Multi-View Partitioning via Tensor Methods

Xinhai Liu, Shuiwang Ji, Wolfgang Glänzel, and Bart De Moor, *Fellow, IEEE*

***Abstract*— Clustering by integrating multi-view representations has become a crucial issue for knowledge discovery in heterogeneous environments. However, most prior approaches assume that the multiple representations share the same dimension, limiting their applicability to homogeneous environments. In this paper, we present a novel tensor-based framework for integrating heterogeneous multi-view data in the context of spectral clustering. Our framework includes two novel formulations; that is multi-view clustering based on the integration of the Frobenius-norm objective function (MC-FR-OI) and that based on matrix integration in the Frobenius-norm objective function (MC-FR-MI). We show that the solutions for both formulations can be computed by tensor decompositions. We evaluated our methods on synthetic data and two real-world data sets in comparison with baseline methods. Experimental results demonstrate that the proposed formulations are effective in integrating multi-view data in heterogeneous environments.**

***Index Terms*— Multi-view clustering, tensor decomposition, spectral clustering, multi-linear singular value decomposition, higher-order orthogonal iteration**

I. Introduction

In many real-world scenarios, each object can be described by multiple sets of features. For example, in scientific literature mining, both the textual content and the citation link between articles are often used in the knowledge discovery processes [25]. In multiplex network analysis, we are given a set of multiple networks that share the same set of nodes but possess network-specific links representing different types of relationships between nodes [29]. A particular instance of this scenario is the social network of university students, which may include symmetrized connections from (i) Facebook friendship, (ii) picture friendship, (iii) roommate relations, and (iv) student housing-group preference. These diverse individual activities result in multiple relationship networks among students. Such a learning scenario is called multi-view learning, since each feature set describes a view of the same set of underlying objects. A simple approach to learn from these multi-view data is to learning from each view separately. However, such approaches fail to account for the complementary information encoded into different views.

Multi-view clustering refers to the clustering of the same set of objects with multi-view features, either from various information sources or from different feature representations. Compared with the clustering that is implemented on single-view data, multi-view clustering is expected to yield robust and novel partition results by exploiting the complementary information in different views. One of the recent developments in clustering is the spectral clustering technique, which has seen an explosive proliferation over the past several years [44]. Among many other factors, such as easy implementation and efficiency, one of the key advantages of spectral clustering is that it is based on the relaxation of a global clustering criterion (i.e., normalized cuts). Spectral clustering has been widely employed in many real applications, from image segmentation to community detection. Although spectral clustering [28] works well on single-view data, it is not well suited for the clustering of multi-view data, since it is inherently based on matrix decompositions.

Recently, several multi-view clustering algorithms have been proposed [1], [3], [5], [25], [26], [37], [40], [47]. These multi-view clustering techniques have been shown to yield better performance in comparison to single-view techniques. However, prior methods have some limitations that prevent their wide applicabilities, as we will discuss in the related work. For instance, some techniques assume that the dimensions of the features in multiple views are the same, limiting their applicability to the homogeneous settings. Some other techniques only concentrate on the clustering of two-view data so that it might be hard to extend them to more than a two-view situation [3]. In addition, an appropriate weighting scheme is lacking for these multiple views although coordinating various information from them is also one crucial step in gaining good clustering results [37], [41]. A unified framework that can integrate various types of multi-view data is lacking to date [26], [40].

Tensors are higher-order generalizations of matrices. They have been successfully applied to several domains, such as chemometrics, signal processing, Web search, data mining, scientific computing and image recognition [10], [21], [22], [34], [38], [45]. Traditionally, tensor-based methods have been used to model multi-view data [21], and tensor methods are very powerful tools to analyze the latent pattern hidden in multi-view data. Tensor decompositions capture multi-linear structures in higher-order data-sets, where the data have more than two modes. Tensor decompositions and multi-way analysis allow for extracting hidden (latent) components (cluster structure) and investigating complex relationship among them.

In this paper, we propose a multi-view clustering framework based on tensor methods. Our formulations model the multi-view data as a tensor and seek a joint latent optimal subspace by tensor analysis. Our framework can leverage the inherent consistency among multi-view data and integrate their information seamlessly. Apart from other multi-view clustering strategies, which are usually devised for ad hoc application, our method provides a general framework in which some limitations of prior methods are overcome systematically. In particular, our framework can be extended to various types of multi-view data.

X. Liu is with the Credit Reference Center & Financial Research Institute, The People's Bank of China, Beijing, 100800, China. E-mail: xinhai.liu@yahoo.com.

X. Liu is also with College of Information Science and Engineering & ERCMAMT, Wuhan University of Science and Technology, 430081, Wuhan, China.

S. Ji is with the Department of Computer Science, Old Dominion University, Norfolk, VA, 235290162, USA.

W. Glänzel is with Center for R &D Monitoring (ECOOM), Dept.MSI, Katholieke Universiteit Leuven, Leuven, B3000, Belgium and Hungarian Academy of Sciences, IRPS, Budapest, Hungry.

B.D. Moor is with Department of Electrical Engineering, ESATSCD and IBBT K.U.Leuven Future Health Department, Katholieke Universiteit Leuven, Leuven, B3001, Belgium.

Almost any multiple similarity matrices of the same entities are allowed to be embedded into our framework. In addition, since our framework can obtain a joint optimal subspace, it can be easily extended to other related machine learning tasks, such as classification, spectral embedding and collaborative filtering. Our framework consists of two novel algorithms: multi-view clustering based on optimization integration of the Frobenius-norm objective function (MC-FR-OI) and that based on matrix integration in the Frobenius-norm objective function (MC-FR-MI). In particular, MC-FR-MI can assign each view a suitable weight to boost the clustering. For each strategy, we provide the relevant tensor based solutions. Similar to other variants of PCA in machine learning applications [46], our strategy can be considered as a multi-view PCA analysis.

Figure 1 illustrates the potential benefit of multi-view clustering. The figure shows two groups of data points in a 3-D space. Suppose that due to limitations of the measurement system (such as 2-D cameras in the real world), only 2-D projections of the data points can be observed (such as, X-Y projection, Y-Z projection and X-Z projection in a 3-D X-Y-Z coordinate system). Each of the three projections yields what we call a single-view data set. The figure shows that separation of the two clusters is not possible from any of the three projections separately. However, the three views together do contain the information that was present in the original data. Combination of the three views does not automatically allow proper clustering. The middle right part of the figure shows the result of spectral projection by means of multiple kernel fusion (MKF). MKF does not yield satisfactory results here. In this paper we present a new class of algorithms for multi-view partitioning. The lower right part of Figure 1 shows the results obtained by our MC-OI-MLSVD algorithm. The latent cluster structure hidden amid the multi-view data has clearly been recovered here.

To the best of our knowledge, our work is the first unified attempt to address multi-view clustering within the framework of tensor methods. The key contributions of our work can be summarized as follows:

- We propose to model multi-view data as a tensor and develop a new framework of multi-view clustering by tensor methods.
- We present two novel multi-view clustering strategies with their tensor solutions.
- We systematically evaluate our methods on both a synthetic data set and two real applications.

The rest of the paper is organized as follows. To start, Section II reviews the related work. Then, Section III introduces the concepts of spectral clustering. Next, Section IV presents our tensor based multi-view clustering algorithms. After that, Section V demonstrates the experimental results on synthetic data and practical applications. The related research issues are discussed in Section VI. Finally, we conclude in Section VII.

Notation: To facilitate the distinction between scalars, vectors, matrices, and higher-order tensors, the type of a given quantity will be reflected by its representation: scalars are denoted by lower-case letters $(a, b, \ldots; \alpha, \beta, \ldots)$, vectors are written as italic capitals$(A, B, \ldots)$, matrices correspond to boldface capitals $(\mathbf{A}, \mathbf{B}, \ldots)$, and tensors are written as calligraphic letters $(\mathcal{A}, \mathcal{B}, \ldots)$. This notation is consistently used for lower-order parts of a given quantity. For instance, a_i, a_{ij} and a_{ijk} denote an entry of a vector *A*, a matrix **A** and a tensor $\mathcal{A}$, respectively. The Kronecker product is denoted by $\otimes$. For $\mathbf{A} \in \mathbb{R}^{I \times J}$, $\text{vec}(\mathbf{A}) = (a_{11}\ a_{21}\ \ldots\ a_{IJ})^T \in \mathbb{R}^{JI}$ is the vector in which the columns of **A** are stacked on top of each other. $\text{diag}(\cdot)$ is the column vector that is given by the diagonal of its matrix argument.

II. Related work

A. Multi-view clustering

Bickel and Scheffere [3] propose a multi-view clustering method that extends *k*-means and hierarchical clustering to deal with data with two conditionally independent views. A multi-view clustering strategy via canonical correlation analysis (CCA) is presented in [5]. This method assumes that the views are uncorrelated given the cluster label. The above algorithms only concentrate on the clustering of two-view data thus it might be hard to extend them to more than two-view situations. Meanwhile our strategy is applicable to any multi-view situation. Long *et al*. [26] formulate a multi-view spectral clustering method while investigating multiple spectral dimension reduction. A clustering method based on linked matrix factorization is introduced to fuse information from multiple graphs in [41]. Zhou *et al*. [47] develop a multi-view clustering strategy via generalizing the normalized cut from a single view to multiple views and subsequently they build a multi-view transductive inference. In the above algorithms, a common problem is that the analysis of inherent relationship among multi-view data might be neglected. While in our tensor based strategy, the multi-linear relationship among multi-view data is taken into account. Furthermore, Long *et al*. propose a general model based on collective factorization of the related matrices for clustering multi-type relational data [27]. The strategy focuses on the clustering of multi-type interrelated data objects, rather than on the clustering of the same objects using multiple representations as in our research.

B. Community detection of multi-view networks

Tang *et al.* propose the concept of feature integration to implement the clustering of multi-view social networks [40]. Based on modularity optimization, Mucha *et al.* [29] develop a generalized framework of network quality functions that allow studies of community structure in a general setting encompassing networks that evolve over time, have multiple types of links (multiplexity), and have multiple scales. These methods are applicable to specific type of data with sparse links while our strategy is devised for general data.

C. Kernel fusion and clustering ensemble

Multiple kernel learning aims at finding a combination of kernels to optimize for classification or clustering [20], [25]. Such a solution might sound natural, but its underlying principal is not clear [47]. In addition, the heavy computation of their convex optimization makes them only applicable to small databases [25]. Meanwhile, with the recent research progress in tensor decomposition [32], our strategy has the potential to tackle large-scale databases. Clustering ensemble is also known as clustering aggregation or consensus clustering, which integrates different partitions into a consolidated partition with a consensus function [1], [37]. However, clustering ensemble methods usually concentrate on single-view data to overcome the drawback of *k*-means. In fact, clustering ensemble is embedded into our strategy to facilitate the final partition.

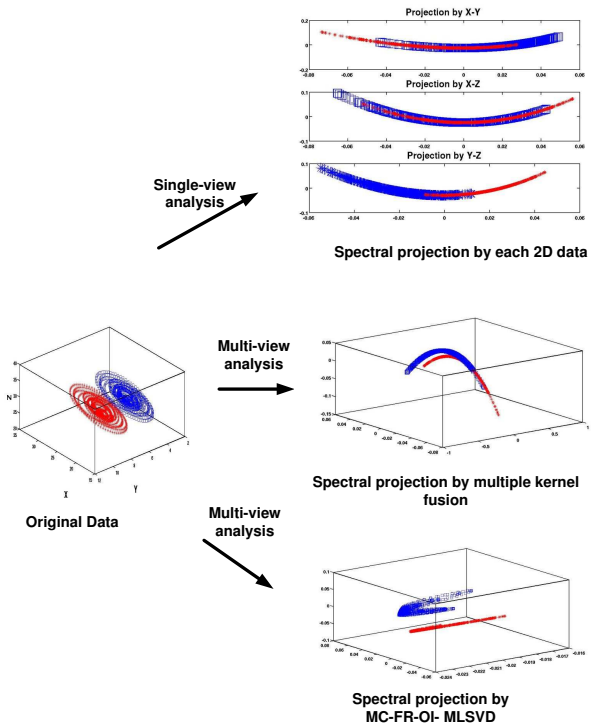

Fig. 1. Comparison of single-view projection versus multi-view projection.

D. Tensor based clustering

Sun *et. al* [38] introduce a dynamic tensor analysis (DTA) algorithm and its variants, and apply them to anomaly detection and multi-way latent semantic indexing. It seems their clustering method is designed for dynamic stream data. Dunlavy *et al.* [10] apply PARAFAC decomposition for analyzing scientific publication data with multiple linkage. Selee *et al.* create a new tensor decomposition called Implicit Slice Canonical Decomposition (IMSCAND) to group information when multiple similarities are known [34]. The last two ideas that integrate multi-view data as a tensor are similar to ours. But our methods rely on a Tucker-type tensor decomposition. Furthermore, in these methods, all single-view data are considered equally important, while we will present a technique that compute weights for the different views.

III. Spectral Clustering

Spectral clustering was originally derived based on relaxation of the normalized cut formulation for clustering [35]. In particular, spectral clustering involves a matrix trace optimization problem [28], [30]. We show in this paper that the spectral clustering formalism can be extended to deal with multi-view problems based on tensor computations.

Given a set of N data points $\{x_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^d$ is the ith data point, a similarity $s_{ij} \geq 0$ can be defined for each pair of data points x_i and x_j based on some similarity measure. An intuitive way to represent this data set is using a graph $G = (V, E)$ in which the vertices V represent the data points and the edges $e_{ij} \in \varepsilon$ characterize the similarity between data points quantified by s_{ij}. Usually, the similarity measure is symmetric, and the graph is undirected. The affinity matrix of the graph G is the matrix $\mathbf{S}$ with entry in row i and column j equal to s_{ij}. The degree of the vertex v_i, defined as

$$d_i = \sum_{j=1}^{N} s_{ij}, \tag{1}$$

is the sum of all the weights of edges connected to v_i. The degree matrix $\mathbf{D}$ is a diagonal matrix containing the vertex degrees $d_1, ..., d_N$ on the diagonal. It follows from the spectral embedding formalism [28], [30], [35] that the Laplacian matrix is defined as $\mathbf{L} = \mathbf{D} - \mathbf{S}$, and the normalized Laplacian matrix, corresponding to the normalized cuts (Ncut), is defined as

$$\mathbf{L}_{\text{Ncut}} = \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{S}_N, \tag{2}$$

where $\mathbf{S}_N$ is the normalized similarity matrix and defined as

$$\mathbf{S}_N = \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}. \tag{3}$$

The matrices $\mathbf{S}_N$ and $\mathbf{L}_{\text{Ncut}}$ have the same eigenvectors, and their eigenvalues are related as $\lambda^{(\mathbf{S}_N)} = 1 - \lambda^{(\mathbf{L}_{\text{Ncut}})}$, where $\lambda^{(\mathbf{S}_N)}$ and $\lambda^{(\mathbf{L}_{\text{Ncut}})}$ are the eigenvalues for $\mathbf{S}_N$ and $\mathbf{L}_{\text{Ncut}}$, respectively.

A. Single-view spectral clustering

We first consider spectral clustering in the single-view setting. Suppose $\mathbf{U} \in \mathbb{R}^{N \times M}$ is the relaxed assignment matrix, where N is the number of data points and M is the number of clusters. The spectral clustering problem can be expressed as

$$\begin{aligned} &\min_{\mathbf{U}} \ \text{trace}(\mathbf{U}^T\mathbf{L}_{\text{Ncut}}\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{4}$$

It follows from the Ky Fan theorem [31] that the optimal solution to the optimization problem in (4) is given by the M dominant

eigenvectors of $\mathbf{L}_{\text{Ncut}}$. Considering the relationship between $\mathbf{S}_N$ and $\mathbf{L}_{\text{Ncut}}$, spectral clustering can equivalently be formulated as

$$\begin{aligned} &\max_{\mathbf{U}} \text{ trace}(\mathbf{U}^T\mathbf{S}_N\mathbf{U}), \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{5}$$

Since $\mathbf{S}_N$ is positive semi-definite, spectral clustering can also be formulated as the following Frobenius norm optimization problem:

$$\begin{aligned} &\max_{\mathbf{U}} \|\mathbf{U}^T\mathbf{S}_N\mathbf{U}\|_F^2, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{6}$$

The objective functions in (5) and (6) are different, but they have the same solution, namely, the columns of the optimal matrix $\mathbf{U}$ span the dominant eigenspace of $\mathbf{S}_N$.

B. Multi-view spectral clustering

We propose different strategies for the integration of multi-view data in the context of spectral clustering.

1) Multi-view clustering by trace maximization (MC-TR-I): The first strategy is to add objective functions of the type in (5), associated with the different views. We consider:

$$\begin{aligned} \max_{\mathbf{U}} \sum_{k=1}^{K} \text{trace}(\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U}) = \text{trace}(\mathbf{U}^T(\sum_{k=1}^{K}\mathbf{S}_N^{(k)})\mathbf{U}), \\ \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}, \end{aligned} \tag{7}$$

where $\mathbf{S}_N^{(k)}$ is the normalized similarity matrix for the kth view and $\mathbf{U}$ is the common factor shared by the views. This corresponds to Multiple Kernel Fusion (MKF) with a linear kernel [20], see Section V-A.

As an alternative, we may optimize a weighted combination of objective functions, where the weights are learnt from the data:

$$\begin{aligned} \max_{\mathbf{U},W} \sum_{k=1}^{K} w_k\text{trace}(\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U}) = \max_{\mathbf{U},W} \text{ trace}(\mathbf{U}^T(\sum_{k=1}^{K} w_k\mathbf{S}_N^{(k)})\mathbf{U}), \\ \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ W \geqslant 0 \text{ and } \|W\|_F = 1. \end{aligned} \tag{8}$$

2) Multi-view clustering by integration of the Frobenius-norm objective function (MC-FR-OI): Note that all terms in the objective function $\sum_{k=1}^{K}\sum_{m=1}^{M}(\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U})_{mm}$ in (7) are nonnegative, since $\mathbf{S}_N^{(k)}$ is positive (semi)definite, $1 \leqslant k \leqslant K$. Instead, we might consider the optimization of $\sum_{k=1}^{K}\sum_{m_1=1}^{M}\sum_{m_2=1}^{M}(\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U})_{m_1m_2}^2$. This corresponds to adding objective functions of the type in (6):

$$\begin{aligned} &\max_{\mathbf{U}} \sum_{k=1}^{K} \|\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U}\|_F^2, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned} \tag{9}$$

3) Multi-view clustering by matrix integration in the Frobenius-norm objective function (MC-FR-MI): As counterpart of (8) we consider:

$$\begin{aligned} &\max_{\mathbf{U},W} \|\mathbf{U}^T(\sum_{k=1}^{K} w_k\mathbf{S}_N^{(k)})\mathbf{U}\|_F^2, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ W \geqslant 0 \text{ and } \|W\|_F = 1. \end{aligned} \tag{10}$$

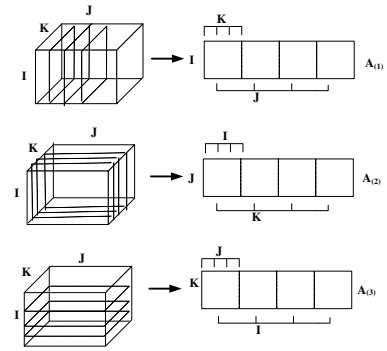

Fig. 3. Matrix unfolding of a third-order tensor

IV. Multi-view spectral clustering via tensor methods

Following the two multi-view clustering strategies discussed above, we present the tensor-based solutions in this Section. Compared to the single-view spectral clustering, which is solved by matrix decomposition, we formulate our multi-view clustering by tensor decomposition. The overview of the tensor-based method is depicted in Figure 2. As shown in the left part of Figure 2, the goal of single-view spectral clustering is to find an optimal latent subspace from single-view data. In contrast, with multi-view data, we want to obtain a joint optimal subspace with the aid of tensor methods.

A. Background on tensors

In this section we provide some basic background on tensors and low multilinear rank approximation. We refer to [6]–[8], [22], [24], [36] for more details. A tensor is a multi-way array. The order of a tensor is the number of modes (or ways). A first-order tensor is a vector, a second-order tensor is a matrix and a tensor of order three or higher is called a higher-order tensor. We only discuss third-order tensor methods that are relevant to our problem.

Matrix unfolding is the process of re-ordering the elements of a tensor into a matrix. The mode-1, mode-2 and mode-3 matrix unfoldings of a tensor $\mathcal{A} \in \mathbb{R}^{I\times J\times K}$ are denoted by $\mathbf{A}_{(1)}, \mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$, respectively. The definition follows from Figure 3.

A tensor can be multiplied by a matrix as follows. Consider matrices $\mathbf{B} \in \mathbb{R}^{I_1\times I}$, $\mathbf{C} \in \mathbb{R}^{J_1\times J}$ and $\mathbf{D} \in \mathbb{R}^{K_1\times K}$, then the mode-1 product $\mathcal{A} \times_1 \mathbf{B}$, mode-2 product $\mathcal{A} \times_2 \mathbf{C}$ and mode-3 product $\mathcal{A} \times_3 \mathbf{D}$ are defined by

$$\begin{aligned} (\mathcal{A} \times_1 \mathbf{B})_{i_1jk} &= \sum_{i=1}^{I} a_{ijk}b_{i_1i},\ \forall i_1, j, k, \\ (\mathcal{A} \times_2 \mathbf{C})_{ij_1k} &= \sum_{j=1}^{J} a_{ijk}c_{j_1j},\ \forall i, j_1, k, \\ (\mathcal{A} \times_3 \mathbf{D})_{ijk_1} &= \sum_{k=1}^{K} a_{ijk}d_{k_1k},\ \forall i, j, k_1, \end{aligned}$$

respectively. The Frobenius norm of $\mathcal{A}$ is defined by

$$\|\mathcal{A}\|_F = (\sum_{ijk} a_{ijk}^2)^{\frac{1}{2}}.$$

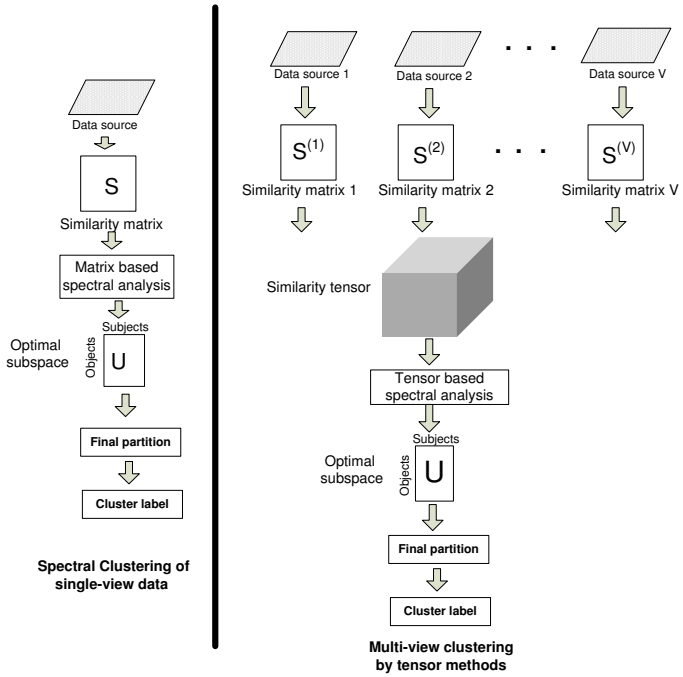

Fig. 2. Comparison between single view (left) and multi-view (right) spectral clustering.

Multilinear singular value decomposition (MLSVD) is one of the possible higher-order extensions of matrix singular value decomposition (SVD) [7], [42], [43]. It decomposes $\mathcal{A}$ as

$$\mathcal{A} = \mathcal{B} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}, \tag{11}$$

in which the factor matrices $\mathbf{U} \in \mathbb{R}^{I \times I}$, $\mathbf{V} \in \mathbb{R}^{J \times J}$ and $\mathbf{W} \in \mathbb{R}^{K \times K}$ are orthogonal and in which the core tensor $\mathcal{B} \in \mathbb{R}^{I \times J \times K}$ satisfies "all-orthogonality" and "ordering" constraints, see [7]. The factor matrices can be thought of as matrices of principal components along each mode. The elements of $\mathcal{B}$ determine the interaction of the factors in the different modes. The matrices $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ can be computed as the matrices of left singular vectors of $\mathbf{A}_{(1)} \in \mathbb{R}^{I \times JK}$, $\mathbf{A}_{(2)} \in \mathbb{R}^{J \times KI}$ and $\mathbf{A}_{(3)} \in \mathbb{R}^{K \times IJ}$, respectively. The columns of $\mathbf{U}, \mathbf{V}$ and $\mathbf{W}$ are the mode-1, mode-2 and mode-3 singular vectors, respectively. The singular values of the unfoldings are the mode-1, mode-2 and mode-3 singular values, respectively.

Consider the following approximation problem:

$$\min_{\mathbf{U},\mathbf{V},\mathbf{W},\mathcal{B}} \|\mathcal{A} - \mathcal{B} \times_1 \mathbf{U} \times_2 \mathbf{V} \times_3 \mathbf{W}\|_F^2, \tag{12}$$

in which now $\mathbf{U} \in \mathbb{R}^{I \times R_1}$, $\mathbf{V} \in \mathbb{R}^{J \times R_2}$ and $\mathbf{W} \in \mathbb{R}^{K \times R_3}$ are column-wise orthonormal with $R_1 \leqslant I$, $R_2 \leqslant J$, $R_3 \leqslant K$, and in which $\mathcal{B} \in \mathbb{R}^{R_1 \times R_2 \times R_3}$. The triplet (R_1, R_2, R_3) is the trilinear rank of the approximand and (12) is a case of what is known as low multilinear rank approximation. It can be shown that the minimization problem is equivalent with the following maximization problem [8], [23]:

$$\max_{\mathbf{U},\mathbf{V},\mathbf{W}} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{V}^T \times_3 \mathbf{W}^T\|_F^2. \tag{13}$$

Analogous to low-rank matrix approximation, one may consider truncated MLSVD for solving (12)–(13), i.e., one may take the columns of $\mathbf{U}$,$\mathbf{V}$,$\mathbf{W}$ in (12)–(13) equal to the dominant multilinear singular vectors of $\mathcal{A}$. Contrary to the matrix case, the approximation is not optimal in general. However, the result is often fairly good and MLSVD truncation is easy to implement. While in the matrix case the sum of the squared discarded singular values give the approximation error, in the tensor case the discarded multilinear singular values yield an upper bound on it [7].

There exist a number of algorithms for the actual optimization in (12)–(13). The most popular technique is the higher-order orthogonal iteration (HOOI), which is an algorithm of the alternating least-squares (ALS) type [8], [23]. In each iteration step, the estimate of one of the matrices $\mathbf{U}, \mathbf{V}, \mathbf{W}$ is optimized, while the other two are kept fixed. It follows from

$$\|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{V}^T \times_3 \mathbf{W}^T\|_F^2 = \|\mathbf{U}^T(\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W}))\|_F^2 \tag{14}$$

that the optimal $\mathbf{U}$, given $\mathbf{V}$ and $\mathbf{W}$, is determined by the R_1-dimensional dominant subspace of the column space of $\mathbf{A}_{(1)}(\mathbf{V} \otimes \mathbf{W})$. The optimization with respect to $\mathbf{V}$ and $\mathbf{W}$ is analogous. In practice the convergence is observed to be linear, with a convergence coefficient that is larger as the problem is better conditioned in the sense of [12]. Alternative algorithms are the trust region method based on truncated conjugate gradient in [18], the quasi-Newton algorithms in [33] and the Newton algorithms in [11], [19]. Truncated MLSVD is often used as initial value. Numerical experiments in [17] suggest that, if there is a gap between the R_nth and the $(R_n + 1)$th mode-n singular values, $n = 1, 2, 3$, one can expect algorithms to find the global optimum. In the same paper it is proved that, if there is a gap and there are nevertheless several local optima, then these are close, both in terms of the cost function value and in terms of the matrices $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{W}$. The absence of a gap may indicate the presence of several local optima for which the cost function value is close. Recent research includes the generalization of numerical algorithms for low-rank approximation of large matrices to low multilinear rank

approximation of large higher-order tensors [32].

B. Tensor construction

There are several options for constructing a tensor from multi-view data. In [15], a tensor is constructed by stacking the object-by-feature matrices derived from multiple views in a tensor as shown in the left part of Figure 4. This construction is only applicable to the scenario of homogeneous data sources, where the dimensions of different feature spaces are the same. In fact, many multi-view applications deal with heterogeneous data sources in which the dimensions of various feature spaces are different. For instance, in the application to scientific publication analysis in Section V-D, the dimension of the citation feature space is 8,305 while the dimension of the text feature space is more than 600,000.

Consequently, in this paper we make a construction that is independent of data dimension, thereby enabling the integration of heterogeneous data sources. We will work with the similarity tensor $\mathcal{A} \in \mathbb{R}^{N\times N\times K}$ obtained by stacking the similarity matrices $\mathbf{S}_N^{(1)}, \mathbf{S}_N^{(2)}, \cdots, \mathbf{S}_N^{(K)}$ associated with the different views. The construction of the similarity tensor is illustrated in the right part of Figure 4. Since the similarity of each view is computed in a different space, normalization is required. In this respect, our definition of similarity matrix in (3) may be regarded as a normalization step.

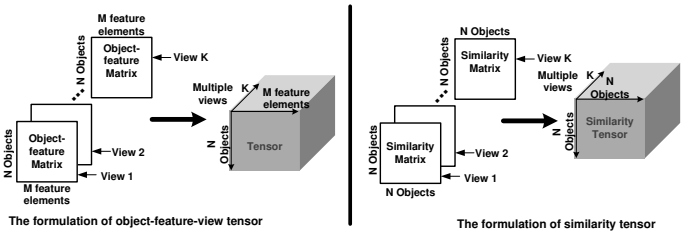

Fig. 4. Comparison of different formulations of multi-view learning using tensor methods.

C. MC-TR-I

The column space of the optimal matrix $\mathbf{U}$ in (7) is the dominant eigenspace of $\sum_{k=1}^{K} \mathbf{S}_N^{(k)}$. The pseudo-code is as follows.

Algorithm IV.1: MC-TR-I-EVD $(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(K)}, M)$

comment: M is the number of clusters

step 1. Build a combined similarity matrix $\sum_{k=1}^{K} \mathbf{S}_N^{(k)}$
step 2. Obtain $\mathbf{U}$ by eigenvalue decomposition
step 3. Normalize the rows of $\mathbf{U}$ to unit length
step 4. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return $(idx:$ the clustering label$)$

The problem in (8) can be written as:

$$\begin{aligned} &\max_{\mathbf{U},W} P(\mathbf{U}) \cdot W, \\ &\text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I} \text{ and } \|W\|_F = 1, \end{aligned} \tag{15}$$

where $P(\mathbf{U}) = \Big(\text{trace}(\mathbf{U}^T\mathbf{S}_N^{(1)}\mathbf{U}) \ \ldots \ \text{trace}(\mathbf{U}^T\mathbf{S}_N^{(K)}\mathbf{U})\Big)$. Note that, compared to (8), the nonnegativity constraint on W has been dropped in (15). Since $\mathbf{S}_N^{(k)}$ is positive (semi)definite, $1 \leqslant k \leqslant K$, all entries of $P(\mathbf{U})$ are nonnegative. Given $\mathbf{U}$, the optimal W is just $P(\mathbf{U})$ scaled to unit-norm, and hence satisfies automatically the nonnegativity constraint. The overall solution can be computed in an alternating fashion by additionally deriving from (8) that the optimal $\mathbf{U}$, given W, follows from the dominant eigenspace of $\sum_{k=1}^{K} w_k\mathbf{S}_N^{(k)}$. The computation of $P(\mathbf{U})$ requires $O(2N^2K)$ flops, the construction of $\sum_{k=1}^{K} w_k\mathbf{S}_N^{(k)}$ also requires $O(2N^2K)$ flops and the computation of its eigenspace $O(6NM^2)$ flops. The pseudo-code is as follows.

Algorithm IV.2: MC-TR-I-EVDIT $(\mathbf{S}_N^{(1)}, \mathbf{S}_N^{(2)}, ..., \mathbf{S}_N^{(K)}, M)$

step 1. Initialize e.g. by MC-TR-I-EVD
while $<$!convergence$>$
do {
iteration step 2.1. Obtain $P(\mathbf{U})$
iteration step 2.2. Calculate the weighting vector W by scaling $P(\mathbf{U})$ to unit-norm
iteration step 2.3. Obtain the relaxed assignment matrix $\mathbf{U}$ from the dominant eigenspace of $\sum_{k=1}^{K}(w_k)\mathbf{S}_N^{(k)}$
step 3. Normalize the rows of $\mathbf{U}$ to unit length
step 4. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return $(idx:$ the clustering label$)$

D. MC-FR-OI

We first discuss the objective function integration approach for multi-view clustering. The problem in (9) can be written as

$$\max_{\mathbf{U}} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 \mathbf{I}\|_F^2, \tag{16}$$

in which $\mathbf{U} \in \mathbb{R}^{N\times M}$ has orthonormal columns. If we take into account the equivalence between (12) and (13), the problem can be visualized as in Figure 5.

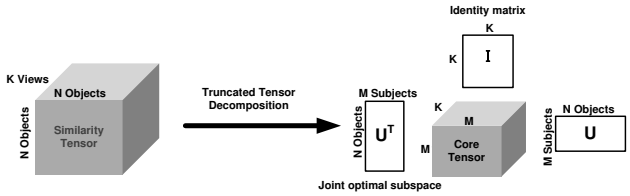

Fig. 5. Illustration of multi-view clustering by objective function integration.

As explained in Section IV-A, an approximate solution to (16) can be obtained from truncated MLSVD. Here, $\mathbf{U}$ is determined by the M dominant mode-1 singular vectors of $\mathcal{A}$, i.e., it follows from the M dominant left singular vectors of $\mathbf{A}_{(1)}$. Because of the partial symmetry of $\mathcal{A}$, $\mathbf{A}_{(2)}$ yields the same $\mathbf{U}$. We call this method MC-FR-OI-MLSVD. Although the approximation is not optimal, the results are often quite good and the algorithm is easy to implement. The computational cost is low, namely O($6NM^2$) flops. The pseudo-code of MC-FR-OI-MLSVD is as follows:

Algorithm IV.3: MC-FR-OI-MLSVD $(\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(K)}, M)$

comment: M is the number of clusters

step 1. Build a similarity tensor $\mathcal{A}$
step 2. Obtain the unfolding matrix $\mathbf{A}_{(1)}$
step 3. Compute $\mathbf{U}$ from the subspace spanned by the M dominant left singular vectors of $\mathbf{A}_{(1)}$
step 4. Normalize the rows of $\mathbf{U}$ to unit length
step 5. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return (*idx*: the clustering label)

We can also look for the optimal solution in (16), for instance by means of the HOOI algorithm. The way one alternates between conditional updates in HOOI makes that the iterates for $\mathbf{U}$ and $\mathbf{V}$ are different, despite the fact that $\mathcal{A}$ is symmetric in its first two modes. Upon convergence, the iterates for $\mathbf{U}$ and $\mathbf{V}$ will match again. Using the estimate of $\mathbf{U}$ for updating in both the first and second mode may lead to divergence [8]. The matrix $\mathbf{W}$ is not updated but set equal to the identity matrix here. The resulting algorithm, called MC-FR-OI-HOOI, is presented as Algorithm IV.4 below. The computation of the product in each of the two steps requires $O(2N^2MK)$ flops. The computation of the subspace additionally requires O($6NK^2M^2$) flops if $N > KM$ and O($2N^2KM$) flops if $N < KM$ [13].

Algorithm IV.4: MC-FR-OI-HOOI ($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(K)}, M$)

step 1. Build a similarity tensor $\mathcal{A}$
step 2. Obtain the unfolding matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$
step 3. Obtain an initial $\mathbf{U}_0$ and $\mathbf{V}_0$ by MLSVD
while <!convergence >
do { iteration step 4.1. $\mathbf{U}_{i+1}$ in dominant subspace of $\mathbf{A}_{(1)}(\mathbf{V}_i \otimes \mathbf{I})$
iteration step 4.2. $\mathbf{V}_{i+1}$ in dominant subspace of $\mathbf{A}_{(2)}(\mathbf{U}_i \otimes \mathbf{I})$
comment: *i* is the counter of iteration
step 5. Normalize the rows of $\mathbf{U}$ to unit length
step 6. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return (idx : the clustering label)

Both MC-FR-OI-MLSVD and MC-FR-OI-HOOI imply a joint matrix compression, as shown in Figure 6. In the case of low multilinear rank approximation, the $(M \times M)$ frontal slices of the core tensor are not necessarily diagonal.

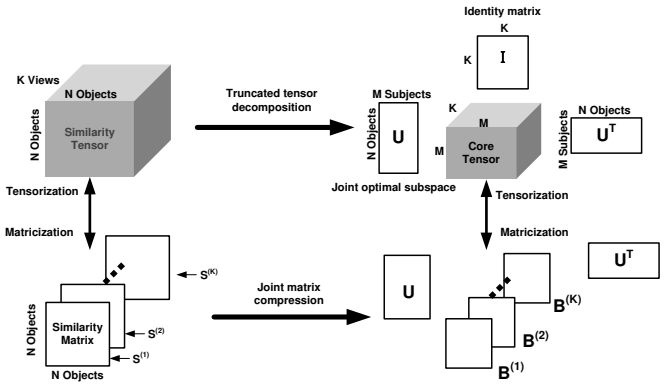

Fig. 6. Joint matrix compression in MC-FR-OI.

E. MC-FR-MI

The problem in (10) can be written as

$$\max_{\mathbf{U},\mathrm{W}} \|\mathcal{A} \times_1 \mathbf{U}^T \times_2 \mathbf{U}^T \times_3 \mathrm{W}^T\|_F^2, \quad \text{s.t. } \mathbf{U}^T\mathbf{U} = \mathbf{I},\ \|\mathrm{W}\|_F^2 = 1. \tag{17}$$

Note that, compared to (10), the nonnegativity constraint on W has been dropped in (17). Since $\mathbf{S}_N^{(k)}$ is positive (semi)definite, $\mathbf{U}^T\mathbf{S}_N^{(k)}\mathbf{U}$ is positive (semi)definite, $1 \leqslant k \leqslant K$. Theorem 1 in the Appendix now implies that, for any $\mathbf{U}$, the entries of the optimal W have the same sign. Since the value of the objective function in (17) is not affected by the sign of W, we can assume that all the weights are nonnegative.

If we take into account the equivalence between (12) and (13), the problem can be visualized as in Figure 7. The matrix $\mathbf{U}$ represents the optimal subspace and the vector W yields the weights of the different views.

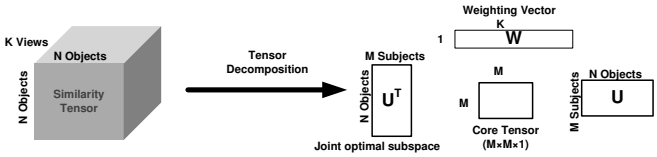

Fig. 7. Multi-view clustering by matrix integration

Using the HOOI algorithm as described earlier, the pseudo code of MC-FR-MI is as follows:

Algorithm IV.5: MC-FR-MI-HOOI-DIRECT ($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(K)}, M$)

step 1. Build a similarity tensor $\mathcal{A}$
step 2. Obtain the unfolding matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$
step 3. Obtain an initial $\mathbf{U}_0$, $\mathbf{V}_0$ and W_0 by MLSVD
while <!convergence >
do { iteration step 4.1. $\mathbf{U}_{i+1}$ in dominant subspace of $\mathbf{A}_{(1)}(\mathbf{V}_i \otimes W_i)$
iteration step 4.2. $\mathbf{V}_{i+1}$ in dominant subspace of $\mathbf{A}_{(2)}(W_i \otimes \mathbf{U}_{i+1})$
iteration step 4.3. W_{i+1} in dominant subspace of $\mathbf{A}_{(3)}(\mathbf{U}_{i+1} \otimes \mathbf{V}_{i+1})$
comment: *i* is the counter of iteration
step 5. Normalize the rows of $\mathbf{U}$ to unit length
step 6. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return (idx : the clustering label)

An equivalent but more efficient implementation is obtained by taking into account that W is not a matrix but a vector. The pseudo code is given as Algorithm IV.6. The matrix $\mathbf{U}_{i+1}$ in step 4.1 of Algorithm IV.5 is just equal to the product $(\sum_{k=1}^{K} \mathbf{S}^{(k)}(W_{i+1})_k)\mathbf{V}_i$. Like-wise, the matrix $\mathbf{V}_{i+1}$ in step 4.2 is equal to $(\sum_{k=1}^{K} \mathbf{S}^{(k)}(W_{i+1})_k)\mathbf{U}_{i+1}$. Alternating until convergence between steps 4.1 and 4.2 of Algorithm IV.5 yields the same matrix for $\mathbf{U}$ and $\mathbf{V}$. The scheme is known as the Orthogonal Iteration for the computation of the dominant eigenspace of $\sum_{k=1}^{K} \mathbf{S}^{(k)}(W_{i+1})_k$ [13]. This yields step 4.3 in Algorithm IV.6, where one may compute the eigenspace by an algorithm of choice. Step 4.1 in Algorithm IV.6 corresponds to the computation of $(\mathcal{A} \times_1 \mathbf{U}_i) \times_2 \mathbf{U}_i$, the cost of which is dominated by the computation of $\mathcal{A} \times_1 \mathbf{U}_i$ since $M \ll N$. This costs $O(2MN^2K)$ flops. The cost of the second step is $O(2N^2K)$ flops. The cost of the third step is O($6NM^2$) flops. Hence, the overall computational cost is O($2(M+1)N^2K$) flops per iteration [13].

Algorithm IV.6: MC-FR-MI-HOOI ($\mathbf{S}^{(1)}, \mathbf{S}^{(2)}, ..., \mathbf{S}^{(K)}, M$)

step 1. Build a similarity tensor $\mathcal{A}$
step 2. Obtain the unfolding matrices $\mathbf{A}_{(1)}$, $\mathbf{A}_{(2)}$ and $\mathbf{A}_{(3)}$
step 3. Obtain an initial $\mathbf{U}_0$ by MLSVD
while <!convergence >
do
- iteration step 4.1. Calculate W_{i+1} as the dominant left singular vector of $\mathbf{A}_{(3)}(\mathbf{U}_i \otimes \mathbf{U}_i)$
- iteration step 4.2. Compute a new integration matrix $\tilde{\mathbf{S}}$ as $\sum_k^K (W_{i+1})_k \mathbf{S}^{(k)}$
- iteration step 4.3. Obtain $\mathbf{U}_{i+1}$ by eigenvalue decomposition of $\tilde{\mathbf{S}}$

comment: i is the counter of iteration
step 5. Normalize the rows of $\mathbf{U}$ to unit length
step 6. Calculate the cluster *idx* with *k*-means on $\mathbf{U}$
return $(idx:$ the clustering label$)$

Remark 1: In the MC-OI framework we discussed two variants, namely MC-FR-OI-MLSVD and MC-FR-OI-HOOI. In the MC-MI framework we have only discussed MC-FR-MI-HOOI. The reason is that tests indicated that here mere truncation of the MLSVD, in which in the third mode only one vector is retained, often yields results that are not satisfactory.

V. Experimental Evaluation

In this Section, we report experimental results of the proposed multi-view partition strategies in comparison with baseline multi-view clustering methods.

A. Baseline methods

We compare with the following six baseline methods.

- Multiple kernel fusion (MKF): Joachims *et al.* [20] integrate different kernels by linear combination for hybrid clustering. The similarity matrix defined in (3) can be regarded as a linear kernel as well. The clustering result of MKF is equal to our MC-TR-I-EVD since the MC-TR-I-EVD is actually the average combination of multiple similarity matrices, so we combine them for the comparsion.
- Feature integration (FI) [40]: With the spectral analysis of each view, their structure features are extracted and then integrated, and SVD is then implemented to obtain the cross-view principal components for clustering.
- Strehl's clustering ensemble algorithm (SA) [37]: Strehl & Ghosh formulate the optimal consensus as the final partition that shares the most information with the partitions of all single-view data to combine. Three heuristic consensus algorithms (cluster-based similarity partition algorithm [C-SPA], hyper-graph partition algorithm [HGPA] and meta-clustering algorithm [MCLA]) based on graph partitioning are employed to obtain the combined partition. In this work, the ensemble consists of single partition from each view. Due to the low computational costs of these techniques, it is quite feasible to use a supra-consensus function that evaluates all three approaches against the objective function and picks the best solution for a given situation [37]. Therefore which exact heuristic consensuses algorithm is adopted relies on each data. In our experiments, MCLA is adopted for all three data sets since it obtains the largest ANMI value for each data respectively. The code of SA is available by the authors [1].
- AdacVote [1]: Ayad & Kamel propose a cumulative vote weighting method (AdacVote) to compute an empirical probability distribution summarizing the clustering ensemble.
- CP-ALS [4], [14]: The CANDECOMP/PARAFAC (CP) decomposition is usually solved by an alternating least squares (ALS) algorithm, for which we use a tensor toolbox for MATLAB [2]. We adopt the default initialization and parameter setting as defined in the toolbox itself.
- Linked matrix factorization (LMF): In Tang's work [41], a quasi-Newton method named Limited memory BFGS (L-BFGS) is adopted for the optimization of LMF. We implement this algorithm with the aid of an optimization toolbox for MATLAB named Poblano [9]. Since LMF is sensitive to initialization, we initialize it by MLSVD that usually provides a good initialization. In addition, the optimization parameters are set as the default setting of the toolbox.

Furthermore, we initialize both MC-FR-OI-HOOI and MC-FR-MI-HOOI by truncated MLSVD. We initialize MC-TR-EVDit with the result of MC-TR-I-EVD (MKF).

B. Performance measures

Regarding clustering evaluation, the data sets used in our experiments are provided with labels. Therefore the clustering performance is evaluated comparing the automatic partitions with the labels using Adaptive Rand Index (ARI) [16] and Normalized Mutual Information (NMI) [37]. To evaluate the ARI and NMI performance, we set the number of clusters for journal data to $M = 7$ and $M = 14$ for disease data.

In order to overcome the drawback of the *k*-means algorithm which is sensitive to various initializations, we adopt the combination of clustering ensemble of SA method and *k*-means for both spectral clustering and multi-view clustering. In particular, we first repeat each clustering method 50 times and use the SA method on the clustering ensemble to obtain the final consensus partition. Consequently, the final partition obtained by each clustering algorithm is unique.

C. Experiment on a synthetic multiplex network

We first evaluate and compare different clustering strategies applied to the synthetic multi-view data. The synthetic data has three communties (clusters), which have 50, 100 and 200 members respectively [39]. We generate various views of interactions among these 350 vertices, that is, each view forms a network that shares the same vertices but has a different interaction pattern. For each view, group members connect with each other following a randomly generated within-group interaction probability. The interaction probability differs with respect to groups at distinct views. After that, we add some noise to the network by randomly connecting any two vertices with low probability. The different views demonstrate different interaction patterns. In this multi-view network that is called a multiplex network according to [29], we construct four interaction matrices, each of whose elements is the interaction strength of a pair of vertices. The visualization of the four adjacent matrices is shown as Figure 8.

In Table I, we list the clustering evaluations of spectral clustering for each single-view data as well as those of multi-view clustering methods. First, it is clear that most multi-view clustering results are better than single-view clustering results. This could be easily explained by the patterns shown in Figure 8.

[1] http://www.lans.ece.utexas.edu/ strehl/soft.html

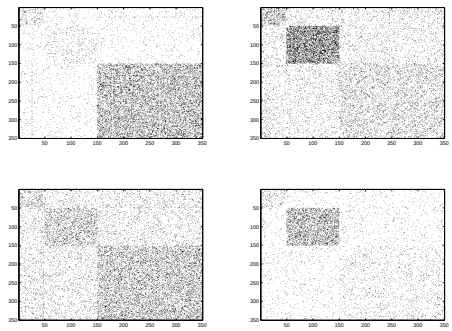

Fig. 8. Visualization of the adjacent matrices of a synthetic multiplex network.

.

	Methods	NMI	ARI
Single View	S-A1	0.7605	0.7995
	S-A2	0.8928	0.9192
	S-A3	0.7198	0.8196
	S-A4	0.6318	0.5599
Multi View	MC-FR-OI-MLSVD	0.9321	0.9508
	MC-FR-OI-HOOI	0.9241	0.9509
	MC-FR-MI-HOOI	*0.9431*	*0.9670*
	MC-TR-I-EVDit	**0.9633**	**0.9717**
	MKF	0.9156	0.9429
	FI	0.8893	0.9243
	SA	0.9251	0.9540
	AdacVote	0.8951	0.9400
	CP-ALS	0.5491	0.1274

TABLE I

EVALUATION OF CLUSTERING METHODS ON A FOUR VIEW SYNTHETIC MULTIPLEX NETWORK.

The first view of the network (left above) only shows one group, and the fourth view (right below) involves another group with the other two groups hidden behind the noise. Thus, using single view is unlikely to recover the inherent cluster structure. This phenomenon is also validated by the low NMI as well as ARI of these two views. Applying multiple views helps reduce the noise and uncover the shared cluster structure. Second, compared with the five other baseline multi-view clustering strategies, our tensor based clustering methods perform better. In particular, MC-FR-OI-MLSVD, MC-FR-MI-HOOI and MC-TR-I-EVDit are obviously superior to others based on both NMI and ARI evaluations. LMF performs wrongly on this data, and thus we omit its comparison.

To evaluate whether the optimized weights assigned to single-view data are correlated with their clustering performance, we compare the ranking of weighting coefficients obtained by MC-FR-MI-HOOI with the ranking of the corresponding clustering performance in Table II, where we list these weighting factors as α_i and we also list weighting factors by MC-TR-I-EVDit as β_i. The ranking of these optimal weights is generally consistent with the ranking of clustering performance. As shown, the top two largest coefficients correctly indicate the top two best single-view data (A_2 and A_3). Although the ranking of the top 2 weighting coefficients is not exactly consistent with the ranking of the corresponding performance, their coefficients are quite near (0.5288 in A2 and 0.5643 in A3).

Sources	β_i	α_i	Ranking of α_i	Performance ranking
A1	0.5036	0.4725	3	3
A2	0.4506	0.5288	2	1
A3	0.5316	0.5643	1	2
A4	0.5106	0.4433	4	4

TABLE II

THE WEIGHTING COEFFICIENTS OF MULTI-VIEW DATA BY MC-MI-HOOI IN SYNTHETIC DATA.

D. Application on scientific documents analysis

In this Section, we apply our algorithms to the scientific analysis of the Web of Science (WoS) journal set. Our objective is to map these journals into different subjects using clustering algorithms.

1) Data description: Historically, bibliometric researchers have focused solely on citation analysis or text analysis, but not on both simultaneously. Recently, many researchers have applied text mining and citation analysis to the journal set analysis. The integration of lexical and citation information is a promising strategy towards better mappings [25]. We adopt a data set obtained from the WoS database by Thomson Scientific which contains articles, letters, notes and reviews from the years 2002 till 2006. To create a balanced benchmark data for evaluation, we select seven categories consisting of 1424 journals. The titles, abstracts and keywords of the journal publications are indexed by a Jakarta Lucene based text mining program using no controlled vocabulary. The weights are calculated by four weighting schemes: TF-IDF, IDF, TF and binary. Therefore, we have obtained four data sources as the lexical information of journals. These four kinds of text data are directly represented as similarity matrices. At the same time, four kinds of citation data represent link-based relationships among journals and consequently, from them, we construct corresponding affinity matrices, denoted as cross-citation, co-citation, bibliographic coupling and binary cross-citation. The details of journal data are presented in **Supplementary material 1**.

We implement the proposed tensor based multi-view clustering methods to integrate multi-view data on journal data. To evaluate the performance, we also apply six popular multi-view clustering methods mentioned in Section V to integrate multi-view data. To verify whether the integration of multi-view data by tensor methods indeed improves the clustering performance, we first systematically compare the performance of all the individual data sources using spectral clustering. As shown in the left part of Table III, the best spectral clustering is obtained on TFIDF data (NMI 0.7280, ARI 0.6601).

Next, we implement our tensor based multi-view clustering on different types of multi-view data integrations detailed in **Supplementary material 2**. Text data and citation data are heterogeneous data because they are generated from various feature spaces (see clustering results of their integration from Table 2 to Table 4). Multi-view data solely from text or citation is homogeneous because it shares the same feature space (see clustering results of homogeneous integration of both text data from Table 5 to Table 7 and citation data from Table 8 to Table 10). As shown, the best multi-view clustering performance is obtained from MC-FR-MI-HOOI by integrating two homogeneous text data of TFIDF and IDF (NMI 0.8201, ARI 0.8229). Moreover, we also find that the clustering performance of different integration schemes varies significantly based on the choice of single-view data. This implies that to some degree, the multi-view clustering

performance depends on the quality of the single-view data involved. For instance, in the best multi-view clustering case above, TFIDF and IDF are the two single-view data sources with the two best clustering performance.

Afterwards, we also investigate the performance of integrating all single-view data using all compared multi-view clustering presented in the right part of Table III. In particular, of all the methods we compared, the best performance is obtained by the MC-FR-OI-HOOI method (NMI 0.7605, ARI 0.7262).

The comparison between the ranking of weighting coefficients by MC-MI-HOOI with the ranking of clustering performance is shown in Table IV, where we list these weighting factors as α_i and we also list weighting factors by MC-TR-I-EVDit as β_i. Because text and citation data are heterogeneous data sources, we separately compare each integration of each type of data in its own feature space. In general, the ranking of these optimal weights is consistent with the ranking of their individual performance. For instance, within the citation feature space, the top two largest coefficients correctly indicate the top two best individual data source (co-citation and cross-citation). In addition, we can see although the values of these weighting factors by MC-FR-MI-HOOI are different from the counterparts by MC-TR-I-EVDit, the ranking of weighting factors by MC-FR-MI-HOOI is the same to that by MC-TR-I-EVDit.

In Figure 9, two confusion matrices of journal data are depicted to illustrate the partition difference between our multi-view clustering result (by MC-FR-OI-HOOI) and the best single-view clustering result (on TFIDF data). The values of the matrices are normalized according to $R_{ij} = C_j/T_i$, where T_i is the total number of journals belonging to standard label of ESI category i and C_j is the number of these T_i journals that are clustered to class j. The results show that the intuitive confusion matrices correspond to the numerical evaluation results. For instance, the quality of clustering obtained by MC-FR-OI-HOOI (NMI 0.7605, ARI 0.7262) is higher than that of spectral clustering on TFIDF. In the confusion matrix of spectral clustering on TFIDF, 15 journals belonging to Agriculture Science (Nr. 1) are mis-clustered to Environment Ecology (Nr. 3), and 60 journals are mis-clustered to Pharmacology and toxicology (Nr. 7). Meanwhile, by MC-FR-OI-HOOI, the number of Agriculture Science (Nr. 1) journals mis-clustered to Environment Ecology is reduced to 7, and the number to Pharmacology and Toxicology is reduced to 26.

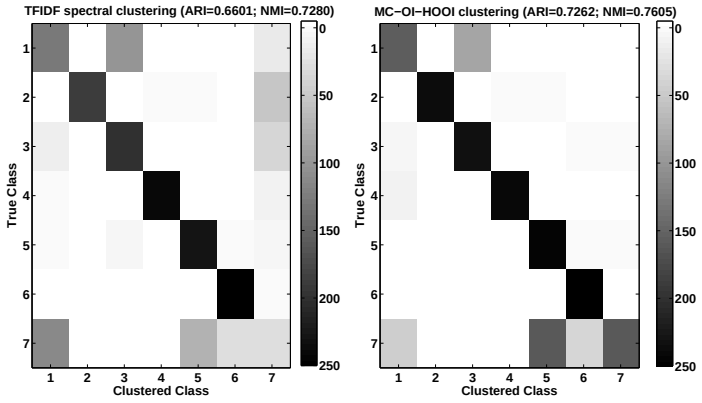

Fig. 9. Confusion matrices of journal data obtained by spectral clustering on TFIDF (left) and MC-FR-OI-HOOI (right). The numbers of cluster labels are consistent with the numbers of ESI journal categories. In each row, the diagonal element represents the fraction of correctly clustered journals and the off-diagonal non-zero element represents the fraction of mis-clustered journals. (Data source: Thomson Reuters, Web of Science)

E. Experiment on disease gene clustering

Text mining helps biologists automatically collect disease-gene associations from large volumes of biological literature. Given a list of genes, we can generate a gene-by-term matrix by the retrieval from the medical literature analysis and retrieval system online (MEDLINE) database. We can also obtain multi-view gene-by-term matrices. The view represents a text mining result retrieved by specific controlled vocabularies, hence multi-view text mining is featured as applying multiple controlled vocabularies to retrieve the gene-centric perspectives from free text publications. The clustering methods can be implemented on these genes to get the group information, which can be utilized for further disease analysis.

The data sets contain ten different gene-by-term text profiles indexed by ten controlled vocabularies. The original disease-related gene data set contains 620 genes that are known to be relevant to 29 diseases. To avoid the effect of imbalanced clusters that may affect the evaluation, we only keep the diseases that have 11 to 40 relevant genes. This step results in 14 genetic diseases and 278 genes. Because the present paper focuses on non-overlapping ("hard") clustering, we additionally remove 16 genes that are relevant to multiple diseases and 17 genes whose term vectors are empty for one of these ten vocabularies. The remaining 245 disease relevant genes are clustered into 14 clusters and biologically evaluated by their disease labels. For each vocabulary based gene-by-term data source, we create a similarity matrix using the value of the cosine similarity for two vectors. The details of the disease gene data analysis are presented in **Supplementary material 3**.

At first, as shown in the left part of Table V, the best clustering performance of individual data sources is obtained on LDDB text mining profile (NMI 0.7088, ARI 0.5942). Next, we also implement 45 types of integration of multi-view text mining data for clustering. The clustering performance is presented in **Supplementary material 4** from Table 13 to Table 15. As shown, the best clustering performance is obtained by MC-FR-OI-HOOI through integrating multi-view data by GO, MeSH, OMIM, NCI, eVOC, KO, LDDB and MP (NMI 0.7687, ARI 0.6364). Afterwards, we also investigate the clustering performance of integrating all single-view data using all the multi-view clustering methods presented in the right part of Table V. In particular, among all the relevant clustering methods, the best performance is still obtained by the MC-FR-OI-HOOI method (NMI 0.7732, ARI 0.6473) as analyzed in the former experiment on journal data. The multi-view clustering strategies with the next two best performance are still our tensor methods, MC-FR-MI-HOOI (NMI 0.7494, ARI 0.6015) and MC-FR-OI-MLSVD (NMI 0.7429, ARI 0.6030). All of our tensor based methods are not only beyond spectral clustering results of any single-view data but also superior to the six baseline multi-view clustering methods, demonstrating the power of our multi-view clustering strategy.

In Table VI, we present the comparison between the ranking of weighting coefficients among multi-view data with the ranking of their corresponding clustering performance, where we list these weighting factors as α_i and we also list weighting factors by

SC-Algorithm	NMI	ARI	MC-Algorithm	NMI	ARI
S-TFIDF	**0.7280**	**0.6601**	MC-FR-OI-MLSVD	0.7331	0.6615
S-IDF	*0.7020*	0.6422	MC-FR-OI-HOOI	**0.7605**	**0.7262**
S-TF	0.6742	0.6305	MC-FR-MI-HOOI	0.7287	0.6756
s-Binary-Text	0.6432	0.6022	MC-TR-I-EVDit	0.7071	0.6612
S-cross-citation	0.6833	0.6057	MKF	0.7327	0.6787
S-co-citation	0.6815	*0.6565*	FI	0.6944	0.6031
S-Bibliographic coupling	0.4398	0.3348	SA	0.7226	0.6952
S-Binary-citation	0.5831	0.5238	AdacVote	*0.7454*	*0.7176*
			CP-ALS	0.7042	0.6377
			LMF	0.5935	0.5058

TABLE III

CLUSTERING PERFORMANCE ON WOS JOURNAL DATABASE

Text data	β_i	α_i	Ranking of α_i	Performance ranking
TFIDF	0.6373	0.5890	1	1
IDF	0.4133	0.4519	3	2
TF	0.5845	0.5580	2	3
Binary-Text	0.2853	0.3708	4	4
Citation	β_i	α_i	Ranking of α_i	Performance ranking
cross-citation	0.5085	0.5372	2	2
co-citation	0.5908	0.5771	1	1
Bibliographic coupling	0.5045	0.5095	3	4
Binary-citation	0.3166	0.3446	4	3

TABLE IV

THE WEIGHTING COEFFICIENTS OF MULTI-VIEW DATA OBTAINED BY MC-MI-HOOI IN JOURNAL DATA.

MC-TR-I-EVDit as β_i. As shown, the largest coefficient correctly indicates the best individual data source (LDDB), while the smallest coefficient correctly indicates the worst individual data source (KO). As a whole, the ranking of these optimal weights are consistent with the ranking of the corresponding performance. In addition, we can see although the values of these weighting factors by MC-MI-HOOI are different from the counterparts by MC-TR-I-EVDit, the ranking of weighting factors by MC-FR-MI-HOOI is almost the same to that by MC-TR-I-EVDit.

In Figure 10, two confusion matrices of disease gene data are depicted to illustrate the partition difference between our multi-view clustering (by MC-FR-OI-HOOI) and the best single-view clustering result (on LDDB). The values of the matrices are normalized according to $R_{ij} = C_j/T_i$, where T_i is the total number of genes belonging to disease category i and C_j is the number of these T_i genes that are clustered to class j. In the first place, it is worth noting that MC-FR-OI-HOOI reduces the number of mis-clustered genes for breast cancer (Nr. 1), mental retardation (Nr. 10), muscular dystrophy (Nr.11) and neuropathy (Nr. 12). Second, there are several diseases where consistent mis-clustering occurs in both methods, such as, cataract (Nr. 3), charcot marie tooth disease (Nr. 4) and diabetes (Nr. 6). The intuitive confusion matrices correspond to the numerical evaluation results. As shown in Table V, the quality of clustering obtained by MC-FR-OI-HOOI (NMI 0.7605, ARI 0.7262) is higher than that of LDDB.

SC-Algorithm	**NMI**	**ARI**	**MC-Algorithm**	**NMI**	**ARI**
S-GO	0.5367	0.3657	MC-FR-OI-MLSVD	0.7429	*0.6030*
S-MeSH	*0.7072*	0.5134	MC-FR-OI-HOOI	**0.7732**	**0.6473**
S-OMIM	0.6971	0.4901	MC-FR-MI-HOOI	*0.7494*	0.6015
S-NCI	0.5153	0.3063	MC-TR-I-EVDit	0.7218	0.5948
S-eVO	0.6048	0.3845	MKF	0.7002	0.5445
S-KO	0.3187	0.1194	FI	0.6743	0.4830
S-LDDB	**0.7088**	**0.5942**	SA	0.7016	0.5495
S-MP	0.6582	0.4962	AdacVote	0.6093	0.5349
S-SNOMED	0.6819	*0.5205*	CP-ALS	0.7241	0.5154
S-Uniprot	0.5692	0.3303	LMF	0.6058	0.4402

TABLE V

CLUSTERING PERFORMANCE ON DISEASE DATA SET.

Sources	β_i	α_i	Ranking of α_i	Performance ranking
GO	0.1818	0.2544	9	8
MeSH	0.2325	0.2842	7	2
OMIM	0.2537	0.2973	4	3
NCI	0.2418	0.2931	6	9
eVO	0.2717	0.3021	3	6
KO	0.1310	0.2216	10	10
LDDB	0.7228	**0.5303**	1	1
MP	0.2725	*0.3113*	2	5
SNOMED	0.2110	0.2713	8	4
Uniprot	0.2405	0.2970	5	7

TABLE VI

THE WEIGHTING COEFFICIENTS OF MULTI-VIEW DATA OBTAINED BY MC-FR-MI-HOOI IN DISEASE DATA.

Q1.1 In spectral clustering, checking the "elbow" of the plot of the eigenvalues of single-view data provides a heuristic estimate of the number of clusters [28]. Analogous, in our tensor approach the plot of mode-1 singular values of the similarity tensor provides a heuristic estimate of the number of clusters. In Figure 11 we plot the 20 dominant mode-1 singular values for our three data sets. The elbow for the synthetic data is between 2 and 4. The real number of clusters is 3. The middle and right parts of Figure 11 show the elbow plots for the journal data and the disease data, respectively. For our analysis we used the cluster numbers indicated by the arrows. Moreover, in Figure 1 and Figure 2 of **Supplementary material 5, we also compare the 1-mode singular value curves using different tensors of journal data and gene-disease data. Those tensors are generated from different numbers of views, for instance, in journal data, we generate different tensors by using various combinations from two to seven views.

As shown, for each data, the 1-mode singular value plot is quite stable w.r.t. the different combinations of multiple views.

To investigate the computational time, we benchmark our tensor

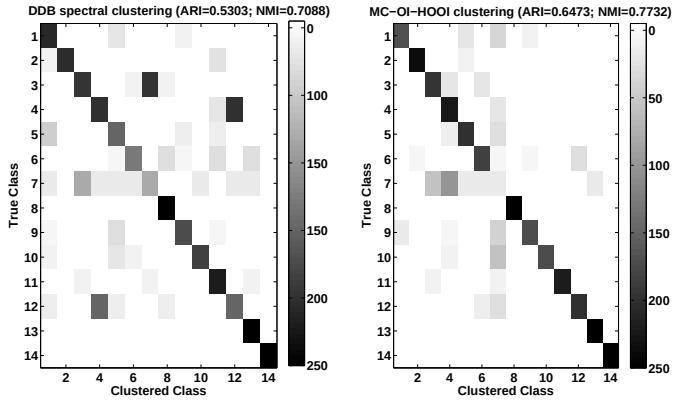

Fig. 10. Confusion matrices of disease gene data obtained by spectral clustering on LDDB (left) and MC-FR-OI-HOOI (right). The numbers of cluster labels are consistent with the numbers of diseases. In each row, the diagonal element represents the fraction of correctly clustered genes and the off-diagonal non-zero element represents the fraction of mis-clustered genes

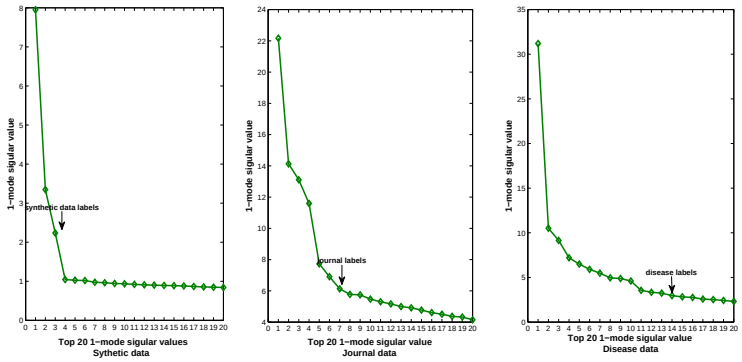

Fig. 11. Plot of the top 20 1-mode singular values of tensors constructed from different multi-view data (synthetic data on the left, journal data on the middle and disease data on the right). All possible views of each data are employed to construct these tensors.

Algorithm	disease data	journal data
MC-FR-OI-MLSVD	4.82	33.34
MC-FR-OI-HOOI	9.32	64.98
MC-FR-MI-HOOI	2.79	41.75
MC-TR-I-EVDit	2.78	8.34
MKF	1.23	3.97
FI	3.94	20.06
SA	37.29	60.94
AdacVote	37.31	44.67
CP-ALS	7.82	127.55
LMF	9.40	203.41

TABLE VII

COMPARISON OF CPU TIME IN SECONDS FOR REAL DATA

based multi-view clustering algorithms with 6 different multi-view clustering methods on the two application data sets. As shown in Table VII, our three tensor based strategies (MC-FR-OI-MLSVD, MC-FR-OI-HOOI and MC-FR-MI-HOOI) are efficient. For instance, they are faster than four multi-view clustering methods (SA, AdacVote, CP-ALS and LMF). Obviously, MC-FR-OI-MLSVD is more efficient. **Q1.3 The difference of the computation time of our three algorithms is caused by the different of their computational complexity. The computational complexity of MC-FR-OI-MLSVD is O$(6N(M)^2)$; the computational complexity of MC-FR-OI-HOOI is O$(6NK^2M^2)$ if $N > KM$ and O$(2N^2KM)$ flops if $N < KM$; and the computational complexity of MC-FR-MI-HOOI is O$(2(M+1)N^2K)$.

On the other hand, although MKF and FI seem more efficient than our three tensor based algorithms, our proposed methods yield much better performance or more enriched information (the weighting factors of the single views).

Meanwhile, the two clustering ensemble methods SA and AdacVote require more computation time since they involve the partition of each single-view data. Consequently, with number of views increasing, the computation of the clustering ensemble method will become more and more intensive.

VI. DISCUSSION

Based on the clustering performance of the multi-view clustering strategies, first, MKF is efficient when compared with tensor based strategies. However, MKF only combines multiple kernels (similarity matrices) in a simple way using the average sum of multiple similarities. Thus, such a simple combination neglects the discriminating capability of each kernel. Second, clustering ensemble methods (SA and AdacVote) rely on discrete hard clustering. Using only the final partition information seems too fragile to integrate.

In addition, because the partition of every single-view data is required, the implementation of clustering ensemble methods is not efficient as shown in Table VII.

Third, considering LMF, we found that the clustering performance relies on the initialization, and hence the partition results are quite unstable. Moreover, its optimization mechanism consumes much time.

Fourth, for CP-ALS, the failure might be due to the un-orthogonal property of the relaxed assignment matrix **U** after tensor decomposition. The reason is that the similarity matrix in (3) we adopted to construct the tensor corresponds to the Ncut based Laplacian matrix that requires the orthogonal partition in spectral clustering.

Meanwhile, our tensor based multi-view spectral clustering can be thought of as a "Multi-view PCA" analysis, which integrates multi-view information seamlessly and forms a joint optimal subspace. Therefore our strategy can extract the latent pattern shared by all views and filter out irrelevant information or noise. The tensor based multi-view clustering of optimization integration strategy (MC-FR-OI-MLSVD and MC-FR-OI-HOOI) leverages the effect of each single-view data in an appropriate way while the tensor based multi-view clustering of matrix integration strategy (MC-FR-MI-HOOI) is able to utilize the linear relationship of multi-view data for joint analysis.

VII. CONCLUSION AND OUTLOOK

We proposed a multi-view clustering framework based on high-order analogues of the matrix Singular Value Decomposition (SVD) and Principal Component Analysis (PCA). Our framework can be regarded as a multi-view extension of spectral clustering. With our tensor formulation, both heterogeneous and homogeneous information can be integrated to facilitate the clustering task.

We presented two new multi-view clustering strategies: multi-view clustering by the integration of the Frobenius-norm objective

function (MC-FR-OI) as well as the matrix integration in the Frobenius-norm objective function (MC-FR-MI). The relevant tensor based solutions are proposed, which are either iterative optimization or efficient approximation. All of them are capable of utilizing the global information of multi-view data while taking the effect of single-view data into consideration. Furthermore, these different methods can be applied to various practical scenarios.

We employed our algorithms to both synthetic data and two real applications. The clustering performance demonstrated that our algorithms are not only superior to single-view spectral clustering methods, but also superior to other baseline multi-view clustering methods.

In later research, we will carry out our work in the following directions: (1) We will investigate other alternative tensor solutions, such as INDSCAL [4], as well as efficient tensor decomposition for scalable application; (2) We will extend our multi-view clustering algorithm to higher-order data (we only use three-order data in this research), such as, adding another temporal order that allows data to vary at different time points; (3) Our framework is not limited to the clustering analysis. Since its core is to seek a joint optimal latent subspace, it can be extended to other multi-view learning tasks: for instance, classification, spectral embedding, collaborative filtering and even information retrieval.

APPENDIX

Theorem 1: Consider positive (semi)definite matrices $\mathbf{S}^{(1)}$, $\mathbf{S}^{(2)}$, $\ldots$, $\mathbf{S}^{(K)} \in \mathbb{R}^{N\times N}$. Let $W \in \mathbb{R}^K$ have unit-norm and consider $f(W) = \|\sum_{k=1}^K w_k \mathbf{S}^{(k)}\|_F$. Then the entries of the vector W that maximizes f, have equal sign.

Proof:

Define $\tilde{\mathbf{S}} = \Big(\text{vec}(\mathbf{S}^{(1)})\ \text{vec}(\mathbf{S}^{(2)})\ \ldots\ \text{vec}(\mathbf{S}^{(K)})\Big) \in \mathbb{R}^{N^2\times K}$. The vector W that maximizes f is the dominant right singular vector of $\tilde{\mathbf{S}}$. This is the dominant eigenvector of $\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}$.

Consider the eigenvalue decomposition $\mathbf{S}^{(k)} = \mathbf{Q}^{(k)}\mathbf{D}^{(k)}{\mathbf{Q}^{(k)}}^T$, in which $\mathbf{Q}^{(k)}$ is orthogonal and $\mathbf{D}^{(k)}$ is diagonal and positive (semi)definite, $1 \leqslant k \leqslant K$. We have

$$\begin{aligned}(\tilde{\mathbf{S}}^T\tilde{\mathbf{S}})_{kl} &= \text{vec}(\mathbf{S}^{(k)})^T\text{vec}(\mathbf{S}^{(l)}) \\ &= \text{vec}(\mathbf{D}^{(k)})^T\text{vec}({\mathbf{Q}^{(k)}}^T\mathbf{S}^{(l)}\mathbf{Q}^{(k)}) \\ &= \text{vec}(\text{diag}(\mathbf{D}^{(k)}))^T\text{vec}(\text{diag}({\mathbf{Q}^{(k)}}^T\mathbf{S}^{(l)}\mathbf{Q}^{(k)})),\end{aligned}$$

$1 \leqslant k,l \leqslant K$. Since $\mathbf{D}^{(k)}$ is positive (semi)definite, $\text{vec}(\text{diag}(\mathbf{D}^{(k)}))$ has only nonnegative entries. Since $\mathbf{S}^{(l)}$ is positive (semi)definite, $\text{vec}(\text{diag}({\mathbf{Q}^{(k)}}^T\mathbf{S}^{(l)}\mathbf{Q}^{(k)}))$ has only nonnegative entries as well. We conclude that the entries of $\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}$ are nonnegative. According to the Perron-Frobenius theorem, the entries of the dominant eigenvector W of $\tilde{\mathbf{S}}^T\tilde{\mathbf{S}}$ have equal sign. ■

ACKNOWLEDGMENT

The authors would like to thank L. De Lathauwer for deriving the version of HOOI with a single vector in one mode and for the theorem and proof in the Appendix. This work was supported by (1) National Natural Science Foundation of China (Grant No. 61105058); (2) The joint postdoctral programme by Credit Reference Center and Financial Research Institute, The People's Bank of China; (3) China Postdoctoral Science Foundation funded project; (4) The related fundings: GOA-AMBioRICS, CoE-EF/05/006(OPTEC), IOF-SCORES4CHEM, FWO-G0452.04, FWO-G.0499.04, FWO-G.0211.05, FWO-G.0226.06, FWO-G.0321.06, FWO-G.0302.07, FWO-ICCoS, FWO-ANMMM, FWO-MLDM, IWT-McKnow-E, Eureka-Fliteplus, IAP-P6/04, ERNSI; (5) Flemish Government: Center for R&D Monitoring.

REFERENCES

[1] H. G. Ayad and M. S. Kamel. Cumulative voting consensus method for partitions with variable number of clusters. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(1):160–173, 2008.

[2] B. W. Bader and T. G. Kolda. MATLAB tensor toolbox version 2.4. http://csmr.ca.sandia.gov/ tgkolda/TensorToolbox/, March 2010.

[3] S. Bickel and T. Scheffer. Multi-view clustering. In *Proceedings of the Fourth IEEE International Conference on Data Mining*, ICDM '04, pages 19–26, Washington, DC, USA, 2004. IEEE Computer Society.

[4] J. D. Carroll and J. J. Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of 'Echart-Young', decomposition. *Psychometricka*, 35:283–319, 1970.

[5] K. Chaudhuri, S. M. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pages 129–136, New York, NY, USA, 2009. ACM.

[6] A. Cichocki, R. Zdunek, A.-H. Phan, and S. Amari. *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis and Blind Source Separation.* John Wiley, 2009.

[7] L. De Lathauwer, B. D. Moor, and J. Vandewalle. A multilinear singular value decomposition. *SIAM J. Matrix Anal. Appl.*, 21(4):1253–1278, 2000.

[8] L. De Lathauwer, B. D. Moor, and J. Vandewalle. On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM J. Matrix Anal. Appl.*, 21(4):1324–1342, 2000.

[9] D. M. Dunlavy, T. G. Kolda, and E. Acar. Poblano v1.0: A MATLAB toolbox for gradient-based optimization. Technical Report SAND2010-1422, Sandia National Laboratories, Albuquerque, NM and Livermore, CA, March 2010.

[10] D. M. Dunlavy, T. G. Kolda, and W. P. Kegelmeyer. Multilinear algebra for analyzing data with multiple linkages. Technical Report SAND2006-2079, Sandia National Laboratories, 2006.

[11] L. Eldén and B. Savas. A Newton–Grassmann method for computing the best multilinear rank-(r_1, r_2, r_3) approximation of a tensor. *SIAM J. Matrix Anal. Appl*, 31:248–271, 2009.

[12] L. Eldén and B. Savas. Perturbation theory and optimality conditions for the best multilinear rank approximation of a tensor. *SIAM. J. Matrix Anal. Appl.*, 32:1422–1450, 2011.

[13] G. H. Golub and C. F. Van Loan. *Matrix Computations, 3rd ed.* Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[14] R. A. Harshman. Foundations of the PARAFAC procedure: Model and conditions for an 'explanatory' multi-modal factor analysis. *UCLA working papers in phonetics*, 16:1–84, 1970.

[15] H. Huang, C. Ding, D. Luo, and T. Li. Simultaneous tensor subspace selection and clustering: the equivalence of high order SVD and k -means clustering. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 327–335, New York, NY, USA, 2008. ACM.

[16] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

[17] M. Ishteva, P.-A. Absil, S. Van Huffel, and L. De Lathauwer. Tucker compression and local optima. *Chemometr. Intell. Lab. Syst.*, 106(1):57–64, 2011.

[18] M. Ishteva, L. De Lathauwer, P.-A. Absil, and S. Van Huffel. Best low multilinear rank approximation of higher-order tensors, based on the Riemannian trust-region scheme. *SIAM Journal on Matrix Analysis and Applications*, 32(1):115–135, 2011.

[19] M. Ishteva, L. D. Lathauwer, P.-A. Absil, and S. V. Huffel. Differential-geometric Newton algorithm for the best rank-(r_1, r_2, r_3) approximation of tensors. *Numerical Algorithms*, 51(2):179–194, 2009.

[20] T. Joachims, N. Cristianini, and J. Shawe-Taylor. Composite kernels for hypertext categorisation. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 250–257, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

[21] T. Kolda and B. Bader. The TOPHITS model for higher-order web link analysis. In *Proceedings of the SIAM Data Mining Conference Workshop on Link Analysis, Counterterrorism and Security*, 2006.

[22] T. G. Kolda and B. W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
[23] P. Kroonenberg and J. de Leeuw. Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, 1980.
[24] P. M. Kroonenberg. *Applied Multiway Data Analysis*. Wiley, 2008.
[25] X. Liu, S. Yu, Y. Moreau, B. De Moor, W. Glänzel, and F. Janssens. Hybrid clustering of text mining and bibliometrics applied to journal sets. In *Proceedings of SIAM International Conference on Data Mining*, pages 49–60, Philadelphia, PA, USA, 2009. SIAM.
[26] B. Long, P. S. Yu, and Z. M. Zhang. A general model for multiple view unsupervised learning. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 822–833, 2008.
[27] B. Long, Z. M. Zhang, X. Wú, and P. S. Yu. Spectral clustering for multi-type relational data. In *Proceedings of the 23rd international conference on Machine learning*, pages 585–592, 2006.
[28] U. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.
[29] P. J. Mucha, T. Richardson, K. Macon, M. A. Porter, and J.-P. Onnela. Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878, 2010.
[30] A. Ng, M. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, pages 849–856. MIT Press, 2001.
[31] M. L. Overton and R. S. Womersley. Optimality conditions and duality theory for minimizing sums of the largest eigenvalues of symmetric matrices. *Math. Program.*, 62(2):321–357, 1993.
[32] B. Savas and L. Eldén. Krylov-type methods for tensor computations. *arXiv:1005.0683v2 [math.NA]*, 2010.
[33] B. Savas and L.-H. Lim. Quasi-Newton methods on Grassmannians and multilinear approximations of tensors. *SIAM Journal on Scientific Computing*, 32:3352–3393, 2010.
[34] T. M. Selee, T. G. Kolda, W. P. Kegelmeyer, and J. D. Griffin. Extracting clusters from large datasets with multiple similarity measures using IMSCAND. In M. L. Parks and S. S. Collis, editors, *CSRI Summer Proceedings 2007, Technical Report SAND2007-7977, Sandia National Laboratories, Albuquerque, NM and Livermore, CA*, pages 87–103, 2007.
[35] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[36] A. Smilde, R. Bro, and P. Geladi. *Multi-way Analysis: Applications in the Chemical Sciences*. Wiley, West Sussex, England, 2004.
[37] A. Strehl and J. Ghosh. Cluster ensembles-a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3:583–617, 2002.
[38] J. Sun, D. Tao, and C. Faloutsos. Beyond streams and graphs: dynamic tensor analysis. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge Discovery and Data mining*, pages 374–383, New York, NY, USA, 2006. ACM.
[39] L. Tang, X. Wang, and H. Liu. Uncovering groups via heterogeneous interaction analysis. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 143–152, Washington, DC, USA, 2009. IEEE Computer Society.
[40] L. Tang, X. Wang, and H. Liu. Community detection in multi-dimensional networks. Technical report, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, 2010.
[41] W. Tang, Z. Lu, and I. S. Dhillon. Clustering with multiple graphs. In *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*, pages 1016–1021, Washington, DC, USA, 2009. IEEE Computer Society.
[42] L. Tucker. The extension of factor analysis to three-dimensional matrices. In H. Gulliksen and N. Frederiksen, editors, *Contributions to mathematical psychology*, pages 109–127. Holt, Rinehart & Winston, NY, 1964.
[43] L. Tucker. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31:279–311, 1966.
[44] D. Verma and M. Meila. A comparison of spectral clustering algorithms. Technical report, Department of CSE University of Washington Seattle, WA, 2003.
[45] J. Ye. Generalized low rank approximations of matrices. *Machine Learning*, 61:167–191, 2005.
[46] J. Ye, R. Janardan, and Q. Li. GPCA: an efficient dimension reduction scheme for image compression and retrieval. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge Discovery and Data mining*, New York, NY, USA, 2004. ACM.
[47] D. Zhou and C. J. C. Burges. Spectral clustering and transductive learning with multiple views. In *Proceedings of the 24th international conference on Machine Learning*, pages 1159–1166, New York, NY, USA, 2007. ACM.

Xinhai Liu received the Ph.D. degree in Electrical Engineering from Katholieke Universiteit Leuven (KU Leuven), Belgium, in 2011. From October, 2011, he has been working as a researcher in People's Bank of China, Beijing, China. His research interests include data mining, credit risk management and information science.

Shuiwang Ji received the Ph.D. degree in Computer Science from Arizona State University, Tempe, AZ, in 2010. Currently, he is an Assistant Professor in the Department of Computer Science, Old Dominion University, Norfolk, VA. His research interests include machine learning, data mining, and bioinformatics. He received the Outstanding Ph.D. Student Award from Arizona State University in 2010.

Wolfgang Glänzel is Professor at KU Leuven and Director of the Centre for R&D Monitoring. He is also Senior Scientist at the Dept. Science Policy & Scientometrics at the Library of the Hungarian Academy of Sciences in Budapest. He holds a doctorate in mathematics from Eötvös Lorand University Budapest (Hungary) as well as a PhD in Social Sciences from University Leiden (Netherlands). In the 1990s he was an Alexander von Humboldt Fellow for two years in Germany. In 1999 he received the international Derek deSolla Price Award for outstanding contributions to the quantitative studies of science.

Bart De Moor is Professor at KU Leuven. In 1983, he obtained his Master (Engineering) Degree in Electrical Engineering at KU Leuven, Belgium, and a PhD in Engineering at the same university in 1988. His work has won him several scientific awards (Leybold-Heraeus Prize (1986), Leslie Fox Prize (1989), Guillemin-Cauer best paper Award of the IEEE Transactions on Circuits and Systems (1990), Laureate of the Belgian Royal Academy of Sciences (1992), bi-annual Siemens Award (1994), best paper award of Automatica (IFAC, 1996), IEEE Signal Processing Society Best Paper Award (1999)). In November 2010, he received the 5-annual FWO Excellence Award out of the hands of King Albert II of Belgium. Since 2004, he is a fellow of the IEEE. His research interests are in numerical linear algebra and optimization, system theory and system identification, quantum information theory, control theory, data-mining, information retrieval and bio-informatics, areas in which he has (co-)authored several books and hundreds of research papers.