Graph dual regularization non-negative matrix factorization for co-clustering

Fanhua Shanga, L.C. Jiaoa, Fei Wanga

a Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi’an 710071, China
b Healthcare Transformation Group, IBM T. J. Watson Research Center at Hawthorne, NY, USA

A R T I C L E   I N F O

Article history:
Received 4 July 2011
Received in revised form 21 November 2011
Accepted 15 December 2011
Available online 24 December 2011

Keywords:
Low-rank matrix factorization
Non-negative matrix factorization (NMF)
Graph Laplacian
Graph dual regularization
Co-clustering

A B S T R A C T

Low-rank matrix factorization is one of the most useful tools in scientific computing, data mining and computer vision. Among of its techniques, non-negative matrix factorization (NMF) has received considerable attention due to producing a parts-based representation of the data. Recent research has shown that not only the observed data are found to lie on a nonlinear low dimensional manifold, namely data manifold, but also the features lie on a manifold, namely feature manifold. In this paper, we propose a novel algorithm, called graph dual regularization non-negative matrix factorization (DNMF), which simultaneously considers the geometric structures of both the data manifold and the feature manifold. We also present a graph dual regularization non-negative matrix tri-factorization algorithm (DNMTF) as an extension of DNMF. Moreover, we develop two iterative updating optimization schemes for DNMF and DNMTF, respectively, and provide the convergence proofs of our two optimization schemes. Experimental results on UCI benchmark data sets, several image data sets and a radar HRRP data set demonstrate the effectiveness of both DNMF and DNMTF.

1. Introduction

Clustering is one of the most important research topics in both machine learning and data mining communities. It aims at partitioning the data into groups of similar objects. An enormous number and variety of methods have been proposed over the past several decades to solve clustering problems [1]. As a flexible class of clustering approaches, spectral clustering such as Normalized cut [2], Ratio cut [3], and Min–max cut [4], has been shown to be more effective than traditional methods such as K-means [5,6]. Unfortunately, when the number of data points (denoted as N) is large, the computational complexity of spectral decomposition can reach O(N3).

Non-negative matrix factorization (NMF) is a recent method for finding two low-rank non-negative matrices whose product provides a good approximation to the original nonnegative matrix. Moreover, the additive reconstruction due to the nonnegative constraints leads to a parts-based representation for human face images, text documents and so on [7,8]. Previous studies such as Biederman’s Recognition-by-Components [9] advocate the suitability of a parts-based approach for recognition in both humans and machines [10]. As a result, NMF has been widely applied in data mining, computer vision, biology and so on. It is worth noting that there is close relationship between K-means, spectral clustering and NMF [11–13].

Recent research has shown that the observed data are found to lie on a nonlinear low dimensional manifold embedded in a high dimensional ambient space [14–18]. A large number of manifold learning techniques have recently been proposed to detect the underlying manifold structure, such as ISOMAP [15], Locally Linear Embedding (LLE) [16], Laplacian Eigenmaps [17]. Moreover, Cai et al. [19] proposed a graph regularized non-negative matrix factorization (GNMF) to find a compact representation which uncovers the hidden semantics and simultaneously respects the intrinsic geometric structure. And it has been shown that learning performance can be greatly improved if the manifold structure information contained in the data is exploited [5,6,18,19].

However, the methods mentioned above focus on one-sided clustering, i.e., clustering the data based on the similarities along the feature. Considering the duality between data points and features, several co-clustering algorithms have been proposed and shown to be superior to traditional one-sided clustering [20–22]. Dhillon et al. [21] proposed an information theoretic co-clustering algorithm. Ding et al. [22] proposed an orthogonal nonnegative matrix tri-factorization to co-clustering words and documents, which achieves an encouraging performance.

Motivated by recent progress in dual regularization [23,24] and matrix factorization [25,26], we propose a novel algorithm called graph dual regularization non-negative matrix factorization (DNMF), which simultaneously considers the geometric
structures of the data manifold as well as the feature manifold. We encode the geometric structure information of data and feature spaces by constructing two nearest neighbor graphs, respectively. Then we present a graph dual regularization non-negative matrix tri-factorization algorithm (DNMTF) as an extension of DNMF. Both DNMF and DNMTF can be optimized by iterative multiplicative updating schemes, and their convergence proofs are been provided.

To summarize, the main contributions of this work include:

1. We propose a novel graph dual regularization non-negative matrix factorization (DNMF) algorithm which simultaneously considers the geometric structure information contained in data points as well as features.
2. We also present a graph dual regularization non-negative matrix tri-factorization algorithm (DNMTF) as an extension of DNMF.
3. We develop two iterative multiplicative updating optimization schemes to solve our proposed algorithms: DNMF and DNMTF, and provide the convergence proofs of two optimization schemes.

The remainder of this paper is organized as follows: Section 2 presents a brief overview of some related works. A novel DNMF algorithm is proposed in Section 3. As an extension of DNMF, a DNMTF algorithm is described in Section 4. Experimental results on many real-world data sets are presented in Section 5. Section 6 is conclusions.

2. Related works

Before we go into the details of our algorithms, first we briefly review some works that are closely related to this paper.

2.1. NMF

Non-negative matrix factorization (NMF) [8] is a low-rank matrix approximation method for finding two low-rank non-negative matrices whose product provides a good approximation to the original non-negative matrix. NMF has been widely applied in many real-world problems such as face analysis [27], document clustering [28], DNA gene expression analysis [29] and spectral data analysis [30]. In addition, its extensions and variants [31–33] have a never increasingly variety of real-world applications.

Given a data matrix \( X \in \mathbb{R}^{M \times N} \), NMF aims to find two low-rank non-negative factors \( U \in \mathbb{R}^{M \times R} \) and \( V \in \mathbb{R}^{N \times R} \), which are the basis factor and low-rank representation factor, respectively. The squared error (Euclidean distance) objective function is formulated as follows:

\[
J_{\text{NMF}} = \sum_{i=1}^{M} \sum_{j=1}^{N} (X_{ij} - UV^T)_{ij}^2 = ||X - UV^T||_F^2
\]

s.t. \( U \geq 0, V \geq 0 \),

(1)

where \( ||.||_F \) is the Frobenius norm. Although the objective function \( J_{\text{NMF}} \) in Eq. (1) is convex in \( U \) only or \( V \) only, it is not convex in both variables together. The optimization problem in Eq. (1) can be optimized by iterative multiplicative updating rules as follows [34]:

\[
U_{ij} \leftarrow \frac{[XV]_{ij}}{[UV^T]_{ij}}, V_{ij} \leftarrow \frac{[X^TU]_{ij}}{[VU^T]_{ij}}.
\]

(2)

Theorem 1. [34] For \( X, U, V \geq 0 \), the objective function \( J_{\text{NMF}} \) in Eq. (1) is non-increasing under each of the above multiplicative updating rules stated in Eq. (2).

The multiplicative updating rules in Eq. (2) can be interpreted as a rescaled gradient descent scheme. And Lee and Seung [34] have proved that the iterative multiplicative updating scheme stated in Eq. (2) will find a local minimum of the objective function \( J_{\text{NMF}} \).

2.2. GNMF

Cai et al. [19] proposed a graph regularized non-negative matrix factorization (GNMF) to find a compact representation which uncovers the hidden semantics and simultaneously respects the intrinsic geometric structure. The objective function of GNMF is formulated as follows:

\[
J_{\text{GNMF}} = ||X - UV^T||_F^2 + \lambda \text{Tr}(V^T LV),
\]

s.t. \( U \geq 0, V \geq 0 \),

(3)

where \( L = D - W \) is graph Laplacian, where \( W \) is the weight matrix and \( D \) is a diagonal matrix whose entries are column sums of \( W \), \( D_{i} = \sum_{j} W_{ij} \).

However, the objective function of GNMF in Eq. (3) is not convex in both \( U \) and \( V \) together. The local minima of the above objective function can be achieved using the following updating rules [19]:

\[
U_{ij} \leftarrow U_{ij} \frac{[XV]_{ij}}{[UV^T]_{ij}}, V_{ij} \leftarrow V_{ij} \frac{[X^TU + \lambda LW]_{ij}}{[VU^T U + \lambda DV]_{ij}}.
\]

(4)

Theorem 2. For \( X, U, V \geq 0 \), the objective function \( J_{\text{GNMF}} \) in Eq. (3) is non-increasing under each of the above multiplicative updating rules stated in Eq. (4).

Cai et al. [19] have proved that the iterative multiplicative updating scheme stated in Eq. (4) will find a local minimum of the objective function \( J_{\text{GNMF}} \).

2.3. DRCC

Gu et al. [24] proposed a dual regularized co-clustering (DRCC) method based on semi-nonnegative matrix tri-factorization with two graph regularizers. The objective function of DRCC is formulated as follows:

\[
J_{\text{DRCC}} = ||X - GSFT^T F^T G^T||_F^2 + \lambda \text{Tr}(F^T LF + \mu \text{Tr}(G^T LG)),
\]

s.t. \( G \geq 0, F \geq 0 \),

(5)

where \( \lambda, \mu \geq 0 \) are the regularization parameters, and \( S \) is a matrix whose entries can take any signs. \( L = D^T - W^T \) is the graph Laplacian of the data graph which reflects the label smoothness of the data points, and \( L_c = D^c - W^c \) is the graph Laplacian of the feature graph which reflects the label smoothness of the feature. Gu et al. [24] presented an alternating scheme to optimize the objective function in Eq. (5):
Theorem 3. [24] For $G$, $F \geq 0$, the objective function $J_{\text{DNMF}}$ in Eq. (5) is non-increasing under each of the above updating rules stated in Eq. (6).

Gu et al. [24] have proved that the iterative multiplicative updating scheme stated in Eq. (6) will find local minima of the objective function $J_{\text{DNMF}}$.

3. Graph dual regularization non-negative matrix factorization

In this section, we first propose a novel graph dual regularization non-negative matrix factorization (DNMF) algorithm, which simultaneously considers the geometric structures of both the data manifold and the feature manifold. Then we present an optimization scheme based on the iterative updating rules of two factor matrices to solve its objective function. Finally, we present the convergence proof of our iterative updating scheme.

3.1. Data and feature graphs

Recent research has shown that not only the observed data are found to lie on a nonlinear low dimensional manifold, namely data manifold, but also features lie on a nonlinear manifold, namely feature manifold [24]. As a result, we introduce two graphs: data graph and feature graph to construct two graphs: data graph and feature graph to effectively model the geometric structures of both the data manifold and the feature manifold.

We first construct a $k$-nearest neighbor graph whose vertices correspond to $X_{1}, ..., X_{N}$. And we use the $0–1$ weighting scheme for constructing the $k$-nearest neighbor graph as in [19], and define the data weight matrix as follows,

$$W_{ij} = \begin{cases} 1, & \text{if } X_{j} \in N(X_{i}); \\ 0, & \text{otherwise}, \end{cases},$$

where $N(X_{j})$ represents the set of $k$-nearest neighbors of $X_{j}$.

The graph Laplacian of the data graph is defined as $L_{D} = D - W_{D}$, where $D$ is a diagonal degree matrix whose entries are given by $D_{ij} = \sum_{j} W_{ij}$. For convenience, we also use the $0–1$ weighting scheme for constructing a $k$-nearest neighbor graph whose vertices correspond to $X_{1}, ..., X_{M}$, and define the feature weight matrix as follows,

$$W_{ij} = \begin{cases} 1, & \text{if } X_{j} \in N(X_{i}); \\ 0, & \text{otherwise}, \end{cases},$$

where $N(X_{j})$ represents the set of $k$-nearest neighbors of $X_{j}$.

The graph Laplacian of the feature graph is also defined as $L_{F} = D - W_{F}$.

3.2. Objective function of DNMF

Based on two graph regularizers of both data manifold and feature manifold, we propose a novel graph dual regularization non-negative matrix factorization (DNMF), whose objective function is formulated as follows:

$$J_{\text{DNMF}} = \|X - UV^{T}\|^{2} + \lambda \text{Tr}(V^{T}LV) + \mu \text{Tr}(U^{T}L_{U}U),$$

subject to $U \geq 0, V \geq 0. \tag{7}$

where $\lambda, \mu \geq 0$ are the regularization parameters which balance the reconstruction error of DNMF in the first term and graph regularizations in the second and third terms. When letting $\mu = 0$, DNMF degenerates to the GNMF method in Eq. (3), and when letting $\lambda = \mu = 0$, DNMF degenerates to the ordinary NMF in Eq. (1).

3.3. Optimization

As we see, the objective function in Eq. (7) is minimized with respect to $U$ and $V$, and we cannot give a closed-form solution. In the following, we introduce an iterative scheme to optimize the model objective function. In other words, we will optimize the objective function with respect to one variable when fixing the other one. This iterative procedure repeats until convergence.

The objective function in Eq. (7) can be rewritten as:

$$J_{\text{DNMF}} = \text{Tr}((X - UV^{T})^{T}(X - UV^{T})) + \lambda \text{Tr}(V^{T}LV) + \mu \text{Tr}(U^{T}L_{U}U),$$

$$= \text{Tr}(XX^{T}) - 2\text{Tr}(XX^{T}V^{T}U^{T}) + \text{Tr}(UV^{T}VU^{T}) + \lambda \text{Tr}(V^{T}LV) + \mu \text{Tr}(U^{T}L_{U}U), \tag{8}$$

where the second equality applies the matrix properties $\text{Tr}(AB) = \text{Tr}(BA)$ and $\text{Tr}(A) = \text{Tr}(A^{T})$. Let $\Psi_{U}$ and $\Phi_{V}$ be the Lagrange multiplier for constraints $U_{ij} \geq 0$ and $V_{ij} \geq 0$, respectively. Then the Lagrange function $\mathcal{L}$ is

$$\mathcal{L} = \text{Tr}(XX^{T}) - 2\text{Tr}(XX^{T}V^{T}U^{T}) + \text{Tr}(UV^{T}VU^{T}) + \lambda \text{Tr}(V^{T}LV) + \mu \text{Tr}(U^{T}L_{U}U) + \text{Tr}(\Psi_{U}U^{T}) + \text{Tr}(\Phi_{V}V). \tag{9}$$

3.3.1. Updating $U$

The partial derivation of $\mathcal{L}$ with respect to $U$ is

$$\frac{\partial \mathcal{L}}{\partial U} = -2XX^{T} + 2UV^{T}V + 2\mu L_{U}U + \Psi_{U}.$$

Using the KKT conditions $\Psi_{U}U_{ij} = 0$, we can get

$$-XX^{T} + UV^{T}V + \mu L_{U}U_{ij} = 0. \tag{10}$$

According to Eq. (10), we present the following updating formula.

$$U_{ij} \leftarrow U_{ij} \frac{[XX^{T} + \mu L_{U}]U_{ij}}{[UV^{T}V + \mu L_{U}]U_{ij}}. \tag{11}$$

3.3.2. Updating $V$

The partial derivation of $\mathcal{L}$ with respect to $V$ is

$$\frac{\partial \mathcal{L}}{\partial V} = -2XX^{T} + 2UV^{T}U + 2\lambda LV + \Phi_{V}.$$

Using the KKT conditions $\Phi_{V}V_{ij} = 0$, we can get

$$-XX^{T} + UV^{T}U + \lambda LV_{ij}V_{ij} = 0. \tag{12}$$

According to Eq. (12), we present the following updating formula.

$$V_{ij} \leftarrow V_{ij} \frac{[XX^{T} + \lambda LV]V_{ij}}{[UV^{T}U + \lambda LV]V_{ij}}. \tag{13}$$

3.4. Convergence analysis

In the following, we will investigate the convergence of the updating rules in Eqs. (11) and (13). And regarding these two updating rules, we have the following theorem:

Theorem 4. For $X$, $U$, $V \geq 0$, the objective function in Eq. (7) is non-increasing under the updating rules in Eqs. (11) and (13).
Please see the Appendix A for the detailed proof for the above theorem. Our proof follows the idea in the proof of two papers [34,19] for NMF and GNMF. Recent studies [35,36] have shown that the alternate updating rules in Eq. (2) do not guarantee the convergence to a stationary point. But a slight modification proposed in [36] achieves this property. Our alternate updating rules in Eqs. (11) and (13) are essentially similar to the updating rules for GNMF, therefore the minor modification can also be applied.

Following the work [19], we know that the multiplicative updating rules in Eqs. (11) and (13) are special cases of gradient descent with an automatic step parameter selection, and can be interpreted as a rescaled gradient descent scheme. Theorem 4 guarantees that the iterative multiplicative updating rules converge to a local optimum.

After our iterative updating procedure converges, we normalize the length of each column vector in the factor $U$ to 1, and adjust accordingly the factor $V$ as follows:

$$U_{ij} \leftarrow \frac{U_{ij}}{\sqrt{\sum_k U_{kj}^2}}, \quad V_{ij} \leftarrow V_{ij} \sqrt{\sum_l U_{lj}^2}. \tag{14}$$

4. Graph dual regularization non-negative matrix tri-factorization

In this section, we propose a novel graph dual regularization non-negative matrix tri-factorization (DNMTF) as an extension of our DNMF algorithm, and simultaneously incorporate two graph regularizers of both data manifold and feature manifold into its objective function. Then we present an optimization scheme based on the iterative updating rules of three factor matrices to solve the objective function. Finally, we present the convergence proof of our iterative updating scheme.

4.1. Objective function of DNMTF

Based on two graph regularizers of both data manifold and feature manifold, we propose a novel graph dual regularization non-negative matrix tri-factorization (DNMTF), whose objective function is formulated as follows:

$$J_{DNMTF} = \|X - USVT\|^2_F + \mu \text{Tr}(U^2L_U) + \mu \text{Tr}(V^2L_V) + \mu \text{Tr}(\Psi^2U^2), \tag{15}$$

s.t. $U \geq 0, \mu \geq 0, V \geq 0,$

where $\mu \geq 0$ are the regularization parameters which balance the reconstruction error of DNMTF in the first term and graph regularizers in the second and third terms. When letting $\mu = 0$, DNMTF degenerates to ordinary nonnegative matrix tri-factorization.

4.2. Optimization

As we see, the objective function in Eq. (15) is minimized with respect to $U, V$ and $S$, and we cannot give a closed-form solution. In the following, we introduce an alternating scheme to optimize the objective function. In other words, we will optimize the objective function with respect to one variable when fixing the other one. This alternating procedure repeats until convergence.

The objective function in Eq. (15) can be rewritten as:

$$J_{DNMTF} = \text{Tr}(X^2 - USVT^2) + \mu \text{Tr}(V^2L_V) + \mu \text{Tr}(\Psi^2U^2) + \mu \text{Tr}(U^2L_U) = \text{Tr}(XX^T) - \text{Tr}(XSVST^T) + \mu \text{Tr}(V^2L_V) \tag{16}$$

Let $\Psi_{ij}$ and $\Phi_{ij}$ be the Lagrange multipliers for constraints $U_{ij} \geq 0$ and $V_{ij} \geq 0$, respectively. Then the Lagrange function $\mathcal{L}$ is

$$\mathcal{L} = \text{Tr}(XX^T) - \text{Tr}(XSVST^T) + \mu \text{Tr}(V^2L_V) + \mu \text{Tr}(U^2L_U) + \mu \text{Tr}(U^2L_U) + \mu \text{Tr}(\Psi^2U^2) + \mu \text{Tr}(\Phi^2V^T) + \mu \text{Tr}(\Omega^2S^T). \tag{17}$$

4.2.1. Updating $S$

The partial derivation of $\mathcal{L}$ with respect to $S$ is

$$\frac{\partial \mathcal{L}}{\partial S} = -2USVT + 2USVTV + \Omega.$$  

Using the KKT conditions $\Omega_{ij}S_{ij} = 0$, we can get

$$-2USVT + 2USVTV + \Omega = 0. \tag{18}$$

According to Eq. (17), we present the following updating formula,

$$S_{ij} \leftarrow S_{ij} \frac{[USVT]_{ij}}{[USVT]_{ij}^2}. \tag{19}$$

4.2.2. Updating $U$

The partial derivation of $\mathcal{L}$ with respect to $U$ is

$$\frac{\partial \mathcal{L}}{\partial U} = -2XTUS + 2USVT^2 + 2\mu LV - \Phi.$$  

Using the KKT conditions $\Phi_{ij}U_{ij} = 0$, we can get

$$-2XTUS + 2USVT^2 + 2\mu LV - \Phi = 0. \tag{20}$$

Since $L_U = D^T - W^T$, then the above function can be rewritten as:

$$-2XTUS + 2USVT^2 + 2\mu LV - \Phi = 0.$$  

According to Eq. (20), we present the following updating formula,

$$U_{ij} \leftarrow U_{ij} \frac{[USVT]^2 + \mu LV_{ij}^2}{[USVT]_{ij}^2 + \mu LV^2_{ij}}. \tag{21}$$

4.2.3. Updating $V$

The partial derivation of $\mathcal{L}$ with respect to $V$ is

$$\frac{\partial \mathcal{L}}{\partial V} = -2XTUS + 2VS^2 + 2\mu LV - \Phi.$$  

Using the KKT conditions $\Phi_{ij}V_{ij} = 0$, we can get

$$-2XTUS + 2VS^2 + 2\mu LV - \Phi = 0. \tag{22}$$

Since $L_V = D^T - W^T$, then the above function can be rewritten as:

$$-2XTUS + 2VS^2 + 2\mu LV - \Phi = 0.$$  

According to Eq. (22), we present the following updating formula,

$$V_{ij} \leftarrow V_{ij} \frac{|US| + \mu LV_{ij}^2}{|VS| + \mu LV^2_{ij}}. \tag{23}$$

4.3. Convergence analysis

In the following, we will investigate the convergence of the updating rules in Eqs. (19), (21) and (23). And regarding these three updating rules, we have the following theorem:

**Theorem 5.** For $X, U, V, S \geq 0$, the objective function in Eq. (15) is non-decreasing under the updating rules stated in Eqs. (19), (21) and (23).

Please see the Appendix B for the detailed proof for the above theorem. Following the works [19,22], we know that the multiplicative updating rules in Eqs. (19), (21) and (23) are special
4.4. Complexity analysis

In this part, we discuss the computational cost of two proposed algorithms: DNMF and DNMTF. For DNMF and DNMTF, the same k-nearest neighbor graph needs $O(N^2 M + NM^2)$ to construct. Since two weight matrices $W^G$ and $W^V$ of DNMF and DNMTF are sparse, based on the updating rules of two algorithms: DNMF and DNMTF, the cost of two iterative multiplicative updating procedures is $O(tMNC)$, where $t$ is the numbers of iterations, and $C$ is the number of clusters. The overall cost for DNMF and DNMTF is $O(N^2 M + NM^2 + tMNC)$.

5. Experiments

In this section, we evaluate the performances of two proposed algorithms: DNMF and DNMTF on a number of real-world data sets, including eight UCI data sets (Section 5.2), five image data sets (Section 5.3), and a radar HRRP data set (Section 5.5), and discuss the sensitivity of our proposed DNMF and DNMTF in relation to their parameters (Section 5.4). All experiments were performed with Matlab 7.1 on a Pentium-IV 3.20 GHz PC running Windows XP with 1-GB main memory.

5.1. Compared algorithms

We compare the performances of two proposed algorithms: DNMF and DNMTF against six existing state-of-the-art clustering algorithms or related methods.

- K-means clustering method (Kmeans).
- Normalized Cut (NCut) [2], one of the typical spectral clustering methods. For NCut, the scale parameter of Gaussian function is searched from the grid of $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$.
- Nonnegative Matrix Factorization based clustering method (NMF) [34].
- Semi-Nonnegative Matrix Factorization based clustering method (SNMF) [37].
- Graph regularized Nonnegative Matrix Factorization (GNMF) [19]. In the implementation of GNMF, we use the 0–1 weighting scheme for constructing the $k$-nearest neighbor graph as in [19]. The number of nearest neighbor $k$ is set by the grid $\{1, 2, 3, \ldots, 10\}$ and the regularization parameter $\lambda$ is set by searching the grid $\{0.1, 1, 10, 100, 500, 1000\}$.
- Following the works [19, 28], we also implement the normalized cut weighted versions of GNMF, DRCC, and two proposed algorithms: DNMF and DNMTF. Let $D = \text{diag}(X^T X e)$, where $e$ is a vector of all ones, and conduct data clustering using the weighted data matrix $X = XD^{-1/2}$.
- Dual regularized co-clustering (DRCC) method based on semi-nonnegative matrix tri-factorization [24]. In the implementation of DRCC, its parameter setting is the same as GNMF. For simplicity, the neighborhood size of the data graph is set to be the same value as that of the feature graph. Furthermore, the regularization parameter $\lambda$ is also set to be the same value as the regularization parameter $\mu$.
- Two proposed algorithms: graph dual regularization non-negative matrix factorization (DNMF) and graph dual

| Table 1 |
| Description of the eight UCI data sets. |
| Data Sets | No. of samples (N) | No. of features (M) | No. of classes (C) |
| Glass | 214 | 9 | 6 |
| Heart | 270 | 13 | 2 |
| Vehicle | 846 | 19 | 4 |
| Wpbc | 198 | 33 | 2 |
| Wine | 178 | 13 | 3 |
| Soybean | 47 | 35 | 4 |
| SPECTF | 267 | 45 | 2 |
| Semeion | 1593 | 256 | 10 |

| Table 2 |
| Clustering accuracy (ACC) on UCI data sets. |
| Data sets | K-means | NCut | NMF | SNMF | GNMF | DRCC | DNMF | DNMTF |
| Glass | 0.5280 | 0.4902 | 0.2682 | 0.3827 | 0.4145 | 0.5047 | 0.5383 | 0.5514 |
| Heart | 0.5900 | 0.6074 | 0.6222 | 0.6222 | 0.6185 | 0.6076 | 0.6259 | 0.6188 |
| Vehicle | 0.4512 | 0.4533 | 0.4241 | 0.3757 | 0.4208 | 0.4349 | 0.4965 | 0.4663 |
| Wpbc | 0.5610 | 0.5758 | 0.5540 | 0.5571 | 0.5657 | 0.5677 | 0.5859 | 0.5768 |
| Wine | 0.6594 | 0.7079 | 0.6090 | 0.6034 | 0.6303 | 0.6853 | 0.7247 | 0.7078 |
| Soybean | 0.7014 | 0.7319 | 0.6915 | 0.6851 | 0.7468 | 0.7489 | 0.7434 | 0.7827 |
| SPECTF | 0.6255 | 0.5019 | 0.5543 | 0.5472 | 0.5880 | 0.6307 | 0.6629 | 0.6599 |
| Semeion | 0.5455 | 0.5061 | 0.4901 | 0.5579 | 0.6218 | 0.6306 | 0.6529 | 0.6728 |

| Table 3 |
| Normalized mutual information (NMI) on UCI data sets. |
| Data sets | Kmeans | NCut | NMF | SNMF | GNMF | DRCC | DNMF | DNMTF |
| Glass | 0.3391 | 0.3535 | 0.0585 | 0.2452 | 0.2945 | 0.3282 | 0.3704 | 0.3902 |
| Heart | 0.0189 | 0.0315 | 0.0372 | 0.0371 | 0.0344 | 0.0319 | 0.0377 | 0.0374 |
| Vehicle | 0.1855 | 0.1791 | 0.1529 | 0.1126 | 0.1690 | 0.1879 | 0.2393 | 0.2184 |
| Wpbc | 0.0270 | 0.0161 | 0.0312 | 0.0302 | 0.0267 | 0.0093 | 0.0279 | 0.0283 |
| Wine | 0.4269 | 0.4315 | 0.2946 | 0.3139 | 0.2633 | 0.4353 | 0.4327 | 0.4359 |
| Soybean | 0.7146 | 0.7463 | 0.6847 | 0.7163 | 0.7521 | 0.7550 | 0.7728 | 0.7570 |
| SPECTF | 0.0898 | 0.1494 | 0.1218 | 0.1251 | 0.1063 | 0.0875 | 0.0848 | 0.0881 |
| Semeion | 0.4989 | 0.4685 | 0.4265 | 0.5060 | 0.6269 | 0.6379 | 0.6476 | 0.6513 |
regularization non-negative matrix tri-factorization (DNMTF). In the implementation of two proposed algorithms: DNMF and DNMTF, their parameter settings are set to be the same as that of GNMF and DRCC.

In all experiments, we set the number of clusters equal to the true number of classes $C$ for all the clustering algorithms. And we use two popular evaluation metrics, the clustering accuracy (ACC) and the normalized mutual information (NMI), to measure the

---

Fig. 1. Convergence curve of two proposed algorithms: DNMF and DNMTF.
performance of all the clustering algorithms. Please see [6] for the detailed definitions of these two metrics.

5.2. UCI data sets

In this part, we perform experiments on eight UCI nonnegative data sets. The basic information of these UCI data sets is summarized in Table 1.

The clustering results of all these clustering algorithms on eight UCI data sets are shown in Tables 2 and 3, in which the best performance for each data set is highlighted. From Tables 2 and 3, we observe that NCut, GNMF, DRCC, and two proposed algorithms: DNMF and DNMTF consider the geometrical structure of the data and usually perform better than the other three algorithms including Kmeans, NMF, and SNMF. Moreover, two proposed algorithms: DNMF and DNMTF, and DRCC commonly outperform the other five methods in terms of ACC and NMI.

Furthermore, we also show the convergence curves of both DNMF and DNMTF on all these UCI data sets to investigate how fast our updating rules can converge, as illustrated in Fig. 1. For each figure, the abscissa represents the iteration number and the ordinate denotes the log-value of objective function (e.g., log(DNMF or log(DNMTF)). We can see that our iterative multiplicative updating rules for both DNMF and DNMTF converge very fast, usually within 100 iterations.

5.3. Image data sets

In the part, we perform experiments on five image data sets: COIL20 image library [38], Optdigits1, UMIST face database,2 ORL face dataset3 and JAFFE facial expression database [39]. The basic information of these five image data sets is summarized in Table 4. The clustering results of all these algorithms on five image data sets are shown in Tables 5–14, in which the best performance for each data set is highlighted. For each given cluster number C, we conduct 50 independent runs and report the mean of the performance in the tables. From the results shown in Tables 5–14, we can observe the following:

- Both Kmeans and NMF perform generally much inferior, and they do not consider the intrinsic geometrical structure of the data. Since the distributions of these data sets, especially image data sets whose dimension are generally high, are commonly much more complicated than mixtures of spherical Gaussians.
- SNMF usually outperforms NMF and Kmeans on these data sets.
- NCut, GNMF, DRCC and two proposed algorithms: DNMF and DNMTF commonly achieve better performance than the other three algorithms including Kmeans, NMF, and SNMF. The former five algorithms consider the geometrical structure information contained in the data. This suggests that the underlying manifold structure of the data is useful in data clustering.
- DRCC and two proposed algorithms: DNMF and DNMTF usually outperform the other five algorithms. And they simultaneously consider the geometric structures of both data manifold and feature manifold, and therefore they have more discriminating power than NMF, SNMF, and GNMF.
- From Tables 5–14, we can see that two proposed algorithms: DNMF and DNMTF consistently perform better than the other six algorithms in terms of ACC and NMI. This indicates that our two proposed algorithms can learn better parts-based representations of data.

Below, we use one specific example to investigate the sparseness of the basis vectors learned in two proposed algorithms: DNMF and DNMTF. Fig. 2 shows the basis vectors learned by NMF, GNMF, DRCC, and two proposed algorithms: DNMF and DNMTF on the COIL20 data set, respectively. We plot these learned basis vectors as gray scale images. From the results shown in Fig. 2, we can see that the basis vectors learned by GNMF and two proposed algorithms: DNMF and DNMTF are sparser than those learned by NMF and DRCC, and this suggests that two proposed algorithms: DNMF and DNMTF, and GNMF can learn better parts-based representations of data than NMF and DRCC.

5.4. Sensitivity in relation to parameters

There are mainly three parameters in our two proposed algorithms, DNMF and DNMTF: the size of neighborhood k, and the regularization parameters λ and μ. We conduct four experiments on the UCI SPECTF and Semeion, COIL20, and ORL data sets to test the sensitivity of two proposed algorithms to the selection

---

Table 4
Description of the five image data sets.

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>No. of samples (N)</th>
<th>No. of features (M)</th>
<th>No. of classes (C)</th>
</tr>
</thead>
<tbody>
<tr>
<td>COIL20</td>
<td>1440</td>
<td>1024</td>
<td>20</td>
</tr>
<tr>
<td>Optdigits</td>
<td>3823</td>
<td>64</td>
<td>10</td>
</tr>
<tr>
<td>UMIST</td>
<td>575</td>
<td>2576</td>
<td>20</td>
</tr>
<tr>
<td>ORL</td>
<td>400</td>
<td>1024</td>
<td>40</td>
</tr>
<tr>
<td>JAFFE</td>
<td>213</td>
<td>4096</td>
<td>10</td>
</tr>
</tbody>
</table>

Table 5
Clustering accuracy (ACC) on COIL20 data set.

<table>
<thead>
<tr>
<th>C</th>
<th>Kmeans</th>
<th>NCut</th>
<th>NMF</th>
<th>SNMF</th>
<th>GNMF</th>
<th>DRCC</th>
<th>DNMF</th>
<th>DNMTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.6719</td>
<td>0.6958</td>
<td>0.6701</td>
<td>0.6876</td>
<td>0.8507</td>
<td>0.8542</td>
<td>0.8576</td>
<td>0.8854</td>
</tr>
<tr>
<td>6</td>
<td>0.4678</td>
<td>0.5475</td>
<td>0.6420</td>
<td>0.5104</td>
<td>0.6923</td>
<td>0.6839</td>
<td>0.7175</td>
<td>0.7338</td>
</tr>
<tr>
<td>8</td>
<td>0.4679</td>
<td>0.5021</td>
<td>0.4583</td>
<td>0.3278</td>
<td>0.7100</td>
<td>0.7361</td>
<td>0.7436</td>
<td>0.7693</td>
</tr>
<tr>
<td>10</td>
<td>0.5360</td>
<td>0.5567</td>
<td>0.4912</td>
<td>0.5763</td>
<td>0.6458</td>
<td>0.6764</td>
<td>0.7222</td>
<td>0.7361</td>
</tr>
<tr>
<td>12</td>
<td>0.5366</td>
<td>0.5854</td>
<td>0.5637</td>
<td>0.6157</td>
<td>0.6609</td>
<td>0.6887</td>
<td>0.7160</td>
<td>0.7245</td>
</tr>
<tr>
<td>14</td>
<td>0.5849</td>
<td>0.6057</td>
<td>0.5925</td>
<td>0.6042</td>
<td>0.7044</td>
<td>0.7113</td>
<td>0.7282</td>
<td>0.7549</td>
</tr>
<tr>
<td>16</td>
<td>0.5825</td>
<td>0.6528</td>
<td>0.5875</td>
<td>0.6076</td>
<td>0.7067</td>
<td>0.7213</td>
<td>0.7543</td>
<td>0.7795</td>
</tr>
<tr>
<td>18</td>
<td>0.6065</td>
<td>0.6689</td>
<td>0.6228</td>
<td>0.6382</td>
<td>0.6862</td>
<td>0.6914</td>
<td>0.7291</td>
<td>0.7508</td>
</tr>
<tr>
<td>20</td>
<td>0.5910</td>
<td>0.6387</td>
<td>0.5570</td>
<td>0.6237</td>
<td>0.6772</td>
<td>0.6756</td>
<td>0.6984</td>
<td>0.7261</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.5606</td>
<td>0.6057</td>
<td>0.5561</td>
<td>0.5991</td>
<td>0.7029</td>
<td>0.7157</td>
<td>0.7409</td>
<td>0.7634</td>
</tr>
</tbody>
</table>

Table 6
Normalized mutual information (NMI) on COIL20 data set.

<table>
<thead>
<tr>
<th>C</th>
<th>Kmeans</th>
<th>NCut</th>
<th>NMF</th>
<th>SNMF</th>
<th>GNMF</th>
<th>DRCC</th>
<th>DNMF</th>
<th>DNMTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>0.5274</td>
<td>0.6113</td>
<td>0.5340</td>
<td>0.5648</td>
<td>0.8400</td>
<td>0.8489</td>
<td>0.8412</td>
<td>0.9008</td>
</tr>
<tr>
<td>6</td>
<td>0.4473</td>
<td>0.6384</td>
<td>0.4787</td>
<td>0.5203</td>
<td>0.7530</td>
<td>0.7433</td>
<td>0.7884</td>
<td>0.8330</td>
</tr>
<tr>
<td>8</td>
<td>0.5468</td>
<td>0.6805</td>
<td>0.5236</td>
<td>0.5504</td>
<td>0.7769</td>
<td>0.8014</td>
<td>0.8095</td>
<td>0.8107</td>
</tr>
<tr>
<td>10</td>
<td>0.6156</td>
<td>0.6569</td>
<td>0.5541</td>
<td>0.6212</td>
<td>0.7586</td>
<td>0.7911</td>
<td>0.7946</td>
<td>0.8120</td>
</tr>
<tr>
<td>12</td>
<td>0.6551</td>
<td>0.7274</td>
<td>0.6315</td>
<td>0.6749</td>
<td>0.7872</td>
<td>0.8365</td>
<td>0.8423</td>
<td>0.8466</td>
</tr>
<tr>
<td>14</td>
<td>0.7097</td>
<td>0.7504</td>
<td>0.6827</td>
<td>0.7043</td>
<td>0.8164</td>
<td>0.8325</td>
<td>0.8567</td>
<td>0.8629</td>
</tr>
<tr>
<td>16</td>
<td>0.7392</td>
<td>0.7787</td>
<td>0.7040</td>
<td>0.7288</td>
<td>0.8489</td>
<td>0.8674</td>
<td>0.8781</td>
<td>0.8822</td>
</tr>
<tr>
<td>18</td>
<td>0.7640</td>
<td>0.8057</td>
<td>0.7411</td>
<td>0.7650</td>
<td>0.8342</td>
<td>0.8598</td>
<td>0.8676</td>
<td>0.8743</td>
</tr>
<tr>
<td>20</td>
<td>0.7334</td>
<td>0.7691</td>
<td>0.7097</td>
<td>0.7436</td>
<td>0.8147</td>
<td>0.8386</td>
<td>0.8502</td>
<td>0.8587</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.6378</td>
<td>0.7132</td>
<td>0.6675</td>
<td>0.6526</td>
<td>0.8033</td>
<td>0.8238</td>
<td>0.8365</td>
<td>0.8535</td>
</tr>
</tbody>
</table>

2 http://images.ee.umist.ac.uk/danny/database.html.

---
From these figures, we can clearly see that:

- The performance of two proposed algorithms: DNMF and DNMTF is very robust to the value of two regularization parameters \( \lambda \) and \( \mu \).

5.5. Radar HRRP data set

Radar target High-Resolution Range Profile (HRRP) is the amplitude of coherent summations of complex time returns from
target scatterers in each range resolution cell, which represents
the projection of complex returned echoes from the target
scattering centers onto the radar line-of-sight (LOS) [40–42].
The selected radar HRRP data set was measured continuously
when the target airplanes including An-26, Yark-42 and
Cessna Citation S/II were flying. The parameters of target air-
planes and radar are shown in Table 15, and the projections of
target trajectories onto the ground plane are shown in Fig. 5,
from which the aspect angle of the airplane can be estimated
according to its relative position to radar. As shown in
Fig. 5, the first, third and fourth segments of the Yark-42, the
first–fourth segments of the An-26, and the first–fifth segments of

<table>
<thead>
<tr>
<th>C</th>
<th>Kmeans NCut NMF SNMF GNMF DRCC DNMF DNMTF</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>0.8625 0.7249 0.7919 0.9652 0.9665 0.9653 1.0000</td>
</tr>
<tr>
<td>4</td>
<td>0.8329 0.9336 0.8828 0.9003 0.9355 0.9328 0.9329 0.9663</td>
</tr>
<tr>
<td>5</td>
<td>0.7744 0.7401 0.8146 0.8768 0.8908 0.9097 0.9252 0.9229</td>
</tr>
<tr>
<td>6</td>
<td>0.8090 0.8597 0.8251 0.8663 0.8678 0.8841 0.9082 0.9121</td>
</tr>
<tr>
<td>7</td>
<td>0.7691 0.8094 0.8596 0.8792 0.8973 0.9092 0.9179 0.9149</td>
</tr>
<tr>
<td>8</td>
<td>0.7965 0.8227 0.8463 0.8596 0.8630 0.8796 0.9072 0.9116</td>
</tr>
<tr>
<td>9</td>
<td>0.8271 0.8445 0.8520 0.8587 0.8853 0.8826 0.9030 0.8972</td>
</tr>
<tr>
<td>10</td>
<td>0.8006 0.8546 0.8133 0.8332 0.8716 0.8768 0.8867 0.8806</td>
</tr>
<tr>
<td>Avg.</td>
<td>0.8090 0.8436 0.8273 0.8583 0.8971 0.9052 0.9183 0.9257</td>
</tr>
</tbody>
</table>

Fig. 2. Basis vectors learned from the COIL20 data set.
Fig. 3. The performance of two proposed algorithms varies with the size of neighborhood.

Fig. 4. The performance of two proposed algorithms varies with the regularization parameters.
the Cessna Citation S/II, are taken as the experimental data. Moreover, because of the property of time-shift alignment of HRRP [43], the $l_2$-norm normalized power spectrum of HRRP is used to perform clustering because of its time-shift invariance.

Each power spectrum for an HRRP sample is a 128-dimensional vector, and an example of HRRP power spectrum is shown in Fig. 6.

The clustering results on the radar HRRP data set are shown in Fig. 7, from which we can see that two proposed algorithms: DNMF and DNMTF, GNMF and DRCC outperform the other four methods in terms of ACC and NMI. The former four algorithms can effectively make use of the geometrical structure information contained in the data set. Furthermore, two proposed algorithms: DNMF and DNMTF, and DRCC achieve better performance than the other five algorithms in terms of ACC and NMI, and they simultaneously consider the manifold structures of both the data manifold and the feature manifold.

6. Conclusions

In this paper, we proposed a novel algorithm, called graph dual regularization non-negative matrix factorization (DNMF), which simultaneously considers the geometric structures of both data manifold and feature manifold. We also presented a graph dual regularization non-negative matrix tri-factorization algorithm (DNMTF) as an extension of DNMF. Since two proposed algorithms: DNMF and DNMTF can effectively make use of the
structure information contained in data as well as features, they have more discriminating power than NMF and GNMF. Then we developed two iterative updating optimization schemes for DNMF and DNMTF, respectively, and provided the convergence proofs of our two optimization schemes. Finally, we provided a variety of experiments on UCI benchmark data sets, several image data sets and our two optimization schemes. Finally, we provided a variety of experiments on UCI benchmark data sets, several image data sets and our two optimization schemes. Finally, we provided a variety of experiments on UCI benchmark data sets, several image data sets and our two optimization schemes. Finally, we provided a variety of experiments on UCI benchmark data sets, several image data sets and our two optimization schemes.

Acknowledgments

We would like to thank the anonymous reviewers for their valuable comments and suggestions to significantly improve the quality of this paper. We also thank Quanquan Gu of University of Illinois at Urbana-Champaign and Dr. Jing Chai of Duke University for numerous discussions and suggestions. This work is supported by the National Natural Science Foundation of China Nos. 60803057, 61003198, 60970067; the National Science and Technology Ministry of China Nos. 9140A07011810D20107, 9140A07021010D20131; the Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) No. B07048; the Fundamental Research Funds for the Central Universities Nos. JY1000902001, K50510020001, JY1000902045.

Appendix A (Proofs of Theorem 1)

To prove Theorem 1, we need to show that the objective function $J_{DNMF}$ in Eq. (7) is non-increasing under the updating rules stated in Eqs. (11) and (13). We make use of an auxiliary function similar to that used in the EM algorithm [44] to prove the convergence of Theorem 1. Furthermore, considering the symmetry of $U$ and $V$ in Eq. (7), we only need to prove the convergence under the updating rule for $U$ in Eq. (11).

Definition $G(u, u')$ is an auxiliary function for $F(u)$ if the conditions $G(u, u') \geq F(u)$, and $G(u, u') = F(u)$ are satisfied.

**Lemma 1.** If $G$ is an auxiliary function of $F$, then $F$ is non-increasing under the following updating formula,

$$u^{(t+1)} = \arg\min_u G(u, u^{(t)}).$$

**(Proof.)**

Next we will show that the updating rule for $U$ in Eq. (11) is exactly the update in Eq. (24) with a proper auxiliary function. Considering any element $U_{ij}$ in $U$, we let

$$F(U) = ||X-UV^T||^2 + \mu \text{Tr}(U^TU).$$

And we get

$$F_{ij} = \frac{\partial F}{\partial U_{ij}} = [-2XV + 2UV^TV + 2\mu U_{ij}]_{ij}, \quad \text{and}$$

$$F_{ij} = 2[V^TV_{ij} + \mu |L_{ij}|].$$

Since our updating rule is essentially element-wise, it is sufficient to show that each $F_{ij}$ is non-increasing under the updating formula in Eq. (11).

**Lemma 2.** The function

$$G(U_{ij}, U_{ij}^{(t)}) = F_{ij}(U_{ij}^{(t)}) + F_{ij}(U_{ij}^{(t)})(U_{ij} - U_{ij}^{(t)}) + \frac{\left[U^TV + \mu D^T U_{ij}^{(t)}(U_{ij} - U_{ij}^{(t)})^2\right]}{U_{ij}^{(t)}}$$

is an auxiliary function for $F_{ij}$.

**(Proof.)**

We first get the Taylor series expansion of $F_{ij}(U_{ij})$

$$F_{ij}(U_{ij}) = F_{ij}(U_{ij}^{(t)}) + F_{ij}(U_{ij}^{(t)})(U_{ij} - U_{ij}^{(t)}) + \left[U^TV_{ij} + \mu |L_{ij}|(U_{ij} - U_{ij}^{(t)})^2\right].$$

With Eq. (25) to find that $G(U_{ij}, U_{ij}^{(t)}) \geq F_{ij}(U_{ij})$ is equivalent to

$$\frac{\left[U^TV + \mu D^T U_{ij}^{(t)}\right]}{U_{ij}^{(t)}} \geq |V^TV_{ij} + \mu |L_{ij}|.$$  

(26)

We can get

$$|V^TV_{ij}| = \sum_{l=1}^{K} U_{ij}^{(t)}|V^TV_{lj}| \geq U_{ij}^{(t)}|V^TV_{lj}|,$$

and

$$\mu D^T U_{ij} = \mu \sum_{l=1}^{M} D_{ij}^{(l)}U_{lj}^{(l)} \geq D_{ij}^{(l)}U_{lj}^{(l)} \geq |D^T - W^{(t)}|U_{ij}^{(t)} = |D|U_{ij}^{(t)}.$$  

(25)
Thus, Eq. (26) holds and $G(U_{ij}^t, V_{ij}^t) \geq F_{ij}(U_{ij})$. Furthermore, $G(U_{ij}, U_{ij}) = F_{ij}(U_{ij})$ is obvious.

**Proof of Theorem 1.** Replacing $G(U_{ij}, V_{ij}^t)$ in Eq. (24) by Eq. (25), we get

$$U_{ij}^{t+1} = U_{ij}^t - \frac{F_{ij}(U_{ij}^t)}{2U_{ij}^T U_{ij} + \mu D_{ii}^2 U_{ij}} = U_{ij}^t \frac{\Sigma V^T + \mu W_{ij} U_{ij}}{U_{ij}^T \Sigma V^T + \mu D_{ii}^2 U_{ij}}.$$  

Since Eq. (25) is an auxiliary function for $F_{ij}$, $F_{ij}$ is non-increasing under the updating rule stated in Eq. (11). Note the symmetry of $U$ and $V$ in Eq. (7), we also can prove the analogous convergence under the updating rule for $V$ in Eq. (13).

**Appendix B (proofs of Theorem 2)**

To prove Theorem 2, we need to show that the objective function $J_{DMMT}$ in Eq. (15) is non-increasing under the updating rules stated in Eqs. (19), (21) and (23). Similarly, the first term of $J_{DMMT}$ in Eq. (15) is only related to $S$, we have exactly the same convergence rule for $S$ as in ONMTF [22]. Thus, we can use the convergence proof of ONMTF to show that $J_{DMMT}$ is monotonically decreasing under the updating rule in Eq. (19). Please see [22,45] for details. Considering the symmetry of $U$ and $V$ in Eq. (15), we only need to prove the convergence rule for the updating rule for $U$ in Eq. (21).

Below, we will make use of an auxiliary function to prove the convergence of the updating rule for $U$. Considering any element $U_{ij}^t$ in $U$, we let

$$F_t(U) = \|X - USVT\|^2_F + \mu Tr(U^T L_U U).$$

And we get

$$F_{ij}^t = \frac{\partial F}{\partial U_{ij}} = \frac{-2XSV^T + 2USVT^2 + 2\mu L_U U_{ij}}{U_{ij}^T \Sigma V^T + \mu D_{ii}^2 U_{ij}}, \text{ and}$$

$$F_{ij}^t = 2SV^T \Sigma V^T + 2\mu L_U U_{ij}.$$  

Since our updating rule is essentially element-wise, it is sufficient to show that each $F_{ij}$ is non-increasing under the updating formula in Eq. (21).

**Lemma 3. Function**

$$G(U_{ij}^t, V_{ij}^t) = F_{ij}(U_{ij}^t) + F_{ij}(U_{ij}^t) (U_{ij} - U_{ij}^t) + \frac{[USVT^2 + \mu D_{ii}^2 U_{ij}]}{U_{ij}^T} (U_{ij} - U_{ij}^t)^2.$$  

is an auxiliary function for $F_{ij}$.

**Proof.** We first get the Taylor series expansion of $F_{ij}(U_{ij})$

$$F_{ij}(U_{ij}) = F_{ij}(U_{ij}^t) + F_{ij}(U_{ij}^t) (U_{ij} - U_{ij}^t) + \frac{1}{2} (U_{ij} - U_{ij}^t)^2.$$  

With Eq. (27) to find that $G(U_{ij}, U_{ij}^t) \geq F_{ij}(U_{ij})$ is equivalent to

$$\|USVT^2 + \mu D_{ii}^2 U_{ij}\|_{U_{ij}^T} \geq \|SV^T \Sigma V^T_{ij} + \mu L_U U_{ij}\|_{U_{ij}^T}.$$  

We can get

$$\|SV^T \Sigma V^T_{ij}\|_{U_{ij}^T} = \sum_{l=1}^k \|U_{ij}^l SV^T \Sigma V^T_{ij} \|_{U_{ij}^l SV^T \Sigma V^T_{ij}} \|V^T \Sigma V^T_{ij} \|_{U_{ij}^l SV^T \Sigma V^T_{ij}},$$

and

$$\mu D_{ii}^2 U_{ij} = \mu \sum_{l=1}^M D_{ii}^2 U_{ij}^l \sum_{l=1}^M D_{ii}^2 U_{ij}^l \geq \mu D_{ii}^2 U_{ij}^l U_{ij}^l = \mu |L_U| U_{ij}^l.$$  

Thus, Eq. (28) holds and $G(U_{ij}, U_{ij}^t) \geq F_{ij}(U_{ij})$. Furthermore, $G(U_{ij}, U_{ij}) = F_{ij}(U_{ij})$ is obvious.

Replacing $G(U_{ij}, U_{ij}^t)$ in Eq. (24) by Eq. (27), we get

$$U_{ij}^{t+1} = U_{ij}^t - \frac{F_{ij}(U_{ij}^t)}{2U_{ij}^T U_{ij}^t + \mu D_{ii}^2 U_{ij}^t} = U_{ij}^t \frac{SV^T + \mu W_{ij} U_{ij}}{USVT^2 + \mu D_{ii}^2 U_{ij}}.$$  

Since Eq. (27) is an auxiliary function for $F_{ij}$, $F_{ij}$ is non-increasing under the updating rule stated in Eq. (21). Note the symmetry of $U$ and $V$ in Eq. (15), we also can prove the analogous convergence under the updating rule for $V$ in Eq. (23).

**References**


Fanhua Shang is currently pursuing the Ph.D. degree in circuits and systems from the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China, Xidian University, Xi’an, China. His current research interests include pattern recognition, machine learning, data mining, and computer vision.

L.C. Jiao received the B.S. degree from Shanghai Jiaotong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees from Xi’an Jiaotong University, Xi’an, China, in 1984 and 1990, respectively. During 1990 and 1991, he was a Postdoctoral Fellow with the Key Lab for Radar Signal Processing, Xidian University, Xi’an, China. Currently, he is the Dean of the School of Electronic Engineering and the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. He is the author or coauthor of more than 200 scientific papers. His current research interests include signal and image processing, nonlinear circuit and systems theory, learning theory and algorithms, optimization problems, wavelet theory, and data mining.

Fei Wang received his Ph.D. degree from Tsinghua University, Beijing, China, in 2008. After that, he came to Florida International University as a postdoctorate research associate till August, 2009. Currently, He is a postdoc researcher in Healthcare Transformation Group, IBM T. J. Watson Research Center at Hawthorne, NY, USA.