K*-NN bases are sparser than NMF-*K*-NNs, this can illuminate that the NMF-F*K*-NN can use the discriminative information and local geometry better.

4.4 Parameter selection

Our algorithm has two essential parameters: the number of NN K and the regularisation parameter α. We set the number of K of fuzzy intra-class and inter-class graph to be the same, and study the parameter K effect on the face recognition accuracy on the YALE database, where the training set is comprised of three images randomly chosen from each individual and the remainder images for test. By definition of the fuzzy adjacent graphs, K varies from 1 to

Table 5 Sparseness of algorithms on ORL and YALE databases

Dataset	NMF	LNMF	NMF-*K*-NN	NMF-F*K*-NN
ORL	0.456	0.854	0.610	0.701
YALE	0.487	0.873	0.663	0.734

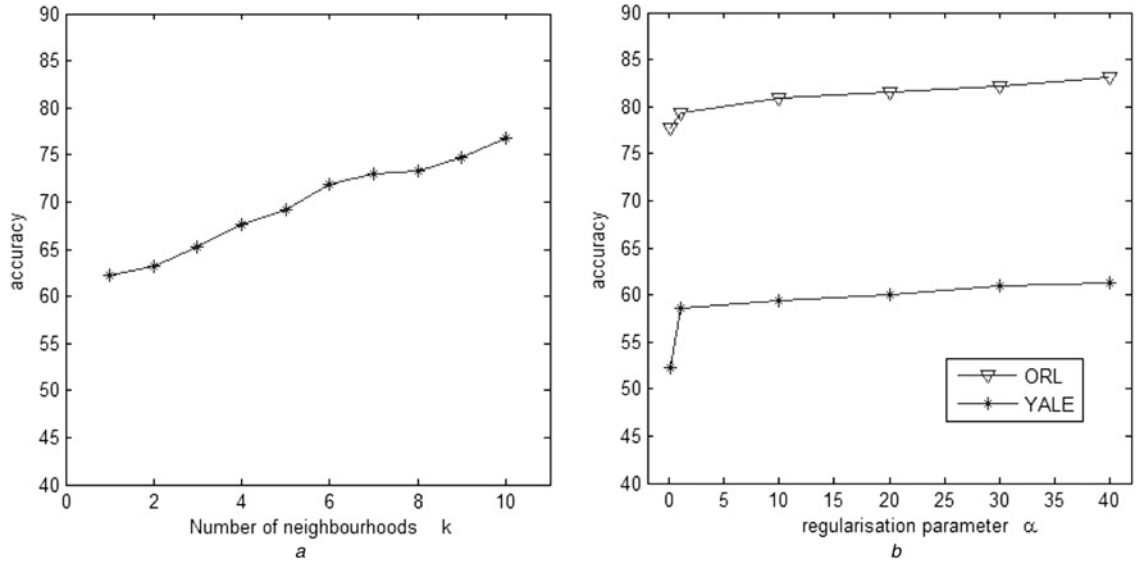

Fig. 3 *Face recognition accuracy for different values of the parameters of the NMF-FK-NN*

a Number of K in the neighbourhood graphs
b α with $K = 5$ fixed

10. Fig. 3*a* presents the average performances of the accuracy against K after running the method 50 times. Moreover, we can conclude that the accuracy increased with the parameter K, but when the parameter K increased, the amount of computation increased enormously. So we had to consider the tradeoff between computation cost and the accuracy to choose the parameter K. The situations of choosing the parameter K on the ORL database are the same with the YALEs. To another parameter α, we set it equal to 0.1, 1, 10, 20, 30, 40, and experiment on ORL and YALE databases, respectively. Also, we randomly selected four images from each identity to form the training set and the remainder images for test. These trails were independently performed ten times, and the average accuracy was reported. Fig. 3*b* presents the average performances of the accuracy against α with K fixed to 5. Moreover, we can see that when α larger than 10, the accuracy was stable gradually.

5 Conclusions

In this paper, we proposed the NMF-F*K*-NN algorithm that incorporates both local geometry and discriminative information by exploiting fuzzy inter-class and intra-class neighbourhoods. By using the fuzzy membership, the influence of the outliers on feature extraction can be effectively reduced and more effective local discriminative features can be obtained. Experimental results of the text clustering on the CLUTO toolkit and face recognition on ORL and YALE datasets show the effectiveness of the proposed method. Also, multiplicative update rules were provided with convergence analysis. In the future, we will make more tests on other types of databases, and further improved the discriminative ability.

6 Acknowledgments

This work is partially supported by the Natural Science Foundation of China under grant nos. 60973098, 60632050, 60705006, 60873151 and Hi-Tech Research and Development Program of China under grant no. 2006AA01Z119.

7 References

1 Kirby, M., Sirovich, L.: 'Application of the Karhunen-Loeve procedure for the characterization of human faces', *IEEE Trans. PAMI*, 1990, **12**, (1), pp. 103–108
2 Chengjun, L., Wechsler, H.: 'Independent component analysis of Gabor features for face recognition', *IEEE Trans. Neural Netw.*, 2003, **14**, (4), pp. 919–928
3 Lee, D.D., Seung, H.S.: 'Learning the parts of objects by nonnegative matrix factorization', *Nature*, 1999, **401**, (21), pp. 788–791
4 Cai, D., He, X., Wu, X., Han, J.: 'Non-negative matrix factorization on manifold'. Proc. IEEE Int. Conf. on Data Mining, 2008, pp. 63–72
5 Belkin, M., Niyogi, P.: 'Laplacian eigenmaps and spectral techniques for embedding and clustering', *Adv. Neural Inf. Process. Syst.*, 2002, **14**, pp. 585–591
6 He, X., Niyogi, P.: 'Locality preserving projections', *Adv. Neural Inf. Process. Syst.*, 2003, **15**, pp. 1–8
7 Guan, N., Tao, D., Luo, Z., Yuan, B.: 'Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent', *IEEE Trans. Image Process.*, 2011, **20**, (7), pp. 2030–2048
8 Li, S.Z., Hou, X., Zhang, H., Cheng, Q.: 'Learning spatially localized, parts-based representation'. Proc. IEEE Int. Conf. on Comput. Vis. Pattern Recognition, 2001, pp. 207–212
9 Hoyer, P.: 'Non-negative matrix factorization with sparseness constraints', *J. Mach. Learn. Res.*, 2004, **5**, pp. 1457–1469
10 Wang, Y., Jia, Y., Hu, C., Turk, M.: 'Fisher non-negative matrix factorization for learning local features'. Proc. Asian Conf. on Computer Vision (ACCV), Jeju Island, Korea, 2004
11 Zafeiriou, S., Tefas, A., Buciu, I., Pitas, I.: 'Exploiting discriminant information in nonnegative matrix factorization with application to frontal face verification', *IEEE Trans. Neural Netw.*, 2006, **17**, (3), pp. 683–695
12 An, S., Yoo, J., Choi, S.: 'Manifold-respecting discriminant nonnegative matrix factorization', *Pattern Recognit. Lett.*, 2011, **32**, pp. 832–837
13 Kwak, K.C., Pedrycz, W.: 'Face recognition using a fuzzy Fisherface classifier', *Pattern Recognit.*, 2005, **38**, (10), pp. 1717–1732
14 Keller, J.M., Gray, M.R., Givens, J.A.: 'A fuzzy k-nearest neighbor algorithm', *IEEE Trans. Syst., Man Cybernet.*, 1985, **15**, (4), pp. 580–585
15 Wan, M., Yang, G., Lai, Z., Jin, Z.: 'Feature extraction based on fuzzy local discriminant embedding with applications to face recognition', *IET Comput. Vis.*, 2011, **5**, (5), pp. 301–308
16 Chen, H.T., Chang, H.W., Liu, T.L.: 'Local discriminant embedding and its variants'. Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition, 2005, pp. 846–853
17 Lee, D.D., Seung, H.S.: 'Algorithms for nonnegative matrix factorization', *Adv. Neural Inf. Process. Syst. (NIPS)*, 2000, pp. 556–562
18 Lovasz, L.: 'Matching Theory (North-Holland Mathematics Studies)' (Elsevier Science Ltd., 1986)

doi: 10.1049/iet-cvi.2013.0055

8 Appendix

Convergence analysis:

By introducing an auxiliary function, we can prove the convergence of NMF-F*K*-NN.

Definition 1: The function $G(v, v')$ is an auxiliary function for $F(v)$, if the $G(v, v') \geq F(v)$ and $G(v, v) = F(v)$ are satisfied.

The auxiliary function is very useful because of the following lemma.

Lemma 1: if $G(v, v')$ is an auxiliary function of $F(v)$, then $F(v)$ is non-increasing under the update

$$v^{t+1} = \arg\min_{v} G\left(v,\ v^t\right) \tag{22}$$

Proof : $F(v^{t+1}) \leq G(v^{t+1}, v^t) \leq G(v^t, v^t) = F(v^t)$.

Since the second term of $J_{\text{NMF-F}K\text{-NN}}$ is only related to $\boldsymbol{V}$, the update rule of $\boldsymbol{U}$ is exactly same as the original NMF. The convergence proof of NMF-F*K*-NN that $J_{\text{NMF-F}K\text{-NN}}$ is non-increasing under update rule (16) is the same as the literature [17]. Here we give the proof that the objective function $J_{\text{NMF-F}K\text{-NN}}$ is non-increasing under the update rule (17).

Considering any element v_{ij} in $\boldsymbol{V}$, we use F_{ij} to denote the part of $J_{\text{NMF-F}K\text{-NN}}$ which is only relevant to v_{ij}. It is easy to check that

$$F_{ij} = \left(-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}} + \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\boldsymbol{V}^{\mathrm{T}}\right)_{ij} + \alpha\mathrm{tr}\left[\boldsymbol{V}^{\mathrm{T}}\left(\boldsymbol{F}\boldsymbol{L}^{W} - \boldsymbol{F}\boldsymbol{L}^{B}\right)\boldsymbol{V}\right]_{ij}$$

$$F'_{ij} = \left(-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + 2\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + 2\alpha\left(\boldsymbol{F}\boldsymbol{L}^{W} - \boldsymbol{F}\boldsymbol{L}^{B}\right)\boldsymbol{V}\right)_{ij},$$

$$F''_{ij} = 2\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{jj} + 2\alpha\left(\boldsymbol{F}\boldsymbol{L}^{W} - \boldsymbol{F}\boldsymbol{L}^{B}\right)_{ii}$$

where F'_{ij} is the first-order partial derivative of $J_{\text{NMF-F}K\text{-NN}}$ w.r.t v_{ij} and F''_{ij} is the second-order partial derivative of $J_{\text{NMF-F}K\text{-NN}}$ w.r.t v_{ij}.

Lemma 2: Function

$$G\left(v,\ v_{ij}^t\right) = F_{ij}\left(v_{ij}^t\right) + F'_{ij}\left(v_{ij}^t\right)\left(v - v_{ij}^t\right) + \frac{\left(\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{ij} + \alpha\left[\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij}}{v_{ij}^t}\left(v - v_{ij}^t\right)^2 \tag{23}$$

is an auxiliary function for F_{ij}.

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

$$F_{ij}(v) = F_{ij}\left(v_{ij}^t\right) + F'_{ij}\left(v_{ij}^t\right)\left(v - v_{ij}^t\right) + \frac{1}{2}F''_{ij}\left(v_{ij}^t\right)\left(v - v_{ij}^t\right)^2$$

For $F_{ij}^{(n)}(v) = 0$with $n \geq 3$, we could just derive the Taylor series expansion of $F_{ij}(v)$ until second order. With (16) to find that $G(v,\ v_{ij}^t) \geq F_{ij}(v)$ is equivalent to

$$\frac{\left(\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{ij} + \alpha\left[\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij}}{v_{ij}^t} \geq \left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{jj} + \alpha\left(\boldsymbol{F}\boldsymbol{L}^{W} - \boldsymbol{F}\boldsymbol{L}^{B}\right)_{ii}$$

We have

$$\left(\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{ij} = \sum_{l=1}^{q} v_{il}^t\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{lj} \geq v_{ij}^t\left(\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U}\right)_{jj}$$

and (see equation at the bottom of the page)

We can easily demonstrate the convergence of Theorem 1.

Proof of Theorem 1: Take the derivative of $G\left(v,\ v_{ij}^t\right)$ w.r.t v to obtain v_{ij}^{t+1} as

$$\frac{\partial G\left(v,\ v_{ij}^t\right)}{\partial v} = 0,\ F'_{ij}\left(v_{ij}^t\right) + \left(v_{ij}^{t+1} - v_{ij}^t\right)\frac{2\left[\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{D}\boldsymbol{F}^{W} + \boldsymbol{W}\boldsymbol{F}^{B}\right)\boldsymbol{V}\right]_{ij}}{v_{ij}^t} = 0$$

Replacing $G(v,\ v_{ij}^t)$ in (22) by (23) results in (see equation at the bottom of the page)

Since (23) is an auxiliary function, F_{ij} is non-increasing under this update rule. For F_{ij} is the part of $J_{\text{NMF-F}K\text{-NN}}$ which is only relevant to v_{ij}, v_{ij} is non-increasing under this update rule (17).

$$\left[\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij} = \sum_{s=1}^{n}\left(\boldsymbol{F}\boldsymbol{D}_{is}^{W} + \boldsymbol{F}\boldsymbol{W}_{is}^{W}\right)v_{sj}^t \geq \left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)_{ii}v_{ij}^t \geq \left[\left(\boldsymbol{F}\boldsymbol{D}^{W} - \boldsymbol{F}\boldsymbol{W}^{W}\right) + \boldsymbol{F}\boldsymbol{W}^{B}\right]_{ii}v_{ij}^t = \left(\boldsymbol{F}\boldsymbol{L}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)_{ii}v_{ij}^t$$
$$\geq \left[\boldsymbol{F}\boldsymbol{L}^{W} + \left(-\boldsymbol{F}\boldsymbol{D}^{B} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\right]_{ii}v_{ij}^t = \left(\boldsymbol{F}\boldsymbol{L}^{W} - \boldsymbol{F}\boldsymbol{L}^{B}\right)_{ii}v_{ij}^t$$

$$v_{ij}^{t+1} = v_{ij}^t - v_{ij}^t\frac{F'_{ij}\left(v_{ij}^t\right)}{2\left[\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{D}\boldsymbol{F}^{W} + \boldsymbol{W}\boldsymbol{F}^{B}\right)\boldsymbol{V}\right]_{ij}}$$
$$= v_{ij}^t\frac{2\left[\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij} - 2\left[-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + \boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{L}\boldsymbol{F}^{W} - \boldsymbol{L}\boldsymbol{F}^{B}\right)\boldsymbol{V}\right]_{ij}}{2\left[\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij}} = v_{ij}^t\frac{\left[\boldsymbol{X}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{F}\boldsymbol{D}^{B} + \boldsymbol{F}\boldsymbol{W}^{W}\right)\boldsymbol{V}\right]_{ij}}{\left[\boldsymbol{V}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} + \alpha\left(\boldsymbol{F}\boldsymbol{D}^{W} + \boldsymbol{F}\boldsymbol{W}^{B}\right)\boldsymbol{V}\right]_{ij}}$$