This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Fine-Grained Bipartite Concept Factorization for Clustering

Chong Peng[1], Pengfei Zhang[1], Yongyong Chen[2,*], Zhao Kang[3], Chenglizhao Chen[4,*], and Qiang Cheng[5]
[1] College of Computer Science and Technology, Qingdao University
[2] School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen)
[3] School of Computer Science and Engineering, University of Electronic Science and Technology of China
[4] College of Computer Science and Technology, China University of Petroleum (East China)
[5] Department of Computer Science, University of Kentucky
{qdcpeng,zpf010108,yongyongchen.cn,sckangz}@gmail.com,cclz123@163.com,qiang.cheng@uky.edu

Abstract

In this paper, we propose a novel concept factorization method that seeks factor matrices using a cross-order positive semi-definite neighbor graph, which provides comprehensive and complementary neighbor information of the data. The factor matrices are learned with bipartite graph partitioning, which exploits explicit cluster structure of the data and is more geared towards clustering application. We develop an effective and efficient optimization algorithm for our method, and provide elegant theoretical results about the convergence. Extensive experimental results confirm the effectiveness of the proposed method.

1. Introduction

Matrix factorization aims at seeking two or more factor matrices, whose product may well approximate the original data [25]. Matrix factorization has been widely studied for dimension reduction [25, 48] and low-dimensional representation [44] of high-dimensional data, which is essentially important in various learning tasks such as clustering [12], classification [24], foreground-background separation in surveillance video [30], community detection [27], link prediction [26], hyperspectral unmixing [8], etc.

Among these learning tasks, nonnegative data are typical due to the fact that they naturally contain only nonnegative elements [15], such as pixels of images [2, 36], connectivity of nodes in a social network [29], etc. For such type of data, the nonnegative matrix factorization (NMF) has been developed for parts-based representation [15, 16], which shows its advantage by mimicking the signal processing mechanism of the human brain [15].

NMF methods have been widely studied during the last decades [19, 39] with various extensions [11, 19, 37]. Among them, the Convex-NMF (CNMF) is quite typical by restricting the basis vectors to be convex combinations of the samples [6, 50]. Let $\mathbf{X} \in \mathbb{R}^{d\times n}$ be the data matrix with d features, n samples, and c clusters, then the CNMF can be represented as $\mathbf{X} \approx \mathbf{XWV}^T$, where $\mathbf{W} \in \mathbb{R}^{n\times c}$ and $\mathbf{V} \in \mathbb{R}^{n\times c}$ denote the so-called score and coefficient matrices [32, 38], respectively. The CNMF is also known as the concept factorization (CF) by revealing the distinct concepts, i.e., the cluster centers of the data, as well as the cluster membership of each sample [50]. The CNMF allows a direct measurement of pairwise similarity of the data [6], which is straightforward to apply the kernel technique for nonlinear relationship learning [6, 32]. Moreover, it has been revealed that the variants of CNMF such as Cluster-NMF, Semi-NMF, and Kernel-NMF, are all soft relaxations of the K-means clustering [6]. With the success of deep learning, deep factorization has been successfully developed for the NMF [28, 51] and CF community [3] from the original one-layer factorization, which helps recognize hierarchical structure of the data [3, 43].

Recently, the CF has been shown closely related with subspace clustering methods in nature [32, 38], where the product of the factor matrices can be treated as the self-expressive representation matrix [32]. These methods integrate the local geometric structure of the data in recovering the score and coefficient factor matrices [32, 38]. It has been shown that local geometric structure is critical and effective in revealing the underlying structure of the data [32], which can be described from the following perspectives: 1) It helps avoid over-fitting issue by recovering local geometric structure of the data [20]; 2) A full kernel matrix does not consider local density of the samples and might be inefficient in the representation capability [17, 21]; 3) The intrinsic information of high-dimensional data needs to be maintained by the low-dimensional representation [45]. For NMF methods, the local geometric structure of the data is often preserved by techniques such as sparse regularization [36] and graph Laplacian [2]. For the former, the lo-

cal relationship of the data is preserved by the sparse representation [46], meanwhile the parts-based representation of NMF naturally leads to sparsity [2]. For the latter, the NMF method recovers the low-dimensional representation on a local manifold that maintains local neighbors [2]. Besides, some other techniques have been developed such as inter-class separability [32] and local centroid learning [4]. All these methods only consider the first-order neighbors, which might be insufficient to fully exploit local geometric structure of the data [13], and more comprehensive information still needs to be exploited.

In fact, high-order neighbors may provide more complicated relationships and reveal latent structure of the data [13,42], which has been rarely considered in the CF or NMF community. In this paper, we fully exploit the high-order neighbor relationships of the data for concept factorization, which provides comprehensive and complementary information of the data. Moreover, we seek factorization with bipartite graph partitioning, which renders the score and coefficient matrices to have clear group structure and thus are more geared towards clustering application.

We summarize the key contributions of this paper as follows: 1) We seek the concept factorization using the fine-grained neighbor graph composed of cross-order neighbors, which provides comprehensive and complementary information of the data; 2) The cluster information is preserved by the bipartite graph partitioning, which renders the coefficient matrix to have a clear group structure; 3) The fine-grained neighbor graph allows the data to have mixed signs, which potentially extends the applicability of NMF methods in broader areas; 4) Efficient optimization algorithm is developed with theoretical convergence guarantee; 5) Extensive experimental results confirm the effectiveness of the proposed method in clustering and data representation.

2. Related Work

We briefly review some closely related methods as follows. Given data set $\mathbf{X} \in \mathbb{R}^{d\times n}$, the CF seeks a factorization with the following minimization problem [50]:

$$\min_{\mathbf{W}\geq 0,\mathbf{V}\geq 0} \|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2, \tag{1}$$

with $\|\cdot\|_F$ being the Frobenius norm, which allows the NMF methods to handle mixed-signed data [6]. It is seen that (1) is closely related with subspace clustering methods [32,33], which assume self-expressiveness of the data and can be expressed as $\mathbf{X} \approx \mathbf{X}\mathbf{Z}$ with $\mathbf{Z} \in \mathbb{R}^{n\times n}$ being the representation matrix [33, 40]. It is straightforward to have nonlinear expansion of (1) using the kernel trick [6, 31].

3. Fine-grained Bipartite CF

The CF methods have been widely studied [1, 38, 50] and have been shown to be closely related with the spectral clustering-based subspace clustering [32]:

$$\min_{\mathbf{W},\mathbf{V}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2+\frac{\alpha}{2}\|\mathbf{W}\mathbf{V}^T\|_F^2, \mathbf{W}\geq 0, \mathbf{V}\geq 0, \tag{2}$$

where $\alpha \geq 0$ is a balancing parameter, the factor $\frac{1}{2}$ is used to facilitate the optimization, and $\mathbf{W}\mathbf{V}^T \in \mathbb{R}^{n\times n}$ can be treated as a low-rank self-expressive representation matrix of subspace clustering methods [33,35]. Although the product of $\mathbf{W}\mathbf{V}^T$ has strict rank constraint, it does not necessarily leads to a proper group structure of $\mathbf{W}\mathbf{V}^T$ nor a strong clustering capability of $\mathbf{V}$. To further enhance the representation capability of $\mathbf{W}\mathbf{V}^T$, as well as $\mathbf{V}$, in revealing underlying group structure of the data, we first define the following symmetric and nonnegative bipartite graph matrix $\mathbf{S} = \begin{bmatrix} \mathbf{0} & (\mathbf{W}\mathbf{V}^\mathbf{T})^T \\ (\mathbf{W}\mathbf{V}^\mathbf{T}) & \mathbf{0} \end{bmatrix}$, and its graph Laplacian matrix $\mathbf{L} = \mathbf{D_S} - \mathbf{S}$, with $\mathbf{0}$ being an $n \times n$ zero matrix, $\mathbf{D_S}$ being diagonal, and $(\mathbf{D_S})_{ii} = \sum_{j=1}^{n} \mathbf{S}_{ij}$. For ease of notation, we define an operator $\mathbb{L}(\cdot)$ such that $\mathbf{L} = \mathbb{L}(\mathbf{W}, \mathbf{V})$. Then, it is desired that $\mathbf{S}$ has exactly c connected components that correspond to c clusters of the data, which is ensured by the following Ky-Fan's Theorem [22]:

Theorem 1 ([22]). *For any $\mathbf{S}^T = \mathbf{S} \geq 0$, let $\mathbf{L}$ be its Laplacian matrix. Then the number of zero eigenvalues of $\mathbf{L}$ equals the number of connected components in $\mathbf{S}$.*

It is straightforward to impose rank constraint of $\mathbf{L}$ in (2) for the desired property. However, the hard rank constraint is usually difficult to solve. Thus, we relax it to minimize the c smallest eigenvalues of $\mathbf{L}$, which leads to

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{V},\mathbf{F}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2+\frac{\alpha}{2}\|\mathbf{W}\mathbf{V}^T\|_F^2+\beta\sum_{i=1}^{c}\boldsymbol{\lambda}_i(\mathbf{L}) \\ &s.t. \quad \mathbf{W}\geq 0, \mathbf{V}\geq 0, \mathbf{L}=\mathbb{L}(\mathbf{W},\mathbf{V}), \end{aligned} \tag{3}$$

where $\beta \geq 0$ is a balancing parameter and $\boldsymbol{\lambda}_i(\cdot)$ returns the i-th smallest eigenvalue of the input matrix. Thus, it is guaranteed that $\sum_{i=1}^{c} \boldsymbol{\lambda}_i(\mathbf{L}) = 0$ with a properly large β. According to [7], we may rewrite (3) as the following:

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{V},\mathbf{F}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2+\frac{\alpha}{2}\|\mathbf{W}\|_F^2+\beta\mathbf{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) \\ &s.t. \mathbf{W}\geq 0, \mathbf{V}\geq 0, \mathbf{F}^T\mathbf{F}=\mathbf{I}_c, \mathbf{L}=\mathbb{L}(\mathbf{W},\mathbf{V}), \end{aligned} \tag{4}$$

where $\mathbf{I}_c$ denotes an identity matrix of size $c \times c$. Moreover, it is desired to make $\mathbf{V}$ a relaxed indicator matrix for immediate clustering interpretation. Therefore, we impose the constraint $\mathbf{V}^T\mathbf{V} = \mathbf{I}_c$ and our model becomes

$$\begin{aligned} &\min_{\mathbf{W},\mathbf{V},\mathbf{F}} \frac{1}{2}\|\mathbf{X}-\mathbf{X}\mathbf{W}\mathbf{V}^T\|_F^2+\frac{\alpha}{2}\|\mathbf{W}\|_F^2+\beta\mathbf{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) \\ &s.t. \mathbf{W}\geq 0, \mathbf{V}\geq 0, \mathbf{F}^T\mathbf{F}=\mathbf{I}_c, \mathbf{V}^T\mathbf{V}=\mathbf{I}_c, \mathbf{L}=\mathbb{L}(\mathbf{W},\mathbf{V}). \end{aligned} \tag{5}$$

To further enhance the nonlinear learning capability, we expand the above model using the kernel trick [31]. In particular, we may expand the first term of (5) and replace $\mathbf{X}^T\mathbf{X}$ with the kernel matrix $\mathbf{K} \in \mathbb{R}^{n\times n}$, which leads to

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{V},\mathbf{F}} \frac{1}{2}\mathbf{Tr}(\mathbf{V}\mathbf{W}^T\mathbf{K}\mathbf{W}\mathbf{V}^T - 2\mathbf{K}\mathbf{W}\mathbf{V}^T)\\ &+ \frac{\alpha}{2}\|\mathbf{W}\|_F^2 + \beta\mathbf{Tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right),\ s.t.\ \mathbf{W}\geq 0, \mathbf{V}\geq 0,\\ &\mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{V}^T\mathbf{V} = \mathbf{I}_c, \mathbf{L} = \mathbb{L}(\mathbf{W},\mathbf{V}),\end{aligned} \tag{6}$$

where $\mathbf{K}$ is obtained by adopting a certain kernel function. It has been revealed that local relationships play an essential role in learning the underlying structures of the data [32,38]. However, such information is omitted in the above model. To combat this issue, we construct a local neighbor graph by element as $\mathbf{K}_{ij} = \mathbb{I}_{\{\mathbf{X}_j\in\mathcal{N}_K(\mathbf{X}_i)\}}$, where $\mathbb{I}_{\{\cdot\}}$ is an indicator function that returns 1 if the conditions in the subscript hold and 0 otherwise, $\mathbf{X}_i$ denotes the i-th sample and $\mathcal{N}_K(\cdot)$ denotes the set of K-nearest neighbors of the input. In this paper, for simplicity and yet without loss of generality, we use the binary kernel similarity to measure the pair-wise neighbor relationships of the data [2], with K being set to 5.

Essentially, the weighted graph is closely related with the random walk and Markov transition probability matrix (TPM) [41,49,52], where the neighbor relationships are directly revealed by the probabilities of the one-step random walks. Inspired by [49], we construct the TPM as $\mathbf{P} = \mathbf{D}_{\mathbf{K}}^{-1}\mathbf{K}$, where $\mathbf{D}_{\mathbf{K}}$ is diagonal and $(\mathbf{D}_{\mathbf{K}})_{ii} = \sum_{j=1}^n \mathbf{K}_{ij}$. We may also treat the probability as the soft neighbor relationship. In practice, there often exist high-order neighbor relationships among the data, which are not directly available in a weighted graph. Intuitively, such relationships can be measured by the probabilities of multi-step random walks [42]. To account for the high-order neighbor information, we define the following a-th order probability matrix of the data as $\mathbf{P}_{[a]} = \mathbf{P}^a = \underbrace{\mathbf{P}\cdot\mathbf{P}\cdot\ \cdots\ \cdot\mathbf{P}}_{a \text{ times}} \in \mathbb{R}^{n\times n}$, which measures the neighbor relationships up to an a-step random walk. To fully exploit such comprehensive cross-order neighbor information, we fuse the high-order probability matrices. Moreover, to ensure the symmetry, which is a natural and desired property for neighbor relationships, we define the following fine-grained probability matrix as $\mathbf{P}_A = \sum_{a=1}^A \frac{\mathbf{P}_{[a]}+\mathbf{P}_{[a]}^T}{2} = \sum_{a=1}^A \frac{\mathbf{P}^a+(\mathbf{P}^a)^T}{2}$, with A being the highest order of the neighbor information. Thus, with $\mathbf{P}_A$, our model can be finally developed as:

$$\begin{aligned}&\min_{\mathbf{W},\mathbf{V},\mathbf{F}} \frac{1}{2}\mathbf{Tr}(\mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T - 2\mathbf{P}_A\mathbf{W}\mathbf{V}^T)\\ &+ \frac{\alpha}{2}\|\mathbf{W}\|_F^2 + \beta\mathbf{Tr}\left(\mathbf{F}^T\mathbf{L}\mathbf{F}\right),\ s.t.\ \mathbf{W}\geq 0, \mathbf{V}\geq 0,\\ &\mathbf{F}^T\mathbf{F} = \mathbf{I}_c, \mathbf{V}^T\mathbf{V} = \mathbf{I}_c, \mathbf{L} = \mathbb{L}(\mathbf{W},\mathbf{V}).\end{aligned} \tag{7}$$

We name model (7) as the **Fi**ne-**g**rain**e**d Bipa**r**tite **C**oncept **F**actorization (Figer-CF). It is noted that $\mathbf{P}_A \geq 0$, which renders the Figer-CF suitable for general data with mixed signs and potentially applicable in broader applications.

Remark 1. In practice, to ensure that $\mathbf{P}_A$ is positive semi-definite, we add an identity matrix with a scaling factor $\tau \geq 0$ to it. This strategy makes sense because it is natural that one sample has the highest similarity with itself. In practice, we may set the scaling factor as $\tau = |\min\{0, \boldsymbol{\lambda}_n(\mathbf{P}_A)\}|$. Moreover, we may provide a general lower bound of τ for any $\mathbf{P}_A$, which only depends on the sample size n and order of neighbors A. In rest of this paper, the adjusted $\mathbf{P}_A + \tau\mathbf{I}_n$ is denoted as $\mathbf{P}_A$ for ease and clarity of representation. The lower bound is provided in the following Theorem 2.

Theorem 2. *It is ensured that* $\mathbf{P}_A + \tau\mathbf{I}_n \succeq 0$ *with* $\tau \geq A\sqrt{n}$.

Proof. According to the definition of $\mathbf{P}_A$, it is clear that $\mathbf{P}_A + \tau\mathbf{I}_n$ is real and symmetric. Thus, $\mathbf{P}_A + \tau\mathbf{I}_n$ is diagonalizable and we only need to show that all its eigenvalues are nonnegative. It is straightforward that $\|\mathbf{P}_A\|_F \leq \sum_{a=1}^A \|\frac{1}{2}(\mathbf{P}_{[a]} + \mathbf{P}_{[a]}^T)\|_F \leq \sum_{a=1}^A \|\mathbf{P}_{[a]}\|_F$. Moreover, $\|\mathbf{P}_{[a]}\|_F^2 = \sum_{a=1}^n \|\mathbf{P}_{[a]}^{(i)}\|_2^2 \leq \sum_{a=1}^n \|\mathbf{P}_{[a]}^{(i)}\|_1^2 \leq \sum_{a=1}^n 1 = n$. Thus, we have $\|\mathbf{P}_A\|_F \leq \sum_{a=1}^A \sqrt{n} = A\sqrt{n}$. According to Schur's inequality, we have $|\boldsymbol{\lambda}_i(\mathbf{P}_A)| \leq \|\mathbf{P}_A\|_F \leq A\sqrt{n}$ for $i = 1, \cdots, n$. Thus, it is straightforward that $\boldsymbol{\lambda}_n(\mathbf{P}_A) \geq -A\sqrt{n}$, and $\boldsymbol{\lambda}_n(\mathbf{P}_A + \tau\mathbf{I}_n) = \boldsymbol{\lambda}_n(\mathbf{P}_A) + \tau \geq 0$, which guarantees that $\mathbf{P}_A + \tau\mathbf{I}_n \succeq 0$.

□

4. Optimization

In this section, we develop an effective iterative optimization algorithm.

4.1. Optimization of F

Keeping the other variables fixed, the optimization problem of $\mathbf{F}$ is as $\min_{\mathbf{F}^T\mathbf{F}=\mathbf{I}_c} \mathbf{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$, which can be effectively solved by the eigenvalue decomposition. As a result, $\mathbf{F}$ is obtained by computing the eigenvectors of $\mathbf{L}$ that are associated with its top c eigenvalues, which is denoted as:

$$\mathbf{F} = \mathbb{EVD}_c(\mathbf{L}). \tag{8}$$

4.2. Optimization of W and V

With fixed $\mathbf{F}$, the subproblem of $\mathbf{W}$ and $\mathbf{V}$ is complicated, since $\mathbf{L}$ is associated with both $\mathbf{W}$ and $\mathbf{V}$. To facilitate the optimization, we reformulate the term $\mathbf{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F})$ to explicitly represent it in $\mathbf{W}$ and $\mathbf{V}$. First, we divide $\mathbf{F}$ into two blocks, i.e., $\mathbf{F}^T = [\mathbf{F}_{(1)}^T, \mathbf{F}_{(2)}^T]$, with $\mathbf{F}_{(1)} \in \mathbb{R}^{n\times c}$ and $\mathbf{F}_{(2)} \in \mathbb{R}^{n\times c}$, respectively. Then, we have $\mathbf{Tr}(\mathbf{F}^T\mathbf{L}\mathbf{F}) = \frac{1}{2}\sum_{i,j=1}^{2n}\|\mathbf{F}^i - \mathbf{F}^j\|_2^2\mathbf{S}_{ij} = \sum_{i,j=1}^n \|\mathbf{F}_{(1)}^i - \mathbf{F}_{(2)}^j\|_2^2(\mathbf{W}\mathbf{V}^T)_{ji} = \sum_{i,j=1}^n \mathbf{T}_{ij}(\mathbf{W}\mathbf{V}^T)_{ji} = \mathbf{Tr}(\mathbf{T}\mathbf{W}\mathbf{V}^T)$,

where $\mathbf{F}^i$, $\mathbf{F}^i_{(1)}$, and $\mathbf{F}^i_{(2)}$ denote the i-th row of $\mathbf{F}$, $\mathbf{F}_{(1)}$, and $\mathbf{F}_{(2)}$, and $\mathbf{T}_{ij} = \|\mathbf{F}^i_{(1)} - \mathbf{F}^j_{(2)}\|_2^2$ is the (i,j)-th element of matrix $\mathbf{T} \in \mathbb{R}^{n\times n}$, respectively. Thus, the joint minimization problem associated with $\mathbf{W}$ and $\mathbf{V}$ becomes

$$\begin{aligned}\min_{\mathbf{W},\mathbf{V}} \frac{1}{2}\mathbf{Tr}(\mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T - 2\mathbf{P}_A\mathbf{W}\mathbf{V}^T) + \frac{\alpha}{2}\|\mathbf{W}\|_F^2 \\ + \beta\mathbf{Tr}\left(\mathbf{T}\mathbf{W}\mathbf{V}^T\right),\ s.t.\ \mathbf{W}\geq 0, \mathbf{V}\geq 0.\end{aligned} \tag{9}$$

Given $\mathbf{P}_A$ and $\mathbf{T}$, we provide the following updating rules of $\mathbf{W}$ and $\mathbf{V}$ for the joint minimization problem (9):

$$\mathbf{W}_{ik} \leftarrow \mathbf{W}_{ik}\sqrt{\frac{(\mathbf{P}_A\mathbf{V})_{ik}}{\beta(\mathbf{T}^T\mathbf{V})_{ik} + (\mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V})_{ik} + \alpha\mathbf{W}_{ik}}}, \tag{10}$$

$$\mathbf{V}_{ik} \leftarrow \mathbf{V}_{ik}\sqrt{\frac{(\mathbf{P}_A\mathbf{W})_{ik} + \beta\left(\mathbf{V}\mathbf{V}^T\mathbf{T}\mathbf{W}\right)_{ik}}{\left(\mathbf{V}\mathbf{V}^T\mathbf{P}_A\mathbf{W}\right)_{ik} + \beta(\mathbf{T}\mathbf{W})_{ik}}}. \tag{11}$$

We denote the objective function as $f(\mathbf{F},\mathbf{W},\mathbf{V})$. We repeat (8), (10) and (11) until convergence of objective value sequence $\{f(\mathbf{F}^{(t)},\mathbf{W}^{(t)},\mathbf{V}^{(t)})\}$ and obtain the solution $\{\mathbf{F}^{(\star)},\mathbf{W}^{(\star)},\mathbf{V}^{(\star)}\}$, where we use the superscript $(\cdot)^{(t)}$ to denote the number of iterations. Then, we perform the K-means to $\mathbf{V}^{(\star)}$ to obtain the final clusters of the data.

5. Convergence Analysis

In this section, we provide the main theoretical results, including correctness of the updating rules and convergence of our algorithm.

5.1. Correctness and Convergence of (10)

For the updating rule of (10), we present the following two main results: 1) When convergent, the limiting solution of (10) satisfies the KKT condition. 2) The iteration of (10) converges. We formally establish the above results in Theorems 3 and 4, respectively.

Theorem 3. *Fixing $\mathbf{V}$ and $\mathbf{T}$, the limiting solution of the updating rule in (10) satisfies the KKT condition.*

Proof. Fixing $\mathbf{V}$, the subproblem for $\mathbf{W}$ is

$$\begin{aligned}\min_{\mathbf{W}} \frac{1}{2}\mathbf{Tr}(-2\mathbf{P}_A\mathbf{W}\mathbf{V}^T + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T) \\ + \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{T}^T\mathbf{V}) + \frac{\alpha}{2}\mathbf{Tr}(\mathbf{W}^T\mathbf{W}),\ s.t.\ \mathbf{W}\geq 0.\end{aligned} \tag{12}$$

We introduce an Lagrangian multiplier $\mathbf{\Psi} \in \mathbb{R}^{n\times c}$ to the above problem and obtain its Lagrangian function:

$$\begin{aligned}L_{\mathbf{W}} = \frac{1}{2}\mathbf{Tr}(-2\mathbf{P}_A\mathbf{W}\mathbf{V}^T + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T) \\ + \frac{\alpha}{2}\mathbf{Tr}(\mathbf{W}^T\mathbf{W}) + \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{T}^T\mathbf{V}) + \mathbf{Tr}(\mathbf{\Psi}\mathbf{W}^T).\end{aligned} \tag{13}$$

First, we may obtain the gradient of $L_{\mathbf{W}}$ as

$$\frac{\partial L_{\mathbf{W}}}{\partial \mathbf{W}} = -\mathbf{P}_A\mathbf{V} + \beta\mathbf{T}^T\mathbf{V} + \mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V} + \alpha\mathbf{W} + \mathbf{\Psi}. \tag{14}$$

By the complementary slackness condition, we can see that

$$(-\mathbf{P}_A\mathbf{V} + \beta\mathbf{T}^T\mathbf{V} + \mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V} + \alpha\mathbf{W})_{ik}\mathbf{W}_{ik} = \mathbf{\Psi}_{ik}\mathbf{W}_{ik} = 0, \tag{15}$$

which provides a fixed point condition that the limiting solution of (10) should satisfy. It is clear that (10) can be reduced to the following equation:

$$(-\mathbf{P}_A\mathbf{V} + \beta\mathbf{T}^T\mathbf{V} + \mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V} + \alpha\mathbf{W})_{ik}\mathbf{W}^2_{ik} = 0, \tag{16}$$

which is identical to (15) because both of them imply the same conditions that either $\mathbf{W}_{ik} = 0$ holds in both (15) and (16) or $(-\mathbf{P}_A\mathbf{V} + \beta\mathbf{T}^T\mathbf{V} + \mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V} + \alpha\mathbf{W})_{ik} = 0$ holds in both (15) and (16), which concludes the proof. □

Before we provide the result about convergence of (10), we introduce the technique of auxiliary function [2, 16] that plays an essential role in the following analysis.

Definition 1. *Given functions $\bar{L}(\mathbf{Q},\mathbf{Q}')$ and $L(\mathbf{Q})$, then $\bar{L}(\mathbf{Q},\mathbf{Q}')$ is called an auxiliary function of $L(\mathbf{Q})$ if*

$$\bar{L}(\mathbf{Q},\mathbf{Q}') \geq L(\mathbf{Q}) \text{ and } \bar{L}(\mathbf{Q},\mathbf{Q}) = L(\mathbf{Q}) \tag{17}$$

hold for any $\mathbf{Q}$ and $\mathbf{Q}'$.

Proposition 1. *Given a function $L(\mathbf{Q})$, let $\bar{L}(\mathbf{Q},\mathbf{Q}')$ be one of its auxiliary functions. Then, the variable sequence $\{\mathbf{Q}^{(t)}\}$, with $t \in \mathbb{N}$ and*

$$\mathbf{Q}^{(t+1)} = \arg\min_{\mathbf{Q}} \bar{L}(\mathbf{Q},\mathbf{Q}^{(t)}), \tag{18}$$

satisfies the following chain of inequalities: $L(\mathbf{Q}^{(t)}) = \bar{L}(\mathbf{Q}^{(t)},\mathbf{Q}^{(t)}) \geq \bar{L}(\mathbf{Q}^{(t+1)},\mathbf{Q}^{(t)}) \geq L(\mathbf{Q}^{(t+1)})$, which implies that the value sequence $\{L(\mathbf{Q}^{(t)})\}_{t=0}^{\infty}$ is monotonically decreasing (nonincreasing).

Proposition 2 ([6]). *For any matrices $\mathbf{\Gamma} \in \mathbb{R}_+^{n\times n}$, $\mathbf{\Omega} \in \mathbb{R}_+^{c\times c}$, $\mathbf{\Delta} \in \mathbb{R}_+^{n\times c}$, and $\mathbf{\Delta}' \in \mathbb{R}_+^{n\times c}$, with $\mathbf{\Gamma}$ and $\mathbf{\Omega}$ being symmetric, the following inequality holds: $\sum_{i=1}^{n}\sum_{s=1}^{c}\frac{(\mathbf{\Gamma}\mathbf{\Delta}'\mathbf{\Omega})_{is}\mathbf{\Delta}^2_{is}}{\mathbf{\Delta}'_{is}} \geq \boldsymbol{Tr}(\mathbf{\Delta}^T\mathbf{\Gamma}\mathbf{\Delta}\mathbf{\Omega})$.*

Next, we provide the main result about the convergence of (10) in the following theorem and proof.

Theorem 4. *For fixed $\mathbf{V}$, the objective in (12) is monotonically decreasing under the updating rule in (10).*

Proof. We define the subproblem of $\mathbf{W}$ as

$$\begin{aligned}P(\mathbf{W}) = \frac{\alpha}{2}\mathbf{Tr}(\mathbf{W}^T\mathbf{W}) + \mathbf{Tr}(-\mathbf{W}^T\mathbf{P}_A\mathbf{V} \\ + \frac{1}{2}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V} + \beta\mathbf{W}^T\mathbf{T}^T\mathbf{V}).\end{aligned} \tag{19}$$

Then, according to Proposition 2, as well as $a \leq \frac{a^2+b^2}{2b}$ and $a \geq 1 + \log a$ for $a, b \geq 0$, it is straightforward that the following inequalities hold for any $\mathbf{W}' \in \mathbb{R}^{n\times c} \geq 0$:

$$\mathbf{Tr}(\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T\mathbf{V}) \leq \sum_{ik} \frac{(\mathbf{P}_A\mathbf{W}'\mathbf{V}^T\mathbf{V})_{ik}\mathbf{W}_{ik}^2}{\mathbf{W}'_{ik}}, \quad (20)$$

$$\mathbf{Tr}(\mathbf{W}^T\mathbf{T}^T\mathbf{V}) \leq \sum_{ik} (\mathbf{T}^T\mathbf{V})_{ik} \frac{\mathbf{W}_{ik}^2 + \mathbf{W}'^2_{ik}}{2\mathbf{W}'_{ik}}, \quad (21)$$

$$\mathbf{Tr}(\mathbf{W}^T\mathbf{P}_A\mathbf{V}) \geq \sum_{ik} (\mathbf{P}_A\mathbf{V})_{ik}\mathbf{W}'_{ik}\Big(1 + \log \frac{\mathbf{W}_{ik}}{\mathbf{W}'_{ik}}\Big), \quad (22)$$

where the equal sign holds when $\mathbf{W} = \mathbf{W}'$. Thus, we may construct an auxiliary function of $P(\mathbf{W})$ as

$$\begin{aligned}\bar{P}(\mathbf{W},\mathbf{W}') = &-\sum_{ik} (\mathbf{P}_A\mathbf{V})_{ik}\mathbf{W}'_{ik}\Big(1 + \log \frac{\mathbf{W}_{ik}}{\mathbf{W}'_{ik}}\Big) \\ &+ \frac{1}{2}\sum_{ik} \frac{(\mathbf{P}_A\mathbf{W}'\mathbf{V}^T\mathbf{V})_{ik}\mathbf{W}_{ik}^2}{\mathbf{W}'_{ik}} \\ &+ \beta\sum_{ik} (\mathbf{T}^T\mathbf{V})_{ik} \frac{\mathbf{W}_{ik}^2 + \mathbf{W}'^2_{ik}}{2\mathbf{W}'_{ik}} + \frac{\alpha}{2}\sum_{ik} \mathbf{W}_{ik}^2,\end{aligned} \quad (23)$$

where $\bar{P}(\mathbf{W}, \mathbf{W}') \geq P(\mathbf{W})$ and $\bar{P}(\mathbf{W}, \mathbf{W}) = P(\mathbf{W})$ hold for any $\mathbf{W}, \mathbf{W}' \geq 0$.

Next, the key is to show that (10) essentially follows (18). It is easy to verify that $\bar{P}(\mathbf{W}, \mathbf{W}')$ is convex in $\mathbf{W}$. Thus, the global optimum is ensured by the first-order optimality condition, which gives rise to

$$\begin{aligned}\frac{\partial \bar{P}(\mathbf{W}, \mathbf{W}')}{\partial \mathbf{W}_{ik}} = &-\frac{(\mathbf{P}_A\mathbf{V})_{ik}\mathbf{W}'_{ik}}{\mathbf{W}_{ik}} + \alpha\mathbf{W}_{ik} \\ &+ \frac{(\mathbf{P}_A\mathbf{W}'\mathbf{V}^T\mathbf{V})_{ik}\mathbf{W}_{ik}}{\mathbf{W}'_{ik}} + \beta\frac{(\mathbf{T}^T\mathbf{V})_{ik}\mathbf{W}_{ik}}{\mathbf{W}'_{ik}} = 0.\end{aligned} \quad (24)$$

With straight algebra, (24) can be reduced to

$$\mathbf{W}_{ik} = \mathbf{W}'_{ik}\sqrt{\frac{(\mathbf{P}_A\mathbf{V})_{ik}}{(\mathbf{P}_A\mathbf{W}'\mathbf{V}^T\mathbf{V})_{ik} + \beta(\mathbf{T}^T\mathbf{V})_{ik} + \alpha\mathbf{W}'_{ik}}}. \quad (25)$$

Let $\mathbf{W}' = \mathbf{W}^{(t)}$ and $\mathbf{W} = \mathbf{W}^{(t+1)}$, then (25) falls back to (10). Thus, (10) essentially follows (18). According to Definition 1 and Proposition 1, we may conclude the proof. □

5.2. Correctness and Convergence of (11)

For the updating rule of (11), we present the following two main results: 1) When convergent, the limiting solution of (11) satisfies the KKT condition. 2) The iteration of (11) converges. We formally establish the above results in Theorems 5 and 6, respectively. Before presenting these results, we first introduce the following Lagrangian function.

Fixing $\mathbf{W}$ and $\mathbf{T}$, we need to solve (26) for $\mathbf{V}$:

$$\begin{aligned}\min_{\mathbf{V}} &\frac{1}{2}\mathbf{Tr}(-2\mathbf{P}_A\mathbf{W}\mathbf{V}^T + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T) \\ &+ \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{T}^T\mathbf{V}),\ s.t.\ \mathbf{V} \geq 0, \mathbf{V}^T\mathbf{V} = \mathbf{I}_c.\end{aligned} \quad (26)$$

We introduce the symmetric Lagrangian multipliers $\mathbf{\Theta} \in \mathbb{R}^{c\times c}$ and obtain the Lagrangian function to be minimized:

$$\begin{aligned}L_{\mathbf{V}} &= \frac{1}{2}\mathbf{Tr}(-2\mathbf{P}_A\mathbf{W}\mathbf{V}^T + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}\mathbf{V}^T) \\ &+ \beta\mathbf{Tr}(\mathbf{W}^T\mathbf{T}^T\mathbf{V}) + \frac{1}{2}\mathbf{Tr}(\mathbf{\Theta}(\mathbf{V}^T\mathbf{V} - \mathbf{I}_c)) \\ &= \mathbf{Tr}(-\mathbf{P}_A\mathbf{W}\mathbf{V}^T + \frac{1}{2}\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^+\mathbf{V}^T \\ &- \frac{1}{2}\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^-\mathbf{V}^T + \beta\mathbf{T}\mathbf{W}\mathbf{V}^T) - \frac{1}{2}\mathbf{Tr}(\mathbf{\Theta}),\end{aligned} \quad (27)$$

where for an arbitrary input matrix $\mathbf{M}$, $(\cdot)^+$ and $(\cdot)^-$ return matrices such that $\mathbf{M}^+, \mathbf{M}^- \geq 0$ and $\mathbf{M}^+ - \mathbf{M}^- = \mathbf{M}$. The correctness and convergence of (11) are ensured by the following Theorems 5 and 6.

Theorem 5. *Fixing $\mathbf{V}$ and $\mathbf{T}$, the limiting solution of (28) satisfies the KKT complementarity condition of (27):*

$$\mathbf{V}_{ik} \leftarrow \mathbf{V}_{ik}\sqrt{\frac{(\mathbf{P}_A\mathbf{W})_{ik} + (\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^-)_{ik}}{\beta(\mathbf{T}\mathbf{W})_{ik} + (\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^+)_{ik}}}. \quad (28)$$

Proof. The gradient of $L_{\mathbf{V}}$ is

$$\frac{\partial L_{\mathbf{V}}}{\partial \mathbf{V}} = -\mathbf{P}_A\mathbf{W} + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \beta\mathbf{T}\mathbf{W} + \mathbf{V}\mathbf{\Theta}. \quad (29)$$

Then the KKT complementarity condition gives

$$(-\mathbf{P}_A\mathbf{W} + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \beta\mathbf{T}\mathbf{W} + \mathbf{V}\mathbf{\Theta})_{ik}\mathbf{V}_{ik} = 0, \quad (30)$$

which provides a fixed point condition that the limiting solution should satisfy. From the updating rule of (28), it is seen that $\mathbf{V}$ satisfies the following condition:

$$\begin{aligned}\big(&-\mathbf{P}_A\mathbf{W} - \mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^- \\ &+ \mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^+ + \beta\mathbf{T}\mathbf{W}\big)_{ik} \cdot \mathbf{V}_{ik}^2 = 0.\end{aligned} \quad (31)$$

Moreover, since

$$\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta} = (\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^+ - (\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \mathbf{\Theta})^-,$$

the condition in (31) can be further reduced to

$$(-\mathbf{P}_A\mathbf{W} + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \beta\mathbf{T}\mathbf{W} + \mathbf{V}\mathbf{\Theta})_{ik}\mathbf{V}_{ik}^2 = 0. \quad (32)$$

It is clear that (30) and (32) are identical because they imply the same conditions that either $\mathbf{V}_{ik} = 0$ holds in both (30) and (32) or $(-\mathbf{P}_A\mathbf{W} + \mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W} + \beta\mathbf{T}\mathbf{W} + \mathbf{V}\mathbf{\Theta})_{ik} = 0$ holds in both (30) and (32), which concludes the proof. □

Theorem 6. *Fixing* $\mathbf{W}$ *and* $\mathbf{T}$*, the limiting solution of (11) satisfies the KKT complementary condition, and the Lagrangian function* $L_{\mathbf{V}}$ *monotonically decreases under (11).*

Proof. For ease of notation, we define $J(\mathbf{V}) = L_{\mathbf{V}}$. Similarly to the proof of Theorem 4, we obtain the following inequalities, which provide some proper upper bounds to the first four terms of $J(\mathbf{V})$ in (27):

$$\mathbf{Tr}(\mathbf{V}^T\mathbf{T}\mathbf{W}) \leq \sum_{ik}(\mathbf{T}\mathbf{W})_{ik}\frac{\mathbf{V}_{ik}^2+\mathbf{V}'^2_{ik}}{2\mathbf{V}'_{ik}}, \tag{33}$$

$$\mathbf{Tr}(\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+\mathbf{V}^T) \leq \sum_{ik}\frac{(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+)_{ik}\mathbf{V}_{ik}^2}{\mathbf{V}'_{ik}} \tag{34}$$

$$\mathbf{Tr}(\mathbf{V}^T\mathbf{P}_A\mathbf{W}) \geq \sum_{ik}(\mathbf{P}_A\mathbf{W})_{ik}\mathbf{V}'_{ik}\Big(1+\log\frac{\mathbf{V}_{ik}}{\mathbf{V}'_{ik}}\Big), \tag{35}$$

$$\mathbf{Tr}(\mathbf{V}(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^-\mathbf{V}^T) \geq \sum_{ikl}(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^-_{kl}\mathbf{V}'_{ik}\mathbf{V}'_{il}\Big(1+\log\frac{\mathbf{V}_{ik}\mathbf{V}_{il}}{\mathbf{V}'_{ik}\mathbf{V}'_{il}}\Big), \tag{36}$$

where the equal sign holds when $\mathbf{V} = \mathbf{V}' \geq 0$. Combining the above inequalities with proper weights, we construct the following auxiliary function for $J(\mathbf{V})$:

$$\begin{aligned}\bar{J}(\mathbf{V},\mathbf{V}') =& -\frac{1}{2}\mathbf{Tr}(\mathbf{\Theta}) - \sum_{ik}(\mathbf{P}_A\mathbf{W})_{ik}\mathbf{V}'_{ik}\Big(1+\log\frac{\mathbf{V}_{ik}}{\mathbf{V}'_{ik}}\Big)\\ &+\beta\sum_{ik}(\mathbf{T}\mathbf{W})_{ik}\frac{\mathbf{V}_{ik}^2+\mathbf{V}'^2_{ik}}{2\mathbf{V}'_{ik}}\\ &+\frac{1}{2}\sum_{ik}\frac{(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+)_{ik}\mathbf{V}_{ik}^2}{\mathbf{V}'_{ik}}\\ &-\frac{1}{2}\sum_{ikl}(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^-_{kl}\mathbf{V}'_{ik}\mathbf{V}'_{il}\Big(1+\log\frac{\mathbf{V}_{ik}\mathbf{V}_{il}}{\mathbf{V}'_{ik}\mathbf{V}'_{il}}\Big),\end{aligned} \tag{37}$$

where $\bar{J}(\mathbf{V},\mathbf{V}') \geq J(\mathbf{V})$ and $\bar{J}(\mathbf{V},\mathbf{V}) = J(\mathbf{V})$ hold for any $\mathbf{V},\mathbf{V}' \geq 0$. It is straightforward to check that $\bar{J}(\mathbf{V},\mathbf{V}')$ is convex in $\mathbf{V}$. According to the first-order optimality condition of $\bar{J}(\mathbf{V},\mathbf{V}')$:

$$\begin{aligned}\frac{\partial\bar{J}(\mathbf{V},\mathbf{V}')}{\partial\mathbf{V}_{ik}} =& -\frac{(\mathbf{P}_A\mathbf{W})_{ik}\mathbf{V}'_{ik}}{\mathbf{V}_{ik}} + \beta\frac{(\mathbf{T}\mathbf{W})_{ik}}{\mathbf{V}'_{ik}}\mathbf{V}_{ik}\\ &+\frac{(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+)_{ik}}{\mathbf{V}'_{ik}}\mathbf{V}_{ik}\\ &-\frac{(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^-)_{ik}\mathbf{V}'_{ik}}{\mathbf{V}_{ik}} = 0,\end{aligned} \tag{38}$$

we obtain the global minimum of $\bar{J}(\mathbf{V},\mathbf{V}')$:

$$\mathbf{V}_{ik} = \mathbf{V}'_{ik}\sqrt{\frac{(\mathbf{P}_A\mathbf{W})_{ik}+(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^-)_{ik}}{\beta(\mathbf{T}\mathbf{W})_{ik}+(\mathbf{V}'(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+)_{ik}}}. \tag{39}$$

Let $\mathbf{V}^{(t)} = \mathbf{V}'$ and $\mathbf{V}^{(t+1)} = \mathbf{V}$, we may recover (28) from (39). By the first-order optimality condition of $L_{\mathbf{V}}$, we have

$$\begin{aligned}&\mathbf{V}^T(-\mathbf{P}_A\mathbf{W}+\mathbf{V}\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\beta\mathbf{T}\mathbf{W}+\mathbf{V}\mathbf{\Theta})\\ =& -\mathbf{V}^T\mathbf{P}_A\mathbf{W}+\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\beta\mathbf{V}^T\mathbf{T}\mathbf{W}+\mathbf{\Theta} = 0,\end{aligned} \tag{40}$$

which implies that $\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta} = \mathbf{V}^T\mathbf{P}_A\mathbf{W} - \beta\mathbf{V}^T\mathbf{T}\mathbf{W}$. Thus, it is proper to define $(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^+ = \mathbf{V}^T\mathbf{P}_A\mathbf{W}$, $(\mathbf{W}^T\mathbf{P}_A\mathbf{W}+\mathbf{\Theta})^- = \beta\mathbf{V}^T\mathbf{T}\mathbf{W}$. Then, it is straightforward to obtain (11) by substituting the above definitions into (28), which concludes the proof. □

5.3. Convergence of Overall Algorithm

The convergence of the overall algorithm is guaranteed by the following Theorem 7.

Theorem 7. *The objective value sequence of* (7) *is monotonically decreasing under the updating rules of* (8), (10) *and* (11)*, and thus converges.*

Proof. Define the objective function as $f(\mathbf{F},\mathbf{W},\mathbf{V})$ and denote the iteration by superscript $(\cdot)^{(t)}$, respectively, then it is clear that the following chain of inequalities hold:

$$\begin{aligned}&f(\mathbf{F}^{(t)},\mathbf{W}^{(t)},\mathbf{V}^{(t)})\\ \geq& f(\mathbf{F}^{(t+1)},\mathbf{W}^{(t)},\mathbf{V}^{(t)}) \geq f(\mathbf{F}^{(t+1)},\mathbf{W}^{(t+1)},\mathbf{V}^{(t+1)}).\end{aligned} \tag{41}$$

Thus, the sequence $\{f(\mathbf{F}^{(t)},\mathbf{W}^{(t)},\mathbf{V}^{(t)})\}_{t=0}^{\infty}$ is monotonically decreasing. Moreover, since $\mathbf{P}_A$ is symmetric and p.s.d., there exists $\mathbf{C}\in\mathbb{R}^{n\times n}$ such that $\mathbf{P}_A = \mathbf{C}^T\mathbf{C}$. Then, it is clear that

$$\begin{aligned}&f(\mathbf{F}^{(t)},\mathbf{W}^{(t)},\mathbf{V}^{(t)})\\ =&\frac{1}{2}\|\mathbf{C}-\mathbf{C}\mathbf{W}^{(t)}(\mathbf{V}^{(t)})^T\|_F^2+\frac{\alpha}{2}\|\mathbf{W}^{(t)}\|_F^2\\ &+\beta\mathbf{Tr}(\mathbf{F}^{(\mathbf{t})^T}\mathbf{L}^{(t)}\mathbf{F}^{(\mathbf{t})}) - \mathbf{Tr}(\mathbf{C}^T\mathbf{C}) \geq -\mathbf{Tr}(\mathbf{C}^T\mathbf{C}).\end{aligned} \tag{42}$$

Therefore, the sequence $\{f(\mathbf{F}^{(t)},\mathbf{W}^{(t)},\mathbf{V}^{(t)})\}_{t=0}^{\infty}$ is decreasing and lower-bounded, and thus converges. □

6. Experiments

In this section, we conduct extensive experiments to verify the effectiveness of the Figer-CF. We compare our method with nine state-of-the-art clustering methods, including the WNMF [14], ONMF [5], CNMF [6], KNMF [6], RMNMF [10], OPMC [18], MKKM-SR [23], EWNMF [47], and GLS-MKNMF [32], on four widely used benckmark data sets, including the Jaffe, PIX, Semeion, and COIL20. Three metrics, including the clustering accuracy (ACC), normalized mutual information (NMI), and

Table 1. Clustering results (mean±standard deviation) of different methods on the Jaffe, PIX, Semeion and COIL20 data sets

	Methods	2 Classes			4 Classes			6 Classes			8 Classes			10 Classes (Overall)		
		ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
Jaffe	WNMF	**0.998±0.008**	**0.986±0.046**	**0.998±0.008**	0.959±0.060	0.915±0.109	0.959±0.060	0.895±0.044	0.856±0.060	0.895±0.044	0.921±0.056	0.912±0.049	0.922±0.052	**0.906**	0.894	0.906
	RMNMF	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**0.984±0.021**	**0.959±0.053**	**0.984±0.021**	**0.949±0.033**	**0.909±0.052**	**0.949±0.033**	0.898±0.073	0.882±0.055	0.907±0.058	0.822	0.831	0.836
	CNMF	**0.998±0.000**	**0.986±0.046**	**0.998±0.008**	0.804±0.174	0.755±0.174	0.833±0.136	0.763±0.135	0.732±0.138	0.783±0.118	0.694±0.101	0.661±0.114	0.704±0.100	0.700	0.707	0.742
	KNMF	**0.998±0.008**	**0.986±0.046**	**0.998±0.008**	0.804±0.174	0.755±0.174	0.833±0.136	0.835±0.081	0.832±0.070	0.848±0.068	0.792±0.073	0.823±0.043	0.815±0.057	0.817	0.824	0.827
	ONMF	**0.946±0.146**	**0.868±0.320**	**0.946±0.146**	0.886±0.160	0.894±0.113	0.916±0.111	0.884±0.126	0.884±0.106	0.899±0.101	0.696±0.060	0.782±0.051	0.746±0.051	0.803	0.838	0.812
	MKKM-SR	0.913±0.138	0.742±0.331	0.913±0.138	0.907±0.154	0.900±0.138	0.924±0.117	0.881±0.112	0.877±0.097	0.890±0.099	**0.940±0.077**	**0.937±0.053**	**0.945±0.064**	0.850	0.866	0.850
	OPMC	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.968±0.078	0.955±0.089	0.968±0.078	0.896±0.088	0.875±0.071	0.906±0.074	0.865±0.066	0.897±0.035	0.884±0.051	0.808	0.854	**0.913**
	EWNMF	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	0.948±0.056	0.900±0.104	0.948±0.056	0.927±0.045	0.890±0.069	0.927±0.045	0.929±0.033	0.908±0.035	0.929±0.033	**0.953**	**0.934**	**0.953**
	GLS-MKNMF	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**0.991±0.020**	**0.978±0.047**	**0.991±0.020**	**0.978±0.011**	**0.957±0.021**	**0.978±0.011**	**0.953±0.047**	**0.938±0.039**	**0.955±0.029**	0.859	**0.896**	0.878
	Figer-CF	**1.000±0.000**	**1.000±0.000**	**1.000±0.000**	**0.994±0.013**	**0.986±0.030**	**0.994±0.013**	**0.991±0.007**	**0.982±0.012**	**0.992±0.007**	**0.984±0.009**	**0.976±0.014**	**0.984±0.009**	**0.977**	**0.970**	**0.977**
PIX	Methods	2 Classes			4 Classes			6 Classes			8 Classes			10 Classes (Overall)		
		ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
	WNMF	0.940±0.102	0.814±0.283	0.940±0.102	0.928±0.078	0.894±0.091	0.928±0.078	0.850±0.116	0.860±0.083	0.873±0.089	0.809±0.040	0.852±0.021	0.839±0.027	0.740	0.839	0.790
	RMNMF	**0.995±0.014**	**0.976±0.076**	**0.995±0.016**	**0.965±0.044**	**0.935±0.071**	**0.965±0.044**	**0.877±0.090**	0.870±0.089	**0.885±0.086**	0.840±0.078	0.853±0.051	**0.865±0.051**	0.750	0.769	0.770
	CNMF	**0.965±0.063**	**0.875±0.215**	**0.965±0.063**	0.880±0.127	0.848±0.112	0.890±0.109	0.775±0.092	0.754±0.085	0.790±0.085	0.785±0.046	0.810±0.034	0.811±0.033	0.800	0.830	0.830
	KNMF	0.945±0.104	0.838±0.288	0.945±0.104	0.898±0.134	0.882±0.117	0.915±0.099	0.863±0.098	0.878±0.060	**0.885±0.070**	0.833±0.086	**0.878±0.047**	0.863±0.060	0.690	0.809	0.740
	ONMF	0.890±0.122	0.676±0.328	0.890±0.122	0.853±0.126	0.831±0.103	0.870±0.099	0.777±0.098	0.813±0.064	0.815±0.069	0.809±0.055	0.833±0.047	0.818±0.053	0.690	0.801	0.760
	MKKM-SR	**0.955±0.083**	0.855±0.254	0.955±0.083	0.835±0.124	0.787±0.126	0.848±0.103	0.790±0.098	0.816±0.074	0.815±0.080	**0.844±0.082**	0.864±0.050	0.854±0.064	0.790	0.855	0.820
	OPMC	0.595±0.142	**0.107±0.318**	0.595±0.142	0.728±0.092	0.621±0.134	0.735±0.089	0.703±0.102	0.671±0.089	0.717±0.089	0.710±0.057	0.712±0.060	0.728±0.056	0.620	0.668	0.640
	EWNMF	0.940±0.102	0.814±0.283	0.940±0.102	**0.948±0.049**	0.907±0.075	**0.948±0.049**	0.853±0.115	**0.879±0.069**	**0.885±0.081**	0.828±0.071	0.855±0.042	0.858±0.048	**0.810**	**0.892**	**0.860**
	GLS-MKNMF	0.945±0.104	0.838±0.288	0.945±0.104	**0.965±0.034**	**0.934±0.071**	**0.965±0.034**	**0.942±0.036**	**0.907±0.050**	**0.942±0.036**	**0.924±0.050**	**0.906±0.028**	**0.929±0.036**	**0.920**	**0.914**	**0.920**
	Figer-CF	**0.955±0.083**	**0.860±0.243**	**0.955±0.083**	**0.978±0.032**	**0.957±0.053**	**0.978±0.032**	**0.917±0.070**	**0.912±0.046**	**0.923±0.054**	**0.894±0.043**	**0.910±0.012**	**0.904±0.023**	**0.940**	**0.950**	**0.940**
Semeion	Methods	2 Classes			4 Classes			6 Classes			8 Classes			10 Classes (Overall)		
		ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
	WNMF	0.876±0.105	0.562±0.283	0.876±0.105	0.720±0.061	0.507±0.049	**0.723±0.058**	0.583±0.057	0.456±0.048	0.609±0.045	0.596±0.075	0.482±0.049	0.606±0.066	0.556	0.448	0.566
	RMNMF	0.878±0.105	0.559±0.284	0.878±0.105	0.668±0.087	0.438±0.083	0.691±0.073	0.537±0.059	0.410±0.062	0.562±0.056	0.501±0.060	0.395±0.045	0.518±0.054	0.413	0.366	0.453
	CNMF	0.882±0.100	0.562±0.287	0.882±0.100	0.703±0.089	0.499±0.062	0.719±0.065	0.575±0.061	0.453±0.061	0.610±0.053	0.535±0.048	0.444±0.035	0.557±0.039	0.452	0.380	0.452
	KNMF	0.879±0.107	0.574±0.283	0.879±0.107	0.679±0.109	0.505±0.073	0.711±0.076	0.617±0.093	0.511±0.050	0.656±0.061	0.603±0.056	0.525±0.028	**0.639±0.037**	0.525	0.474	0.525
	ONMF	0.871±0.116	0.555±0.305	0.871±0.116	0.685±0.084	0.468±0.067	0.703±0.053	0.606±0.074	0.470±0.062	0.630±0.063	0.556±0.068	0.456±0.039	0.578±0.056	0.508	0.428	0.529
	MKKM-SR	0.797±0.175	0.430±0.358	0.797±0.175	0.686±0.062	0.474±0.059	0.695±0.053	0.607±0.068	0.476±0.052	0.629±0.052	0.600±0.041	0.505±0.023	0.612±0.040	0.418	0.383	0.449
	OPMC	0.852±0.163	0.554±0.358	0.852±0.163	0.681±0.132	**0.519±0.116**	0.701±0.107	**0.627±0.061**	**0.524±0.064**	**0.659±0.061**	**0.646±0.051**	**0.560±0.035**	**0.671±0.038**	**0.654**	**0.546**	**0.655**
	EWNMF	**0.892±0.088**	**0.590±0.260**	**0.892±0.088**	**0.718±0.064**	0.510±0.056	0.722±0.061	0.605±0.070	0.462±0.042	0.619±0.030	0.605±0.058	0.492±0.033	0.617±0.045	0.510	0.430	0.538
	GLS-MKNMF	**0.889±0.105**	**0.607±0.303**	**0.889±0.105**	**0.753±0.112**	**0.583±0.090**	**0.765±0.088**	**0.650±0.086**	**0.553±0.078**	**0.661±0.075**	**0.663±0.074**	**0.582±0.055**	0.619±0.057	**0.615**	**0.553**	**0.568**
	Figer-CF	**0.894±0.105**	**0.629±0.301**	**0.894±0.106**	**0.776±0.118**	**0.642±0.096**	**0.788±0.092**	**0.673±0.088**	**0.609±0.081**	**0.702±0.074**	**0.682±0.054**	**0.649±0.036**	**0.716±0.035**	**0.635**	**0.626**	**0.682**
COIL20	Methods	4 Classes			8 Classes			12 Classes			16 Classes			20 Classes (Overall)		
		ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR	ACC	NMI	PUR
	WNMF	0.724±0.114	0.621±0.151	0.727±0.111	0.668±0.101	0.694±0.076	0.699±0.087	0.671±0.085	0.718±0.045	0.690±0.076	0.638±0.037	0.720±0.031	0.655±0.035	**0.669**	0.740	**0.680**
	RMNMF	0.736±0.120	0.610±0.153	0.742±0.113	0.638±0.054	0.657±0.058	0.676±0.056	0.593±0.053	0.649±0.044	**0.619±0.051**	0.552±0.040	0.657±0.034	0.583±0.043	0.564	0.671	0.592
	CNMF	0.732±0.077	0.604±0.112	0.732±0.077	0.667±0.100	0.686±0.090	0.691±0.086	0.582±0.035	0.634±0.033	0.605±0.039	0.578±0.039	0.657±0.034	0.601±0.038	0.526	0.653	0.565
	KNMF	0.715±0.110	0.637±0.108	0.745±0.081	0.684±0.111	0.709±0.085	0.712±0.095	0.625±0.062	0.704±0.044	**0.668±0.057**	0.641±0.034	**0.730±0.030**	0.671±0.032	0.616	0.710	0.653
	ONMF	0.639±0.123	0.534±0.156	0.669±0.108	0.673±0.099	0.687±0.106	0.700±0.096	0.612±0.057	0.687±0.046	**0.652±0.053**	0.584±0.042	0.708±0.031	0.620±0.041	0.570	0.713	0.594
	MKKM-SR	0.606±0.126	0.723±0.051	0.647±0.089	0.690±0.083	0.646±0.103	0.708±0.078	0.654±0.042	0.655±0.078	0.669±0.041	0.626±0.039	0.582±0.026	0.650±0.038	0.631	**0.744**	0.659
	OPMC	0.723±0.051	**0.649±0.050**	0.741±0.040	0.646±0.103	**0.715±0.087**	0.691±0.089	0.655±0.078	**0.729±0.055**	0.686±0.065	0.582±0.026	0.718±0.030	0.626±0.022	0.522	0.682	0.563
	EWNMF	**0.746±0.077**	0.639±0.108	**0.747±0.077**	**0.706±0.108**	0.704±0.090	**0.727±0.097**	**0.663±0.058**	**0.715±0.044**	0.684±0.059	**0.662±0.026**	**0.732±0.033**	**0.683±0.032**	**0.690**	**0.762**	**0.690**
	GLS-MKNMF	**0.741±0.126**	**0.654±0.154**	**0.756±0.110**	**0.762±0.110**	**0.757±0.093**	**0.772±0.099**	**0.677±0.094**	**0.729±0.071**	0.703±0.076	**0.645±0.043**	0.725±0.035	**0.668±0.041**	0.569	0.700	0.613
	Figer-CF	**0.751±0.123**	**0.677±0.148**	**0.764±0.108**	**0.745±0.085**	**0.765±0.079**	**0.767±0.074**	**0.708±0.046**	**0.769±0.042**	**0.725±0.050**	**0.663±0.051**	**0.763±0.026**	**0.689±0.048**	**0.652**	**0.760**	**0.665**

The top three performances are boldfaced and highlighted in **red**, **blue**, and **orange**, respectively.

purity (PUR) are used in our experiments for evaluation, whose details can be found in [36]. The details of the experiments are presented in rest of this section. For purpose of reproducibility, we release the codes of our paper at https://www.researchgate.net/profile/Chong-Peng-8.

6.1. Experimental settings

As a common step of all methods, we apply the K-means method to the coefficient matrix to obtain the final clustering results. Since the K-means is nonconvex, we follow [32] and conduct 200 independent trials, among which the one with the minimum objective value is used as the final result. For each data set, we follow [2, 32] and use its subsets that consist of different classes for a more comprehensive evaluation, where the size information can be found in Table 1. For a certain size, we randomly select the classes to construct 10 subsets, respectively, on which the averaged performance is recorded and reported in Table 1. Since the OPMC is a multi-view method, we use the original features and the extracted LBP features to construct multi-view features. For the multi-kernel methods, i.e., the MKKM-SR and GLS-MKNMF, we follow their original parameter settings [23,32]. For all the other methods, we tune the balancing parameters within $\mathbb{S} = \{10^{-3}, 10^{-2}, \cdots, 10^{3}\}$. For all the single kernel-based methods, the RBF kernel is used and the radial parameter is searched within $\mathbb{S}$. For the Figer-CF, we set $A = 3$ throughout our experiments unless otherwise clarified. The baseline methods cover various types, such as local learning, single- or multi-kernel learning, multi-view learning, etc, which provide a comprehensive comparison.

6.2. Comparison of Clustering Performance

We report the clustering performance of all methods in Table 1. In general, the proposed method has the best performance on all data sets. In particular, the Figer-CF obtains the top first, second, and third results in 47, 8, and 5 out of a total number of 60 cases, respectively, which is the only method that is among the top three in all cases. We discuss the detailed experimental results as follows. On the Jaffe data, almost all methods obtain rather promising performance on the smallest subsets. However, the baseline methods have significantly degraded performance on larger subsets, whereas the Figer-CF is still promising and obtains the best results in all cases. Similar observations can be found on the PIX data, where the Figer-CF and GLS-MKNMF are the only two that have satisfying performance on all subsets. On the Semeion data, the Figer-CF has the

best performance in almost all cases, with improvements, in general, by about 2-3% in ACC and PUR and 6-7% in NMI, respectively, compared with the top second method. On the COIL20 data, methods such as the EWNMF and GLS-MKNMF have relatively better performance among the baselines. However, both of them are inferior to the Figer-CF and it is difficult to tell which of them is better. In general, the Semeion and COIL20 data appear to be more difficult, since they contain pairs of clusters such as {"0", "O"} and {"2", "Z"}, which are quite similar and confusing. In this case, it may help to reveal the underlying relationships of the data by seeking the local geometric structures of the data since the global relationships may provide misleading information. All these observations confirm the effectiveness and superiority of the Figer-CF.

6.3. Convergence Study

To empirically verify the convergent property of our method, we show curves of the objective value sequences on the Semeion and COIL20 data sets in Fig. 1, where the parameters are set according to Section 6.1. It is seen that the objective value sequences of both data generally converge within about 50 iterations, which confirms the effectiveness and efficiency of the optimization algorithm.

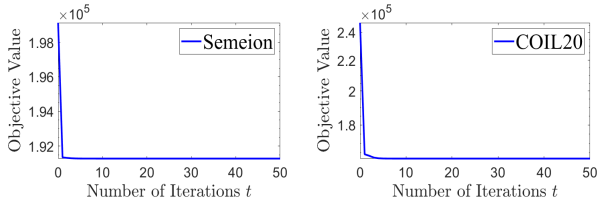


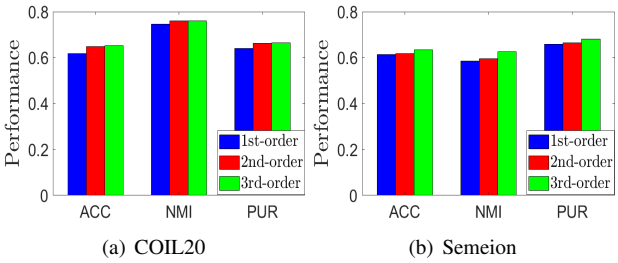
Figure 1. Convergence curve of the objective value sequence.

6.4. Effects of High-order Neighbors

In this test, we empirically show the effects of high-order neighbors using the COIL20 and Semeion data. In particular, we show the performance of the Figer-CF using $\mathbf{P}_A$ of orders from 1 to 3, respectively, and report the performance in Fig. 2, where the other parameters are tuned to the best accordingly. It is observed that the Figer-CF has significantly improved performance in all metrics when a higher-order $\mathbf{P}_A$ is used, which confirms the effectiveness of exploiting comprehensive and complementary neighbor information. If the order tends to be even higher, the improvement tends to be tight. Considering the difficulty in searching a proper A for the Figer-CF in practical sense, it is effective and fairly convincing to suggest $A = 3$ in real-world applications for both effectiveness and efficiency.

6.5. Parameter Sensitivity

We test how the balancing parameters affect the performance of the Figer-CF. Without loss of generality and due



(a) COIL20 (b) Semeion

Figure 2. Effects of the high-order neighbors on the clustering capability of Figer-CF.

to space limit, we report the results on the Jaffe data. We report the performance of the Figer-CF with respect to all combinations of $\{\alpha, \beta\} \in \mathbb{S} \times \mathbb{S}$ in Fig. 3. It is seen that the Figer-CF generally has high performance in all metrics with a broad range of parameter values. Similar observations can be found on other data sets as well, which confirms that the Figer-CF is rather insensitive to parameters. This property is desired by unsupervised learning and is essentially important for potential applicability in real applications.

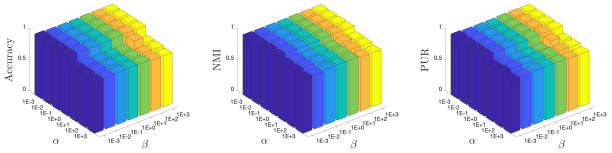


Figure 3. Performance changes with respect to different combinations of parameters on the Jaffe data.

7. Conclusion

In this paper, we propose a novel concept factorization method that uses the cross-order neighbor information of the data to learn score and coefficient matrices with bipartite graph partitioning, which fully exploits comprehensive and complementary neighbor information, as well as explicit cluster structure of the data. Our algorithm admits both elegant theoretical guarantee for convergent property and promising experimental performance in clustering, which confirms the effectiveness and efficiency of the new method.

Acknowledgement

Yongyong Chen and Chenglizhao Chen are corresponding authors. This work is supported in part by the National Natural Science Foundation of China (NSFC) under Grants 62276147, 62172246, and 62106063; in part by the Shandong Province Colleges and Universities Youth Innovation Technology Plan Innovation Team Project under Grants 2022KJ149, 2021KJ062, and 2020KJN011; and in part by the Guangdong Major Project of Basic and Applied Basic Research under Grant 2023B0303000010.

References

[1] Deng Cai, Xiaofei He, and Jiawei Han. Locally consistent concept factorization for document clustering. *IEEE Transactions on Knowledge & Data Engineering*, 23(6):902–913, 2011. 2

[2] Deng Cai, Xiaofei He, Jiawei Han, and Thomas S Huang. Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1548–1560, 2011. 1, 2, 3, 4, 7

[3] Shuai Chang, Jie Hu, Tianrui Li, Hao Wang, and Bo Peng. Multi-view clustering via deep concept factorization. *Knowledge-Based Systems*, 217:106807, 2021. 1

[4] Mulin Chen and Xuelong Li. Concept factorization with local centroids. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11):5247–5253, 2020. 2

[5] Chris Ding, Tao Li, Wei Peng, and Haesun Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135. ACM, 2006. 6

[6] Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2008. 1, 2, 4, 6

[7] Ky Fan. On a theorem of weyl concerning eigenvalues of linear transformations i. *Proceedings of the National Academy of Sciences*, 35(11):652–655, 1949. 2

[8] Xin-Ru Feng, Heng-Chao Li, Rui Wang, Qian Du, Xiuping Jia, and Antonio Plaza. Hyperspectral unmixing based on nonnegative matrix factorization: A comprehensive review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:4414–4436, 2022. 1

[9] Baoxiang Huang, Linyao Ge, Ge Chen, Milena Radenkovic, Xiaopeng Wang, Jinming Duan, and Zhenkuan Pan. Nonlocal graph theory based transductive learning for hyperspectral image classification. *Pattern Recognition*, 116:107967, 2021.

[10] Jin Huang, Feiping Nie, Heng Huang, and Chris Ding. Robust manifold nonnegative matrix factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 8(3):11, 2014. 6

[11] Yuheng Jia, Hui Liu, Junhui Hou, and Sam Kwong. Semisupervised adaptive symmetric non-negative matrix factorization. *IEEE transactions on cybernetics*, 51(5):2550–2562, 2020. 1

[12] Kehan Kang, Chenglizhao Chen, and Chong Peng. Consensus low-rank multi-view subspace clustering with cross-view diversity preserving. *IEEE Signal Processing Letters*, 30:1512–1516, 2023. 1

[13] Zhao Kang, Zhanyu Liu, Shirui Pan, and Ling Tian. Fine-grained attributed graph clustering. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 370–378. SIAM, 2022. 2

[14] Yong-Deok Kim and Seungjin Choi. Weighted nonnegative matrix factorization. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 1541–1544. IEEE, 2009. 6

[15] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999. 1

[16] Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001. 1, 4

[17] Jiyuan Liu, Xinwang Liu, Jian Xiong, Qing Liao, Sihang Zhou, Siwei Wang, and Yuexiang Yang. Optimal neighborhood multiple kernel clustering with adaptive local kernels. *IEEE Transactions on Knowledge and Data Engineering*, 34(6):2872–2885, 2022. 1

[18] Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Li Liu, Siqi Wang, Weixuan Liang, and Jiangyong Shi. One-pass multi-view clustering for large-scale data. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12324–12333, 2021. 6

[19] Kai Liu, Xiangyu Li, Zhihui Zhu, Lodewijk Brand, and Hua Wang. Factor-bounded nonnegative matrix factorization. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(6):1–18, 2021. 1

[20] Xinwang Liu, Lei Wang, Jian Zhang, Jianping Yin, and Huan Liu. Global and local structure preservation for feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, 25(6):1083–1095, 2014. 1

[21] Xinwang Liu, Sihang Zhou, Li Liu, Chang Tang, Siwei Wang, Jiyuan Liu, and Yi Zhang. Localized simple multiple kernel k-means. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9273–9281, 2021. 1

[22] Canyi Lu, Jiashi Feng, Zhouchen Lin, Tao Mei, and Shuicheng Yan. Subspace clustering by block diagonal representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):487–501, 2019. 2

[23] Jitao Lu, Yihang Lu, Rong Wang, Feiping Nie, and Xuelong Li. Multiple kernel k-means clustering with

simultaneous spectral rotation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4143–4147. IEEE, 2022. 6, 7

[24] Yuwu Lu, Chun Yuan, Wenwu Zhu, and Xuelong Li. Structurally incoherent low-rank nonnegative matrix factorization for image classification. *IEEE Transactions on Image Processing*, 27(11):5248–5260, 2018. 1

[25] Jiaqi Ma, Yipeng Zhang, and Lefei Zhang. Discriminative subspace matrix factorization for multiview data clustering. *Pattern Recognition*, 111:107676, 2021. 1

[26] Xiaoke Ma, Penggang Sun, and Guimin Qin. Nonnegative matrix factorization algorithms for link prediction in temporal networks using graph communicability. *Pattern Recognition*, 71:361 – 374, 2017. 1

[27] Xiaoke Ma, Benhui Zhang, Changzhou Ma, and Zhiyu Ma. Co-regularized nonnegative matrix factorization for evolving community detection in dynamic networks. *Information Sciences*, 528:265–279, 2020. 1

[28] Yang Meng, Ronghua Shang, Fanhua Shang, Licheng Jiao, Shuyuan Yang, and Rustam Stolkin. Semi-supervised graph regularized deep nmf with bi-orthogonal constraints for data representation. *IEEE transactions on neural networks and learning systems*, 31(9):3245–3258, 2019. 1

[29] Yulong Pei, Nilanjan Chakraborty, and Katia Sycara. Nonnegative matrix tri-factorization with graph regularization for community detection in social networks. In *Proceedings of the 24th International Conference on Artificial Intelligence*, pages 2083–2089, 2015. 1

[30] Chong Peng, Chenglizhao Chen, Zhao Kang, Jianbo Li, and Qiang Cheng. Res-pca: A scalable approach to recovering low-rank matrices. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7309–7317, 2019. 1

[31] Chong Peng and Qiang Cheng. Discriminative ridge machine: A classifier for high-dimensional data or imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, 32(6):2595–2609, 2021. 2, 3

[32] Chong Peng, Xingrong Hou, Yongyong Chen, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Global and local similarity learning in multi-kernel space for nonnegative matrix factorization. *Knowledge-Based Systems*, 279:110946, 2023. 1, 2, 3, 6, 7

[33] Chong Peng, Zhao Kang, Huiqing Li, and Qiang Cheng. Subspace clustering using log-determinant rank approximation. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 925–934. ACM, 2015. 2

[34] Chong Peng, Yang Liu, Kehan Kang, Yongyong Chen, Xinxing Wu, Andrew Cheng, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Hyperspectral image denoising using nonconvex local low-rank and sparse separation with spatial-spectral total variation regularization. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022.

[35] Chong Peng, Qian Zhang, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Kernel two-dimensional ridge regression for subspace clustering. *Pattern Recognition*, 113:107749, 2021. 2

[36] Chong Peng, Yiqun Zhang, Yongyong Chen, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Log-based sparse nonnegative matrix factorization for data representation. *Knowledge-Based Systems*, 251:109127, 2022. 1, 2, 7

[37] Chong Peng, Zhilu Zhang, Chenglizhao Chen, Zhao Kang, and Qiang Cheng. Two-dimensional semi-nonnegative matrix factorization for clustering. *Information Sciences*, 590:106–141, 2022. 1

[38] Chong Peng, Zhilu Zhang, Zhao Kang, Chenglizhao Chen, and Qiang Cheng. Nonnegative matrix factorization with local similarity learning. *Information Sciences*, page 325346, Jul 2021. 1, 2, 3

[39] Siyuan Peng, Wee Ser, Badong Chen, and Zhiping Lin. Robust semi-supervised nonnegative matrix factorization for image clustering. *Pattern Recognition*, 111:107683, 2021. 1

[40] Xi Peng, Zhang Yi, and Huajin Tang. Robust subspace clustering via thresholding ridge regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015. 2

[41] Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000. 3

[42] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, and Qiaozhu Mei. Line: Large-scale information network embedding. In *Proceedings of the 24th international conference on world wide web*, pages 1067–1077, 2015. 2, 3

[43] Ming Tong, Yiran Chen, Mengao Zhao, Haili Bu, and Shengnan Xi. A deep discriminative and robust nonnegative matrix factorization network method with soft label constraint. *Neural Computing and Applications*, 31:7447–7475, 2019. 1

[44] Fusheng Wang, Chenglizhao Chen, and Chong Peng. Essential low-rank sample learning for group-aware

subspace clustering. *IEEE Signal Processing Letters*, 30:1537–1541, 2023. 1

[45] Haixian Wang, Sibao Chen, Zilan Hu, and Wenming Zheng. Locality-preserved maximum information projection. *IEEE Transactions on Neural Networks*, 19(4):571–585, 2008. 1

[46] Yu-Xiang Wang, Huan Xu, and Chenlei Leng. Provable subspace clustering: When lrr meets ssc. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. 2

[47] Jiao Wei, Can Tong, Bingxue Wu, Qiang He, Shouliang Qi, Yudong Yao, and Yueyang Teng. An entropy weighted nonnegative matrix factorization algorithm for feature representation. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–11, 2022. 6

[48] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 1

[49] Jianlong Wu, Zhouchen Lin, and Hongbin Zha. Essential tensor learning for multi-view spectral clustering. *IEEE Transactions on Image Processing*, 28(12):5910–5922, 2019. 3

[50] Wei Xu and Yihong Gong. Document clustering by concept factorization. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 202–209, 2004. 1, 2

[51] Yang Zhao, Huiyang Wang, and Jihong Pei. Deep non-negative matrix factorization architecture based on underlying basis images learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(6):1897–1913, 2021. 1

[52] Dengyong Zhou, Jiayuan Huang, and Bernhard Schölkopf. Learning from labeled and unlabeled data on a directed graph. In *Proceedings of the 22nd international conference on Machine learning*, pages 1036–1043, 2005. 3