arXiv:1709.01212v1 [cs.CV] 5 Sep 2017

Multi-View Spectral Clustering via Structured Low-Rank Matrix Factorization

Yang Wang[†] **and Lin Wu**[‡]
[†]The University of New South Wales, Kensington, Sydney, Australia
[‡] The University of Queensland, Brisbane, Australia
wangy@cse.unsw.edu.au, lin.wu@uq.edu.au

Abstract

Multi-view data clustering attracts more attention than their single view counterparts due to the fact that leveraging multiple independent and complementary information from multi-view feature spaces outperforms the single one. Multi-view Spectral Clustering aims at yielding the data partition agreement over their local manifold structures by seeking eigenvalue-eigenvector decompositions. Among all the methods, Low-Rank Representation (LRR) is effective, by exploring the multi-view consensus structures beyond the low-rankness to boost the clustering performance. However, as we observed, such classical paradigm still suffers from the following stand-out limitations for multi-view spectral clustering of (1) overlooking the flexible local manifold structure, caused by (2) aggressively enforcing the low-rank data correlation agreement among all views, such strategy therefore cannot achieve the satisfied between-views agreement; worse still, (3) LRR is not intuitively flexible to capture the latent data clustering structures. In this paper, we present the structured LRR by factorizing into the latent low-dimensional data-cluster representations, which characterize the data clustering structure for each view. Upon such representation, (b) the laplacian regularizer is imposed to be capable of preserving the flexible local manifold structure for each view. (c) We present an iterative multi-view agreement strategy by minimizing the divergence objective among all factorized latent data-cluster representations during each iteration of optimization process, where such latent representation from each view serves to regulate those from other views, such intuitive process iteratively coordinates all views to be agreeable. (d) We remark that such data-cluster representation can flexibly encode the data clustering structure from any view with adaptive input cluster number. To this end, (e) a novel non-convex objective function is proposed via the efficient alternating minimization strategy. The complexity analysis are also presented. The extensive experiments conducted against the real-world multi-view datasets demonstrate the superiority over state-of-the-arts.

1 Introduction

Spectral clustering [Ng *et al.*, 2001; Zelnik-Manor and Perona, 2004; Cai and Chen, 2015; Nie *et al.*, 2011] aims at exploring the local nonlinear manifold (spectral graph)[1] structure [Hou *et al.*, 2015; Tao *et al.*, 2016], attracting great attention within recent years. With the development of information technology, multi-view spectral clustering, due to the fact of outperforming the single view counterparts by leveraging the complementary information from multi-view spaces. As implied by multi-view research [Xu *et al.*, 2015; Xu *et al.*, 2013; Wang *et al.*, 2016a], an individual view is not capable of being faithful for effective multi-view learning. Therefore, exploring multi-view information is necessary, and has been demonstrated by a wide spectrum of applications *e.g.,* similarity search [Liu *et al.*, 2015; Wang *et al.*, 2015b; Wu *et al.*, 2013b; Wu *et al.*, 2015; Wu *et al.*, 2013a; Wang *et al.*, 2014a; Wang *et al.*, 2015a; Wang *et al.*, 2017a; Wang *et al.*, 2013b], human action recognition [Jones and Shao, 2014; Shao *et al.*, 2016; Wu *et al.*, 2017b; Wang *et al.*, 2016b; Wang *et al.*, 2014b; Wu *et al.*, 2017a; Wu and Wang, 2017; Wang *et al.*, 2012] *etc.*.

Essentially, given the complementary information from multi-views, the critical issue of multi-view clustering is to achieve the multi-view clustering agreement/consensus [Xu *et al.*, 2013; Gui *et al.*, 2014; Gao *et al.*, 2015] to yield a substantial superior clustering performance over the single view paradigm. Numerous multi-view based methods are proposed for spectral clustering. [Huang *et al.*, 2012; Bickel and Scheffer., 2004; Wang *et al.*, 2013a] performs multi-view information incorporation into the clustering process by optimizing certain objective loss function. *Early fusion* strategy can also be developed by concatenating the multi-view features into a uniform one [Huang *et al.*, 2010], upon which the similarity matrix is calculated for further multi-view spectral clustering. As mentioned by [Wang *et al.*, 2015c; Wang *et al.*, 2017b], such strategy will be more likely to destroy the inherent property of original feature representations within each view, hence resulting into a worse

[1] In the rest of this paper, we will alternatively use nonlinear manifold structure or spectral graph structure

performance; worse still, sometimes, as indicated by the experimental reports from our previous research [Wang *et al.*, 2015c], it may even be inferior to the clustering performance with a single view. In contrast, *Late fusion* strategy [Greene and Cunningham, 2009] conducts spectral clustering performance for each view, and then combining multiple them afterwards, which, however, cannot achieve the multi-view agreement, without collaborating with each other.

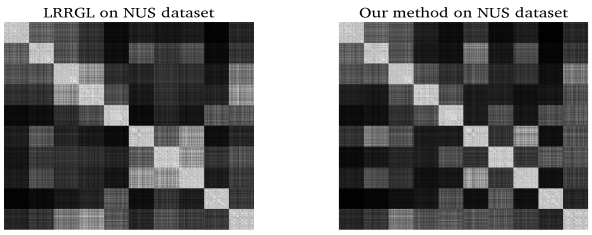
Canonical Correlation Analysis (CCA) based methods [Blaschko and Lampert., 2008; Chaudhuri *et al.*, 2009] for multi-view spectral clustering project the data from multi-view feature spaces onto one common lower dimensional subspace, where the spectral clustering is subsequently conducted. One limitation of such method lies in the fact that such common lower-dimensional subspace cannot flexibly characterize the local manifold structures from heterogeneous views, resulting into an inferior performance. Kumar *et al.* [Kumar *et al.*, 2011] proposed a state-of-the-art co-regularized spectral clustering for multi-view data. Similarly, a co-training [Blum and Mitchell, 1998; Wang and Zhou, 2010] model is proposed for this problem [Kumar and Daume, 2011].

One assumption for above work [Kumar *et al.*, 2011; Kumar and Daume, 2011] is the scenario with noise corruption free for each view. However, it is not easily met. To this end, Low-Rank Representation (LRR) [Xia *et al.*, 2014; Liu *et al.*, 2010; Wang *et al.*, 2015c; Liu and Yan, 2011; Liu *et al.*, 2013] is proposed. As summarized in [Wang *et al.*, 2016a], where the basic idea is to decompose data representation into a view-dependent noise corruption term and a common low rank based representation shared by all views, leading to common data affinity matrix for clustering; The effectiveness of low-rank model also leads to numerous research on multi-view subspace learning [Ding and Fu, 2016; Dingg and Fu, 2014] applied to the pattern recognition field.

LRR tries a common multi-view low-rank representation, but overlooks the distinct manifold structures. To remedy the limitations, inspired by the latest development of graph regularized LRR [Yin *et al.*, 2016; Yin *et al.*, 2015], we recently proposed another iterative views agreement strategy [Wang *et al.*, 2016a] with graph regularized Low-rank Representation for multi-view spectral clustering, named **LRRGL** for short, to characterize the non-linear manifold structure from each view, **LRRGL** couples LRR with multi-graph regularization, where each one can characterize the view-dependent non-linear local data manifold structure [Zhuang *et al.*, 2016]. A novel iterative view agreement process is proposed of optimizing the proposed, where, during each iteration, the low-rank representation yielded from each view serves as the constraint to regulate the representation learning from other views, to achieve the consensus, implemented by applying Linearized Alternating Direction Method with Adaptive Penalty [Lin *et al.*, 2011].

Despite the effectiveness of **LRRGL**, we still identify the following non-trivial observations that are not addressed by **LRRGL** to obtain the further improvement

- It is less flexible for Z_i yielded by low-rank constraint to capture the flexible latent data similarity that can encode the more rich similarity information than Z_i over X_i; that can be better solved by matrix factorization.
- **LRRGL** mainly focused on yielding the low-rank primal data similarity matrix Z_i derived from X_i. However, such primal Z_i is less intuitive to understand and less effective to reveal the ideal data clustering structure for the ith view, as well as multi-views. Hence, it will prevent that achieving the better multi-view spectral clustering performance. The structured consensus loss term imposed over $Z_i (i \in V)$ may not effectively achieve the consensus regarding the multi-view spectral clustering.



(a) (b)

Figure 1: The visualization results of the multi-view (please refer to section 4 for specific multi-view features) affinity matrix between ours and **LRRGL** over NUS data; The more whiter for diagonal blocks, the more ideally the cluster is to characterize the data objects within the larger similarity, meanwhile, the more blacker for non-diagonal blocks, the more reasonable the non-similarity data objects are unlikely to cluster together. For such result, we can see the diagonal blocks from 3rd to the 8th of our method are more whiter than those of **LRRGL**, leading to the result that the surrounding black non-diagonal blocks of our method are more salient than those of **LRRGL**, which demonstrate the advantages of our method via a latent factorized data-cluster representation over **LRRGL**

1.1 Our Contributions

This paper is the extension of our recent work [Wang *et al.*, 2016a], upon that, we deliver the following novel contributions to achieve the further improvement over multi-view spectral clustering

- Instead of focusing on primal low-rank data similarity matrix Z_i such that $i = 1, \ldots, V$, we perform a symmetric matrix factorization over Z_i into the data-cluster indicator matrix, so that such latent factorization provides the better chance to preserve the ideal cluster structure besides flexible manifold structure for each view.
- We impose the laplacian regularizer over factorized data-clustered representation to further characterize the nonlinear local manifold structure for each view. We remark that the factorized data-cluster matrix can effectively encode the clustering structure, we provide an example to illustrate this in Fig.1. *To reach the multi-view clustering agreement, we set the same clustering number*

Table 1: The Notations Summarization

Notations	Explanation
$X_i \in \mathbb{R}^{d_i \times n}$	Feature Representation Matrix for the ith view.
$E_i \in \mathbb{R}^{d_i \times n}$	Feature noise matrix for the ith view.
$Z_i \in \mathbb{R}^{n \times n}$	Self-expressive similarity matrix for the ith view.
$U_i \in \mathbb{R}^{n \times d}$	Data-cluster matrix for the ith view.
$W_i \in \mathbb{R}^{n \times n}$	Data similarity matrix over X_i for the ith view.
rank(A)	The rank of the matrix A.
$A[k]$	Updated matrix A at the kth iteration.
AB	Matrix Multiplication between A and B.
n	The number of data objects.
d_i	The dimension of the feature space for the ith view.
$\|\| \cdot \|\|^T$	The matrix transpose.
$\|\| \cdot \|\|_F$	Frobenius norm.
$\|\| \cdot \|\|_*$	Nuclear norm.
$\|\| \cdot \|\|_1$	ℓ_1 norm of matrix seen as a long vector.
$(\cdot)^{-1}$	The matrix inverse computation.
I_d	Identity matrix with the size of $d \times d$.
$\|\| \cdot \|\|_2$	ℓ_2 norm of a vector.
Tr($\cdot$)	Trace operator over the square matrix.
$\langle \cdot, \cdot \rangle$	inner product.
V	The set of all views.
C_k	The kth data cluster.
$\|C_k\|$	The cardinality of C_k.
$\|V\|$	Cardinality of the set V.
$A(l, \cdot)$	The lth row of the matrix A.
$A(\cdot, m)$	The mth column of the matrix A.
$A(l, m)$	The $(l, m)^{th}$ entry of the matrix A.

for all views to the data-clustering representation for all views.

- We impose the consensus loss term to minimize the divergence among all the latent data-cluster matrix instead of Z_i to achieve the multi-view spectral clustering agreement.
- To implement all the above insights, we propose a novel objective function, and an efficient alternating optimization strategy together with the complexity analysis to solve the objective function; moreover, we deliver the intuitions of iterative multi-view agreement over the factorized latent data-cluster representation during each iteration of our optimization strategy, that will eventually lead to the multi-view clustering agreement.
- Extensive experiments over real-world multi-view data sets demonstrate the advantages of our technique over the state-of-the-arts including our recently proposed **LRRGL** [Wang *et al.*, 2016a].

Recently another elegant graph based PCA method [Tang *et al.*, 2017] is proposed spectral clustering with out-of-sample case. Unlike this effective technique, we study the multi-view case to address the effective consensus for spectral clustering. We summarize the main notations in Table 1.

2 Structured Low-Rank Matrix Factorization to Spectral Clustering

We get started from each single view *e.g.,* the ith view as

$$\min_{Z_i, E_i} \frac{\theta}{2}||X_i - X_i Z_i - E_i||_F^2 + ||Z_i||_* + \beta||E_i||_1, \quad (1)$$

where θ and β are the trade-off parameters, as aforementioned, we always adopt D_i to be X_i, so that X_i can be decomposed as clean component $X_i Z_i$ and another corrupted component E_i for the ith view. One can easily verify that rank$(X_i Z_i) \leq$ rank(Z_i), hence minimizing rank(Z_i) is equivalent to bounding the low-rank structure of clean component $X_i Z_i$.

Now we are ready to deeply investigate $||Z_i||_*$ for the ith view. Following [Recht *et al.*, 2010], we reformulate the nuclear norm $||Z_i||_*$ as

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

where $U_i \in \mathbb{R}^{n \times d}$ and $V_i \in \mathbb{R}^{n \times d}$; d is always less than d_i since high-dimensional data objects always characterize the low-rank structure.

2.1 Notes regarding U_i and V_i for multi-view spectral clustering

Before further discussing the low-rank matrix factorization, one may consider the following notes that the factorized U_i and V_i may need to satisfy in the context of both the **within-view** data structure preserving and **multi-view** spectral clustering agreement:

1. The low-rank data structure should be characterized by the factorized U_i or V_i for the ith view, especially to characterize the underlying data clustering structure.
2. The factorized latent factors should well encode the manifold structure for the ith view, which, as previously mentioned, is critical to the spectral clustering performance.
3. Either the row based matrix U_i or column based matrix V_i is considered to meet the above two notes? if so, which one? One may claim both to be considered, which, however, may inevitably raise more parameters to be tuned.
4. Not only the factorized latent low-dimensional factors *e.g.,* U_i or V_i, should meet the above notes within each view *e.g.,* the ith view, but also need the same scale to unify all views to reach possible agreement.

To address all the above notes, in what follows, we will present our technique of data-cluster based structured low-rank matrix factorization.

2.2 Data-cluster (landmark) based Structured Low-Rank Matrix factorization

We aim at factorizing Z_i as an approximate symmetric low-rank data-cluster matrix to minimize the reconstruction error

$$\min_{Z_i} ||X_i - X_i Z_i||_F^2, \quad (3)$$

where we assume the rank of Z_i is k_i, such that k_i is related to the data cluster number for the ith view. As indicated by [Liu *et al.*, 2013], minimizing the Eq.(3) is equivalent to finding the optimal rank k_i approximation relying on skinny singular value decomposition of $X_i = V \Sigma U^T$ to yield the following optimal solution

$$Z_i^* = U_i U_i^T, \quad (4)$$

where $U_i \in \mathbb{R}^{n \times k_i}$, such that k_i denotes the top k_i principle basis of X_i. Here we follow the assumption in [Kuang

et al., 2012] to see k_i as the cluster number of data objects within the ith view, and the data-cluster symmetric matrix factorization has been widely adopted by the numerous existing research including semi-supervised learning [Wang *et al.*, 2014c; Wu *et al.*, 2016], metric fusion [Wang *et al.*, 2017b] and clustering [Cai and Chen, 2015]. We aim at solving the following equivalent low-rank minimization over Z_i via the clustered symmetric matrix factorization below

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

where we often minimize the following for derivative convenience with respect to U_i

$$||Z_i||_* = \min_{U_i, Z_i = U_i U_i^T} \frac{1}{2}||U_i||_F^2 \tag{6}$$

Remark. Following Eqs. (3) and (4), we initialize the $U_i \in \mathbb{R}^{n \times k_i}$ via a k means clustering over X_i and normalize $U_i(j,k) = \frac{1}{|C_k|}$ provided $X_i(\cdot, j)$ *i.e.,* the jth data object is assigned to C_k. By such normalization, all the columns of U_i are orthonormal; moreover, they are within the same magnitude so as to perform the agreement minimization. Furthermore, such factorization can well address the aforementioned challenges, it is worthwhile to summarize them below

- The data cluster structure can be well encoded by such low-rank data-cluster representation within each view. The setting $U_i = V_i$ can avoid the more parameters and importance weight discussion provided $U_i \neq V_i$.
- More importantly, inspired by the reasonable assumption hold by all the multi-view clustering research [Kumar *et al.*, 2011; Kumar and Daume, 2011; Bickel and Scheffer., 2004; Wang *et al.*, 2015c]. As indicated by [Wang *et al.*, 2015c], the ideal multi-view clustering performance is that the common underlying data clustering structure is shared by all the views; we naturally set all the U_i with the same size by adopting the same value for $k_i = d(i = 1, \ldots, V)$ *i.e.,* the clustering number, upon the same data objects number n for all views, so that the feasible loss functions can be developed to seek the multi-view clustering agreement with the same clustering number for all views.

For spectral clustering from each view, we preserve the nonlinear local manifold structure of X_i via such low-rank data-cluster representation U_i for the ith view, which can be formulated as

$$\begin{aligned} &\frac{1}{2}\sum_{j,k}^{n} ||u_j^i - u_k^i||_2^2 W_i(j,k) \\ &= \sum_{j=1}^{N} (u_j^i)^T u_j^i H_i(j,j) - \sum_{j,k}^{N} (u_k^i)^T u_j^i W_i(j,k) \\ &= \text{Tr}(U_i^T H_i U_i) - \text{Tr}(U_i^T W_i U_i) = \text{Tr}(U_i^T L_i U_i), \end{aligned} \tag{7}$$

where $u_k^i \in \mathbb{R}^d$ is the kth row vector of $U_i \in \mathbb{R}^{n \times d}$ representing the linear correlation between x_k and $x_j (j \neq k)$ in the ith view; $W_i(j,k)$ encodes the similarity between x_j and x_k for the ith view; H_i is a diagonal matrix with its kth diagonal entry to be the summation of the kth row of W_i, and $L_i = H_i - W_i$ is the graph laplacian matrix for the ith view.

Following [Wang *et al.*, 2016a], we choose Gaussian kernel to define W_{jk}^i

$$W_i(j,k) = e^{-\frac{||x_j^i - x_k^i||_2^2}{2\sigma^2}}. \tag{8}$$

We aim to minimize the difference of low-rank based data-cluster representations for all views via a mutual consensus loss function term to coordinate all views to reach clustering agreement, while structuring such representation with laplacian regularizer to encode the local manifold structure for each view.

Unlike the traditional LRR to achieve the common data similarity by all views, we propose to learn a variety of factorized low-rank data-cluster representations for different views to preserve the flexible local manifold structure while achieving the data cluster structure for each view, upon which, the consensus loss term is imposed to achieve the multi-view consensus, leading to our iterative views agreement in the next section.

2.3 The Objective Function with structured low-rank Matrix factorized representation

We propose the objective function with structured low-rank representation U_i for each view *e.g.,* the ith view with factorized low-rank via Eq. (6) data-clustered representation via Eq.(3). Then we have the following

$$\begin{aligned} \min_{U_i, E_i (i \in V)} \sum_{i \in V} (& \underbrace{\frac{1}{2}||U_i||_F^2}_{\text{minimize } ||Z_i||_* \text{ via Eq.(6)}} + \underbrace{\lambda_1 ||E_i||_1}_{\text{noise and corruption robustness}} \\ & + \underbrace{\lambda_2 \text{Tr}(U_i^T L_i U_i)}_{\text{Graph Structured Regularization}} + \underbrace{\frac{\beta}{2} \sum_{j \in V, j \neq i} ||U_i - U_j||_F^2}_{\text{Views-agreement}}) \\ \text{s.t.} \quad & i = 1, \ldots, V, \;\; X_i = X_i U_i U_i^T + E_i, U_i \geq 0, \end{aligned} \tag{9}$$

where

- $U_i \in \mathbb{R}^{n \times d}$ denotes the factorized low-rank data-cluster representation of X_i for the ith view. $\text{Tr}(U_i^T L_i U_i)$ makes U_i to be structured with local manifold structure for the ith view. $||E_i||_1$ is responsible for possible noise with X_i. $\lambda_1, \lambda_2, \beta$ are all trade-off parameters.
- One reasonable assumption hold by a lot of multi-view clustering research [Gao *et al.*, 2013; Kumar *et al.*, 2011; Bickel and Scheffer., 2004; Kumar and Daume, 2011] is that all the views should share the similar underlying clustering structure. $\sum_{i,j \in V} ||U_i - U_j||_F^2$ aims to achieve the views-agreement regarding the factorized low-rank representations U_i from all $|V|$ views; unlike the traditional LRR method to enforce an identical representation, we construct different U_i for each view, then further minimize their divergence to generate a view-agreement.
- $U_i \geq 0$ is a non-negative constraint, through $X_i = X_i Z_i + E_i = X_i U_i U_i^T + E_i$ for the ith view.

Eq.(9) is non-convex, we hence alternately optimize each variable while fixing the others; that is, updating all the U_i and $E_i (i \in \{1, \ldots, V\})$ in an alternative way until the convergence is reached. As solving all the $\{U_i, E_i\}(i \in V)$ pairs shares the similar optimization strategy, only the ith view is presented. To this end, we introduce two auxiliary variables D_i and G_i, then solving the Eq.(9) with respect to U_i, E_i, D_i and G_i that can be written as follows

$$\begin{aligned}&\min_{U_i,E_i,D_i,G_i} \frac{1}{2}||U_i||_F^2 + \lambda_1||E_i||_1 \\ &+ \lambda_2 \text{Tr}(U_i L_i U_i^T) + \frac{\beta}{2}\sum_{j\in V, j\neq i}||U_i - U_j||_F^2 \\ &\text{s.t. } X_i = D_i U_i^T + E_i, D_i = X_i U_i, G_i = U_i, G_i \geq 0,\end{aligned} \tag{10}$$

where $D_i \in \mathbb{R}^{d_i \times d}$, we will show the intuition for the auxiliary variable relationship $D_i = X_i U_i$ by introducing the augmented lagrangian function based on Eq.(10) below

$$\begin{aligned}&\mathcal{L}(U_i, E_i, D_i, G_i, K_1^i, K_2^i, K_3^i) \\ &= \frac{1}{2}||U_i||_F^2 + \lambda_1||E_i||_1 + \lambda_2 \text{Tr}(U_i^T L_i U_i) \\ &+ \frac{\beta}{2}\sum_{j\in V, j\neq i}||U_i - U_j||_F^2 + \langle K_1^i, X_i - D_i U_i^T - E_i\rangle \\ &+ \langle K_2^i, U_i - G_i\rangle + \langle K_3^i, D_i - X_i U_i\rangle \\ &+ \frac{\mu}{2}(||X_i - D_i U_i^T - E_i||_F^2 + ||U_i - G_i||_F^2 + ||D_i - X_i U_i||_F^2),\end{aligned} \tag{11}$$

where $K_1^i \in \mathbb{R}^{d_i \times n}$, $K_2^i \in \mathbb{R}^{n \times d}$ and $K_3^i \in \mathbb{R}^{d_i \times d}$ are Lagrange multipliers, $\mu > 0$ is a penalty parameter.

From Eq.(11), we can easily show the intuition on $D_i = X_i U_i$, that is,

- minimizing $||X_i - D_i U_i^T - E_i||_F^2$ w.r.t. D_i is similar as dictionary learning, while pop out the U_i^T as corresponding representations learning, both of them reconstruct the X_i for the ith view. Besides the above intuition, it is quite simple to optimize only single U_i^T by merging the other into D_i.

3 Optimization Strategy

We minimize Eq.(11) by updating each variable while fixing the others.

3.1 Solve U_i

Minimizing U_i is to resolve Eq.(12)

$$\begin{aligned}&\mathcal{L}_1 = \frac{1}{2}||U_i||_F^2 + \lambda_2 \text{Tr}(U_i^T L_i U_i) \\ &+ \frac{\beta}{2}\sum_{j\in V, j\neq i}||U_i - U_j||_F^2 + \langle K_1^i, X_i - D_i U_i^T - E_i\rangle \\ &+ \langle K_2^i, U_i - G_i\rangle + \langle K_3^i, D_i - X_i U_i\rangle \\ &+ \frac{\mu}{2}(||X_i - D_i U_i^T - E_i||_F^2 + ||U_i - G_i||_F^2 + ||D_i - X_i U_i||_F^2),\end{aligned} \tag{12}$$

We set the derivative of Eq.(12) w.r.t. U_i to be the zero matrix, which yields the Eq.(13) below

$$\begin{aligned}\frac{\partial \mathcal{L}_1}{\partial U_i} &= U_i + 2\lambda_2 L_i U_i + \beta \sum_{j\in V, j\neq i}(U_i - U_j) \\ &- (K_1^i)^T D_i + K_2^i - X_i^T K_3^i \\ &+ \mu U_i D_i^T D_i + \mu E_i^T D_i + \mu(U_i - G_i) \\ &- \mu X_i^T X_i U_i = \mathbf{0},\end{aligned} \tag{13}$$

where $\mathbf{0} \in \mathbb{R}^{n \times d}$ shares the same size as U_i. Rearranging the other terms further yields the following

$$U_i = \underbrace{(2\lambda_2 L_i + (1 + \beta(|V| - 1) + \mu)I_n - \mu X_i^T X_i)^{-1}}_{\text{with computational complexity } \mathcal{O}(n^3)} S, \tag{14}$$

$$\begin{aligned}S = &\sum_{j\in V, j\neq i} U_j + ((K_1^i)^T - \mu U_i D_i^T - \mu E_i^T) D_i \\ &+ X_i^T K_3^i + \mu X_i^T X_i U_i\end{aligned} \tag{15}$$

The bottleneck of computing Eq.(14) lies in the inverse matrix computation over the matrix of the size $\mathbb{R}^{n \times n}$ causing the computational complexity $\mathcal{O}(n^3)$, which is computationally prohibitive provided that n is large. Therefore, we turn to update each row of U_i; without loss of generality, we present the derivative with respect to $U_i(l, \cdot)$ as

$$\begin{aligned}&U_i(l,\cdot) + U_i(l,\cdot)\left(\sum_{k=1}^{n}(2\lambda_2 L_i(k,l) - \mu(X_i^T X_i)(k,l))\right) \\ &+ (K_1^i)^T(l,\cdot) D_i + \mu U_i(l,\cdot) D_i^T D_i + K_2^i(l,\cdot) - X_i^T(l,\cdot) K_3^i \\ &+ \beta \sum_{j\in V, j\neq i}(U_i(l,\cdot) - U_j(l,\cdot)) \\ &+ \mu\left(U_i(l,\cdot) + E_i^T(l,\cdot) D_i - G_i(l,\cdot)\right) = \mathbf{0},\end{aligned} \tag{16}$$

where $\mathbf{0} \in \mathbb{R}^d$ denotes the vector of the size d with all entries to be 0, $U_i(l, \cdot) \in \mathbb{R}^d$ represents the lth row of U_i; we rearrange the terms to yield the following

$$\begin{aligned}&U_i(l,\cdot) \\ &= \left(T_i^l + \beta \underbrace{\sum_{j\neq i, j\in V} U_j(l,\cdot)}_{\text{Influences from other views}}\right) \\ &\underbrace{\left((1 + \mu + \sum_{k=1}^{n}(2\lambda_2 L_i(k,l) - \mu(X_i^T X_i)(k,l)))I_d + D_i^T D_i\right)^{-1}}_{\text{with computational complexity } \mathcal{O}(d^3)}\end{aligned} \tag{17}$$

where

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

$I_d \in \mathbb{R}^{d \times d}$ is the identity matrix.

Complexity discussion for the row updating strategy for U_i

Unlike the closed form regarding U_i, it is apparent that the major computational complexity lies in the inverse matrix computation over the size of $\mathbb{R}^{d\times d}$, which leads to $\mathcal{O}(d^3)$ according to Eq.(17), which is much smaller than $\mathcal{O}(n^3)$. Besides, as d is set as the cluster number across all views; moreover, aforementioned, it should be less than the inherent rank of X_i, and hence a small value. Upon the above facts, it is tremendously efficient via $\mathcal{O}(d^3)$ to sequentially update each row of U_i.

Intuitions for views agreement

The iterative views clustering agreement can be immediately captured via the terms underlined in Eq.(17). Specifically, during each iteration, the $U_i(l,\cdot)$ is updated via the influence from others view, while served as the constraint to generate $U_j(l,\cdot)(j\neq i)$, the divergence among all $U_i(l,\cdot)$ is decreased gradually towards an agreement for all views, such process repeats until the convergence is reached.

Unlike the existing LRR method by directly imposing the common representation, our iterative multi-view agreement can better preserve the flexible manifold structure for each view meanwhile achieve the multi-view agreement, which will be critical to final multi-view spectral clustering.

Remark. After the whole U_i is updated for the ith view, we simply perform a K-means clustering over it to assign each data object to one cluster exclusively. Then normalized each column of U_i to form an orthonormal matrix.

3.2 Solve D_i

The optimization process regarding D_i is equivalent to the following

$$\min_{D_i} < K_i^i, X_i - D_iU_i^T - E_i > + < K_3^i, D_i - X_iU_i > + \frac{\mu}{2}\left(||X_i - D_iU_i - E_i||_F^2 + ||D_i - X_iU_i||_F^2\right) \tag{18}$$

We get the derivative with respect to D_i, then it yields the following closed form updating rule

$$D_i = \left(K_1^iU_i - K_3^i + \mu(2X_i - E_i)U_i\right)\frac{\left(I_d + U_i^TU_i\right)^{-1}}{\mu}, \tag{19}$$

where the major computational complexity lies in the inverse computation over matrix $(I_d + U_i^TU_i) \in \mathbb{R}^{d\times d}$, resulting into $\mathcal{O}(d^3)$, as aforementioned, that is the same as updating each row of U_i, and hence quite efficient.

3.3 Solve E_i

it is equivalent to solving the following:

$$\min_{E_i} \lambda_1||E_i||_1 + \frac{\mu}{2}||E_i - (X_i - D_iU_i^T + \frac{1}{\mu}K_1^i)||_F^2, \tag{20}$$

where the following closed form solution can be yielded for E_i according to [Cai *et al.*, 2008]

$$E_i = S_{\frac{\lambda_1}{\mu}}(X_i - D_iU_i^T + \frac{1}{\mu}K_1^i) \tag{21}$$

3.4 Solve G_i

It is equivalent for the following:

$$< K_2^i, U_i - G_i > + \frac{\mu}{2}||G_i - U_i||_F^2 \tag{22}$$

where the following closed form solution of G_i can be derived as

$$G_i = U_i + \frac{K_2^i}{\mu} \tag{23}$$

Algorithm 1: Alternating optimization strategy for Eq.(9).

Input: $X_i(i=1,\ldots,V), d, \lambda_1, \lambda_2, \beta$
Output: $U_i, D_i, E_i, G_i(i\in V)$

Initialize: $U_i[0], L_i(i=1,\ldots,V)$ computation, set all entries of $K_1^i[0], G_i[0], K_2^i[0]$ to be 0, initialize $E_i[0]$ with sparse noise as 20% entries corrupted with uniformly distributed noise over [-5,5], $\mu[0]=10^{-3}$, $\epsilon_1 = 10^{-3}$, $\epsilon_2 = 10^{-1}$
$k=0$
for $i\in V$ **do**
  Solve U_i:
  Sequentially update each row of U_i according to Eq.(17).
  Orthonormalized each column of U_i.
  Update E_i:
  $E_i[k+1] = S_{\frac{\lambda_1}{\mu[k]}}(X_i - D_iU_i^T[k] + \frac{1}{\mu[k]}K_1^i[k])$
  Update G_i:
  $G_i[k+1] = U_i[k] + \frac{K_2^i[k]}{\mu[k]}$
  Update K_1^i**,** K_2^i**,** K_3^i **and** μ:
  $K_1^i[k+1] = K_1^i[k] + \mu(X_i - D_iU_i^T[k] - E_i[k])$
  $K_2^i[k+1] = K_2^i[k] + \mu(U_i[k] - G_i[k])$
  $K_3^i[k+1] = K_3^i[k] + \mu(D_i[k] - X_iU_i[k])$
  Update μ according to [Lin *et al.*, 2011]
  whether converged
  if $||X_i - D_iU_i^T[k+1] - E_i[k+1]||/||X_i|| < \epsilon_1$ *and* *max*$\{\xi||U_i[k+1]-U_i[k]||, \mu[k]||G_i[k+1] - G_i[k]||, \mu[k]||E_i[k+1]-E_i[k]||\} < \epsilon_2$ **then**
    Remove the ith view from the view set as $V = V - i$
    $U_i[N] = U_i[k+1]$, s.t. N is any positive integer.
  else
    $k = k+1$
Return $U_i[k+1], D_i[k+1], E_i[k+1], G_i[k+1]$ $(i=1,\ldots,V)$

3.5 Updating K_1^i, K_2^i, K_3^i and μ

We update Lagrange multipliers K_1^i, K_2^i and K_3^i via

$$K_1^i = K_1^i + \mu(X_i - D_iU_i - E_i) \tag{24}$$

$$K_2^i = K_2^i + \mu(U_i - G_i) \tag{25}$$

$$K_3^i = K_3^i + \mu(D_i - X_iU_i) \tag{26}$$

Following [Wang *et al.*, 2016a], μ is tuned using the adaptive updating strategy [Lin *et al.*, 2011] to yield a faster convergence. The optimization strategy alternatively updates each variable while fixing others until the convergence, which is summarized by Algorithm 1.

3.6 Notes regarding Algorithm 1

It is worthwhile to highlight some critical notes regarding the Algorithm 1 below

- We initialize the $U_i[0] \in \mathbb{R}^{n\times d}$ for all views, such that each entry of $U_i[0]$ represents similarity between each data object and one of the d anchors (cluster representatives), which can be seen as the centers from the clusters generated from the k-means or spectral clustering.
- For our initialization, we adopt the spectral clustering outcome with the clustering number to be d, where the similarity matrix is calculated via the original X_i feature representation within each view, then the $U_i[0](i,j)$ entry *i.e.,* the similarity between the ith data object and the jth anchor is yielded via Eq.(8). The laplacian matrix $L_i(i = 1, \ldots, V)$ are computed once offline also within the original X_i feature representation.
- More importantly, we set the identical value of d(the cluster number) to the column size of $U_i[0](i = 1, \cdot\cdot, V) \in \mathbb{R}^{n\times d}$ for all the views. We remark that the above initial setting for $U_i[0]$ with the same d is reasonable, as stated before all the views should share the similar underlying data clustering structure. This fact also implies that the initialized $U_i[0]$ is reasonably not divergent a lot among all views.

3.7 Convergence discussion

Often, the above alternating minimization strategy can be seen as the coordinate descent method. According to [Bertsekas, 1999], the sequences (U_i, D_i, E_i, G_i) above will eventually converge to a stationary point. However, we are not sure whether the converged stationary point is a global optimum, as it is not jointly convex to all the variables above.

3.8 Clustering

Following [Wang *et al.*, 2016a], once the converged $U_i(i = 1, \ldots, V)$ are ready, all column vectors of $U_i(i = 1, \ldots, V)$ while set small entries under given threshold τ to be 0. Afterwards, the similarity matrix for the ith view between the jth and k^{th} data objects as

$$W_i(j,k) = (U_i U_i^T)(j,k) \tag{27}$$

Following [Wang *et al.*, 2016a], The final data similarity matrix can be defined as

$$W = \frac{\sum_i^V W_i}{|V|} \tag{28}$$

The clustering is carried out against W via Eq.(28) to yield final outcome of d data groups.

Table 2: Data sets.

Features	UCI digits	AwA	NUS	VOC
1	FC (76)	CQ (2688)	CH(65)	Color (1500)
2	PC (216)	LSS (2000)	CM(226)	HOG (250)
3	-	PHOG (252)	CORR(145)	-
4	-	SIFT(2000)	EDH(74)	-
5	-	RGSIFT(2000)	WT(129)	-
6	-	SURF(2000)	-	-
# of data	2000	4000	26315	5600
# of classes	10	50	31	20

4 Experiments

We adopt the data sets mentioned in [Wang *et al.*, 2016a] below:

- UCI handwritten Digit set[2]: consists of features for hand-written digits (0-9), with 6 features and contains 2000 samples with 200 in each category. Analogous to [Lin *et al.*, 2011; Wang *et al.*, 2016a], we choose two views as 76 Fourier coefficients (FC) of the character shapes and the 216 profile correlations.
- Animal with Attribute (AwA)[3]: 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). Following [Wang *et al.*, 2016a], 80 images for each category and get 4000 images in total.
- NUS-WIDE-Object (NUS) [Chua *et al.*, 2009]: 30000 images from 31 categories. 5 views are adopted using 5 features as provided by the website [4]: 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).
- PASCAL VOC 2012[5]: we select 20 categories with 11530 images, two views are constructed with Color features (1500-dim) and HOG features (250 dim). Among them, 5600 images are selected by removing the images with multiple categories.

We summarize the above throughout Table 2.

4.1 Baselines

The following state-of-the-art baselines used in [Wang *et al.*, 2016a] are compared:

- **MFMSC**: concatenating multi-features to be the multi-view representation for similarity matrix, the spectral clustering is then conducted [Huang *et al.*, 2010].
- Multi-feature representation similarity aggregation for spectral clustering (**MAASC**) [Huang *et al.*, 2012].
- Canonical Correlation Analysis (CCA) model (**CCAMSC**) [Chaudhuri *et al.*, 2009]: Projecting multi-view data into a common subspace, then perform spectral clustering.

[2] http://archive.ics.uci.edu/ml/datasets/Multiple+Features
[3] http://attributes.kyb.tuebingen.mpg.de
[4] lms.comp.nus.edu.sg/research/NUS-WIDE.html
[5] http://host.robots.ox.ac.uk/pascal/VOC/voc2012/

- Co-regularized multi-view spectral clustering (**CoMVSC**) [Kumar *et al.*, 2011]: It regularizes the eigenvectors of view-dependent graph laplacians and achieve consensus clusters across views.
- **Co-training** [Kumar and Daume, 2011]: Alternately modify one view's Laplacian eigenspace by learning from the other views 's eigenspace, the spectral clustering is then conducted.
- Robust Low-Rank Representation method (**RLRR**) [Xia *et al.*, 2014], after obtaining the data similarity matrix, upon which, the spectral clustering is performed to be the final multi-view spectral clustering result.
- Low-rank Representation with Graph laplacian (**LRRGL**) [Wang *et al.*, 2016a] regularizer over the non-factorized low-rank representations, with each of which corresponds to one view to preserve the individual manifold structure, while iteratively boost all these low-rank representations to reach agreement. The final multi-view spectral clustering is performed upon the similarity representations

4.2 Experimental Settings and Parameters Study

We implement these competitors under the experimental setting as mentioned in [Wang *et al.*, 2016a]. Following [Wang *et al.*, 2016a], σ in Eq.(8) is learned via [Zelnik-Manor and Perona, 2004], and $s = 20$ to construct s-nearest neighbors for Eq.(8). We adopt two standard metrics: clustering accuracy (ACC) and normalized mutual information (NMI) as the metric defined as Eq.(29)

$$\mathsf{ACC} = \frac{\sum_{i=1}^{n} \delta(\text{map}(r_i), l_i)}{n}, \tag{29}$$

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 $\text{map}(r_i)$ is the permutation mapping function that maps each cluster label r_i to the equivalent label from the database. Meanwhile the NMI is formulated below

$$\mathsf{NMI} = \frac{\sum_{i=1}^{c}\sum_{j=1}^{c} n_{i,j} \log \frac{n_{i,j}}{n_i \hat{n}_j}}{\sqrt{(\sum_{i=1}^{c} n_i \log \frac{n_i}{n})(\sum_{j=1}^{c} \hat{n}_j \log \frac{\hat{n}_j}{n})}}, \tag{30}$$

where n_i is the sample number in cluster C_i ($1 \leqslant i \leqslant c$), $\hat{n}_j$ is the sample number from class L_j ($1 \leqslant j \leqslant c$), and $n_{i,j}$ denotes the sample number in the intersection between C_i and L_j.

Remark. Following [Wang *et al.*, 2016a], *we repeated the running 10 times, and their averaged mean value for multi-view spectral clustering for all methods is reported. For each method including ours, we input the clustering number as the number of ground-truth classes from all data sets.*

Feature noise modeling for robustness: Following [Siyahjani *et al.*, 2015; Wang *et al.*, 2016a], 20% feature elements are corrupted with uniform distribution over the range [5,-5], which is consistent to the practical setting while matching with **LRRGL**,**RLRR** and our method.

Following [Wang *et al.*, 2016a], We set $\lambda_1 = 2$ in Eq.(9) for sparse noise term. We test ACC and NMI over different value of λ_2 and β in Eq.(9) in the next subsection.

Table 3: ACC results.

ACC (%)	UCI digits	AwA	NUS	VOC
MFMSC [Huang *et al.*, 2010]	43.81	17.13	22.81	12.98
MAASC [Huang *et al.*, 2012]	51.74	19.44	25.13	13.64
CCAMSC[Chaudhuri *et al.*, 2009]	73.24	24.04	27.56	12.05
CoMVSC[Kumar *et al.*, 2011]	80.27	29.93	33.63	14.03
Co-training[Kumar and Daume, 2011]	79.22	29.06	34.25	14.92
RLRR[Xia *et al.*, 2014]	83.67	31.49	35.27	17.13
LRRGL[Wang *et al.*, 2016a]	86.39	37.22	41.02	18.07
Ours	**89.64**	**41.76**	**43.14**	**18.85**

Table 4: NMI results.

NMI (%)	UCI digits	AwA	NUS	VOC
MFMSC [Huang *et al.*, 2010]	41.57	11.48	12.21	9.16
MAASC [Huang *et al.*, 2012]	47.85	12.93	11.86	9.84
CCAMSC[Chaudhuri *et al.*, 2009]	56.51	15.62	14.56	8.42
CoMVSC[Kumar *et al.*, 2011]	63.82	17.30	7.07	9.97
Co-training[Kumar and Daume, 2011]	62.07	18.05	8.10	10.88
RLRR[Xia *et al.*, 2014]	81.20	25.57	18.29	11.65
LRRGL[Wang *et al.*, 2016a]	85.45	31.74	20.61	12.03
Ours	**87.81**	**34.03**	**23.43**	**12.97**

4.3 Validation over factorized low-rank latent data-cluster representation

First, we will would like to validate our method regarding the multi-graph regularization and iterative views agreement over factorized latent data-cluster representation.

Following [Wang *et al.*, 2016a], we test λ_2 and β within the interval [0.001,10], with one parameter while fixing the value of the other parameter, the ACC results are shown in Fig. 2, where we have

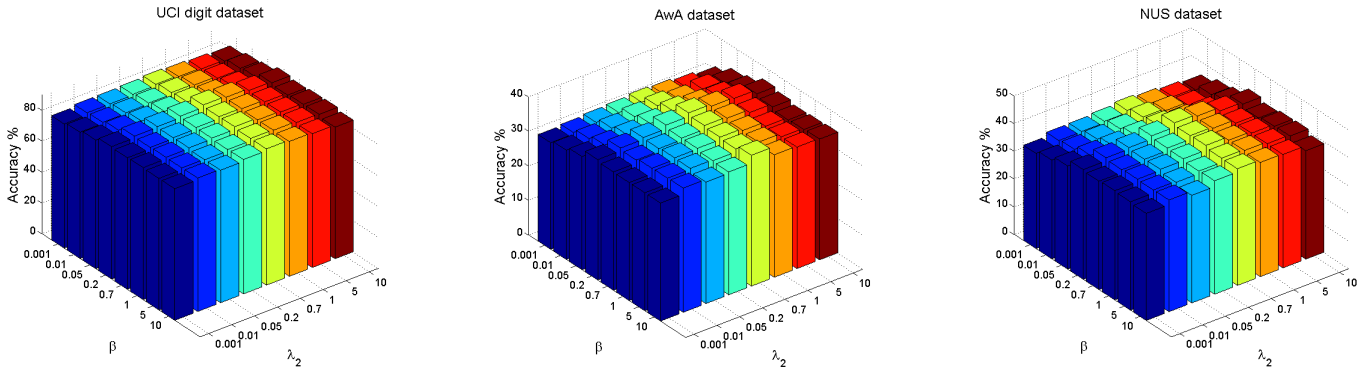
- Increasing β will improve the performance, and vice versa; that is, increasing λ_2 will improve the performance.
- The clustering metric ACC increases when both λ_2 and β increase.

Based on the above, we choose a balance pair values: $\lambda_2 = 0.7$ and $\beta = 0.2$ for our method.

4.4 Results

According to Table 3 and Table 4, the following identification can be drawn, note that we mainly deliver the analysis between our method and **LRRGL**, as the analysis over other competitors have been detailed in [Wang *et al.*, 2016a].

- First, our method outperforms **LRRGL**, implying the effectiveness of the factorized latent data-cluster representation, as it can better encode the data-cluster representation for each view as well as all views. We provide more insights about that in Fig. 3.
- Second, both our method and **LRRGL** outperforms the model of learning a common low-dimensional subspace among multi-view data, as indicated by [Wang *et al.*, 2016a] it is incapable of encoding local graph structures within a single subspace.
- Our method and **LRRGL** are more effective under noise corruptions than other methods. More analysis can be referred to our conference version [Wang *et al.*, 2016a].



(a) (b) (c)

Figure 2: Study over λ_2 and β over latent factorized data-cluster representation on three datasets.

- Our method achieves the best performance over PASCAL VOC 2012 under the selected two views via the tuned the parameters.

We present Fig. 3 to show more intuitions on why our method with the multi-view affinity matrix yielded from factorized data-cluster representation outperforms the primal similarity matrix for **LRRGL**. For example,

- For UCI dataset, *i.e.,* the multi-view affinity matrix illustrated in Fig. 3(a) and 3(d), we can see the both 4th and 5th diagonal blocks of our method in Fig. 3(d) are more whiter than those of **LRRGL** illustrated in Fig. 3(a); meanwhile the surrounding non-diagonal black blocks *e.g.,* $(4,5)^{th}$ and $(5,4)^{th}$ are more black than those of textbfLRRGL.
- For AwA dataset, the diagonal blocks of our method from the 2nd to the 6th are whiter than those of **LRRGL**, leading to a slight deeper black color over the surrounding non-diagonal blocks than **LRRGL**.
- The similar conclusions also hold for NUS dataset, we can see the diagonal blocks from 3rd to the 8th of our method are more whiter than those of **LRRGL**, leading to the result that the surrounding black non-diagonal blocks of our method are more salient than those of **LRRGL**.

From the above observations, we can safely infer the advantages of the affinity matrix representation yielded by our factorized latent data-cluster representation over the primal affinity matrix of **LRRGL** for Multi-view spectral clustering.

5 Conclusion

In this paper, we propose to learn a clustered low-rank representation via structured matrix factorization for multi-view spectral clustering. Unlike the existing methods, we propose an iterative strategy of intuitively achieving the multi-view spectral clustering agreement by minimizing the between-view divergences in terms of the factorized latent data-clustered representation for each view. Upon that, we impose the graph Laplacian regularizer over such low-dimensional data-cluster representation, so as to adapt to the multi-view spectral clustering, as demonstrated by the extensive experiments.

The future work includes the following directions: The graph regularized low-rank embedding out-of-sample case has been researched [Nie *et al.*, 2011], and will be applied for multi-view out-of-sample scenario. Unlike the pre-defined graph similarity value, inspired by [Nie *et al.*, 2017], we will simultaneously learn and achieve the consensus graph clustering result and graph structure *i.e.,* graph similarity. Besides, the latest non-parametric graph construction model [Nie *et al.*, 2016b] will also be incorporated for multi-view spectral clustering. The practice of our method can be improved by reducing the tuned parameters further. Upon that, we will also investigate the problem of learning the weight [Nie *et al.*, 2017; Nie *et al.*, 2016a; Cai *et al.*, 2013] for each view.

References

[Bertsekas, 1999] D. P. Bertsekas. Nonlinear programming. *Athena Scientific*, 1999.

[Bickel and Scheffer., 2004] S. Bickel and T. Scheffer. Multi-view clustering. In *IEEE International Conference on Data Mining*, pages 19–26, 2004.

[Blaschko and Lampert., 2008] M. Blaschko and C. Lampert. Correlational spectral clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[Blum and Mitchell, 1998] Avrim Blum and Tom M. Mitchell. Combining labeled and unlabeled sata with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pages 92–100, 1998.

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

[Cai *et al.*, 2008] Jian-Feng Cai, Emmanuel J. Candes, and Zuowei Shen. A singular value thresholding algorithm

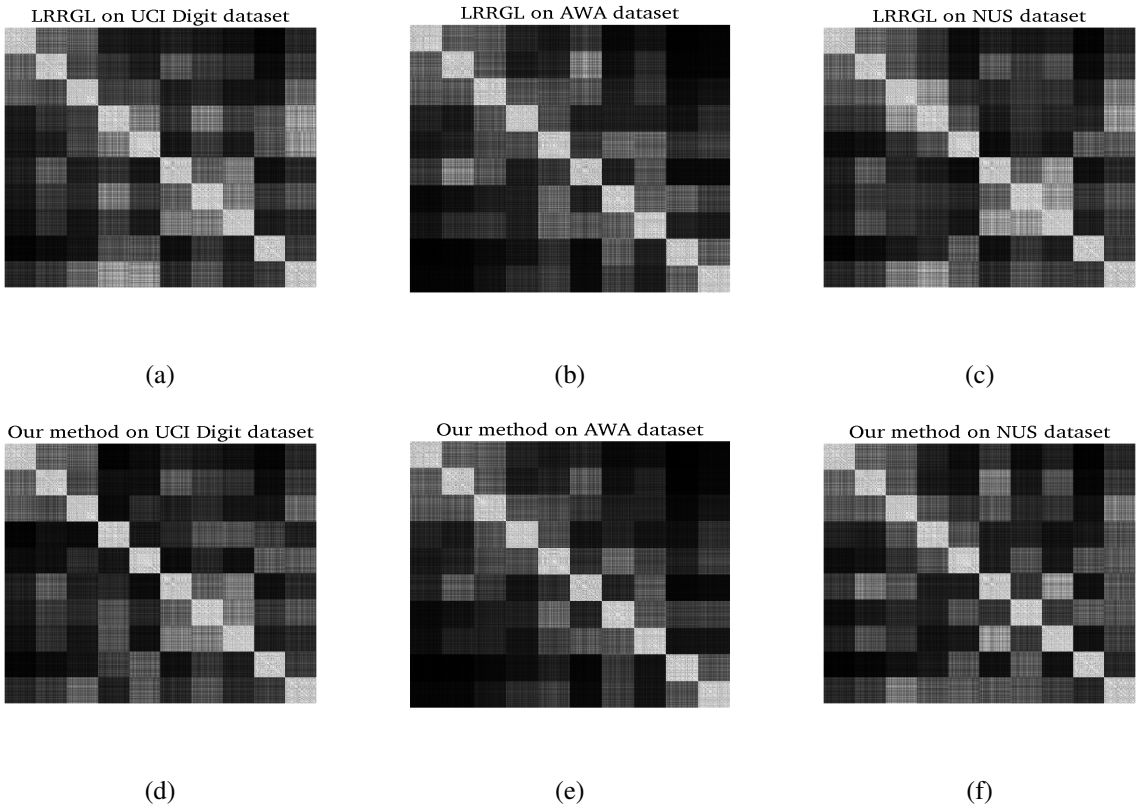


Figure 3: Recovered multi-view based consensus affinity matrix over both our proposed method and **LRRGL** on three multi-view data sets with noise corruption. For UCI digit dataset, we plot the affinity matrix over all data sample. For AWA and NUS datasets, we randomly select 10 classes, where 80 samples are randomly selected for each of them. The 10 diagonal block represents the data samples within the 10 clusters w.r.t ground truth classes, where more white the color is, the ideally larger affinity value will be to better reveal the data samples clusters within the same classes. Meanwhile, for non-diagonal blocks, the more black the color is, the ideally smaller affinity will be to reveal the data samples within different clusters w.r.t ground truth classes.

for matrix completion. *Society for Industrial and Applied Mathematics Journal on Optimization.*, 20(4):1956–1982, 2008.

[Cai *et al.*, 2013] X. Cai, F. Nie, and H. Huang. Multi-view k-means clustering on big data. In *AAAI*, 2013.

[Chaudhuri *et al.*, 2009] K. Chaudhuri, S. Kakade, K. Livescu, and K. Sridharan. Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th International Conference on Machine Learning*, pages 129–136, 2009.

[Chua *et al.*, 2009] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yan-Tao Zheng. Nus-wide: A real-world web image database from national university of singapore. In *Proceedings of the 8th ACM International Conference on Image and Video Retrieval*, 2009.

[Ding and Fu, 2016] Z Ding and Yun Fu. Robust multi-view subspace learning through dual low-rank decompositions. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1181–1187, 2016.

[Dingg and Fu, 2014] Zheng Dingg and Yun Fu. Low-rank common subspace for multi-view learning. In *14th IEEE International Conference on Data Mining*, pages 110–119, 2014.

[Gao *et al.*, 2013] Jing Gao, Jiawei Han, Jialu Liu, and Chi Wang. Multi-view clustering via joint nonnegative matrix factorization. In *SDM*, pages 252–260, 2013.

[Gao *et al.*, 2015] Hongchang Gao, Feiping Nie, Xuelong Li, and Heng Huang. Multi-view subspace clustering. In *IEEE International Conference on Computer Vision*, pages 4238–4246, 2015.

[Greene and Cunningham, 2009] D. Greene and P. Cunningham. A matrix factorization approach for integrating multiple data views. In *Machine Learning and Knowledge Discovery in Databases, European Conference, ECML PKDD*, pages 423–438, 2009.

[Gui *et al.*, 2014] Jie Gui, Dacheng Tao, Zhenan Sun, Yong Luo, Xinge You, and Yuan Yan Tang. Group sparse multiview patch alignment framework with view consistency

for image classification. *IEEE Transactions on Image Processing*, 23(7):3126–3137, 2014.

[Hou *et al.*, 2015] Chenping Hou, Feiping Nie, Dongyun Yi, and Dacheng Tao. Discriminative embedded clustering: A framework for grouping high-dimensional data. *IEEE Trans. Neural Netw. Learning Syst*, 26(6):1287–1299, 2015.

[Huang *et al.*, 2010] Y. Huang, Q. Liu, S. Zhang, and D. N. Metaxas. Image retrieval via probabilistic hypergraph ranking. In *IEEE Conf. Comput. Vis. Pattern Recognit*, pages 3376–3383, 2010.

[Huang *et al.*, 2012] Hsin-Chien Huang, Yung-Yu Chuang, and Chu-Song Chen. Affinity aggregation for spectral clustering. In *IEEE International Conference on Computer Vision and Pattern Recognition*, pages 773–780, 2012.

[Jones and Shao, 2014] Simon Jones and Ling Shao. A multigraph representation for improved unsupervised/semi-supervised learning of human actions. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 820–826, 2014.

[Kuang *et al.*, 2012] Da Kuang, Chris Ding, and Haesun Park. Symmetric nonnegative matrix factorization for graph clustering. In *SDM*, 2012.

[Kumar and Daume, 2011] Abhishek Kumar and Hal Daume. A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th International Conference on Machine Learning*, pages 393–400, 2011.

[Kumar *et al.*, 2011] Abhishek Kumar, Piyush Rai, and Hal Daume. Co-regularized multi-view spectral clustering. In *25th Annual Conference on Neural Information Processing Systems*, pages 1413–1421, 2011.

[Lin *et al.*, 2011] Zhouchen Lin, Risheng Liu, and Zhixun Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *25th Annual Conference on Neural Information Processing Systems*, pages 612–620, 2011.

[Liu and Yan, 2011] Guangcan Liu and Shuicheng Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *IEEE International Conference on Computer Vision*, 2011.

[Liu *et al.*, 2010] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning*, pages 663–670, 2010.

[Liu *et al.*, 2013] Guangcan Liu, Zhuochen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(1):171–184, 2013.

[Liu *et al.*, 2015] L. Liu, M. Yu, and L. Shao. Multiview alignment hashing for efficient image search. *IEEE Transactions on Image Processing*, 24(3):956–966, 2015.

[Ng *et al.*, 2001] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*, 2001.

[Nie *et al.*, 2011] Feiping Nie, Zinan Zeng, Ivor W. Tsang, Dong Xu, and Changshui Zhang. Spectral embedded clustering: A framework for in-sample and out-of-sample spectral clustering. *IEEE Trans Neural Networks*, 22(11):1796–1808, 2011.

[Nie *et al.*, 2016a] F. Nie, J. Li, and X. Li. Parameter-free auto-weighted multiple graph learning: A framework for multiview clustering and semi-supervised classification. In *IJCAI*, 2016.

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

[Nie *et al.*, 2017] Feiping Nie, Guohao Cai, and Xuelong Li. Multi-view clustering and semi-supervised classification with adaptive neighbours. In *AAAI*, pages 2408–2414, 2017.

[Recht *et al.*, 2010] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 45(8):1669–1680, 2010.

[Shao *et al.*, 2016] Ling Shao, Li Liu, and Mengyang Yu. Kernelized multiview projection for robust action recognition. *International Journal of Computer Vision*, 118(2):115–129, 2016.

[Siyahjani *et al.*, 2015] Farzad Siyahjani, Ranya Almohsen, Sinan Sabri, and Gianfranco Doretto. A supervised low-rank method for learning invariant subspace. In *IEEE International Conference on Computer Vision*, pages 4220–4228, 2015.

[Tang *et al.*, 2017] Mengfan Tang, Feiping Nie, and Ramesh Jain. A graph regularized dimension reduction method for out-of-sample data. *Neurocomputing*, 225:58–63, 2017.

[Tao *et al.*, 2016] Hong Tao, Chenping Hou, Feiping Nie, Yuanyuan Jiao, and Dongyun Yi. Effective discriminative feature selection with nontrivial solution. *IEEE Trans. Neural Netw. Learning Syst*, 27(4):796–808, 2016.

[Wang and Zhou, 2010] Wei Wang and Zhi-Hua Zhou. A new analysis of co-training. In *Proceedings of the 27th International Conference on Machine Learning*, pages 1135–1142, 2010.

[Wang *et al.*, 2012] Yang Wang, Lin Wu, Xiaodi Huang, and Xuemin Lin. Human action recognition from video sequences by enforcing tri-view constraints. *The Computer Journal*, 55(9):1030–1040, 2012.

[Wang *et al.*, 2013a] Yang Wang, Xiaodi Huang, and Lin Wu. Clustering via geometric median shift over riemannian manifolds. *Information Sciences*, 220:292–305, 2013.

[Wang *et al.*, 2013b] Yang Wang, Xuemin Lin, and Qing Zhang. Towards metric fusion on multi-view data: a cross-view based graph random walk approach. In *ACM CIKM*, pages 805–810, 2013.

[Wang *et al.*, 2014a] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. Exploiting correlation

consensus: Towards subspace clustering for multi-modal data. In *ACM Multimedia*, pages 981–984, 2014.

[Wang *et al.*, 2014b] Yang Wang, Xuemin Lin, Qing Zhang, and Lin Wu. Shifting hypergraphs by probabilistic voting. In *PAKDD*, pages 234–246, 2014.

[Wang *et al.*, 2014c] Yang Wang, Jian Pei, Xuemin Lin, Qing Zhang, and Wenjie Zhang. An iterative fusion approach to graph-based semi-supervised learning from multiple views. In *PAKDD*, 2014.

[Wang *et al.*, 2015a] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions:robust landmark retrieval. In *ACM Multimedia*, pages 79–88, 2015.

[Wang *et al.*, 2015b] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, and Qing Zhang. Lbmch: Learning bridging mapping for cross-modal hashing. In *ACM SIGIR*, pages 999–1002, 2015.

[Wang *et al.*, 2015c] Yang Wang, Xuemin Lin, Lin Wu, Wenjie Zhang, Qing Zhang, and Xiaodi Huang. Robust subspace clustering for multi-view data by exploiting correlation consensus. *IEEE Transactions on Image Processing*, 24(11):3939–3949, 2015.

[Wang *et al.*, 2016a] Y. Wang, W. Zhang, L. Wu, X. Lin, M. Fang, and S. Pan. Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In *The 25th International Joint Conference on Artificial Intelligence, New York, USA*, pages 2153–2159, 2016.

[Wang *et al.*, 2016b] Yang Wang, Xuemin Lin, Lin Wu, Qing Zhang, and Wenjie Zhang. Shifting multi-hypergraphs via collaborative probabilistic voting. *Knowledge and Information Systems*, 46(3):515–536, 2016.

[Wang *et al.*, 2017a] Yang Wang, Xuemin Lin, Lin Wu, and Wenjie Zhang. Effective multi-query expansions:collaborative deep networks for robust landmark retrieval. *IEEE Transactions on Image Processing*, 26(3):1393–1404, 2017.

[Wang *et al.*, 2017b] Yang Wang, Wenjie Zhang, Lin Wu, Xuemin Lin, and Xiang Zhao. Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. *IEEE Transactions on Neural Networks and Learning Systems*, 28(1):57–70, 2017.

[Wu and Wang, 2017] Lin Wu and Yang Wang. Robust hashing for multi-view data: Jointly learning low-rank kernelized similarity consensus and hash functions. *Image and Vision Computing*, 57:58–66, 2017.

[Wu *et al.*, 2013a] Lin Wu, Yang Wang, and John Shepherd. Co-ranking images and tags via random walks on a heterogeneous graph. In *MMM*, pages 228–238, 2013.

[Wu *et al.*, 2013b] Lin Wu, Yang Wang, and John Shepherd. Efficient image and tag co-ranking: a bregman divergence optimization method. In *ACM Multimedia*, pages 593–596, 2013.

[Wu *et al.*, 2015] Lin Wu, Xiaodi Huang, John Shepherd, and Yang Wang. Multi-query augmentation-based web landmark photo retrieval. *The Computer Journal*, 58(9):2120–2134, 2015.

[Wu *et al.*, 2016] Lin Wu, Yang Wang, and Shirui Pan. Exploiting attribute correlations: A novel trace lasso based weakly supervised dictionary learning method. *IEEE Transactions on Cybernetics*, 2016.

[Wu *et al.*, 2017a] Lin Wu, Yang Wang, Junbin Gao, and Xue Li. Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition*, 2017.

[Wu *et al.*, 2017b] Lin Wu, Yang Wang, Xue Li, and Junbin Gao. What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. *arXiv:1707.07074*, 2017.

[Xia *et al.*, 2014] Rongkai Xia, Yan Pan, Lei Du, and Jian Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, pages 2149–2155, 2014.

[Xu *et al.*, 2013] Chang Xu, Dacheng Tao, and Chao Yu. A survey on multi-view learning. *arxiv.org 1304.5634*, 2013.

[Xu *et al.*, 2015] Chang Xu, Dacheng Tao, and Chao Xu. Multi-view intact space learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, (12):2531–2544, 2015.

[Yin *et al.*, 2015] M. Yin, J. Gao, Z. Lin, Qinfeng Shi, and Yi Guo. Dual graph regularized latent low-rank representation for subspace clustering. *IEEE Trans. Image Processing*, 24(12):4918–4933, 2015.

[Yin *et al.*, 2016] M. Yin, J. Gao, and Z. Lin. Laplacian regularized low-rank representation and its applications. *IEEE Trans. Pattern Anal. Mach. Intell*, 38(3):504–517, 2016.

[Zelnik-Manor and Perona, 2004] Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems*, 2004.

[Zhuang *et al.*, 2016] Liansheng Zhuang, Jingjing Wang, Zhouchen Lin, Allen Y. Yang, Yi Ma, and Nenghai Yu. Locality-preserving low-rank representation for graph construction from nonlinear manifolds. *Neurocomputing*, 175:715–722, 2016.