Accepted Manuscript

Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering

Yang Wang, Lin Wu

Please cite this article as: Wang, Y., Wu, L., Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. *Neural Networks* (2018), https://doi.org/10.1016/j.neunet.2018.03.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Beyond Low-Rank Representations: Orthogonal Clustering Basis Reconstruction with Optimized Graph Structure for Multi-view Spectral Clustering

Yang Wang[a,b], Lin Wu[c]

[a] *Dalian University of Technology, Dalian, China*
[b] *The University of New South Wales, NSW 2052, Australia*
[c] *The University of Queensland, Queensland 4072, Australia*

Abstract

Low-Rank Representation (LRR) is arguably one of the most powerful paradigms for Multi-view spectral clustering, which elegantly encodes the multi-view local graph/manifold structures into an intrinsic low-rank self-expressive data similarity embedded in high-dimensional space, to yield a better graph partition than their single-view counterparts. In this paper we revisit it with a fundamentally different perspective by discovering LRR as essentially a latent clustered orthogonal projection based representation winged with an optimized local graph structure for spectral clustering; each column of the representation is fundamentally a cluster basis orthogonal to others to indicate its members, which intuitively projects the view-specific feature representation to be the one spanned by all orthogonal basis to characterize the cluster structures. Upon this finding, we propose our technique with the followings: (1) We decompose LRR into latent clustered orthogonal representation via low-rank matrix factorization, to encode the more flexible cluster structures than LRR over primal data objects; (2) We convert the problem of LRR into that of simultaneously learning orthogonal clustered representation and optimized local graph structure for each view; (3) The learned orthogonal clustered representations and local graph structures enjoy the same magnitude for multi-view, so that the ideal multi-view

Email addresses: yeungwangresearch@gmail.com, wangy@cse.unsw.edu.au (Yang Wang), lin.wu@uq.edu.au (Lin Wu)

consensus can be readily achieved. The experiments over multi-view datasets validate its superiority, especially over recent state-of-the-art LRR models.

1. Introduction

Spectral clustering [1], which partitions the data objects via their local graph/manifold structure relying on the Laplacian eigenvalue-eigenvector decomposition, is one fundamental clustering problem. Unlike K-Means clustering [2], the data objects within the same group characterize not only the large data similarity but also the similar local graph/manifold structure. With the rapid development of information technology, the data are largely available with the multi-view feature representations (*e.g.,* images can be featured by a color histogram view or a texture view), which naturally paves the way to multi-view spectral clustering. As extensively claimed by the multi-view research [3, 4, 5, 6], the information encoded by multi-view features describe different properties; thus leveraging the multi-view information can outperform the single-view counterparts. One critical issue on a successful multi-view incorporation implied by the existing work [7, 8, 9, 10, 11, 12], lies in how to achieve the multi-view consensus/agreement.

Following such principle, a lot of multi-view clustering methods [13, 14] claim that similar data objects should be within the same group across all views. Based on that, the consensus multi-view local manifold structure is further explored with great efforts [15, 16, 17, 8] for multi-view spectral clustering. Among all these methods, Low-Rank Representation (LRR) [18] coupled with sparse decomposition based model has been emerged as a substantially elegant solution, due to its strength of exploring their intrinsic low-dimensional manifold structure encoded by the data correlations embedded in high-dimensional space, while exhibiting strong robustness to feature noise corruptions addressed by sparse noise modeling, hence attracting great attention.

2

Before proceeding further, some notations that are used throughout the paper are shown below.

1.1. Notations

For Matrix M, the trace of M is denoted as $Tr(M)$; $||M||_F = \sqrt{\sum_{i,j} M_{i,j}^2}$ (or $||\cdot||_2$ for vector space) denotes the Frobenius norm; $||M||_1(\sum_{i,j} |M_{i,j}|)$ is the ℓ_1 norm, and M^T denotes the transpose of M, and its unclear norm as $||M||_*$ (sum of all singular values); $M(i,\cdot)$ and $M(\cdot,i)$ as the ith row and column of M. $M \succeq 0$ means all entries of M are nonnegative. I is the identity matrix with adaptive size. $\mathbf{1}$ indicates the vector of adaptive length with all entries to be 1. $|\cdot|$ indicates the cardinality of the set.

1.2. Motivation: LRR Revisited for Multi-View Spectral Clustering

Specifically, the typical LRR model for multi-view spectral clustering stems from the formulation below:

$$
\min_{Z,E_i} ||Z||_* + \lambda \sum_{i \in V} ||E_i||_1
$$
$$
\text{s.t.} \quad X_i = X_i Z + E_i, \quad i \in V, \quad Z \succeq 0, \tag{1}
$$

where $X_i \in \mathbb{R}^{d_i \times n}$ is the data representation for the ith view with d_i as its feature dimension, n as the number of data objects identical for each view, λ is the balance parameter, and V is the view set. $Z \in \mathbb{R}^{n \times n}$ is the self-expressive low-rank similarity representation shared by all $|V|$ views, constrained with $||Z||_*$ based on $X_i(i \in V)$, which can also be substituted by the other specific dictionaries; $||E_i||_1$ is modeled to address the noise-corruption for the ith view-specific feature representation. $Z \succeq 0$ ensures the nonnegativity for all its entries. Based on such optimized low-rank Z, the spectral clustering is finally conducted. One significant limitation of Eq.(1) pointed out by [16] is that, *only one common Z is learned to* *preserve the flexible local manifold structures for all views*, hence fails to achieve the ideal spectral clustering result. To this end, various low-rank Z_i are learned to preserve the ith view-specific local manifold structures, meanwhile minimize their divergence via an iterative-views-agreement strategy for

3

multi-view consensus, followed by a final spectral clustering stage. Despite its encouraging performance, the following standout limitations are inattentively overlooked for LRR model: **(1)** The low-rank data similarity may not well encode the flexible latent cluster structures over primal view-specific feature space; worse still for the non-ideal local graph construction over such representation for spectral clustering; **(2)** The low-rank data similarities coming from multi-views may not be within the same magnitude, so that the divergence minimization may not achieve the ideal multi-view clustering consensus.

Our new perspective. The above facts motivate us to revisit the low-rank representation Z_i to help $X_i Z_i$ reconstruct X_i below for the ith view

$$\min_{Z_i \in \mathbb{S}} ||X_i - X_i Z_i||_F^2, \tag{2}$$

where $\mathbb{S}$ denotes the set of $Z_i \in \mathbb{R}^{n \times n}$ with low-rankness *e.g.,* cluster number c far less than d_i; **Instead of narrowing Low-Rank Z_i as self-expressive data similarity from the conventional viewpoint, it is essentially seen as a special case of a generalized Low-Rank projection, to map feature representation to a low-dimensional space to reconstruct X_i with minimum error.** As discussed, the self-expressive similarity projection equipped with LRR models still suffer from the aforementioned non-trivial limitations.

Here we ask a question: Is there a superior low-rank projection Z_i to minimize Eq.(2), meanwhile address the limitations over the existing LRR models. Our answer to this question is positive. Specifically, we propose to consider Z_i as a latent clustered orthogonal projection, via $Z_i = U_i U_i^T$, where

1. **Clustered orthogonal projection**: $U_i \in \mathbb{R}^{n \times c}$, where each column indicates one cluster to characterize its belonging data objects. Compared with LRR over original feature space, the latent factor U_i can better preserve the flexible latent cluster structure.

2. **Feature reconstruction with cluster basis**: Instead of low-rank data similarity, Z_i essentially serves as a mapping to reconstruct the view-

4

specific features via the column of U_i to encode the latent cluster structures.

3. **Rethinking** $X_i Z_i$: We revisit the intuition of $X_i Z_i$ via $(X_i U_i) U_i^T$ throughout two stages, remind that $X_i \in \mathbb{R}^{d_i \times n}$ where

   - $X_i U_i$ is performed to obtain the new projection value for all d_i features over c orthogonal columns of U_i;

   - $X_i U_i U_i^T$ is subsequently the projected representation for all d_i features spanned by c clustered orthogonal column basis of U_i.

4. **Same magnitude for multi-view consensus**: All $U_i (i \in V)$ enjoy the same magnitude due to their orthonormal columns. Hence, the feasible divergence minimization will facilitate the multi-view consensus.

Before shedding light on our technique, we review the typical related work for multi-view spectral clustering

1.3. Prior Arts

The prior arts can be classified as per the strategy at which the multi-view fusion takes place for spectral clustering.

The most straightforward method goes to the *Early fusion* [19] by concatenating the multi-view feature vectors with equal or varied weights into an unified one, followed by the spectral clustering over such unified space. However, such method ignores the statistical property belonging to an individual view. *Late fusion* [20] may address the limitation to some extents by aggregating the spectral clustering result from each individual view, which follows the assumption that all views are independent to each other. Such assumption is not effective for multi-view spectral clustering as they assume the views to be dependent so that the multi-view consensus information can be exploited for promising performance.

Canonical Correlation Analysis (CCA) is applied for multi-view spectral clustering [21] by learning a common low-dimensional representations for all views, upon which the spectral clustering is performed. One salient drawback

5

lies in the failure of preserving the flexible local manifold structures for different views via such common subspace. *Co-training* based model [8] learned the Laplacian eigenmap for each view over its projected data representation throughout the laplacian eigenmaps from other views, such process repeated till the convergence, the final similarity are then aggregated for spectral clustering. A similar method [17] is also proposed to coordinate multi-view laplacian eigenmaps consensus for spectral clustering. Despite their effectiveness, they have to follow the scenario of noise free for the feature representations. Unfortunately, it cannot be met in practice. The Low-Rank Representation and sparse decomposition models [16, 15] well tackle the problem, meanwhile exhibits the robustness to feature noise corruptions. However, they still suffer from the aforementioned limitations. To this end, we make the following orthogonal contributions to typical LRR model for multi-view spectral clustering.

1.4. Our Contributions

- We revisit the classical Low-Rank Representation (LRR) for multi-view spectral clustering with a fundamentally novel viewpoint of finding it as essentially the latent clustered orthogonal projection based representation with optimized graph structure, to better encode the flexible latent cluster structures than LRR over primal data objects.

- We convert the problem of learning LRR into that of simultaneously learning the clustered orthogonal representation and its optimized local graph structure for each view, rather than directly rely on the local graph construction over original data objects.

- The learned multi-view latent clustered representations and local graph structures enjoy the same magnitude, so as to facilitate a feasible divergence minimization to achieve superior multi-view consensus for spectral clustering.

Extensive experiments over multi-view datasets validate the superiority of our method.

6

2. Learning Clustered Orthogonal projection with Optimized Graph Structure

In this section, we formally discuss our technique. We get started by revisiting the conventional Low-rank representation model by Eq.(1) of yielding a low-rank Z_i, and revisit that from a novel perspective of clustered orthogonal projection with optimized graph structure.

2.1. Problem Formulation

As previously defined in section 1.2, X_i is the data representation for the ith view. $Z_i \in \mathbb{R}^{n \times n}$ is low-rank data similarity representation for X_i. The Eq.(2) is equivalent to computing $U_i \in \mathbb{R}^{n \times c}$ such that $Z_i = U_i U_i^T$, where U_i has orthonormal columns with its gth column representing the relevance each data object belongs to the gth cluster, and c indicates latent cluster number. We then arrive at the following

$$\min_{U_i} ||X_i - X_i U_i U_i^T||_F^2 \tag{3}$$

As discussed in section 1.2, $X_i U_i U_i^T$ reveals the new projection representation for all d_i features spanned by the orthogonal basis of U_i to reconstruct X_i.

Optimizing Eq.(3) w.r.t. U_i is equivalent to computing the principle components of X_i to constitute the orthogonal columns of U_i using the principle eigenvectors. Inspired by this, we exploit the latent cluster structures of X_i to form non-overlapping clusters with each characterized by one orthogonal column basis of U_i, which is equivalent to exploit the $||Z_i||_*$ in Eq.(1), under the condition of $Z_i = U_i U_i^T$. Before that, we revisit one nice property of low-rank matrix factorization. Thanks to [22] on low-rank matrix factorization formulated as

$$||Z_i||_* = \min_{U_i, V_i, Z_i = U_i V_i^T} \frac{1}{2}(||U_i||_F^2 + ||V_i||_F^2), \tag{4}$$

where $U_i \in \mathbb{R}^{n \times c}$ and $V_i \in \mathbb{R}^{n \times c}$ are latent factors from Z_i. Based on that, we approximate $||Z_i||_*$ via the clustered orthogonal projection factorization $Z_i = U_i U_i^T$, and convert the problem Eq.(1) of minimizing $||Z_i||_*$ to that of learning

7

clustered projection representation U_i below

$$||Z_i||_* = \min_{U_i, Z_i = U_i U_i^T} ||U_i||_F^2 \qquad (5)$$

Another issue for spectral clustering is the local data manifold structure modeling, we discuss that by the following remark.

Remark 1. *Unlike the data similarity over raw data objects, U_i via the low-rank matrix factorization can achieve the flexible latent cluster structures. Another crucial issue left to be addressed lies in its local manifold/graph structure modeling over U_i, which is crucial for spectral clustering.* **One may directly refer to the local graph construction over X_i. However, as previously stated, it cannot effectively encode the local graph structure over U_i.**

Towards this end, we propose to learn an optimized local graph structure over U_i by solving the following

$$\frac{1}{2} \min_{\forall j \sum_k^n W_i(j,k)=1, W_i \succeq 0} \sum_{j,k}^n ||U_i(j,\cdot) - U_i(k,\cdot)||_2^2 W_i(j,k) \qquad (6)$$
$$= Tr(U_i^T L_i U_i),$$

where $L_i = B_i - \frac{W_i^T + W_i}{2}$ is Laplacian matrix, B_i is the diagonal matrix with its gth diagonal entry equaled to the sum of the gth row of $\frac{W_i^T + W_i}{2}$. The ideal W_i reveals the probability of jth and k^{th} data points within the same cluster according to cluster projection representation U_i. We impose the constraint that $\forall j, \sum_k^n W_i(j,k) = 1$ and $W_i(j,\cdot) \succeq 0$ to meet the probability nature of W_i. Following [23], we will impose the regularization $||W_i||_F^2$ to avoid that only the nearest neighbor of each data point is assigned 1 with others 0.

With all the above collected, beyond conventional Low-rank Representation model as Eq.(1), we finally propose a novel formulation of clustered orthogonal representation with optimized local graph/manifold structure for multi-view spectral clustering as Eq.(7)

$$\min_{U_i, E_i, W_i(i \in V)} \sum_{i \in V} (\underbrace{||E_i||_1}_{\text{sparse noise modeling}} + \underbrace{\lambda_1 ||W_i||_F^2}_{\text{regularized graph structure}} \qquad (7)$$
$$+ \underbrace{\frac{\lambda_2}{2} Tr(U_i^T L_i U_i)}_{\text{structuring } U_i \text{ with optimized local manifold structure}}$$

$$+ \qquad \underbrace{\frac{\beta}{2} \sum_{j \in V, j \neq i} ||U_i - U_j||_F^2}_{\text{modeling the multi-view consensus over the } U_i \text{ within the same magnitude}} \qquad)$$

$$\text{s.t.} \quad i = 1, \ldots, V, \quad X_i = D_i U_i^T + E_i, D_i = X_i U_i$$

$$U_i^T U_i = I, G_i = U_i, G_i \geq 0, \forall j \sum_{k=1}^{n} W_i(j,k) = 1; W_i \succeq 0,$$

where $||U_i||_F^2$ is omitted due to constraint $U_i^T U_i = I$; all the $U_i \in \mathbb{R}^{n \times c} (i \in V)$ share the same cluster number c for multi-view clustering consensus. λ_1, λ_2 and β are non-negative weights related to learning the clustered orthogonal representation, its local graph structure and multi-view consensus modeling, and will be studied in Section 5. The constraint $U_i^T U_i = I$ ensures the orthonormal columns of U_i.

Remark 2. *We introduce two auxiliary variables $G_i = U_i$ and $D_i = X_i U_i \in \mathbb{R}^{d_i \times c}$. As will be shown later, the intuition of introducing D_i lies in minimizing $||X_i - D_i U_i^T - E_i||_F^2$ w.r.t. D_i, where*

- *it is similar as dictionary learning, while popping up U_i^T as the corresponding sparse representation learning; moreover, it also enjoys the optimization of the isolated U_i^T after merging the other U_i into D_i.*

3. Optimization

Solving Eq.(7) is equivalent to be a unified process of simultaneously learning U_i and W_i for the $i^{th}(i \in V)$ view. As will be shown later, learning either of them will promote the other. Optimizing Eq.(7) is not jointly convex to U_i, W_i and E_i, we hence alternately optimize each of them with the others fixed. Following [24], we deploy the Augmented Lagrange Multiplier (ALM) together with Alternating Direction Minimization (ADM) strategy, which is widely known as an effective and efficient solver. As the optimization process for the above variables within each view is similar, we only present the optimization process for the ith view, the same process holds for other $j^{th}(j \neq i)$ views. The augmented lagrangian

9

function can be written below

$$\min_{\forall j \sum_{k=1}^n W_i(j,k)=1, 0 \leq W_i(j,k) \leq 1, U_i^T U_i = I} \mathcal{L}(U_i, E_i, D_i, G_i, W_i) \qquad (8)$$

$$= ||E_i||_1 + \frac{\lambda_2}{2} Tr(U_i^T L_i U_i) + \lambda_1 ||W_i||_F^2$$

$$+ \frac{\beta}{2} \sum_{j \in V, j \neq i} ||U_i - U_j||_F^2 + \Phi(K_1^i, X_i - D_i U_i^T - E_i)$$

$$+ \Phi(K_2^i, U_i - G_i) + \Phi(K_3^i, D_i - X_i U_i)$$

$$+ \Phi(\mu, ||X_i - D_i U_i - E_i||_F^2 + ||U_i - G_i||_F^2 + ||D_i - X_i U_i||_F^2),$$

where $K_1^i \in \mathbb{R}^{d_i \times n}$, $K_2^i \in \mathbb{R}^{n \times c}$, and $K_3^i \in \mathbb{R}^{d_i \times c}$ are Lagrange multipliers. $\Phi(\cdot, \cdot)$ indicates element-wise multiplication. $\mu > 0$ is a penalty parameter.

Solving U_i: We calculate the partial derivative of Eq.(8) w.r.t. U_i, $\frac{\partial \mathcal{L}}{\partial U_i}$ to be $\mathbf{0} \in \mathbb{R}^{n \times c}$, while fixing others to be constant. After rearranging the terms, it has

$$U_i = \underbrace{(\lambda_2 L_i + (\mu + \beta(|V| - 1))I - \mu X_i^T X_i)^{-1}}_{\text{With } \mathcal{O}(n^3) \text{ computational complexity}} S, \qquad (9)$$

where

$$S = \sum_{j \in V, j \neq i} U_j + ((K_1^i)^T - \mu U_i D_i^T - \mu E_i^T)D_i$$

$$+ X_i^T K_3^i + \mu X_i^T X_i U_i$$

Efficient Row updating strategy of U_i. As shown in Eq.(9), the bottleneck of updating U_i lies in the high computational complexity of $\mathcal{O}(n^3)$ caused by the matrix inverse operation against the $\mathbb{R}^{n \times n}$. To resolve it, we propose to update each row of U_i. Without loss of generality, we set the derivative w.r.t. $U_i(l, \cdot), 1 \leq l \leq n$ to be $\mathbf{0} \in \mathbb{R}^c$. It then yields the following

$$U_i(l, \cdot) = (T_i^l + \beta \underbrace{\sum_{j \neq i, j \in V} U_j(l, \cdot)}_{\text{Influences from other views}}) \underbrace{(R_i + D_i^T D_i)^{-1}}_{\text{computational complexity } \mathcal{O}(c^3)}, \qquad (10)$$

10

where

$$R_i = ((1 + \mu + \sum_{k=1}^{n} (\lambda_2 L_i(k,l) - \mu(X_i^T X_i)(k,l)))I \in \mathbb{R}^{c \times c} \quad (11)$$

$$T_i^l = X_i^T(l,\cdot)K_3^i + \mu \left(G_i(l,\cdot) - E_i^T(l,\cdot)D_i \right)$$
$$- K_2^i(l,\cdot) - (K_1^i)^T(l,\cdot)D_i$$

Orthonormalize U_i: After obtaining the whole U_i by updating all rows for each iteration, the clustering algorithm *e.g.,* fast k-means is performed, which yields the cluster indicator for each data point/each row, leading to orthogonal columns then normalize each entry of U_i via the rules as: $U_i(j,k) = \frac{1}{\sqrt{|C_k|}}$ if x_j is assigned with the kth cluster C_k, it is 0 otherwise. According to the processing above, it successfully achieves the orthonormal columns of U_i ($i = 1, \ldots, |V|$).

Remark 3. *As per the row-update strategy for U_i in Eq.(10), we remark the followings:*

1. *We dramatically reduces the computational complexity from $\mathcal{O}(n^3)$ by Eq.(9) to $\mathcal{O}(c^3)$, due to $c \ll n$.*

2. *Another note goes to the process of multi-view consensus of U_i via the row update. Specifically, during each iteration, the $U_i(l,\cdot)$ is updated via the influence from other views, while served as a constraint to guide the $U_j(l,\cdot)(j \neq i)$ updating, among all of which the divergence is decreased towards a consensus, **which is based on the same magnitude among** $U_i(i \in V)$ **with orthonormal columns.***

Solving D_i: We get the partial derivative of Eq.(8) w.r.t. D_i, then yields the following closed form:

$$D_i = (K_1^i U_i - K_3^i + \mu(2X_i - E_i)U_i)\frac{(I + U_i^T U_i)^{-1}}{\mu} \quad (12)$$

The major computational burden lies in $(I + U_i^T U_i)^{-1} \in \mathbb{R}^{c \times c}$, resulting into $\mathcal{O}(c^3)$, which is identical to that for row-updating of U_i, hence efficient.

Solving E_i : Optimizing Eq.(8) w.r.t. E_i is equivalent to solving the following

$$\min_{E_i} ||E_i||_1 + \frac{\mu}{2}||E_i - (X_i - D_i U_i^T + \frac{1}{\mu}K_1^i)||_F^2. \tag{13}$$

According to [25], the following closed form can be obtained

$$E_i = S_{\frac{1}{\mu}}(X_i - D_i U_i^T + \frac{1}{\mu}K_1^i), \tag{14}$$

where $S_{\frac{1}{\mu}}(x) = sign(x)max(||x|| - \frac{1}{\mu}, 0)$, $sign(x) = 1$ if x is positive, it is 0 otherwise.

Solving G_i: Optimizing Eq.(8) w.r.t. G_i is equivalent to the following

$$\min_{G_i} \Phi(K_2^i, U_i - G_i) + \frac{\mu}{2}||G_i - U_i||_F^2 \tag{15}$$

Based on that, we enjoy the following closed form

$$G_i = U_i + \frac{K_2^i}{\mu} \tag{16}$$

Solving W_i: The problem of optimizing W_i can be converted to the following

$$\min_{W_i} \sum_{j,k} \left(\lambda_2 ||U_i(j,\cdot) - U_i(k,\cdot)||_2^2 W_i(j,k) + \lambda_1 W_i(j,k)^2 \right) \tag{17}$$

$$\text{s.t. } \forall j \sum_{k=1}^{n} W_i(j,k) = 1, 0 \leq W_i(j,k) \leq 1$$

As the similarity vector for each sample is independent, we only study the jth sample.

$$\min \sum_{k} \left(\lambda_2 ||U_i(j,\cdot) - U_i(k,\cdot)||_2^2 W_i(j,k) + \lambda_1 W_i(j,k)^2 \right) \tag{18}$$

$$\text{s.t. } \sum_{k=1}^{n} W_i(j,k) = 1, 0 \leq W_i(j,k) \leq 1$$

We convert Eq.(18) to the following

$$\min_{\sum_{k=1}^{n} W_i(j,k)=1, 0 \leq W_i(j,k) \leq 1} ||W_i(j,\cdot) + m_i^j||_2^2, \tag{19}$$

12

where $m_i^j \in \mathbb{R}^{n \times 1}$ is a vector, with its kth entry $m_i^j(k) = \frac{\lambda_2 ||U_i(j,\cdot) - U_i(k,\cdot)||_2^2}{4\alpha}$, leading to the following closed form:

$$W_i(j,\cdot) = \left(\frac{1 + \sum_{l=1}^s m_i^j(l)}{s} \mathbf{1} - m_i^j \right)_+ , \tag{20}$$

where $(v)_+$ turns the negative entries in v to 0 while with positive entries remained. s denotes the number of data points that have nonzero weight connected to the jth sample. We empirically set $s = 5$ for all views. Once the $\forall j$ $W_i(j,\cdot)$ is obtained, we may update that to be a balanced undirected graph as $\frac{W_i + W_i^T}{2}$.

Consensus $W_i(i \in V)$: As W_i is solely determined by U_i according to Eq.(20), the consensus on $U_i(i \in V)$ in Remark 3 naturally leads to the consensus over $W_i(i \in V)$.

Multiplier updating: The lagrange multipliers K_1^i, K_2^i and K_3^i are automatically updated as

$$\begin{aligned}
K_1^i &= K_1^i + \mu(X_i - D_i U_i - E_i) \\
K_2^i &= K_2^i + \mu(U_i - G_i) \\
K_3^i &= K_3^i + \mu(D_i - X_i U_i)
\end{aligned} \tag{21}$$

Besides, μ is tuned via the adaptive updating rule according to [24].

Algorithm convergence: It is worth nothing that ADM strategy converges to a stationary point yet no guaranteed to be global optimum. Upon that, we define the convergence when $\forall i \in V, ||X_i - X_i U_i - E_i||_F \leq \theta ||X||_F$ with $\theta = 10^{-6}$ or maximum iteration number is reached, which is set to be 25 for our method. **The optimization process is conducted regarding each variable alternatively within each view, the entire process is terminated until the convergence rule is met for all views.**

Multi-view clustering output: After the above updating rule is converged, we got the final multi-view clustered representation $U = \sum_{i \in V} U_i \in \mathbb{R}^{n \times c}$; and multi-view optimized local graph structure $W = \sum_{i \in V} W_i \in \mathbb{R}^{n \times n}$. The normalized graph cut is applied to generate the c clusters as the multi-view spectral clustering output.

We summarize the whole updating process in Algorithm 1.

4. Computational Complexity Analysis

We clarify its computational feasibility based on Algorithm 1. For each view (i^{th} view) within one iteration, the computational complexity for all major operations is listed below

1. Updating U_i: According to the Remark 3 about Eq. (10), it is $\mathcal{O}(c^3)$, and efficient due to $c \ll n$.

2. Updating W_i: According to Eq.(20), it is $\mathcal{O}(s)$ to update each row of $W_i(j, \cdot)$, s is the number of data points having positive similarity for j^{th} data, far less than n. We empirically set $s = 5$ for the datasets.

3. Updating D_i: According to Eq.(12), the major computational complexity lies in the matrix inverse computation $(I + U_i^T U_i)^{-1} \in \mathbb{R}^{c \times c}$, resulting into $\mathcal{O}(c^3)$.

4. Updating E_i: According to Eq.(14), computational complexity is $\mathcal{O}(nd)$. As observed, the essential operation is to make comparison between each entry against $\frac{1}{\mu}$, therefore it is quite efficient in term of implementation.

Suppose the total iteration number is T, leading to the total computational complexity of our method as $\mathcal{O}(T(c^3 + nd))$.

5. Experimental Validation

The following multi-view data sets and their view-specific features are selected according to [15, 16].

- UCI handwritten Digit set[1]: It consists of features of hand-written digits (0-9). The dataset is described by 6 features and contains 2000 samples with 200 in each category. Analogous to [24], we choose 76 Fourier coefficients (FC) of the character shapes and the 216 profile correlations (PC) as two views.

---

[1]http://archive.ics.uci.edu/ml/datasets/Multiple+Features

14

- Animal with Attribute (AwA)[2]: It consists of 50 kinds of animals described by 6 features (views): Color histogram (CQ, 2688-dim), local self-similarity (LSS, 2000-dim), pyramid HOG (PHOG, 252-dim), SIFT (2000-dim), Color SIFT (RGSIFT, 2000-dim), and SURF (2000-dim). We randomly sample 80 images for each category and get 4000 images in total.

- NUS-WIDE-Object (NUS) [26]: The data set consists of 30000 images from 31 categories. We construct 5 views: 65-dimensional color histogram (CH), 226-dimensional color moments (CM), 145-dimensional color correlation (CORR), 74-dimensional edge estimation (EDH), and 129-dimensional wavelet texture (WT).

The following typical multi-view baselines are compared for spectral clustering, covering *Early fusion*, *Late fusion*, *CCA*, *Co-training* strategy and *LRR models* as reviewed in Section 1.3. All the parameters are tuned to their best performance.

- **MFMSC**: Concatenating multi-view features to perform spectral clustering.

- Multi-view affinity aggregation for multi-view spectral clustering (**MAASC**) [27].

- Canonical Correlation Analysis (CCA) based multi-view spectral clustering (**CCAMSC**) [21] by learning a common subspace for multi-view data, then perform spectral clustering.

- **Co-training** [8]: Learning multi-view Laplacian eigenspace via a co-training fashion over each individual one.

- Robust Low-Rank Representation Method (**RLRR**) [15], as formulated in Eq.(1).

- Low-Rank Representation with Multi-Graph Learning (**LRRGL**) [16].

Clustering accuracy (ACC) and normalized mutual information (NMI) are adopted as the evaluation metric for clustering performance. Pleaser refer to [28, 29] for their detailed descriptions. They are defined below:

$$ACC = \frac{\sum_{i=1}^{n} \delta(map(r_i), l_i)}{n}, \tag{22}$$

where r_i denotes the cluster label of x_i, and l_i denotes the true class label, n is the total number of images, $\delta(x, y)$ is the function that equals one if $x = y$ and equals zero otherwise, and $map(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the database.

$$NMI = \frac{\sum_{i=1}^{c} \sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}, \tag{23}$$

where n_i denotes the number of images contained in the cluster C_i ($1 \leqslant i \leqslant c$), $\hat{n}_j$ is the number of images belonging to the class L_j ($1 \leqslant j \leqslant c$), and $n_{i,j}$ denotes the number of images that are in the intersection between cluster C_i and class L_j. To demonstrate the robustness superiority over non-LRR methods, following [16], we set the feature corruption noise for each view is with sparse noise as 20% entries with uniformly noise over $[-5, 5]$ for **RLRR**, **LRRGL** and our method, with $\lambda_1 = 0.8$ in Eq.(7) for our method. All experiments are repeated 10 times, the average clustering results are shown in Tables 1 and 2, where our method outperforms the others, especially better than **RLRR** and **LRRGL**, due to its strengthes of

- encoding more flexible latent cluster structures, along with the more ideal optimized local graph structure based on such latent clustered representation.

- The superior multi-view consensus in terms of both latent clustered representation and optimized local graph structure for all views.

To penetrate the first finding, we illustrate the visualized consensus multi-view affinity matrix over NUS data set between our method and **LRRGL** in Fig. 1,

---

[2]http://attributes.kyb.tuebingen.mpg.de

16

which validates the advantages of our clustered orthogonal representation over low-rank similarity yielded by **LRRGL**.

To further showcase the superior consensus of our method, we define the following consensus metric over the multi-view data similarities yielded once the algorithm converged

$$\frac{1}{\mathbb{C}_{|V|}^2} \sum_{i,j\neq i}^{|V|} \max\{\frac{W_i}{W_j}, \frac{W_j}{W_i}\}, \tag{24}$$

where $\mathbb{C}_{|V|}^2 = \frac{|V|(|V|-1)}{2}$ implies Eq.(24) to be averaged pair-wise-views similarity divergence. $\frac{W_i}{W_j} = \sum_{k,l}^n \frac{W_i(k,l)}{W_j(k,l)}$; we choose the *maximum* value to reveal the magnitude divergence. The results are shown in Table 3, where our method enjoys a same magnitude for multi-view similarities by the value nearly to be 1 for all data sets.

Parameter Study: We further study the parameter λ_2 (clustered orthogonal representations and optimized local graph structure) and β (multi-view consensus term) in Eq.(7), and against the clustering accuracy (ACC) over AwA and NUS data sets; the parameter is tuned via grid search against $[0.001, 10]$ for λ_2 and β, we varied one parameter while fixed the others, and the results are illustrated in Fig. 2, where increasing either of them can improve the clustering accuracy until meet the optimal pair-wise values, followed by a slight performance decreasing. To balance Figs.2(a) and (b), we finalize $\lambda_2 = 0.7$ and $\beta = 0.25$.

6. Conclusion

In this paper, we revisit the classical Low-Rank Representation (LRR) for multi-view spectral clustering, by viewing LRR as essentially a latent clustered orthogonal projection winged with its optimized local graph structure. Following this, we propose to simultaneously learn clustered orthogonal projection and optimized local graph structure for each view, while enjoy the same magnitude over them both for all views, leading to a superior multi-view spectral clustering consensus. The complexity analysis is also delivered to show the computational feasibility. Extensive experiments validate its strength.

17

(a)　　　　　　　　　　　(b)
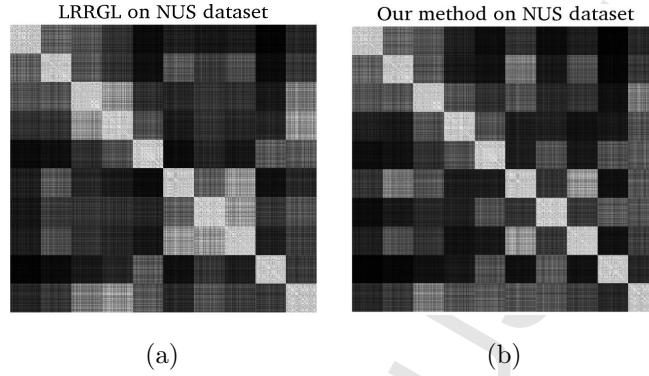
Figure 1: The visualized multi-view data similarity consensus result between ours and **LRRGL** over NUS data set. We randomly select 10 classes, where 80 samples are randomly selected for each of them. The 10 diagonal block represents the data samples within the 10 ideal clusters, where the whiter the color is, the ideally affinity value will be. Meanwhile, for non-diagonal blocks, the more black the color is, the smaller affinity will be to reveal the different clusters.

References

[1] A. Y. Ng, M. I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: NIPS, 2001.

[2] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, D. Steinberg, Top 10 algorithms in data mining, Knowledge and Information Systems. 14 (2008) 1–37.

[3] Y. Li, F. Nie, H. Huang, J. Huang, Large-scale multi-view spectral clustering via bipartite graph, in: AAAI, 2015.

[4] F. Nie, G. Cai, X. Li, Multi-view clustering and semi-supervised classification with adaptive neighbours, in: AAAI, 2017.

[5] C. Xu, D. Tao, C. Xu, Multi-view self-paced leanring for clustering, in: IJCAI, 2015.

[6] C. Deng, Z. Lv, W. Liu, J. Huang, D. Tao, X. Gao, Multi-view matrix
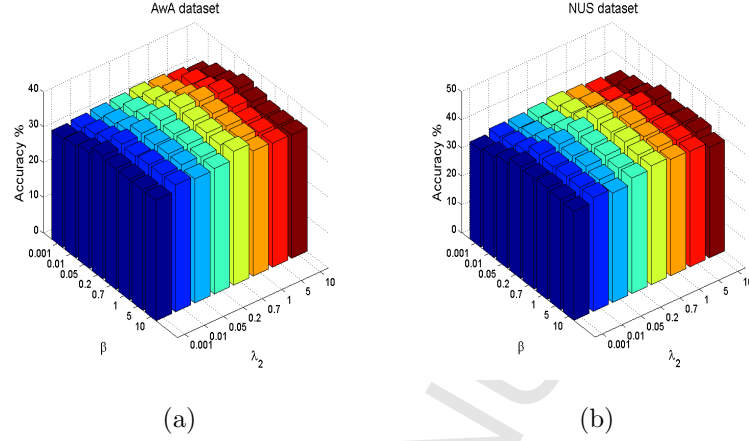
Figure 2: Parameters study of λ_2 and β for our method over NUS and AwA data sets evaluated with clustering accuracy.

decomposition:a new scheme for exploring discriminative information, in: IJCAI, 2015.

[7] J. Gui, D. Tao, Z. Sun, Y. Luo, X. You, Y. Y. Tang, Group sparse multiview patch alignment framework with view consistency for image classification, IEEE Transactions on Image Processing 23 (7) (2014) 3126–3137.

[8] A. Kumar, H. Daume, A co-training approach for multi-view spectral clustering, in: ICML, 2011.

[9] Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang, X. Huang, Robust subspace clustering for multi-view data by exploiting correlation consensus, IEEE Trans. Image Processing 24 (11) (2015) 3939–3949.

[10] Y. Wang, W. Zhang, L. Wu, X. Lin, X. Zhao, Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion, IEEE Trans. Neural Networks and Learning Systems. 28 (1) (2017) 57–70.

[11] Y. Wang, X. Lin, L. Wu, W. Zhang, Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval, IEEE Trans. Image Processing 26 (3) (2017) 1393–1404.

[12] Y. Wang, L. Wu, X. Lin, J. Gao, Multi-view spectral clustering via structured low-rank matrix factorization, IEEE Trans. Neural Networks and Learning Systems`doi:10.1109/TNNLS.2017.2777489`.

[13] J. Gao, J. Han, J. Liu, C. Wang, Multi-view clustering via joint nonnegative matrix factorization, in: SDM, 2013, pp. 252–260.

[14] H. Gao, F. Nie, X. Li, H. Huang, Multi-view subspace clustering, in: ICCV, 2015, pp. 4238–4246.

[15] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: AAAI, 2014.

[16] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, S. Pan, Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering, in: IJCAI, 2016.

[17] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: NIPS, 2011.

[18] G. Liu, Z. Lin, Y. Yu, Robust subspace segmentation by low-rank representation, in: ICML, 2010.

[19] Y. Huang, Q. Liu, S. Zhang, D. N. Metaxas, Image retrieval via probabilistic hypergraph ranking, in: CVPR, 2010.

[20] D. Greene, P. Cunningham, A matrix factorization approach for integrating multiple data views, in: ECMLPKDD, 2009.

[21] K. Chaudhuri, S. Kakade, K. Livescu, K. Sridharan, Multi-view clustering via canonical correlation analysis., in: ICML, 2009.

[22] B. Recht, M. Fazel, P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM review 45 (8) (2010) 1669–1680.

[23] F. Nie, X. Wang, H. Huang, Clustering and projected clustering with adaptive neighbors, in: KDD, 2014, pp. 977–986.

[24] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: NIPS, 2011.

[25] B. Recht, M. Fazel, P. A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, SIAM Journal on Optimization 20 (4) (2008) 1956–1982.

[26] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, Y.-T. Zheng, Nus-wide: A real-world web image database from national university of singapore, in: ACM CIVR, 2009.

[27] H.-C. Huang, Y.-Y. Chuang, C.-S. Chen, Affinity aggregation for spectral clustering, in: CVPR, 2012.

[28] D. Cai, X. Chen, Large scale spectral clustering via landmark-based sparse representation 45 (8) (2015) 1669–1680.

[29] W.-Y. Chen, Y. Song, H. Bai, C.-J. Lin, E. Y. Chang, Parallel spectral clustering in distributed systems, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 568–586.

Algorithm 1: Solving Eq.(7) for multi-view spectral clustering.

Input: $X_i(i = 1, \ldots, V), \lambda_1, \lambda_2, \beta$

Output: Multi-view spectral clustering result

1 Initialize: $U_i[0](i = 1, \ldots, V)$ computation, all entries of

$W_i[0], K_1^i[0], G_i[0], K_2^i[0]$ to be 0, $E_i[0]$ with sparse noise as 20% entries with

uniformly noise over $[-5, 5]$, $\mu[0] = 10^{-3}$, $k = 0$

2 for $i \in V$ **do**

3   **Solve** $U_i[k + 1]$:

4   Sequentially update each row of $U_i[k + 1]$ via Eq.(10).

5   Orthonormalizing $U_i[k + 1]$.

6   **Sequentially update each row of** $W_i[k + 1]$ via Eq.(20).

7   **Update** $E_i[k + 1]$ via Eq.(14).

8   **Update** $D_i[k + 1]$ via Eq.(12).

9   **Update** $G_i[k + 1]$ via Eq.(16).

10   **Update** $K_1^i[k + 1]$**,** $K_2^i[k + 1]$**,** $K_3^i[k + 1]$ via Eq.(21).

11   **Update** μ according to [24].

12   **if** Meet the Convergence Rule **then**

13   Remove the ith view from the view set as $V = V - i$

14 $U_i[N] = U_i[k + 1]$, s.t. N is any positive integer.

15   **end**

16   **else**

17 $k = k + 1$

18   **end**

19 end

20 $U = \sum_{i \in V} U_i[k + 1], W = \sum_{i \in V} W_i[k + 1]$ $(i = 1, \ldots, V)$

21 Return Multi-view spectral clustering results based on U and W via

normalized graph cut.

22

Table 1: Averaged clustering results in terms of ACC on three benchmark data sets.

ACC (%)	UCI digits	AwA	NUS
MFMSC	43.81	17.13	22.81
MAASC	51.74	19.44	25.13
CCAMSC	73.24	24.04	27.56
Co-training	79.22	29.06	34.25
RLRR	83.67	31.49	35.27
LRRGL	86.39	37.22	41.02
Ours	**92.22**	**44.55**	**45.78**

Table 2: Averaged clustering results in terms of NMI on three benchmark data sets.

NMI (%)	UCI digits	AwA	NUS
MFMSC	41.57	11.48	12.21
MAASC	47.85	12.93	11.86
CCAMSC	56.51	15.62	14.56
Co-training	62.07	18.05	18.10
RLRR	81.20	25.57	18.29
LRRGL	85.45	31.74	20.61
Ours	**89.61**	**36.67**	**26.42**

Table 3: Multi-view consensus ratio metric as per Eq.(24) between our method and **LRRGL** over three data sets. Smaller value means similar magnitude.

Method	UCI digits	AwA	NUS
LRRGL	17.39	25.78	34.21
Ours	**1.15**	**1.18**	**1.21**