Spectral Perturbation Meets Incomplete Multi-view Data

Hao Wang[1,3], **Linlin Zong**[2], **Bing Liu**[3], **Yan Yang**[1*] and **Wei Zhou**[1]

[1]School of Information Science and Technology, Southwest Jiaotong University, Chengdu, China
[2]School of Software, Dalian University of Technology, Dalian, China
[3]Department of Computer Science, University of Illinois at Chicago, Chicago, USA
hwang@my.swjtu.edu.cn, llzong@dlut.edu.cn, liub@uic.edu, yyang@swjtu.edu.cn

Abstract

Beyond existing multi-view clustering, this paper studies a more realistic clustering scenario, referred to as *incomplete multi-view clustering*, where a number of data instances are missing in certain views. To tackle this problem, we explore spectral perturbation theory. In this work, we show a strong link between perturbation risk bounds and incomplete multi-view clustering. That is, as the similarity matrix fed into spectral clustering is a quantity bounded in magnitude $\mathcal{O}(1)$, we transfer the missing problem from data to similarity and tailor a matrix completion method for incomplete similarity matrix. Moreover, we show that the minimization of perturbation risk bounds among different views maximizes the final fusion result across all views. This provides a solid fusion criteria for multi-view data. We motivate and propose a Perturbation-oriented *Incomplete multi-view Clustering* (PIC) method. Experimental results demonstrate the effectiveness of the proposed method.

1 Introduction

Many applications face the situation where each data instance in a set $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ is sampled from multiple views. Here each $\mathbf{x}_i|_{i=1}^n$ is denoted by multiple views, e.g., m views $\{\mathbf{x}_i^1, ..., \mathbf{x}_i^m\}$. Such forms of data are referred to as multi-view data. Multi-view clustering aims to provide a more accurate and stable partition than single view clustering by considering data from multi-views [Chao *et al.*, 2017; Yang and Wang, 2018]. To date, most existing multi-view clustering methods, even the most recent methods such as [Tao *et al.*, 2018; Zong *et al.*, 2018; Nie *et al.*, 2018; Wang *et al.*, 2019; Huang *et al.*, 2019] work under the assumption that every data instance is sampled from all views. We call this assumption the *complete sampling* assumption. However, the complete sampling assumption is too strong as it frequently happens that some data instances are not sampled in certain views because of sensor faults or machine malfunctions. This leads to the result that the collected multi-view data are incomplete in some views. We call such data *incomplete multi-view data*.

The problem of clustering incomplete multi-view data is known as *incomplete multi-view clustering* (or *partial multi-view clustering*) [Hu and Chen, 2018; Li *et al.*, 2014]. Based on the existing works, the main challenge of this problem is 2-fold: (1) how to partition each instance with m views into its group, and (2) how to deal with incomplete views. To address these two challenges, existing incomplete multi-view clustering methods built upon non-negative matrix factorization, kernel learning or spectral clustering to learn a consensus representation for all views and tackled incomplete views by exploring two main directions. The first direction is to project each incomplete view data into a common subspace or a specific subspace [Li *et al.*, 2014; Zhao *et al.*, 2016; Yin *et al.*, 2017; Zhao *et al.*, 2018; Cai *et al.*, 2018; Wang *et al.*, 2018]. However, these methods only work for two-view data. The second direction is to fill the missing instances using matrix completion [Xu *et al.*, 2015; Zhu *et al.*, 2018; Wen *et al.*, 2018; Liu *et al.*, 2018; Shao *et al.*, 2015; Hu and Chen, 2018; Zhou *et al.*, 2019]. Most of them still fill the missing features with average feature values. However, such a filling method is naive as the features in both inter-class and intra-class may have large variances. In addition, most existing approaches only evaluate their clustering performance on toy (randomly generated) incomplete multi-view data.

To address the limitations discussed above, this paper builds a strong link between the spectral perturbation theory and incomplete multi-view clustering. Specifically, we propose a new approach, denoted by Perturbation-oriented Incomplete multi-view Clustering (PIC). It transfers *feature-value missing* to *similarity-value missing* and reduces the spectral perturbation risk among different views to generate the final clustering results by exploiting the key characteristics of spectral clustering. The proposed approach consists of two main phases.

Phase 1 is similarity matrix completion. Given the data matrix of each view, it first generates a similarity matrix (or affinity matrix) for each view. Then it completes the missing similarity entries using average similarity values of other views which have those missing instances.

Phase 2 is consensus matrix learning. It first computes the Laplacian matrix of each completed similarity matrix, and then weights each Laplacian matrix using the perturbation theory to learn a consensus Laplacian matrix. Finally, it performs clustering on the consensus Laplacian matrix.

*Corresponding author.

The proposed method PIC can work because spectral clustering partitions data instances according to their similarities, where the similarity value of any two data instances is a quantity bounded in magnitude $\mathcal{O}(1)$. Another crucial point is that the perturbation of spectral clustering is determined by the eigenvectors of the Laplacian matrix, which can be measured by the canonical angle between the subspaces of different eigenvectors. Thus, we can reduce the perturbations among different views by optimizing the canonical angle. We will discuss the details in the subsequent sections.

In summary, this paper makes the following contributions. (1) It proposes a novel incomplete multi-view clustering method by exploiting the spectral perturbation theory. The proposed method transfers feature missing to similarity missing and weights the Laplacian matrix of each view based on perturbation theory to learn a consensus Laplacian matrix for the final clustering. To our knowledge, this is the first such formulation. (2) It provides an upper bound of the spectral perturbation risk among different views and formulates a key task in the proposed model into a standard quadratic programming problem. (3) It experimentally evaluates the proposed method on both toy/synthetic incomplete multi-view data and real-life incomplete multi-view data. The experimental results show that the proposed method makes considerable improvement over the state-of-the-art baselines.

Before going further, we explain some notational conventions used throughout the paper. We will use boldface capital letters (e.g., $\mathbf{X}$), boldface lowercase letters (e.g., $\mathbf{x}$) and lowercase letters (e.g., x) to denote matrices, vectors and scalars, respectively. Further, $\mathbf{I}$ denotes the identity matrix, and $\mathbf{1}$ denotes a column vector with all the entries as one. For a matrix $\mathbf{X} \in \mathbb{R}^{n_1 \times n_2}$, the j-th column vector and the ij-th entry are denoted by $\mathbf{x_j}$ and x_{ij}, respectively. The trace and the Frobenius norm of $\mathbf{X}$ are denoted by $Tr(\mathbf{X})$ and $\|\mathbf{X}\|_F$, respectively. For a column vector $\mathbf{x} \in \mathbb{R}^{n_1 \times 1}$, the j-th entry is denoted by x_j, and l_p-norm is denoted by $\|\mathbf{x}\|_p$.

2 Preliminaries

In this paper, we build upon the work of [Ng *et al.*, 2001] (denoted as NgSC), which analyzed the spectral clustering algorithm using the top k eigenvectors of the Laplacian matrix of the similarity matrix to partition data. Given a single view data matrix $\mathbf{X} \in \mathbb{R}^{d \times n}$, where d is the dimension of features and n is the number of data instances, NgSC partitions the n data instances into c clusters as follows:

Step 1. Construct the data similarity matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$, where each entry a_{ij} in $\mathbf{A}$ denotes the relationship between $\mathbf{x}_i$ and $\mathbf{x}_j$;

Step 2. Compute the normalized graph Laplacian matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}^T\mathbf{D}^{-1/2}$, where $\mathbf{D}$ is a diagonal matrix whose i-th diagonal element is $\sum_j a_{ij}$;

Step 3. Let $\lambda_1 \geq ... \geq \lambda_k$ be the k largest eigenvalues of $\mathbf{L}$ and $\mathbf{u}_1, ..., \mathbf{u}_k$ denote the corresponding eigenvectors. Normalize all eigenvectors to have unit length and form the matrix $\mathbf{U} = [\mathbf{u}_1, ..., \mathbf{u}_k]$ by stacking the eigenvectors in columns;

Step 4. Form the matrix $\mathbf{Y}$ from $\mathbf{U}$ by normalizing each of $\mathbf{U}$'s rows to have unit length;

Step 5. Treat each row of $\mathbf{Y}$ as a data instance, and partition them using K-means to produce the final clustering results.

To explain why the eigenvectors of spectral clustering can work, [Ng *et al.*, 2001] gave an "ideal" case to explain it according to the following proposition.

Proposition 1. *[Ng* et al.*, 2001]*[1] *Given n data instances with c clusters of sizes $\hat{n}_1, ..., \hat{n}_c$ respectively, let the off-diagonal blocks $\hat{\mathbf{A}}^{(ij)}$ be zero. Also assume that each cluster is connected. Then there exist k ($k = c$) orthogonal vectors $\mathbf{u}_1, ..., \mathbf{u}_k$ ($\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$, 0 otherwise) so that each row of $\hat{\mathbf{Y}}$ satisfies $\hat{\mathbf{y}}_j^{(i)} = \mathbf{u}_i$ for all $i = 1, ..., k$ and $j = 1, ..., n_i$.*

Proposition 1 states that there are k ($k = c$) mutually orthogonal points on the surface of the unit k-sphere around which $\hat{\mathbf{Y}}$'s rows will cluster. These clusters correspond exactly to the true clustering results of the original data.

However, in a general case, the off-diagonal blocks $\mathbf{A}^{(ij)}$ are non-zero. Suppose $\mathbf{E} = \mathbf{A} + \hat{\mathbf{A}}$ as perturbations to the "ideal" $\hat{\mathbf{A}}$ that makes $\mathbf{A} = \hat{\mathbf{A}} + \mathbf{E}$. Earlier results [Hunter and Strohmer, 2010] have shown that small perturbations in the similarity matrix can affect the spectral coordinates and clustering ability. The results are based on the following proposition by [Hunter and Strohmer, 2010].

Proposition 2. *Suppose $\|a_{ij} - \hat{a}_{ij}\| \leq \epsilon$, then*

$$\|\mathbf{A} - \hat{\mathbf{A}}\| \leq n\epsilon.$$

Proposition 2 is a reformulation of the Corollary 10 in [Hunter and Strohmer, 2010]. It is easy to prove this proposition as follows

$$\|\mathbf{A} - \hat{\mathbf{A}}\| = \sqrt{\sum_{ij}(a_{ij} - \hat{a}_{ij})} \leq \sqrt{\sum_{ij}\epsilon^2} = \sqrt{n^2\epsilon^2} = n\epsilon.$$

As can be seen, if n is very large (even ϵ is small), then $n\epsilon$ cannot be ignored. This problem is more acute in multi-view setting because the constructed similarity matrices of different views may vary greatly. Given all the above, in an incomplete multi-view data clustering setting, we ask the following questions:

- How to handle incomplete multi-view data?
- How to find a consensus matrix $\mathbf{Y}^*$ for all views?
- How to make the resulting rows of $\mathbf{Y}^*$ to cluster similarly to the rows of $\hat{\mathbf{Y}}^*$?

The first one is our key question. We will propose our solutions to these questions in the next section.

3 Proposed Method

This section presents the proposed PIC method together with its optimization algorithm.

3.1 Similarity Matrix Generation

Given a set of unlabeled data instances $\{\mathbf{x}_1, ..., \mathbf{x}_n\}$ sampled from m views, let $\mathbf{X}^1, ..., \mathbf{X}^m$ be the data matrices of the m views and $\mathbf{X}^v = \{\mathbf{x}_1^v, ..., \mathbf{x}_{n_v}^v\} \in \mathbb{R}^{d_v \times n_v}$ be the v-th view data matrix, where d_v is the dimension of the features and

[1] Here we denote $\mathbf{A}$ and $\mathbf{Y}$ as $\hat{\mathbf{A}}$ and $\hat{\mathbf{Y}}$ respectively as it is an "ideal" case.

n_v ($n_v \leq n$) is the number of data instances. Like most existing work, we make the assumption that at least one view is available for each data instance in the data matrix. Now we generate each view's similarity matrix from each view's data matrix respectively. As each view is independent in this phase, we take the v-th view as an example.

The intuition here is that if two data instances are close, they should be also close to each other in the similarity graph. Thus, we propose to learn a similarity matrix as follows

$$\min_{\mathbf{A}} \sum_{i,j=1}^{n} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \, a_{ij} + \alpha \sum_{i=1}^{n} \|\mathbf{a}_i\|_2^2 \\ s.t.\ a_{ii} = 0,\ 0 \leq a_{ij} \leq 1,\ \mathbf{1}^T \mathbf{a}_i^v = 1. \tag{1}$$

The above optimization, i.e., Eq. (1), is able to learn a similarity matrix (whose size is $n \times n$ as there are n instances) from a complete data matrix (whose size is also $n \times n$). However, it cannot learn such a similarity matrix from an incomplete data matrix (whose size is not $n \times n$). Our $\mathbf{X}^v$ falls in this case as some instances may be missing in view v, resulting in $n_v \leq n$. To handle missing instances, we define a missing operator on each instance $\mathbf{x}_i$ as below

$$\mathcal{P}_M(\mathbf{x}_i^v) \overset{def}{=} \begin{cases} \mathbf{x}_i^v, & \text{if } \mathbf{x}_i \text{ is sampled in the } v\text{-th view;} \\ NaN, & \text{otherwise;} \end{cases}$$

where NaN denotes "not a number", which can be seen as an invalid number. Based on $\mathcal{P}_M(\mathbf{x}_i^v)|_{i=1}^n$, we formulate our similarity matrix generation task as follows

$$\min_{\mathbf{A}^v} \sum_{i,j=1}^{n} \|\mathcal{P}_M(\mathbf{x}_i^v) - \mathcal{P}_M(\mathbf{x}_j^v)\|_2^2 \, a_{ij}^v + \alpha \sum_{i=1}^{n} \|\mathbf{a}_i^v\|_2^2 \\ s.t.\ a_{ii}^v = 0,\ 0 \leq a_{ij}^v \leq 1,\ \mathbf{1}^T \mathbf{a}_i^v = 1. \tag{2}$$

In such a way, Eq. (2) can learn a similarity matrix $\mathbf{A}^v \in \mathbb{R}^{n \times n}$ with adaptive neighbors for each view. More precisely, it assigns NaN to a_{ij}^v if either $\mathbf{x}_i$ or $\mathbf{x}_j$ is missing in view v; otherwise it assigns a similarity value to a_{ij}^v using the following solution.

We denote $d_{ij}^v = \|\mathcal{P}_M(\mathbf{x}_i^v) - \mathcal{P}_M(\mathbf{x}_j^v)\|_2^2$ and further denote $\mathbf{d}_i^v$ as a vector with j-th element as d_{ij}^v. Here we assign NaN to d_{ij}^v if either $\mathcal{P}_M(\mathbf{x}_i^v) = NaN$ or $\mathcal{P}_M(\mathbf{x}_j^v) = NaN$. Then we rewrite Eq. (2) in a vector form as follows,

$$\min_{\mathbf{a}_i^v} \left\| \mathbf{a}_i^v + \frac{\mathbf{d}_i^v}{2\alpha} \right\|_2^2,\ s.t.\ a_{ii}^v = 0,\ 0 \leq a_{ij}^v \leq 1,\ \mathbf{1}^T \mathbf{a}_i^v = 1. \tag{3}$$

This problem can be solved with a closed form solution as introduced in [Nie *et al.*, 2016]. So, we generate a similarity matrix $\mathbf{A}^v \in \mathbb{R}^{n \times n}$ for each view v ($v = 1, ..., m$). We show that the similarity value of any two data instances (except for missing instances) is a quantity bounded in magnitude $\mathcal{O}(1)$ because we make the constraint $0 \leq a_{ij}^v \leq 1$. This allows for completing those NaNs using average similarity values to reduces to perturbations from missing instances.

3.2 Similarity Matrix Completion

Given the learned similarity matrices $\mathbf{A}^1, ..., \mathbf{A}^m$, we now focus on completing those NaNs in each similarity matrix. Specifically, we complete those NaNs using the average similarity values of the valid view(s). We also take the v-th view as an example. Similar to $\mathcal{P}_M(\mathbf{x}_i^v)$, we define a completion operator on each $\mathbf{a}_i^v$ as

$$\mathcal{P}_\Omega(\mathbf{a}_i^v) \overset{def}{=} \begin{cases} \mathbf{a}_i^v, & \text{if every item in } \mathbf{a}_i^v \text{ is not a } NaN; \\ \mathbf{a}_i^{ave}, & \text{otherwise;} \end{cases}$$

where $\mathbf{a}_i^{ave} = \sum_j \mathbf{a}_i^j / N_v$. Here $\mathbf{a}_i^j$ denotes the similarity value vector in the view j (which has valid similarity value vector for the i-th entry) and N_v is the number of such views.

According to the following theorem, we discuss why our completion scheme is stable and effective.

Theorem 1. *[Candes and Plan, 2010] Let $\mathbf{Z} \in \mathbb{R}^{t_1 \times t_2}$ be a fixed rank matrix with strong incoherence parameter μ. Suppose there are $\hbar$ observed entries of $\mathbf{Z}$ with locations sampled uniformly at random with noise $\|P_\Omega(\mathbf{Z}) - \hat{\mathbf{Z}}\|_F \leq \delta$. Then with high probability, the resulting completion $\hat{\mathbf{Z}}$ obeys*

$$\|\mathbf{Z} - \hat{\mathbf{Z}}\|_F \leq 4\sqrt{\frac{(2+\hbar) min(t_1, t_2)}{\hbar}}\delta + 2\delta.$$

The details of Theorem 1 are introduced in [Candes and Plan, 2010; Hunter and Strohmer, 2010]. The theorem provides an upper bound (which is proportional to the noise level δ) on the recovery error from matrix completion. It states the following: when perfect noiseless recovery occurs, then matrix completion is stable vis-à-vis perturbations. Our $\mathbf{A}^v$ is a special case of $\mathbf{Z}$ with $t_1 = t_2 = n$. As discussed early, similarity value is a quantity bounded in magnitude $\mathcal{O}(1)$, resulting in a small δ. Thus, our completion scheme using the average similarity values makes completion stable and effective.

Here we have proposed our solution to the key question, i.e., how to handle incomplete multi-vew data. Next, we discuss how to find a consensus matrix $\mathbf{Y}^*$ for all views.

3.3 Consensus Learning

Recall the framework of NgSC, see Section 2. $\mathbf{Y}$ is generated from $\mathbf{U}$ by normalizing each of $\mathbf{U}$'s rows to have unit length. $\mathbf{U}$ is formed by the eigenvectors of Laplacian matrix $\mathbf{L}$. As can be seen, the above processes from $\mathbf{L}$ to $\mathbf{Y}$ are simple yet solid but nothing can be changed. Thus, we transfer learning a consensus $\mathbf{Y}^*$ to learning a consensus $\mathbf{L}^*$. Another crucial reason is that the perturbation of spectral clustering is determined by eigenvector of Laplacian matrix [Hunter and Strohmer, 2010]. However, small perturbations in the entries of a Laplacian matrix can lead to large perturbations in the eigenvectors. We will detail this in the next subsection.

Suppose we have computed the normalized Laplacian matrix $\mathbf{L}^v \in \mathbb{R}^{n \times k}$ for each completed similarity matrix $\mathbf{A}^v$ using $\mathbf{L}^v = (\mathbf{D}^v)^{-1/2} (\mathbf{A}^v)^T (\mathbf{D}^v)^{-1/2}$, then we propose to solve our consensus learning task as below

$$\mathbf{L}^* = \sum_{v=1}^{m} \omega_v \mathbf{L}^v \quad s.t.\ \sum_{v=1}^{m} \omega_v = 1, \omega \geq 0 \tag{4}$$

where ω_v is the weight of the v-th view. Note that each $\omega_v|_{v=1}^m$ is determined automatically by reducing perturbation risk among different views, which will be clear shortly.

3.4 Perturbation Risk

Now we respond to the previous subsection and answer the last question, i.e., how to make the resulting rows of $\mathbf{L}^*$ to cluster similarly to the rows of "ideal" $\hat{\mathbf{L}}^*$ as in Proposition 2.

The study in [Hunter and Strohmer, 2010] shows that small perturbations in the entries of Laplacian matrix can lead to large perturbations in the eigenvectors. Matrix perturbation theory [Stewart and Sun, 1990] indicates that the perturbations can be captured by the closeness of the subspaces spanned by the eigenvectors. Let $\mathbf{u}_1^v, ..., \mathbf{u}_k^v$ and $\mathbf{u}_1^*, ..., \mathbf{u}_k^*$ denote the first k eigenvectors of $\mathbf{L}^v$ and $\mathbf{L}^*$, respectively. The subspaces spanned by the eigenvectors $\mathbf{u}_1^v, ..., \mathbf{u}_k^v$ and $\mathbf{u}_1^*, ..., \mathbf{u}_k^*$ are formed as $[\mathbf{u}_1^v, ..., \mathbf{u}_k^v]$ and $[\mathbf{u}_1^*, ..., \mathbf{u}_k^*]$. Following [Stewart and Sun, 1990; Hunter and Strohmer, 2010], we define closeness of these supspaces using canonical angles.

Definition 1. *Let $\gamma_1 \leq ... \leq \gamma_k$ be the singular values of $[\mathbf{u}_1^v, ..., \mathbf{u}_k^v]^T[\mathbf{u}_1^*, ..., \mathbf{u}_k^*]$. Then the values,*

$$\theta_i|_{i=1}^k = \arccos \gamma_i$$

*are called the **canonical angles** between these subspaces.*

The largest canonical angle indicates the perturbation level. Next we make $\mathbf{L}^*$ close to the "ideal" $\hat{\mathbf{L}}^*$ according to the following theorem of canonical angle.

Theorem 2. *[Hunter and Strohmer, 2010] Let $\lambda_i^v, \mathbf{u}_i^v, \lambda_i^*, \mathbf{u}_i^*$ be the i-th eigenvalue and eigenvector of $\mathbf{L}^v$ and $\mathbf{L}^*$ respectively. Let $\mathbf{\Theta} = diag(\theta_1, ..., \theta_k)$ be the diagonal matrix of canonical angles between the subspaces of $\mathbf{U}^v = [\mathbf{u}_1^v, ..., \mathbf{u}_k^v]$ and $\mathbf{U}^* = [\mathbf{u}_1^*, ..., \mathbf{u}_k^*]$. If there is a gap ξ such that*

$$|\lambda_k^v - \lambda_{k+1}^*| \geq \xi \quad and \quad \lambda_k^v \geq \xi$$

then

$$\|\sin \mathbf{\Theta}\|_F \leq \frac{1}{\xi}\|\mathbf{L}^*\mathbf{U}^v - \mathbf{U}^v\mathbf{\Sigma}^v\|_F$$

where $\sin \mathbf{\Theta}$ is taken entry-wise and $\mathbf{\Sigma}^v = diag(\lambda_1^v, ..., \lambda_k^v)$.

This is a reformulation of the Theorem 3 in [Hunter and Strohmer, 2010]. Based on Proposition 2 and Theorem 2, we further give an upper bound of $\sin \mathbf{\Theta}$.

Corollary 1. *With the notations of Theorem 2, suppose $\|l_{ij}^* - l_{ij}^v\| \leq \epsilon$. Then*

$$\|\sin \mathbf{\Theta}\|_F \leq \frac{1}{\xi}\sqrt{k}n\epsilon.$$

Proof. Recall $\|\sin \mathbf{\Theta}\|_F \leq \frac{1}{\xi}\|\mathbf{L}^*\mathbf{U}^v - \mathbf{U}^v\mathbf{\Sigma}^v\|_F$.

As the diagonal elements of $\mathbf{\Sigma}^v$ and the column vectors of $\mathbf{U}^v$ are exactly the first k eigenvalues and eigenvectors of $\mathbf{L}^v$ respectively, we have $\mathbf{L}^v\mathbf{U}^v = \mathbf{U}^v\mathbf{\Sigma}^v$. Then

$$\begin{aligned}\|\sin \mathbf{\Theta}\|_F &\leq \frac{1}{\xi}\|\mathbf{L}^*\mathbf{U}^v - \mathbf{U}^v\mathbf{\Sigma}^v\|_F = \frac{1}{\xi}\|\mathbf{L}^*\mathbf{U}^v - \mathbf{L}^v\mathbf{U}^v\|_F \\ &\leq \frac{1}{\xi}\|\mathbf{U}^v\|_F\|\mathbf{L}^* - \mathbf{L}^v\|_F.\end{aligned}$$

According to the orthogonality of $\mathbf{U}^v$, we have

$$\|\mathbf{U}^v\|_F = \sqrt{Tr((\mathbf{U}^v)^T\mathbf{U}^v)} = \sqrt{k}.$$

Based on Proposition 2, we have $\|\mathbf{L}^* - \mathbf{L}^v\|_F \leq n\epsilon$.

Then, we conclude the proof as

$$\|\sin \mathbf{\Theta}\|_F \leq \frac{1}{\xi}\sqrt{k}n\epsilon.$$

□

Now the key task is to minimize the upper bound of $\sin \mathbf{\Theta}$ as it is equivalent to reduce the perturbation risk. Considering the "ideal" Laplacian matrix $\hat{\mathbf{L}}^*$, we have the following theorem by [Mohar *et al.*, 1991].

Theorem 3. *[Mohar* et al.*, 1991] The multiplicity of the eigenvalue 0 of the Laplacian matrix $\hat{\mathbf{L}}^*$ is exactly equal to the number of clusters c.*

Theorem 3 indicates that the "ideal" Laplacian matrix $\hat{\mathbf{L}}^*$ has c positive eigenvalues and $n - c$ zero eigenvalues. Let $k = c$. As we aim to make our $\mathbf{L}^*$ approximate $\hat{\mathbf{L}}^*$, the rank of $\mathbf{L}^*$ is expected to k, $\lambda_{k+1}^* \approx 0$ and $|\lambda_k^v - \lambda_{k+1}^*| \approx \lambda_k^v$. Thus, given the Laplacian matrix of each view, ξ in Theorem 2 can be seen as a constant. This motivates us to rewrite Eq. (4) into the following objective function to reduce perturbation risk.

$$\begin{aligned}\min_{\mathbf{L}^*,\, \omega} &\sum_{v=1}^m \|\mathbf{L}^*\mathbf{U}^v - \mathbf{U}^v\mathbf{\Sigma}^v\|_F^2 \\ &s.t.\ \mathbf{L}^* = \textstyle\sum_{v=1}^m \omega_v\mathbf{L}^v,\ \sum_{v=1}^m \omega_v = 1,\ \omega \geq 0.\end{aligned} \tag{5}$$

Here another motivation is that views having similar clustering ability should be assigned similar weights. According to Definition 1 and Theorem 2, we conclude that the largest canonical angle between subspaces spanned by the eigenvectors indicates the similarity of the clustering ability. Thus, the difference in weights between the views should be small if the largest canonical angle between corresponding subspaces is small. This is exactly what manifold learning aims to do [Cai *et al.*, 2008]. Let $\psi_{ij} \in [0, \pi]$ be the largest canonical angle between subspaces of the i-th view and the j-th view and $s_{ij} = \pi - \psi_{ij}$. We propose to perform our task using manifold learning as follows

$$\min_{\omega} \frac{1}{2}\sum_{i,j}^m s_{ij}(\omega_i - \omega_j)^2 = \min_{\omega} \omega^T\mathbf{H}\omega \tag{6}$$

where $\mathbf{H} = \breve{\mathbf{D}} - \mathbf{S}$ and $\breve{\mathbf{D}} \in \mathbb{R}^{m\times m}$ is a diagonal matrix with each diagonal element as $\breve{d}_{ii} = \sum_{j=1}^m s_{ij}$.

Plugging the right item of Eq. (6) into Eq. (5), formally, our objective function is formulated as

$$\begin{aligned}\min_{\mathbf{L}^*,\, \omega} &\sum_{v=1}^m \|\mathbf{L}^*\mathbf{U}^v - \mathbf{U}^v\mathbf{\Sigma}^v\|_F^2 + \beta\omega^T\mathbf{H}\omega \\ &s.t.\ \mathbf{L}^* = \textstyle\sum_{v=1}^m \omega_v\mathbf{L}^v,\ \sum_{v=1}^m \omega_v = 1,\ \omega \geq 0\end{aligned} \tag{7}$$

where β is a trade-off parameter. We will analyze β in the experiment section. Given β, here we can rewrite Eq. (7) as

$$\begin{aligned}\min_{\mathbf{L}^*,\omega} &\omega^T(\sum_{v=1}^m \mathbf{Q}^v + \beta\mathbf{I})\omega - 2\omega^T(\sum_{v=1}^m \mathbf{f}^v) \\ &s.t.\ \mathbf{L}^* = \textstyle\sum_{v=1}^m \omega_v\mathbf{L}^v,\ \sum_{v=1}^m \omega_v = 1,\ \omega \geq 0.\end{aligned} \tag{8}$$

Note, each entry q_{ij}^v in $\mathbf{Q}^v$ and each entry f_i^v in $\mathbf{f}^v$ come shortly on the next page. It is easy to see that Eq. (8) is a

Algorithm 1: The proposed overall algorithm.

Input : Data matrices with m views $\mathbf{X}^1, ..., \mathbf{X}^m$, the number of clusters c, and parameter β.

begin

Generate similarity matrix $\mathbf{A}^v$ from each data matrix $\mathbf{X}^v$ by solving Eq. (2);

Complete similarity matrix $\mathbf{A}^v$ using completion operator $\mathcal{P}_\Omega(\mathbf{a}_i^v)|_{i=1}^n$;

Compute normalized Laplacian matrix $\mathbf{L}^v$ of each completed similarity matrix $\mathbf{A}^v$;

Calculate the eigendecomposition $\mathbf{U}^v$ and $\mathbf{\Sigma}^v$ of each normalized Laplacian matrix $\mathbf{L}^v$;

Calculate ω by solving Eq. (8);

Calculate consensus Laplacian matrix $\mathbf{L}^*$ using Eq. (4);

Produce the final clustering results by performing spectral clustering algorithm (e.g., NgSC) on the learned consensus Laplacian matrix $\mathbf{L}^*$;

end

Output: The clustering results with c clusters.

standard quadratic programming problem with respect to ω and can be solved by a classic technique, e.g., the technique called *quadprog* in MATLAB. We used MATLAB because all baselines used it.

In Eq. (8), each entry q_{ij}^v in $\mathbf{Q}^v$ and each entry f_i^v in $\mathbf{f}^v$ is respectively defined as

$$q_{ij}^v = Tr(\mathbf{L}^i\mathbf{U}^v(\mathbf{U}^v)^T(\mathbf{L}^v)^T), f_i^v = Tr(\mathbf{L}^i\mathbf{U}^v(\mathbf{\Sigma}^v)^T(\mathbf{U}^v)^T).$$

Hereto, we presented the proposed PIC approach with four tasks. We now couple overall solutions into a joint framework and optimize the joint framework using Algorithm 1. The convergence of our algorithm involves two parts, i.e., generating similarity matrix $\mathbf{A}^v|_{v=1}^m$ using Eq. (2) and calculating the weights ω using Eq. (8). Eq. (2) is clearly a convex function as its second order derivative w.r.t. $\mathbf{a}_i^v$ is a positive value. Eq. (8) is a standard quadratic programming problem w.r.t. the weights ω. Thus, the convergence of the proposed algorithm is guaranteed. Compared to single view spectral clustering algorithm, PIC needs to optimize Eq. (8). The computational complexity of optimizing Eq. (8) is $O(m^3n^2k + m^3)$ in total, where $m \ll n$ and $k \ll n$. Thus, PIC does not increase the computational complexity of spectral clustering, i.e., $O(n^3)$. For large-scale data, data sampling as in [Cai and Chen, 2015; Li *et al.*, 2015] is a potential way to speed up our method. In addition, the proposed algorithm can be implemented with data on disk as it runs without iterative optimization. Thus, our algorithm can be deployed on a small memory machine.

4 Experiments

4.1 Datasets and Baselines

Datasets. We perform evaluation using four complete multi-view datasets and three natural incomplete multi-view datasets. The datasets are summarized in Table 1, where the first four datasets are complete and the last three datasets are naturally incomplete. For the dataset Mfeat, we collected it from two Handwritten Digits sources, i.e., MNIST and USPS.

Dataset	m	c	n	n_v $(v = 1, ..., m)$	d_v $(v = 1, ..., m)$
100Leaves [2]	3	100	1600	1600, 1600, 1600	64, 64, 64
Flowers17 [3]	7	17	1360	1360, 1360, ..., 1360	1360, 1360, ..., 1360
Mfeat [4]	2	10	10000	10000, 10000	784, 256
ORL [5]	4	40	400	400, 400, 400, 400	256, 256, 256, 256
3Sources [6]	3	6	416	352, 302, 294	3560, 3631, 3068
BBC [7]	4	5	2225	1543, 1524, 1574, 1549	4659, 4633, 4665, 4684
BBCSport [7]	2	5	737	644, 637	3183, 3203

Table 1: Summary of the datasets. $\{m, c, n, n, n_v, d_v\}$ denotes the number of {views, clusters, instances, observed instances, features} in each view, respectively.

Baselines. We consider the following algorithms as the baselines: **BSV** [8] (Best Single View) [Ng *et al.*, 2001], **PVC** [Li *et al.*, 2014], **IMG** [Zhao *et al.*, 2016], **MIC** [Shao *et al.*, 2015] and **DAIMC** [Hu and Chen, 2018]. Note that BSV only works for complete single view data. Following [Shao *et al.*, 2015; Zhao *et al.*, 2016], we first fill the missing instance in each incomplete view using the average feature values of that incomplete view. PVC and IMG only work for two-view data. Following [Hu and Chen, 2018], we evaluate PVC and IMG on all two-view combinations and report the best results. Since DAIMC works for multi-view data, we use it as it is.

4.2 Experimental Settings and Results

Experimental Settings

To generate incomplete multi-view datasets from complete multi-view datasets, we follow all the above baselines and use a general setting. Same as the baselines, we first randomly select partial examples under different Partial Example Ratio (PER). Then we additionally generate a random binary vector $\mathbf{b} = (b_1, ..., b_m)$ for each partial example (e.g., $\mathbf{x}_j$). If $b_i = 0$, we delete/remove example $\mathbf{x}_j$ from view i.

For the baselines, we obtained the original systems from their authors and used their default parameter settings. For PIC, we set the parameter β using $\beta = \tilde{\beta} \times \|\sum_v \mathbf{Q}^v\|_F / \|\mathbf{I}\|_F$ to balance Eq. 5 and Eq. 6. Then we empirically set $\tilde{\beta} = 0.1$ in evaluation. The parameter study will come shortly.

Following the baselines, two metrics, accuracy (ACC) and normalized mutual information (NMI) are used to measure the clustering performance. In order to randomize the experiments, we run each algorithm 20 times and report the average values of the performance measures.

Experimental Results

We first perform evaluation using four toy incomplete multi-view datasets (generated from complete multi-view datasets as introduced above). In this experiment, PER varies from 0.1 to 0.9 with an interval of 0.2, same as baselines PVC and IMG. We also set PER = 0, i.e., every data instance is sampled

[2] https://archive.ics.uci.edu/ml/datasets/One-hundred+plant+species+leaves+data+set

[3] http://www.robots.ox.ac.uk/~vgg/data/flowers/17/index.html

[4] https://cs.nyu.edu/~roweis/data.html

[5] www.cad.zju.edu.cn/home/dengcai/Data/FaceData.html

[6] http://mlg.ucd.ie/datasets/3sources.html

[7] http://mlg.ucd.ie/datasets/segment.html

Method	Clustering performance in terms of ACC						Clustering performance in terms of NMI					
	BSV	PVC	IMG	MIC	DAIMC	PIC	BSV	PVC	IMG	MIC	DAIMC	PIC
3Sources	22.50±0.63	26.45±0.39	25.59±0.13	43.73±6.28	58.68±8.12	**88.08±1.22**	5.30±0.27	1.77±0.25	2.00±0.12	38.94±5.58	48.80±7.27	**73.50±1.45**
BBC	40.79±1.02	37.80±0.97	29.92±0.01	57.07±9.74	51.34±7.44	**87.03±0.05**	25.99±2.33	14.72±0.11	6.24±0.01	39.19±6.54	37.74±7.23	**70.12±0.02**
BBCSport	40.99±0.81	44.33±1.46	37.86±0.07	58.66±9.39	75.16±8.82	**76.02±5.28**	26.60±0.48	13.77±1.67	7.55±0.03	46.26±6.74	57.80±9.53	**75.12±2.05**

Table 2: Clustering performance on three natural incomplete multi-view datasets.

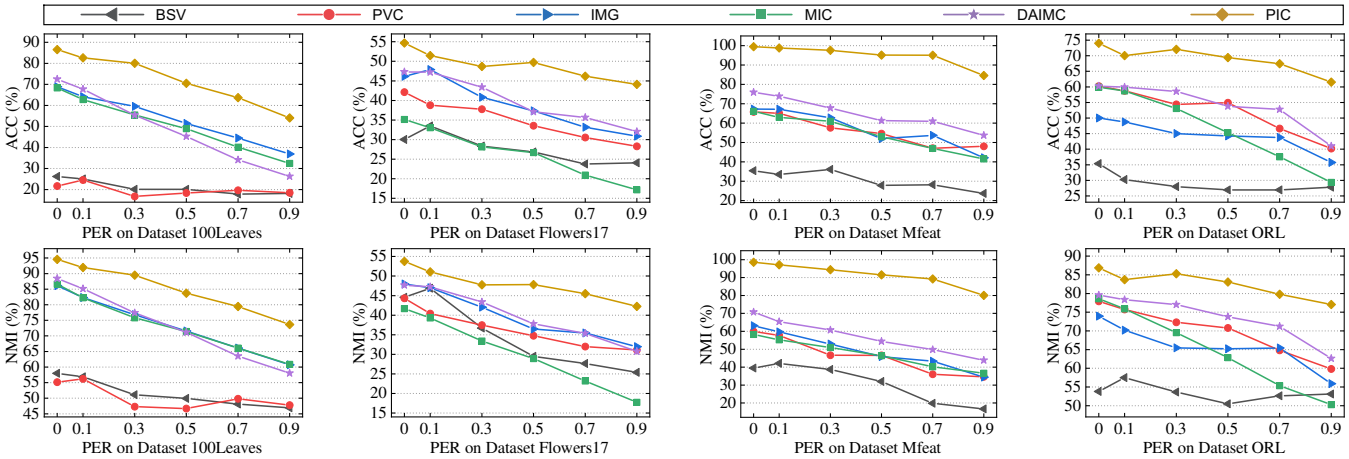

Figure 1: Clustering performance results with different PER settings on four toy incomplete multi-view datasets. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

in all views. Figure 1 shows the performance results in terms of ACC and NMI. From Figure 1, we make the following observations:

- Our PIC significantly outperforms the baselines in all the PER settings. As the PER increases, the clustering performance of all the methods drops.
- All baselines are inferior to our model. One reason is that they complete the missing instances with the average feature values, which results in a large deviation, especially when the PER is large. In this work, we transfer feature missing to similarity missing, then complete the missing similarity entries (marked with NaN) using the average similarity values of all the valid views. This shows that our completion scheme is powerful in dealing with missing instances.
- The recent baseline DAIMC performs better than other baselines, but worse than our method. This shows that our method presents a new margin to beat.

We then perform evaluation using three natural incomplete multi-view data. The results, i.e., the average values and the standard deviations (denoted as ave±std), are shown in Table 2. From the table, we make the following observations:

- Our PIC again outperforms the baselines markedly. PIC achieves the best ACC and NMI on each dataset. The results clearly show that our PIC is a promising incomplete multi-view clustering method.
- Multi-view clustering methods (MIC, DAIMC and PIC) are superior to two-view clustering methods (PVC and IMG), but two-view clustering methods (PVC and IMG) are not always superior to the single-view clustering method (BSV). All of them are inferior to our PIC. This indicates that multi-view data boost clustering results with multi-view clustering techniques.

4.3 Parameter Study

In the above experiments, parameter $\tilde{\beta}$ is set to 0.1 for PIC. Here we explore the effect of parameter $\tilde{\beta}$. Due to the space limit, we only show the results on three natural incomplete multi-view data. Figure 2 shows how the average performance of PIC varies with different $\tilde{\beta}$ values. From Figure 2, we can see that PIC achieves consistently good performance when $\tilde{\beta}$ is around 0.1 (i.e., 1e-1) on three datasets. As introduced early, we use a balance scheme, i.e., $\beta = \tilde{\beta} \times \|\sum_v \mathbf{Q}^v\|_F / \|\mathbf{I}\|_F$. This is the reason why we can use the same parameter $\tilde{\beta}$ (i.e., 0.1) for all datasets.

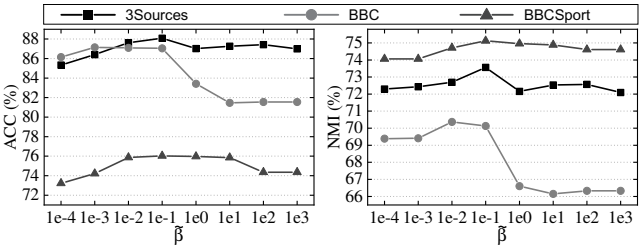

Figure 2: Parameter study on three natural incomplete datasets.

5 Conclusions

This paper built a bridge between spectral perturbation and incomplete multi-view clustering. We explored spectral perturbation theory and proposed a novel Perturbation-oriented Incomplete multi-view Clustering (PIC) method. The key idea is to transfer the missing problem from data matrix to similarity matrix and reduce the spectral perturbation risk among different views while balancing all views to learn a consensus representation for the final clustering results. Both theoretical results and experimental results showed the effectiveness of the proposed method.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (No. 61572407), and the Project of National Science and Technology Support Program (No. 2015BAH19F02). Hao Wang would like to thank the China Scholarship Council (No. 201707000064).

References

[Cai and Chen, 2015] Deng Cai and Xinlei Chen. Large scale spectral clustering via landmark-based sparse representation. *IEEE Trans. Cybern.*, 45(8):1669–1680, 2015.

[Cai *et al.*, 2008] Deng Cai, Xiaofei He, Xiaoyun Wu, and Jiawei Han. Non-negative matrix factorization on manifold. In *ICDM*, pages 63–72, 2008.

[Cai *et al.*, 2018] Yang Cai, Yuanyuan Jiao, Wenzhang Zhuge, Hong Tao, and Chenping Hou. Partial multi-view spectral clustering. *Neurocomputing*, 311:316–324, 2018.

[Candes and Plan, 2010] Emmanuel J Candes and Yaniv Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[Chao *et al.*, 2017] Guoqing Chao, Shiliang Sun, and Jinbo Bi. A survey on multi-view clustering. *arXiv preprint arXiv:1712.06246*, 2017.

[Hu and Chen, 2018] Menglei Hu and Songcan Chen. Doubly aligned incomplete multi-view clustering. In *IJCAI*, pages 2262–2268, 2018.

[Huang *et al.*, 2019] Shudong Huang, Zhao Kang, Ivor W Tsang, and Zenglin Xu. Auto-weighted multi-view clustering via kernelized graph learning. *Pattern Recognit.*, 88:174–184, 2019.

[Hunter and Strohmer, 2010] Blake Hunter and Thomas Strohmer. Performance analysis of spectral clustering on compressed, incomplete and inaccurate measurements. *arXiv preprint arXiv:1011.0997*, 2010.

[Li *et al.*, 2014] Shao-Yuan Li, Yuan Jiang, and Zhi-Hua Zhou. Partial multi-view clustering. In *AAAI*, pages 1968–1974, 2014.

[Li *et al.*, 2015] Yeqing Li, Feiping Nie, Heng Huang, and Junzhou Huang. Large-scale multi-view spectral clustering via bipartite graph. In *AAAI*, pages 2750–2756, 2015.

[Liu *et al.*, 2018] Xinwang Liu, Xinzhong Zhu, Miaomiao Li, Lei Wang, Chang Tang, Jianping Yin, Dinggang Shen, Huaimin Wang, and Wen Gao. Late fusion incomplete multi-view clustering. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1–14, 2018.

[Mohar *et al.*, 1991] Bojan Mohar, Y Alavi, G Chartrand, and OR Oellermann. The Laplacian spectrum of graphs. *Graph Theory, Combinatorics, and Applications*, 2(12):871–898, 1991.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, pages 849–856, 2001.

[Nie *et al.*, 2016] Feiping Nie, Xiaoqian Wang, Michael I. Jordan, and Heng Huang. The constrained Laplacian rank algorithm for graph-based clustering. In *AAAI*, pages 1969–1976, 2016.

[Nie *et al.*, 2018] Feiping Nie, Lai Tian, and Xuelong Li. Multiview clustering via adaptively weighted procrustes. In *KDD*, pages 2022–2030, 2018.

[Shao *et al.*, 2015] Weixiang Shao, Lifang He, and Philip S. Yu. Multiple incomplete views clustering via weighted nonnegative matrix factorization with $l_{2,1}$ regularization. In *ECML-PKDD*, pages 318–334, 2015.

[Stewart and Sun, 1990] G. W. Stewart and Jiguang Sun. *Matrix perturbation theory*. Academic Press, 1990.

[Tao *et al.*, 2018] Hong Tao, Chenping Hou, Xinwang Liu, Tongliang Liu, Dongyun Yi, and Jubo Zhu. Reliable multi-view clustering. In *AAAI*, pages 4123–4130, 2018.

[Wang *et al.*, 2018] Qianqian Wang, Zhengming Ding, Zhiqiang Tao, Quanxue Gao, and Yun Fu. Partial multi-view clustering via consistent GAN. In *ICDM*, pages 1290–1295, 2018.

[Wang *et al.*, 2019] Hao Wang, Yan Yang, and Bing Liu. Gmc: Graph-based multi-view clustering. *IEEE Trans. Knowl. Data Eng.*, pages 1–14, 2019.

[Wen *et al.*, 2018] Jie Wen, Yong Xu, and Hong Liu. Incomplete multiview spectral clustering with adaptive graph learning. *IEEE Trans. Cybern.*, pages 1–12, 2018.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view learning with incomplete views. *IEEE Trans Image Process.*, 24(12):5812–5825, 2015.

[Yang and Wang, 2018] Yan Yang and Hao Wang. Multi-view clustering: A survey. *Big Data Mining and Analytics*, 1(2):83–107, 2018.

[Yin *et al.*, 2017] Qiyue Yin, Shu Wu, and Liang Wang. Unified subspace learning for incomplete and unlabeled multi-view data. *Pattern Recognit.*, 67:313–327, 2017.

[Zhao *et al.*, 2016] Handong Zhao, Hongfu Liu, and Yun Fu. Incomplete multi-modal visual data grouping. In *IJCAI*, pages 2392–2398, 2016.

[Zhao *et al.*, 2018] Liang Zhao, Zhikui Chen, Yi Yang, Z Jane Wang, and Victor CM Leung. Incomplete multi-view clustering via deep semantic mapping. *Neurocomputing*, 275:1053–1062, 2018.

[Zhou *et al.*, 2019] Wei Zhou, Hao Wang, and Yan Yang. Consensus graph learning for incomplete multi-view clustering. In *PAKDD*, pages 529–540, 2019.

[Zhu *et al.*, 2018] Xinzhong Zhu, Xinwang Liu, Miaomiao Li, En Zhu, Li Liu, Zhiping Cai, Jianping Yin, and Wen Gao. Localized incomplete multiple kernel k-means. In *IJCAI*, pages 3271–3277, 2018.

[Zong *et al.*, 2018] Linlin Zong, Xianchao Zhang, Xinyue Liu, and Hong Yu. Weighted multi-view spectral clustering based on spectral perturbation. In *AAAI*, pages 4621–4628, 2018.