Latent Multi-view Subspace Clustering

Changqing Zhang[1], Qinghua Hu[1]*, Huazhu Fu[2], Pengfei Zhu[1], Xiaochun Cao[3,4]
[1]School of Computer Science and Technology, Tianjin University
[2]Institute for Infocomm Research, Agency for Science, Technology and Research
[3]State Key Laboratory of Information Security, IIE, Chinese Academy of Sciences
[4]School of Cyber Security, University of Chinese Academy of Sciences
{zhangchangqing,huqinghua}@tju.edu.cn

Abstract

In this paper, we propose a novel Latent Multi-view Subspace Clustering (LMSC) method, which clusters data points with latent representation and simultaneously explores underlying complementary information from multiple views. Unlike most existing single view subspace clustering methods that reconstruct data points using original features, our method seeks the underlying latent representation and simultaneously performs data reconstruction based on the learned latent representation. With the complementarity of multiple views, the latent representation could depict data themselves more comprehensively than each single view individually, accordingly makes subspace representation more accurate and robust as well. The proposed method is intuitive and can be optimized efficiently by using the Augmented Lagrangian Multiplier with Alternating Direction Minimization (ALM-ADM) algorithm. Extensive experiments on benchmark datasets have validated the effectiveness of our proposed method.

1. Introduction

Subspace clustering is a fundamental and important technique in many applications, especially for the high dimensional data. Generally, subspace clustering methods [7, 18, 12] hold the assumption that data points are drawn from multiple subspaces corresponding to different clusters. Recently, the subspace clustering based on self-representation has been proposed, where each data point can be expressed with a linear combination of the data points themselves. The general formulation can be presented as

$$\min_{\mathbf{Z}} L(\mathbf{X}, \mathbf{XZ}) + \alpha\Omega(\mathbf{Z}), \tag{1}$$

where the scalar $\alpha > 0$ balances the reconstruction error and the regularization for subspace representation $\mathbf{Z}$. $L(\cdot,\cdot)$ and $\Omega(\cdot)$ denote the loss function and regularization term, respectively, which are usually defined based on different assumptions. For example, Sparse Subspace Clustering (SSC) [7] searches a sparsest representation among the infinitely many possible representations based on ℓ_1-norm. Low-Rank Representation Clustering (LRR) [18] tries to reveal cluster structure with a low-rank representation. SMooth Representation clustering (SMR) [12] analyzes the grouping effect of self-representation based methods in depth. Based on the self-representation matrix $\mathbf{Z}$, the similarity matrix is often constructed with $\mathbf{S} = abs(\mathbf{Z}) + abs(\mathbf{Z}^T)$, where $abs(\cdot)$ is the element-wise absolute operator. Finally, based on the similarity matrix $\mathbf{S}$, spectral clustering algorithm is usually performed for the final clustering result [7, 18, 12].

These subspace clustering methods achieve promising performances, however, they are usually affected by the quality of original features, especially under the condition that the observations are insufficient and/or grossly corrupted. Therefore, the multi-view subspace clustering methods have been proposed [3, 31, 9], in which each data point is described with information from multiple sources of features. These multi-view representations hold rich information from multiple cues, which could be beneficial to clustering task. With proper multi-view constraints, these subspace clustering methods have shown their power. They usually reconstruct the data points on the original view directly, and generate the individual subspace representation for each view. However, each single view alone is usually not sufficient to describe data points, which makes the reconstruction by using only one view itself risky. Moreover, the data collection may be noisy, which further increases the difficulty of clustering.

To address these issues, in this paper, we introduce a latent representation to explore the relationships among data points and handle the possible noise. As agreed by [11, 26], we assume that multiple views are originated from one underlying latent representation, which could depict the data

*Corresponding Author (Qinghua Hu)

in essence and reveal the common latent structure shared by different views. Based on this assumption, we propose the *Latent Multi-view Subspace Clustering* (LMSC) method. Our proposed method learns the latent representation based on multi-view features, and generates a common subspace representation rather than that of individual view. Moreover, our method integrates the latent representation learning and multi-view subspace clustering in a unified framework, which is optimized effectively by using the Augmented Lagrangian Multiplier with Alternating Direction Minimization strategy. Extensive experiments in comparison with several state-of-the-art methods are performed to assess the performance of our LMSC.

1.1. Related Work

Generally, most existing multi-view clustering methods belong to the category of *graph-based model*. The early methods (e.g., [6]) usually concentrate on the 2-view case. Some methods [20, 24] utilize *matrix factorization* technique for multi-view clustering. The subspace clustering methods [3, 31, 9] describe each data point with the data collection itself on the original view directly. Under the spectral clustering framework, the methods [16, 15] *co-regularize* the hypothesis to be consistent across these different views. To address the large scale issue, a robust *large-scale* multi-view clustering method [20] is proposed under the framework of K-means algorithm. Another nature way to integrate different views is *Multiple Kernel Learning (MKL)*. The work in [4] has demonstrated the effectiveness of direct combination of different kernels. Furthermore, the researchers in [25] proposed a more general way based on MKL to learn the weights of different kernels. It is noteworthy that, our method performs data reconstruction with the comprehensive multi-view latent representation, instead of each original single view [3, 31, 9].

Recently, researchers have extended the subspace clustering methods [7, 18] to latent representation based subspace clustering. The method Latent Space Sparse Subspace Clustering (LS3C) [22] simultaneously performs dimensionality reduction and sparse coding for SSC. Latent Low-Rank Representation (LatLRR) [19] is built on the top of LRR [18], and constructs the dictionary by using both observed and hidden data. For multi-view representation, some methods [11, 26] explicitly learns a common representation based on multiple views as a joint optimization problem with a common subspace representation matrix. Different from LS3C which performs dimensionality reduction on the original single view data, our method recovers the latent multi-view representation, and the projections corresponding to different views are learned simultaneously under this latent representation. There are also some recent methods focusing on other topics, e.g., dimensionality reduction [30] and feature selection [23].

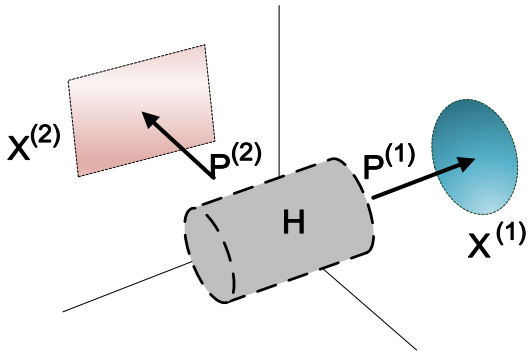


Figure 1: Illustration of multi-view latent representation. Observations $\{\mathbf{X}^{(v)}\}_{v=1}^{V}$ ($V \geq 2$) corresponding to different views are partially projected by $\{\mathbf{P}^{(v)}\}_{v=1}^{V}$ from one underlying latent representation $\mathbf{H}$.

2. Proposed Approach

In this work, we consider subspace clustering with multi-view latent representation. Given N multi-view observations $\{[\mathbf{x}_i^{(1)}; ...; \mathbf{x}_i^{(V)}]\}_{i=1}^{N}$ which consist of V different views, our goal is to infer a shared latent representation, $\mathbf{h}$, for each data point. Our method assumes that these different views are all originated from one underlying latent representation. Specifically, as shown in Fig. 1, the observations from different views can be reconstructed by their respective models $\{\mathbf{P}^{(1)}, ..., \mathbf{P}^{(V)}\}$ with the shared latent representations $\mathbf{H} = \{\mathbf{h}_i\}_{i=1}^{N}$. Accordingly, we have $\mathbf{x}_i^{(v)} = \mathbf{P}^{(v)}\mathbf{h}_i$. Considering the noise, it is

$$\mathbf{x}_i^{(v)} = \mathbf{P}^{(v)}\mathbf{h}_i + \mathbf{e}_i^{(v)}, \tag{2}$$

where $\mathbf{e}_i^{(v)}$ denotes the reconstruction error corresponding to the v^{th} view. The objective function to infer the multi-view latent representation is as follows

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{H}} L_h(\mathbf{X}, \mathbf{PH}), \\ &\text{with } \mathbf{X} = \begin{bmatrix} \mathbf{X}^{(1)} \\ \cdots \\ \mathbf{X}^{(V)} \end{bmatrix} \text{ and } \mathbf{P} = \begin{bmatrix} \mathbf{P}^{(1)} \\ \cdots \\ \mathbf{P}^{(V)} \end{bmatrix}, \end{aligned} \tag{3}$$

where $\mathbf{X}$ and $\mathbf{P}$ are the multi-view observations and reconstruction models aligned, respectively. $L_h(\cdot, \cdot)$ denotes the loss functions associated with the latent (hidden) representation. Generally, with the help of complementarity from multiple views, the latent representation $\mathbf{H}$ is more comprehensive than the representation corresponding to each single view individually.

Then, based on the latent representation $\mathbf{H}$, the objective function of self-representation based subspace clustering of Eq. (1) is reformulated as

$$\min_{\mathbf{Z}} L_r(\mathbf{H}, \mathbf{HZ}) + \alpha\Omega(\mathbf{Z}), \tag{4}$$

where $L_r(\cdot,\cdot)$ denotes the loss functions associated with the data reconstruction and $\mathbf{Z}$ is the reconstruction coefficient matrix.

We integrate the latent representation learning in Eq. (3) and subspace clustering in Eq. (4) into one unified objective function, shown as follows

$$\min_{\mathbf{P},\mathbf{H},\mathbf{Z}} L_h(\mathbf{X},\mathbf{PH}) + \lambda_1 L_r(\mathbf{H},\mathbf{HZ}) + \lambda_2 \Omega(\mathbf{Z}), \tag{5}$$

where $\lambda_1 > 0$ and $\lambda_2 > 0$ balance the three terms. The subspace clustering is guaranteed by the reasonable latent representation and the constraint of subspace reconstruction, while the latent representation is guaranteed by the complementarity of multiple views and improved by the subspace reconstruction. Considering the robustness for outliers, our final objective function is as follows

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{H},\mathbf{Z},\mathbf{E}_h,\mathbf{E}_r} \|\mathbf{E}_h\|_{2,1} + \lambda_1 \|\mathbf{E}_r\|_{2,1} + \lambda_2 \|\mathbf{Z}\|_* \\ &s.t.\ \mathbf{X} = \mathbf{PH} + \mathbf{E}_h, \mathbf{H} = \mathbf{HZ} + \mathbf{E}_r \text{ and } \mathbf{PP}^T = \mathbf{I}, \end{aligned} \tag{6}$$

where $||\cdot||_*$ is the matrix nuclear norm, which enforces the subspace representation to be low-rank. $||\cdot||_{2,1}$ is called $\ell_{2,1}$-norm which encourages the columns of a matrix to be zero [18], and the definition for the $\ell_{2,1}$- norm for a matrix $(\mathbf{A})$ is: $\|\mathbf{A}\|_{2,1} = \sum_{j=1}^{D} \sqrt{\sum_{i=1}^{C} \mathrm{A}_{ij}^2}$ with $\mathbf{A} \in \mathbb{R}^{C\times D}$. The underlying assumption is that the corruptions are sample-specific. We constrain $\mathbf{P}$ since without constraint $\mathbf{H}$ can be pushed arbitrarily close to zero only by re-scaling $\mathbf{H}/s$ and $\mathbf{P}s$ $(s > 0)$ while preserving the same loss. The first term is utilized to assure the learned latent representations $\mathbf{H}$ and reconstruction models $\mathbf{P}^{(v)}$ associated to different views to be good for reconstructing the observations, while the second one penalizes the reconstruction error in the latent multi-view subspaces. The last term prevents the trivial solution by enforcing the subspace representation to be low-rank. The robustness of our method benefits from two aspects. Firstly, due to the complementary information of multiple views, the latent multi-view representation can depict data more comprehensively than the single view and accordingly leads to subsequent more promising clustering result. Secondly, the $\ell_{2,1}$-norm on the first two term is a matrix block norm, which is more robust to outliers than the Frobenius norm.

Furthermore, we vertically concatenate together along the column of errors corresponding to the latent representation and the subspace representation. In the way of integration, it will enforce the columns of $\mathbf{E}_h$ and $\mathbf{E}_r$ to have jointly consistent magnitude values, and the effectiveness of which has been widely proved. Then, the objective function of our proposed method has the following form

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{H},\mathbf{Z},\mathbf{E}_h,\mathbf{E}_r} \|\mathbf{E}\|_{2,1} + \lambda \|\mathbf{Z}\|_* \\ &s.t.\ \mathbf{X} = \mathbf{PH} + \mathbf{E}_h, \mathbf{H} = \mathbf{HZ} + \mathbf{E}_r, \\ &\mathbf{E} = [\mathbf{E}_h; \mathbf{E}_r] \text{ and } \mathbf{PP}^T = \mathbf{I}. \end{aligned} \tag{7}$$

Then, there is one parameter $\lambda > 0$ which balances the error and regularization.

3. Optimization

Our objective function in Eq. (7) simultaneously learns the latent representations from multiple views and finds the meaningful similarity matrix with respect to the latent representations. Although the objective function is not jointly convex with respect to all the variables $\mathbf{P}$, $\mathbf{H}$, $\mathbf{Z}$, $\mathbf{E}_h$ and $\mathbf{E}_r$, each of them can be solved efficiently by fixing the others. The Augmented Lagrange Multiplier (ALM) with Alternating Direction Minimizing (ADM) strategy [17] is an efficient and effective solver for our problems. To adopt ADM strategy to our problem, we need to make our objective function separable. Therefore, we introduce one auxiliary variable $\mathbf{J}$ to replace $\mathbf{Z}$ in the nuclear term of our objective function. Then we have the following equivalent problem

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{H},\mathbf{Z},\mathbf{E}_h,\mathbf{E}_r,\mathbf{J}} \|\mathbf{E}\|_{2,1} + \lambda \|\mathbf{J}\|_* \\ &s.t.\ \mathbf{X} = \mathbf{PH} + \mathbf{E}_h,\ \ \mathbf{H} = \mathbf{HZ} + \mathbf{E}_r, \\ &\mathbf{E} = [\mathbf{E}_h; \mathbf{E}_r],\ \ \mathbf{PP}^T = \mathbf{I} \text{ and } \mathbf{J} = \mathbf{Z}. \end{aligned} \tag{8}$$

The above objective function can be solved by minimizing the following ALM problem

$$\begin{aligned} &\mathcal{L}(\mathbf{P},\mathbf{H},\mathbf{Z},\mathbf{E}_h,\mathbf{E}_r,\mathbf{J}) \\ &= \|\mathbf{E}\|_{2,1} + \lambda \|\mathbf{J}\|_* \\ &+ \Phi(\mathbf{Y}_1, \mathbf{X} - \mathbf{PH} - \mathbf{E}_h) \\ &+ \Phi(\mathbf{Y}_2, \mathbf{H} - \mathbf{HZ} - \mathbf{E}_r) + \Phi(\mathbf{Y}_3, \mathbf{J} - \mathbf{Z}) \\ &s.t.\ \mathbf{PP}^T = \mathbf{I}. \end{aligned} \tag{9}$$

Note that, for convenience, we give the following definition: $\Phi(\mathbf{C},\mathbf{D}) = \frac{\mu}{2}||\mathbf{D}||_F^2 + \langle \mathbf{C},\mathbf{D}\rangle$, where $\langle\cdot,\cdot\rangle$ defines the matrix inner product and μ is a positive penalty scalar. To optimize our problem with ALM-ADM, we separate our problem into the following subproblems.

1. P-subproblem: To update $\mathbf{P}$, we solve the following optimization problem by fixing the other variables

$$\begin{aligned} &\mathbf{P}^* = \arg\min \Phi(\mathbf{Y}_1, \mathbf{X} - \mathbf{PH} - \mathbf{E}_h) \\ &s.t.\ \mathbf{PP}^T = \mathbf{I}. \end{aligned} \tag{10}$$

Theorem 1. *[13] Given the objective function* $\min_{\mathbf{R}} ||\mathbf{Q} - \mathbf{GR}||_F^2$ *s.t.* $\mathbf{R}^T\mathbf{R} = \mathbf{RR}^T = \mathbf{I}$*, the optimal solution is* $\mathbf{R} = \mathbf{UV}^T$*, where* $\mathbf{U}$ *and* $\mathbf{V}$ *are left and right singular values of SVD decomposition of* $\mathbf{G}^T\mathbf{Q}$.

It is not difficult to show that, the optimal solution of $\mathbf{P}$-subproblem is $\mathbf{P}^T = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular values of SVD of $\mathbf{H}(\mathbf{Y}_1 + \mathbf{X} - \mathbf{E}_h)^T$, since we have

$$\begin{aligned}\mathbf{P}^* &= \arg\min \Phi(\mathbf{Y}_1, \mathbf{X} - \mathbf{P}\mathbf{H} - \mathbf{E}_h) \\ &= \arg\min \frac{\mu}{2}||\mathbf{X} - \mathbf{P}\mathbf{H} - \mathbf{E}_h + \mathbf{Y}_1||_F^2 \\ &= \arg\min \frac{\mu}{2}||(\mathbf{X} + \mathbf{Y}_1/\mu - \mathbf{E}_h) - \mathbf{P}\mathbf{H}||_F^2 \\ &= \arg\min \frac{\mu}{2}||(\mathbf{X} + \mathbf{Y}_1/\mu - \mathbf{E}_h)^T - \mathbf{H}^T\mathbf{P}^T||_F^2.\end{aligned}$$

Based on Theorem 1, if we constrain $\mathbf{P}$ to be an orthonormal matrix (i.e., $\mathbf{P}\mathbf{P}^T = \mathbf{P}^T\mathbf{P} = \mathbf{I}$), the optimal solution of $\mathbf{P}$-subproblem is $\mathbf{P}^T = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular values of SVD decomposition of $\mathbf{H}(\frac{\mathbf{Y}_1}{\mu} + \mathbf{X} - \mathbf{E}_h)^T$. For efficiency, we can relax $\mathbf{P}$ to be row orthogonal in practice (i.e., $\mathbf{P}\mathbf{P}^T = \mathbf{I}$, where $\mathbf{P} \in \mathbb{R}^{k\times d}, k \ll d$), and the promising performance and convergence are also achieved in practice.

2. H-subproblem: By fixing the others variables, we update $\mathbf{H}$ by the following rule

$$\begin{aligned}\mathbf{H}^* = \arg\min \Phi(\mathbf{Y}_1, &\mathbf{X} - \mathbf{P}\mathbf{H} - \mathbf{E}_h) \\ &+ \Phi(\mathbf{Y}_2, \mathbf{H} - \mathbf{H}\mathbf{Z} - \mathbf{E}_r).\end{aligned} \tag{11}$$

Taking the derivative with respect to $\mathbf{H}$ and setting it to zero, we get

$$\begin{aligned}&\mathbf{A}\mathbf{H} + \mathbf{H}\mathbf{B} = \mathbf{C} \\ \text{with } &\mathbf{A} = \mu\mathbf{P}^T\mathbf{P}, \mathbf{B} = \mu(\mathbf{Z}\mathbf{Z}^T - \mathbf{Z} - \mathbf{Z}^T + \mathbf{I}), \\ &\mathbf{C} = (\mathbf{P}^T\mathbf{Y}_1 + \mathbf{Y}_2(\mathbf{Z}^T - \mathbf{I})) \\ &+ \mu(\mathbf{P}^T\mathbf{X} + \mathbf{E}_r^T - \mathbf{P}^T\mathbf{E}_h - \mathbf{E}_r\mathbf{Z}^T).\end{aligned} \tag{12}$$

The above equation is a Sylvester equation [1]. To avoid instability issue, we ensure $\mathbf{A}$ to be strictly positive definite by $\hat{\mathbf{A}} = \mathbf{A} + \epsilon\mathbf{I}$, where $\mathbf{I}$ is a identity matrix and $0 < \epsilon \ll 1$.

Proposition 1. *The Sylvester equation (12) has a unique solution.*

Proof. The Sylvester equation $\mathbf{A}\mathbf{H} + \mathbf{H}\mathbf{B} = \mathbf{C}$ has a unique solution for $\mathbf{H}$ exactly when there are no common eigenvalues of $\mathbf{A}$ and -$\mathbf{B}$ [1]. Since $\hat{\mathbf{A}}$ is a positive definite matrix, so all of its eigenvalues are positive: $\alpha_i > 0$. While since $\mathbf{B}$ is a positive semi-definite matrix, so all of its eigenvalues are nonnegative: $\beta_i \geq 0$. Hence, for any eigenvalues of $\mathbf{A}$ and $\mathbf{B}$, $\alpha_i + \beta_j > 0$. Accordingly, the Sylvester equation (12) has a unique solution. □

Remark: For solving the Sylvester equation, Bartels-Stewart algorithm [1] is employed. The algorithm firstly transforms the coefficient matrices into Schur forms by QR decomposition. Then it solves the obtained triangular system by back-substitution. It also noteworthy that under $\mathbf{P}\mathbf{P}^T = \mathbf{P}\mathbf{P}^T = \mathbf{I}$, our method could be exactly solved, i.e., $\mathbf{A} = \mathbf{P}^T\mathbf{P}$ is strictly positive definite without introducing a perturbation (for $\mathbf{H}$-subproblem) and Theorem 1 is exactly suitable for $\mathbf{P}$-subproblem. For efficiency in practice use, we can learn a low dimensional latent representation by relaxing P to be row orthogonal, i.e., $\mathbf{P}\mathbf{P}^T = \mathbf{I}$, then we should introduce a small perturbation to ensure it strictly positive definite.

3. Z-subproblem: Fix the other variables, we update $\mathbf{Z}$ by solving the following problem

$$\mathbf{Z}^* = \arg\min_{\mathbf{Z}} \Phi(\mathbf{Y}_3, \mathbf{J} - \mathbf{Z}) + \Phi(\mathbf{Y}_2, \mathbf{H} - \mathbf{H}\mathbf{Z} - \mathbf{E}_r). \tag{13}$$

Taking the derivative with respect to $\mathbf{Z}$ and setting it to zero, we get

$$\begin{aligned}\mathbf{Z}^* = (\mathbf{H}^T\mathbf{H} + \mathbf{I})^{-1}[&(\mathbf{J} + \mathbf{H}^T\mathbf{H} - \mathbf{H}^T\mathbf{E}_r) \\ &+ (\mathbf{Y}_3 + \mathbf{H}^T\mathbf{Y}_2)/\mu].\end{aligned} \tag{14}$$

4. E-subproblem: The reconstruction error $\mathbf{E}$ is updated by solving the following problem

$$\begin{aligned}\mathbf{E}^* &= \arg\min_{\mathbf{E}} \|\mathbf{E}\|_{2,1} + \Phi(\mathbf{Y}_1, \mathbf{X} - \mathbf{P}\mathbf{H} - \mathbf{E}_h) \\ &\qquad + \Phi(\mathbf{Y}_2, \mathbf{H} - \mathbf{H}\mathbf{Z} - \mathbf{E}_r) \\ &= \arg\min_{\mathbf{E}} \frac{1}{\mu}\|\mathbf{E}\|_{2,1} + \frac{1}{2}\|\mathbf{E} - \mathbf{G}\|_F^2,\end{aligned} \tag{15}$$

where $\mathbf{G}$ is formed by vertically concatenating the matrices $\mathbf{X} - \mathbf{P}\mathbf{H} + \mathbf{Y}_1/\mu$ and $\mathbf{H} - \mathbf{H}\mathbf{Z} + \mathbf{Y}_2/\mu$. This subproblem can be efficiently solved by Lemma 3.2 in [18].

5. J-subproblem: Fix the others, the Lagrange function with respect to $\mathbf{J}$ can be written as

$$\begin{aligned}\mathbf{J}^* &= \arg\min_{\mathbf{J}} \lambda\|\mathbf{J}\|_* + \Phi(\mathbf{Y}_3, \mathbf{J} - \mathbf{Z}) \\ &= \frac{\lambda}{\mu}\|\mathbf{J}\|_* + \frac{1}{2}\|\mathbf{J} - (\mathbf{Z} - \mathbf{Y}_3/\mu)\|_F^2.\end{aligned} \tag{16}$$

The above problem can be efficiently solved by the singular value thresholding operator [2].

6. Updating Multipliers: We update the multipliers by

$$\begin{cases}\mathbf{Y}_1 = \mathbf{Y}_1 + \mu(\mathbf{X} - \mathbf{P}\mathbf{H} - \mathbf{E}_h) \\ \mathbf{Y}_2 = \mathbf{Y}_2 + \mu(\mathbf{H} - \mathbf{H}\mathbf{Z} - \mathbf{E}_r) \\ \mathbf{Y}_3 = \mathbf{Y}_3 + \mu(\mathbf{J} - \mathbf{Z}).\end{cases} \tag{17}$$

Intuitively, the multipliers are updated proportionally to the violation of the equality constraints.

Note that, simply initializing the block variables $\mathbf{H}$ with zero is not appropriate, since in this way, the optimal $\mathbf{H}$ (see $\mathbf{H}$-subproblem in Eq. (11)) will be zero. Then, subsequent optimizations for all the other subproblems (e.g., $\mathbf{Z}$-subproblem in Eq. (13)) will be trivial. Based on this, we randomly initialize $\mathbf{H}$ in practice and we can also initialize $\mathbf{H}$ with other preprocessing ways (e.g., PCA) to avoid unstable results.

Algorithm 1: Optimization Algorithm for LMSC

Input: Multi-view matrices: $\{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(V)}\}$, hyperparameter λ and the dimension K of latent representation $\mathbf{H}$.

Initialize: $\mathbf{P} = 0$, $\mathbf{E}_r = 0$, $\mathbf{E}_h = 0$, $\mathbf{J} = \mathbf{Z} = 0$, $\mathbf{Y}_1 = 0$,$\mathbf{Y}_2 = 0$,$\mathbf{Y}_3 = 0$, $\mu = 10^{-6}$, $\rho = 1.1$, $\varepsilon = 10^{-4}$, $\max_\mu$ =10^6; Initialize $\mathbf{H}$ with random values.

while *not converged* **do**

Update variables $\mathbf{P}, \mathbf{H}, \mathbf{Z}, \mathbf{E}_h, \mathbf{E}_r, \mathbf{J}$ according to *subproblems 1-5*;

Update multipliers $\mathbf{Y}_1, \mathbf{Y}_2, \mathbf{Y}_3$ according to *subproblems 6*;

Update the parameter μ by μ = min($\rho\mu$; $\max_\mu$);

Check the convergence conditions: $||\mathbf{X} - \mathbf{PH} - \mathbf{E}_h||_\infty < \epsilon, ||\mathbf{H} - \mathbf{HZ} - \mathbf{E}_r)||_\infty < \epsilon$ and $||\mathbf{J} - \mathbf{Z}||_\infty < \epsilon$.

end

Output: $\mathbf{Z}$, $\mathbf{H}$, $\mathbf{P}$, $\mathbf{E}$.

3.1. Complexity and Convergence

Our method is composed of six sub-problems. The complete algorithm is shown in Algorithm 1. The complexity of updating $\mathbf{P}$ is $O(k^2d + d^3)$, where k, d and n are the dimension of the latent representation, the total dimensions of multi-view features, and the sample number of data, respectively. The complexities of the other subproblems are as follows: For updating $\mathbf{J}$ (the nuclear norm proximal operator), the complexity is $O(n^3)$. For updating $\mathbf{H}$, the classical algorithm for the Sylvester equation is the Bartels Stewart algorithm[1], whose complexity is $O(k^3)$. For updating $\mathbf{Z}$, the main complexity is the matrix inversion, which is $O(n^3)$. For updating $\mathbf{E}$ and the multipliers, the main complexity is the matrix multiplication, which is $O(dkn + kn^2)$. Overall, the total complexity is $O(k^2d+d^3+k^3+n^3+dkn+kn^2)$ for each iteration. Under the condition $k \ll d$, the total complexity is basically $O(d^3 + n^3)$. It is difficult to generally prove the convergence for our algorithm. Fortunately, empirical evidence on both synthesized and real data presented suggests that the proposed algorithm has very strong and stable convergence behavior even with initializing $\mathbf{H}$ randomly.

Remarks. 1) Linear projection employed in our model is a simple but effective technique for high-dimensional data, and actually it is easy to resolve in practice. The nonlinearity could be introduced into our model based on the kernel technique, which will be considered in our future work. 2) For $\mathbf{P}$-subproblem, although the strict correctness is given under the complete case (i.e., with $\mathbf{P}$ being a square matrix), the promising results are observed with low-dimensional projection in practice. Moreover, given other constraints (e.g., $||\mathbf{P}(:, j)||^2 \leq 1$) instead of $\mathbf{PP}^T = \mathbf{I}$, ADMM can be used to solve this subproblem [10]. Though similar performance achieved, the inner ADMM makes the algorithm more complex.

4. Experiments

4.1. Experiment Setting

We employ both synthetic data and real-world datasets for evaluation. Synthetic data is used to validate the effectiveness of using multiple views. These real-world datasets are as follows: **MSRCV1** [28] consists of 240 images and 8 object classes. We select 7 classes, i.e., tree, building, airplane, cow, face, car and bicycle, and extract 6 types of features: CENT (view1), CMT (view2), GIST (view3), HOG (view4), LBP (view5), (SIFT (view6) from each image to construct different view features. **Scene-15** [8] dataset contains 15 scene categories with both indoor and outdoor environments, 4485 images in total. Three common image features GIST (view1), PHOG (view2), and LBP (view3) are used similar to [5]. **ORL**[1] contains 10 different images of each of 40 distinct subjects. For Yale and ORL, three types of features: intensity (view1), LBP (view2) and Gabor (view3) are used. Each category has 200 to 400 images. **LandUse-21** [29] consists of satellite images from 21 categories, 100 images each. The features used are same to Scene-15. **Still DB** [14] consists of 467 images with 6 classes of actions. Three features are extracted, i.e., Sift Bow (view1), Color Sift Bow (view2) and Shape context Bow (view3). **BBCSport**[2] consists the documents from the BBC Sport website corresponding to sports news in 5 topical areas, which is associated with 2 views [27].

We compare our method with the following baselines. **SPC$_{\text{BestSV}}$** is the standard spectral clustering with the best single view. **LRR$_{\text{BestSV}}$** is the LRR [18] with the best single view. **Min-Disagreement** [6] creates a bipartite graph and is based on the minimizing-disagreement idea. We report the 2-view best results due to its limitation. **Co-Reg SPC** [16] co-regularizes the clustering hypothesis to enforce different views to be consistent. **RMSC** [27] recovers a shared low-rank transition probability matrix as input to the standard Markov chain.

For evaluation metrics, we use NMI (normalized mutual information), ACC (accuracy), F-measure and RI (rand index) to comprehensively evaluate the performance. Note that, higher values indicate better performance for all metrics. For the compared methods, we tune all the parameters to best performances. For our method, we set the dimension of the latent representation $K = 100$ and tune the parameter λ from $\{0.001, 0.01, 0.1, 1, 10, 100, 1000\}$ for all datasets.

[1] http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[2] http://mlg.ucd.ie/datasets/

Table 1: Performance comparison of clustering.

Datasets	Methods	NMI	ACC	F-measure	RI
MSRCV1	SPC_{BestSV}	57.42 ± 3.16	66.82 ± 5.08	53.54 ± 4.31	86.35 ± 0.63
	LRR_{BestSV}	49.21 ± 1.14	59.34 ± 0.13	45.31± 0.44	84.71 ± 0.08
	Min-Disagreement	60.64 ± 0.32	69.21 ± 3.43	57.48 ± 0.47	88.18 ± 0.09
	Co-Reg	56.92 ± 1.25	65.30 ± 1.66	53.71 ± 2.15	89.21 ± 0.30
	RMSC	58.57 ± 0.65	69.10 ± 0.71	57.63 ± 1.62	87.97 ± 0.21
	Ours	**65.34 ± 1.05**	**80.55 ± 1.28**	**65.17 ± 1.71**	**90.40 ± 0.22**
ORL	SPC_{BestSV}	90.35 ± 1.62	77.73 ± 3.36	71.12 ± 4.34	98.61 ± 0.23
	LRR_{BestSV}	88.56 ± 0.72	77.65± 0.41	73.12± 0.68	97.10± 0.07
	Min-Disagreement	81.66 ± 0.14	73.45 ± 4.08	66.30 ± 0.31	96.65 ± 0.17
	Co-Reg	84.56 ± 1.41	60.89 ± 1.88	63.38 ± 2.37	97.30 ± 0.22
	RMSC	89.76 ± 1.98	75.20 ± 2.84	70.11 ± 4.34	97.53 ± 0.23
	Ours	**93.10 ± 1.16**	**81.94 ± 1.71**	**75.83 ± 0.91**	**98.83 ± 0.22**
Scene-15	SPC_{BestSV}	28.32 ± 1.43	30.15 ± 1.29	21.23 ± 1.42	89.14 ± 0.13
	LRR_{BestSV}	30.02 ± 0.19	30.72± 0.26	20.98± 0.23	88.35 ± 0.13
	KernelAddition	27.51 ± 0.42	31.01 ± 0.75	20.35 ± 0.42	88.89 ± 0.07
	Min-Disagreement	36.99 ± 1.67	**41.38 ± 2.57**	26.09 ± 1.31	**90.03 ± 0.24**
	Co-Reg	36.41 ± 0.05	37.39 ± 0.00	25.92 ± 0.00	89.66 ± 0.03
	RMSC	25.41 ± 0.95	27.79 ± 1.47	17.95 ± 0.85	88.70 ± 0.30
	Ours	**38.20 ± 0.65**	37.55 ± 0.44	**28.15 ± 0.71**	89.68 ± 0.06
LandUse-21	SPC_{BestSV}	33.63 ± 0.93	29.71± 0.89	19.19± 0.61	92.04± 0.64
	LRR_{BestSV}	32.16 ± 0.53	28.76 ± 0.23	19.04 ± 0.41	91.21 ± 0.06
	Min-Disagreement	27.95 ± 0.05	25.39 ± 1.40	15.25 ± 0.04	82.39 ± 0.01
	Co-Reg	29.63 ± 0.06	25.52 ± 0.00	16.78 ± 0.00	88.26 ± 0.05
	RMSC	32.88 ± 0.26	28.96 ± 0.49	18.92 ± 0.48	91.19 ± 0.06
	Ours	**35.29 ± 0.30**	**31.00 ± 0.53**	**20.46 ± 0.50**	91.54 ± 0.06
Still DB	SPC_{BestSV}	10.45± 0.78	29.42± 0.94	22.14± 0.64	73.29± 0.58
	LRR_{BestSV}	10.91 ± 0.30	30.62 ± 0.39	24.01 ± 0.52	72.39 ± 0.01
	Min-Disagreement	9.67 ± 0.05	**33.62 ± 1.40**	22.30 ± 0.04	73.48 ± 0.04
	Co-Reg	9.93 ± 0.16	26.31 ± 0.24	22.61 ± 0.35	73.16 ± 0.02
	RMSC	10.57 ± 0.56	28.54 ± 2.03	23.17 ± 2.12	72.59 ± 0.46
	Ours	**13.59 ± 0.32**	32.76 ± 0.29	**26.92 ± 0.55**	**74.11 ± 0.01**
BBCSport	SPC_{BestSV}	71.54 ± 0.60	83.60 ± 3.56	76.78 ± 0.38	89.10 ± 0.09
	LRR_{BestSV}	69.02 ± 0.19	78.72± 0.26	76.98± 0.23	87.35 ± 0.13
	Min-Disagreement	77.61 ± 0.19	79.71 ± 4.92	26.09 ± 1.31	90.03 ± 0.24
	Co-Reg	71.76 ± 0.05	73.31 ± 0.58	76.64 ± 0.14	89.14 ± 0.03
	RMSC	81.28 ± 0.95	85.78 ± 1.47	86.62 ± 0.85	92.19 ± 0.30
	Ours	**82.59 ± 0.65**	**90.07 ± 0.44**	**88.65 ± 0.71**	**94.53 ± 0.06**

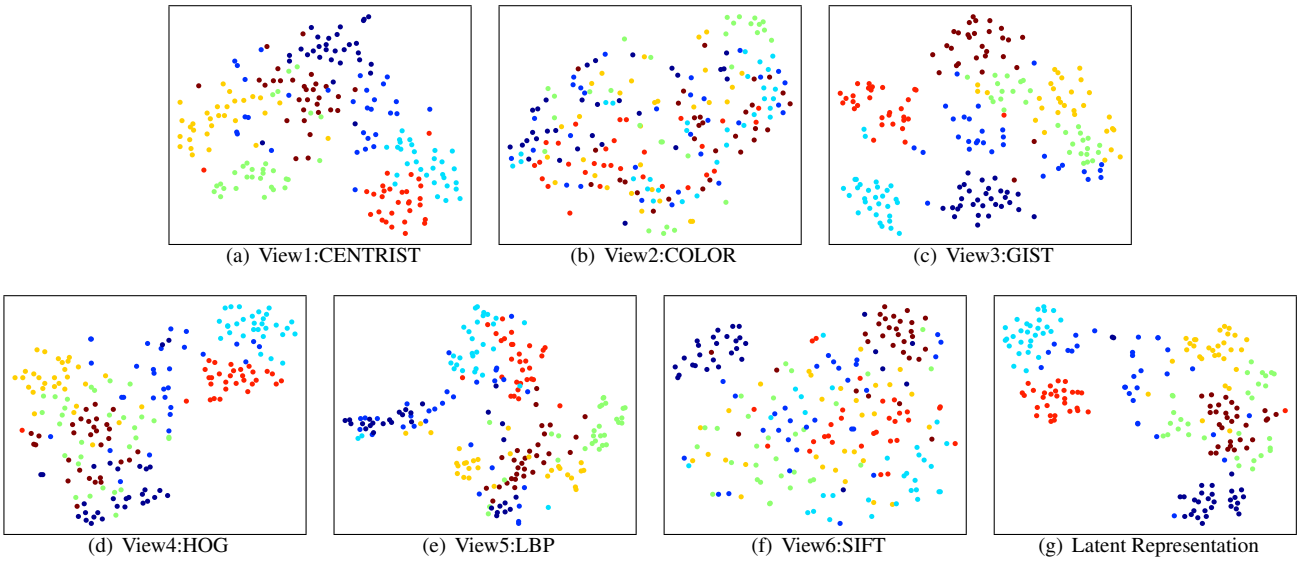
We run 30 times for each method and report the mean values and standard deviations.

4.2. Results on Synthetic Data

We firstly evaluate our method on synthetic data. Each matrix is firstly generated with each element independently sampled from a uniform distribution on the [0, 1] interval. We generate the synthetic data which consists of 6 clusters (or subspaces). The numbers of samples in these subspaces are {25,30,35,40,45,50}, respectively. We firstly generate the matrix $\mathbf{H} \in R^{K\times N}$ uniformly as the latent representation, with the dimensionality $K = 90$ and the data point number $N = 225$. The subspaces have disjoint features with 10,12,14,16,18 and 20 features, respectively. Then two different views are generated according to the latent representation matrix $\mathbf{H}$ with $\mathbf{X}^{(v)} = \mathbf{P}^{(v)}\mathbf{H} + \mathbf{E}^{(v)}$. We consider two types of noise for $\mathbf{E}^{(v)}$, i.e., the global noise $\mathbf{E}_g^{(v)}$ and the sample-specific noise $\mathbf{E}_s^{(v)}$. Formally, we have $\mathbf{E}^{(v)} = \mathbf{E}_s^{(v)} + \alpha\mathbf{E}_g^{(v)}$. For the sample-specific noise, $\mathbf{E}_s^{(v)}$, we randomly generate a matrix and then select randomly a few columns (20 in experiments), setting the other columns with zeros. While for the global noise, we randomly generate an matrix $\mathbf{E}_u^{(v)}$ and multiply it with a coefficient α to control the noise magnitude. As shown in Fig. 2(a), under different degrees of noise, with the help

Table 2: Comparison between single view and the learned multi-view latent representation.

Datasets	Methods	NMI	ACC	F-measure	RI
MSRCV1	View1	51.95 ± 3.12	54.00 ± 5.94	47.91 ± 4.30	83.80 ± 0.41
	View2	15.27 ± 2.14	27.15 ± 2.87	19.77 ± 1.65	75.52 ± 0.28
	View3	62.03 ± 0.72	70.42 ± 0.51	58.95 ± 0.85	88.39 ± 0.13
	View4	53.45 ± 1.41	60.63 ± 1.69	49.79 ± 2.13	85.57 ± 1.46
	View5	43.67 ± 0.60	49.90 ± 0.90	37.77 ± 0.56	80.89 ± 0.26
	View6	38.03 ± 2.15	52.42 ± 1.98	38.95 ± 1.55	82.39 ± 0.81
	LatentRepresentation	**71.67 ± 1.56**	**80.76 ± 1.71**	**68.92 ± 1.09**	**90.64 ± 0.06**
Scene-15	View1	**28.60 ± 0.31**	30.03 ± 0.42	20.92 ± 0.66	89.17 ± 0.12
	View2	25.06 ± 0.35	26.44 ± 0.24	18.18 ± 0.32	88.53 ± 0.07
	View3	16.84 ± 0.33	27.40 ± 0.36	14.44 ± 0.45	88.29 ± 0.03
	LatentRepresentation	25.56 ± 0.60	**33.71 ± 0.49**	**21.99 ± 0.29**	**90.32 ± 0.07**
ORL	View1	77.83 ± 0.86	56.68 ± 1.59	44.90 ± 1.90	97.05 ± 0.16
	View2	**86.13 ± 0.91**	68.38 ± 3.35	59.79 ± 2.86	97.92 ± 0.22
	View3	79.50 ± 1.69	60.68 ± 2.45	48.47 ± 3.45	97.29 ± 0.28
	LatentRepresentation	85.83 ± 0.81	**68.46 ± 2.77**	**59.85 ± 2.78**	**97.95 ± 0.20**
LandUse-21	View1	27.60 ± 1.34	20.95 ± 1.58	14.14 ± 2.07	88.67 ± 0.21
	View2	26.78 ± 1.91	20.81 ± 2.38	12.95 ± 2.11	87.75 ± 0.12
	View3	27.95 ± 1.26	21.76 ± 1.47	14.07 ± 1.56	89.96 ± 0.07
	LatentRepresentation	**31.27 ± 1.11**	**23.68 ± 1.34**	**15.81 ± 0.60**	**90.85 ± 0.68**
Still DB	View1	11.68 ± 1.19	29.79 ± 0.77	23.70 ± 1.01	71.27 ± 1.42
	View2	6.53 ± 0.10	28.31 ± 0.28	23.69 ± 1.69	67.09 ± 2.51
	View3	5.97 ± 0.15	27.52 ± 0.50	22.16 ± 0.67	68.78 ± 0.85
	LatentRepresentation	**12.09 ± 0.40**	**31.11 ± 0.93**	**23.79 ± 0.18**	**73.04 ± 0.58**
BBCSport	View1	59.64 ± 17.04	64.17 ± 15.26	62.43 ± 13.46	73.73 ± 15.41
	View2	23.17 ± 16.86	44.85 ± 8.47	44.47 ± 7.80	42.69 ± 12.36
	LatentRepresentation	**62.18 ± 12.42**	**66.66 ± 12.15**	**63.96 ± 12.18**	**76.72 ± 11.30**

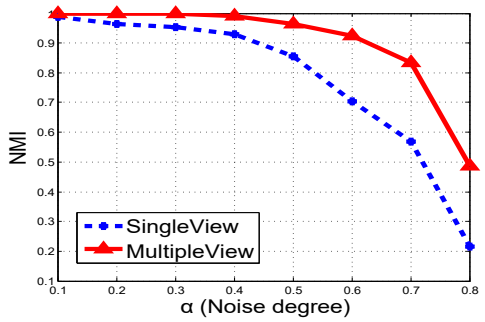


(a) View1:CENTRIST (b) View2:COLOR (c) View3:GIST

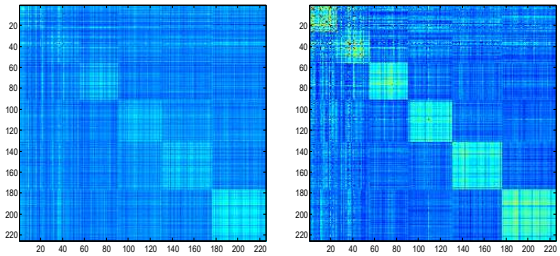
(d) View4:HOG (e) View5:LBP (f) View6:SIFT (g) Latent Representation

Figure 3: Visualization of different views and latent representation with t-SNE.

of multiple views our method achieves much more promising results compared with the result that only using single view. Fig. 2(b) is a visualization example of similarity matrices corresponding to single view (left) and multiple views



(a) Performance with different degree of noise.



(b) Visualization of similarity matrices: single feature (left) and multiple features (right).

Figure 2: Robustness experiment on synthetic data.

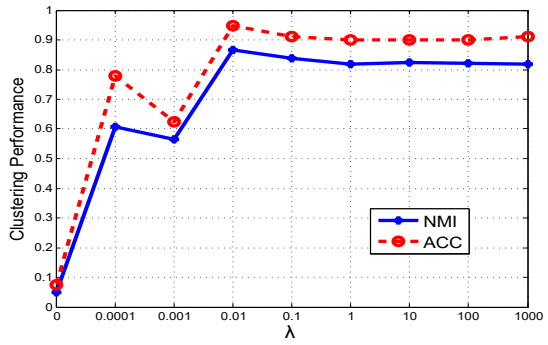


Figure 4: Parameter tuning with respect to λ.

(right) respectively with $\alpha = 0.5$. Clearly, the similarity matrix corresponding to multiple views better reveals the underlying cluster structure.

4.3. Results on Real Datasets

Table 1 gives the clustering results for different clustering methods. In a big picture, our approach outperforms all the baselines with a large margin. Take the MSRCV1 dataset for example, our method outperforms the second performer, Min-Disagreement, about 4.7% and 11.3% in terms of NMI and accuracy, respectively. It should be noted that, the performances of most compared methods are not robust on different datasets. For example, Min-Disagreement achieves the best performance on Still DB in terms of ACC due to the combination of the best two views. However, the performances on the other datasets are not such promising.

To further investigate the improvement of our method, we conduct k-means on each single view and the learned latent representation, respectively. According to the results in Table 2, the clustering performances with latent representation are usually better than those of each single view, which empirically proves that the latent representation is more reasonable than each single view. To be more intuitive, we visualize different views and the learned latent representation with t-Distributed Stochastic Neighbor Embedding (t-SNE) [21] for the dataset MSRCV1 as shown in Fig. 3. It is observed that the figure is well consistent with the clustering results in Table 2. Specifically, Fig. 3(c)-(d) (corresponding to view-3 and view-4) reveals the underlying cluster structure much better and the clustering performances are much higher on view-3 and view-4 than the other views. Fig. 3(g) (corresponding to latent representation) clearly demonstrates the advantage of the learned latent representation, for example, the clusters in red, dark blue, cyan and yellow are more compact than those of each single view.

We also give the parameter tuning experiment (on BBCSport) as shown in Fig. 4. It is observed that the performance of our method is relatively stable and promising since a relatively large value ($\lambda \geq 0.01$) is sufficient.

5. Conclusions

We introduce multi-view latent representation to explore multiple views of data, based on which the subspace clustering is improved. Our main novelty is making use of the complementarity among different views for subspace clustering, and the multi-view latent representation encodes the complementarity under the assumption that each view is originated from one underlying latent representation. This is different from most of existing methods which reconstruct data points directly within each single view. The learned multi-view latent representation and the self-representation based clustering improve each other well. Our method is relatively robust due to the latent representation based on multiple views and structure sparsity. In the future, large scale data will be considered and nonlinearity by kernel technique will be introduced into our model.

Acknowledgment

This work was partly supported by National Program on Key Basic Research Project (2013CB329304), National High-tech R&D Program of China (2014BAK11B03), National Key Research and Development Plan (No.2016YFB0800603) and National Natural Science Foundation of China (61602337, 61502332, 61432011, 61422213).

References

[1] R. H. Bartels and G. Stewart. Solution of the matrix equation AX+ XB= C. *Communications of the ACM*, 15(9):820–826, 1972.

[2] J.-F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[3] X. Cao, C. Zhang, H. Fu, S. Liu, and H. Zhang. Diversity-induced multi-view subspace clustering. In *CVPR*, pages 586–594, 2015.

[4] C. Cortes, M. Mohri, and A. Rostamizadeh. Learning non-linear combinations of kernels. In *NIPS*, pages 396–404, 2009.

[5] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *ICCV*, 2013.

[6] V. R. de Sa. Spectral clustering with two views. In *ICML workshop on learning with multiple views*, pages 20–27, 2005.

[7] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[8] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*.

[9] H. Gao, F. Nie, X. Li, and H. Huang. Multi-view subspace clustering. In *ICCV*, pages 4238–4246, 2015.

[10] S. Gu, L. Zhang, W. Zuo, and X. Feng. Projective dictionary pair learning for pattern classification. In *NIPS*, 2014.

[11] Y. Guo. Convex subspace representation learning from multi-view data. In *AAAI*, 2013.

[12] H. Hu, Z. Lin, J. Feng, and J. Zhou. Smooth representation clustering. In *CVPR*, pages 3834–3841, 2014.

[13] J. Huang, F. Nie, and H. Huang. Spectral rotation versus k-means in spectral clustering. In *AAAI*, 2013.

[14] N. Ikizler, R. G. Cinbis, S. Pehlivan, and P. Duygulu. Recognizing actions from still images. In *ICPR*, 2008.

[15] A. Kumar and H. Daumé. A co-training approach for multi-view spectral clustering. In *ICML*, pages 393–400, 2011.

[16] A. Kumar, P. Rai, and H. Daume. Co-regularized multi-view spectral clustering. In *NIPS*, pages 1413–1421, 2011.

[17] Z. Lin, R. Liu, and Z. Su. Linearized alternating direction method with adaptive penalty for low-rank representation. In *NIPS*, pages 612–620, 2011.

[18] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma. Robust recovery of subspace structures by low-rank representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

[19] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *ICCV*, pages 1615–1622, 2011.

[20] J. Liu, C. Wang, J. Gao, and J. Han. Multi-view clustering via joint nonnegative matrix factorization. In *Proc. of SDM*, volume 13, pages 252–260. SIAM, 2013.

[21] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

[22] V. M. Patel, H. Van Nguyen, and R. Vidal. Latent space sparse subspace clustering. In *ICCV*, pages 225–232, 2013.

[23] J. Tang, X. Hu, H. Gao, and H. Liu. Unsupervised feature selection for multi-view data in social media. In *SDM*, pages 270–278, 2013.

[24] W. Tang, Z. Lu, and I. S. Dhillon. Clustering with multiple graphs. In *ICDM*, pages 1016–1021, 2009.

[25] G. Tzortzis and A. Likas. Kernel-based weighted multi-view clustering. In *ICDM*, pages 675–684, 2012.

[26] M. White, X. Zhang, D. Schuurmans, and Y.-l. Yu. Convex multi-view subspace learning. In *NIPS*, pages 1673–1681, 2012.

[27] R. Xia, Y. Pan, L. Du, and J. Yin. Robust multi-view spectral clustering via low-rank and sparse decomposition. In *AAAI*, pages 2149–2155, 2014.

[28] J. Xu, J. Han, and F. Nie. Discriminatively embedded k-means for multi-view clustering. In *CVPR*, pages 5356–5364, 2016.

[29] Y. Yang and S. Newsam. Bag-of-visual-words and spatial extensions for land-use classification. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 270–279, 2010.

[30] C. Zhang, H. Fu, Q. Hu, P. Zhu, and X. Cao. Flexible multi-view dimensionality co-reduction. *IEEE Transactions on Image Processing*, 26(2):648–659, 2017.

[31] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao. Low-rank tensor constrained multiview subspace clustering. In *ICCV*, pages 1582–1590, 2015.