From Joint Feature Selection and Self-Representation Learning to Robust Multi-view Subspace Clustering

Hui Yan*, Siyu Liu
School of Computer Science and Engineering
Nanjing University of Science and Technology
Nanjing, Jiangsu 210094, China
yanhui@njust.edu.cn

Philip S. Yu
Department of Computer Science
University of Illinois at Chicago
Chicago, Illinois 60607, USA
psyu@uic.edu

***Abstract*—In era of big data, we have easier access to the data with multi-view representations from heterogeneous feature spaces, where each view is often unlabeled, partial and even full of noises. These unique challenges and properties motivate us to develop a novel robust multi-view subspace clustering framework (RMSC), which learns a consensus affinity matrix with the ideal subspace structure, by extending our joint feature selection and self-representation model (JFSSR). Concretely, RMSC learns the consensus graph across diverse views with exactly k connected components (k is the number of clusters), which is encoded by a block diagonal self-representation matrix. Besides, we emphasize $\ell_{2,1}$-norm minimization on the loss function to reduce redundant and irrelevant features, and implicitly assign an adaptive weight to each view without introducing additional parameters. Lastly, an alternating optimization algorithm is derived to solve the nonconvex formulated objective. Extensive empirical results on both synthetic data and real-world benchmark data sets show that RMSC consistently outperforms several representative multi-view clustering approaches.**

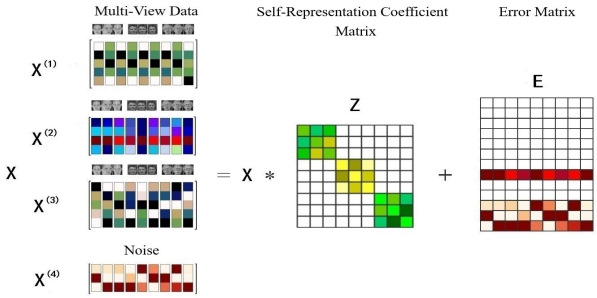
***Index Terms*—multi-view, clustering, robust, sparsity**

I. Introduction

SUBSPACE clustering, which assumes that high dimensional data are drawn from the union of multiple low-dimensional linear subspaces, aims to assign data to their respective subspace corresponding to a cluster, a class or category. As a kind of representative subspace clustering method, the spectral clustering attracts more attention due to its simplicity and outstanding performance. Various spectral-type methods primarily differ in the approaches to learning the affinity matrix [12], [15].

Traditional clustering methods only use a single set of features or one information window of the subjects. Nowadays more and more data have been collected from multiple sources or different views. Therefore, numerous multi-view clustering methods emerge, including co-training [12], multiple kernel learning [5], and multi-view subspace learning [16], [17]. In this paper, we focus on multi-view subspace clustering.

Some multi-view subspace clustering methods work on the assumption that all the views are reliable and therefore maximize the consensus of the cluster structure shared by multiple views [3], [16], [17]. In reality, the views of data could be either inherently strong or weak, which means the final performance will be severely deteriorated if we ignore to distinguish different views. Accordingly, [8] computes roughly the weights for different views by combining the prior knowledge. Different from this manually intervening approach, [1] explicitly define these view weights as variables and then learn them by optimizing the corresponding objective function.



$$\mathbf{X} = \mathbf{XZ} + \mathbf{E}$$

Fig. 1. Schematic of our Robust Multi-view Subspace Clustering framework. Suppose data samples with 3 views $\mathbf{X}^{(1)}$, $\mathbf{X}^{(2)}$, $\mathbf{X}^{(3)}$ and 1 view of noisy features $\mathbf{X}^{(4)}$, our framework pursues a diagonal block self-representation coefficient matrix $\mathbf{Z}$ by removing redundant and irrelevant features shown in red elements in $\mathbf{E}$. Note that the last several rows in $\mathbf{E}$ shows that our framework potentially learns view weights where most of non-zero rows correspond to insignificant view.

Although existing subspace multi-view clustering algorithms achieve promising performances, they have the following two main drawbacks. (1) **Robustness**: The nature of multi-view information acquisition techniques determine that each view is often redundant, noisy or even incomplete. Unfortunately, the aforementioned methods treat the features in the same source as equally vital, which means the final results will be degraded in the presence of irrelevant noisy features. (2) **Consensus-preservation**: It is not reasonable to learn a consensus subspace representation matrix or pair-wise

* Hui Yan is corresponding author.

2374-8486/19/$31.00 ©2019 IEEE
DOI 10.1109/ICDM.2019.00183

similarity matrix across multiple views, since the magnitude of element values in such matrix can be dramatically different. Comparatively, a consensus cluster indicator matrix seems to be a more feasible choice. Unfortunately, the corresponding binary optimization is NP-hard. It seems plausible that learning a consensus cluster structure could have advantages in both measure rationality and optimization efficiency.

To circumvent the issues mentioned above, we propose a joint feature selection and self-representation framework (JFSSR), emphasizing $\ell_{2,1}$-norm minimization on loss term and block diagonal regularizer on self-representation coefficient matrix. We also further extend the proposed single view representation learning model for multi-view learning in the presence of error dubbed robust multi-view subspace clustering (RMSC). Eventually, corresponding optimization algorithms based on the Augmented Lagrange Multiplier (ALM) [6] are proposed. Extensive experiments demonstrate the effectiveness and competitiveness of the proposed methods compared with several state-of-the-art clustering approaches. Please notice that the word "error" in this paper focus on the deviation between model assumption (i.e., subspaces) and feature-specic corruptions.

Fig. 1 illustrates the proposed RMSC framework. We aim to recover the consensus subspace structure reflected by exactly k non-lapping blocks in the diagonal of self-representation coefficient matrix $\mathbf{Z}$ of multi-view data.

In summary, the main contributions of this paper can be delivered as follows:

- We propose a novel joint feature selection and self-representation framework by combining feature reduction and direct diagonal block self-representation learning, which is further extended to a robust multi-view subspace clustering method dubbed RMSC.
- We take both view diversity and feature-specific error into account by introducing reasonable relaxations and regularizers. As a result, the learned self-representation discovers consistency and specificity hidden in multiple views, which contributes to its robustness to noisy features and non-independent subspaces.
- Experiments are conducted on real-world benchmark data sets show so as to demonstrate the super performance of the proposed frameworks.

The rest of this paper is organized as follows. The next section reviews related works briefly. Section III and Section IV introduce the proposed joint feature selection and self-representation learning framework, and robust multi-view subspace clustering method, respectively. Comprehensive experimental results and discussions are provided in Section V. Finally, we give the conclusions in Section VI.

Notations. We define matrices by boldface capital letters, e.g., $\mathbf{A}$, vectors by boldface lowercase letters, e.g., $\mathbf{a}$, and scalars by lowercase letters, e.g., a. For $\mathbf{A}$, its (i,j)-th entry is denoted as $\mathbf{A}_{ij}$; for $\mathbf{a}$, its i-th entry is denoted as $\mathbf{a}_i$. The absolute matrix of $\mathbf{A}$ is denoted by $|\mathbf{A}|$. We define $diag(\mathbf{A})$ as a vector with its i-th entry being the i-th diagonal element of $\mathbf{A}$, and $Diag(\mathbf{a})$ as a diagonal matrix with its i-th entry on the diagonal being $\mathbf{a}_i$. The all one vector and all zero vector are denoted as $\mathbf{1}$ and $\mathbf{0}$, respectively. The identity matrix is denoted as $\mathbf{I}$. If $\mathbf{A}$ is positive semi-definite, we note it as $\mathbf{A} \succeq 0$. If all entries in $\mathbf{A}$ are nonnegative, we note it as $\mathbf{A} \geq 0$. The trace of a square matrix $\mathbf{A}$ is denoted as $Tr(\mathbf{A})$ and its transposition is denoted as $\mathbf{A}^T$. We note it as $[\mathbf{A}]_+ = \max(0, \mathbf{A})$. Some norms will be used, including Frobenius norm (or ℓ_2-norm of a vector) $||\mathbf{A}||_F = \sqrt{\sum_{ij} \mathbf{A}_{ij}^2}$ and $\ell_{2,1}$-norm $||\mathbf{A}||_{2,1} = \sum_i \sqrt{\sum_j \mathbf{A}_{ij}^2}$. For symmetric matrices $\mathbf{A}, \mathbf{B}$, we denote $\mathbf{A} \preceq \mathbf{B}$ if $\mathbf{B} - \mathbf{A} \succeq 0$. $< \cdot, \cdot >$ defines the matrix inner product. For ease of presentation, the horizontal (resp. vertical) concatenation of a collection of matrices along row (resp. column) is denoted by $[\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_n]$ (resp. $[\mathbf{X}_1; \mathbf{X}_2; ...; \mathbf{X}_n]$).

II. RELATED WORK

Given data $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_k] \in \mathbb{R}^{d \times n}$ drawn from k subspaces corresponding to different clusters, where d indicates the dimension of the samples, $\mathbf{X}_i \in \mathbb{R}^{d \times n_i}$ means the submatrix in $\mathbf{X}$ that belongs to the i-th cluster, and $n (n = \sum_{i=1}^{k} n_i)$ is the total number of samples. Subspace clustering aims to cluster the n samples into k classes.

Self-representation based subspace clustering looks for a linear representation $\mathbf{Z}$, whose general formulation can be presented as

$$\begin{aligned} \min_{\mathbf{Z}} \quad & loss(\mathbf{X}, \mathbf{XZ}) + \lambda \Omega(\mathbf{Z}) \\ s.t., \quad & diag(\mathbf{Z}) = \mathbf{0} \end{aligned} \tag{1}$$

where $\mathbf{Z} \in \mathbb{R}^{n \times n}$ is the self-representation coefficient matrix, and it is expected not to be an identity matrix under the constraint in Eq. (1). $loss(\cdot)$ and $\Omega(\cdot)$ denote the loss function and regularization terms, respectively. And the scalar $\lambda > 0$ balances the reconstruction error and the regularization for $\mathbf{Z}$.

In the ideal case, each sample is represented as a linear combination of samples belonging to the same subspace. In this case, $\mathbf{Z}$ has the k-block diagonal structure [2].

III. JOINT FEATURE SELECTION AND SELF-REPRESENTATION LEARNING

A. Preliminary

Definition 1 (k-block diagonal matrix [2]). For any matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$, $\mathbf{B}$ is k-block diagonal if it has k connected components (or submatrices, blocks).

Definition 2 (Laplacian matrix). In graph theory, the Laplacian matrix of the affinity matrix $\mathbf{B} (\mathbf{B} \geq 0, \mathbf{B} = \mathbf{B}^T)$ is defined as

$$\mathbf{L_B} = Diag(\mathbf{B1}) - \mathbf{B}$$

Definition 3 (k-block diagonal regularizer [2]). For any affinity matrix $\mathbf{B}$, its k-block diagonal regularizer denoted by $||\mathbf{B}||_k$ is defined as the sum of the smallest k eigenvalues of the corresponding Laplacian matrix $\mathbf{L_B}$.

Theorem 1 [18]. For any affinity matrix $\mathbf{B}$, the multiplicity k of the eigenvalue 0 of the corresponding Laplacian matrix $\mathbf{L_B}$ equals to the number of connected components (blocks) in $\mathbf{B}$.

It can be seen that if the affinity matrix $\mathbf{B}$ is k-block diagonal, $||\mathbf{B}||_k = 0$. In contrast, if $||\mathbf{B}||_k = 0$, at least $\mathbf{B}$ has k blocks.

B. Formulations

Considering the feature specific corruption and leveraging the k-block diagonal regularizer prior, we propose joint feature selection and self-representation learning framework:

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{E}} ||\mathbf{E}||_{2,1} + \lambda||\mathbf{Z}||_k \\ s.t., \quad &\mathbf{X} = \mathbf{XZ} + \mathbf{E} \\ &diag(\mathbf{Z}) = \mathbf{0}, \mathbf{Z} \geq 0, \mathbf{Z} = \mathbf{Z}^T \end{aligned} \tag{2}$$

where $\lambda > 0$ is a penalty parameter, and $\mathbf{E} \in \mathbb{R}^{d\times n}$ denotes the reconstruction error. Here we emphasize $\ell_{2,1}$-norm on the error term $\mathbf{E}$ because: (1) $\ell_{2,1}$-norm minimization encourages the rows of $\mathbf{E}$ to be zero. Therefore, row-sparse regularized loss function can eliminate the redundant or irrelative features. (2) Such $\ell_{2,1}$-norm regularized model can be further extended to multi-view subspace learning. Note that different from existing $\ell_{2,p}$-norm $(0 \leq p \leq 1)$ regularization based feature selection methods [7], we enforce the row-sparsity on error matrix, not the feature weight matrix. Besides, we want to model the feature-specific corruptions via the row-sparsity property, while the previous self-representation method [15] models the sample-specific noise away from the underlying subspaces as the error term with column-sparsity supports.

We require the representation matrix $\mathbf{Z}$ in Eq. (2) to be non-negative and symmetric, which are necessary properties for defining the block diagonal regularizer. But these restrictions on $\mathbf{Z}$ will limit its representation capability. Introducing an intermediate term $\mathbf{S}$, we get

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{S},\mathbf{E}} ||\mathbf{E}||_{2,1} + \lambda||\mathbf{S}||_k + \rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ s.t., \quad &\mathbf{X} = \mathbf{XZ} + \mathbf{E} \\ &diag(\mathbf{S}) = \mathbf{0}, \mathbf{S} \geq 0, \mathbf{S} = \mathbf{S}^T \end{aligned} \tag{3}$$

The above two models are equivalent when $\rho > 0$ is sufficiently large. As will be seen in Section Optimization, another advantage of the relaxation term $||\mathbf{Z} - \mathbf{S}||_F^2$ is that it makes the objective function separable. More importantly, the subproblems for updating $\mathbf{Z}$ and $\mathbf{S}$ are strongly convex, leading to the final solutions are unique and stable.

C. Optimization

Note that $||\mathbf{S}||_k$ is a nonconvex term, and we introduce a property about the sum of eigenvalues to reformulate $||\mathbf{S}||_k$.

Theorem 2 [20]. Let $\mathbf{L} \in \mathbb{R}^{n\times n}$ and $\mathbf{L} \succeq 0$. Then

$$\begin{aligned} &\sum_{i=n-k+1}^{n} \lambda_i(\mathbf{L}) = \min_{\mathbf{W}} < \mathbf{L}, \mathbf{W} > \\ &s.t. \quad 0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned}$$

where $\lambda_i(\mathbf{L})$ are the eigenvalues of $\mathbf{L}$ in the decreasing order. So Eq. (3) is equivalent to

$$\begin{aligned} &\min_{\mathbf{Z},\mathbf{S},\mathbf{E},\mathbf{W}} ||\mathbf{E}||_{2,1} + \lambda < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > +\rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ s.t., \quad &\mathbf{X} = \mathbf{XZ} + \mathbf{E} \\ &diag(\mathbf{S}) = \mathbf{0}, \mathbf{S} \geq 0, \mathbf{S} = \mathbf{S}^T \\ &0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \tag{4}$$

The ALM with Alternating Direction Minimizing (ADM) strategy [19] is an efficient and effective solver for our problems. Then the problem in Eq. (4) is transformed into the equivalent ALM problem as follows:

$$\begin{aligned} &L(\mathbf{E}, \mathbf{S}, \mathbf{Z}, \mathbf{W}) \\ =&||\mathbf{E}||_{2,1} + \lambda < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > +\rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ &+ < \mathbf{Y}, \mathbf{X} - \mathbf{XZ} - \mathbf{E} > + \frac{\mu}{2}||\mathbf{X} - \mathbf{XZ} - \mathbf{E}||_F^2 \\ s.t., \quad &diag(\mathbf{S}) = \mathbf{0}, \mathbf{S} \geq 0, \mathbf{S} = \mathbf{S}^T \\ &0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \tag{5}$$

where μ is a positive penalty scalar.

We divide the problem in Eq. (5) into four subproblems, and develop an alternative and iterative algorithm to solve them.

Z sub-problem: To update $\mathbf{Z}$, we minimize the following objective function by fixing other variables

$$\begin{aligned} L(\mathbf{Z}) = &< \mathbf{Y}, \mathbf{X} - \mathbf{XZ} - \mathbf{E} > + \frac{\mu}{2}||\mathbf{X} - \mathbf{XZ} - \mathbf{E}||_F^2 \\ &+ \rho||\mathbf{Z} - \mathbf{S}||_F^2 \end{aligned}$$

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

$$\begin{aligned} \mathbf{Z}^* = &(\mu\mathbf{X}^T X + 2\rho\mathbf{I})^{-1} \\ &(\mu\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{Y} + 2\rho S - \mu\mathbf{X}^T\mathbf{E}) \end{aligned} \tag{6}$$

S sub-problem: Fixing the other variables, we update $\mathbf{S}$ by solving the following problem

$$\begin{aligned} \mathbf{S}^* = &\arg\min \frac{1}{2}||\mathbf{S} - (\mathbf{Z} - \frac{\lambda}{2\rho}(diag(\mathbf{W})\mathbf{1}^T - \mathbf{W}))||_F^2 \\ s.t., \quad &\mathbf{S} \geq \mathbf{0}, \mathbf{S} = \mathbf{S}^T, diag(\mathbf{S}) = 0 \end{aligned}$$

This problem has a closed form solution given by

$$\mathbf{S}^* = [(\mathbf{A} + \mathbf{A}^T)/2]_+ \tag{7}$$

where $\mathbf{A} = (\mathbf{Z} - \frac{\lambda}{2\rho}(diag(\mathbf{W})\mathbf{1}^T - \mathbf{W}))$.

W sub-problem: Fixing the others variables, we update $\mathbf{W}$ by the following rule

$$\begin{aligned} \mathbf{W}^* = &\arg\min < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > \\ s.t., \quad &0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \tag{8}$$

Note that the above subproblem is convex and has closed form solutions, i.e., $\mathbf{W} = \mathbf{U}\mathbf{U}^T$, where $\mathbf{U} \in \mathbb{R}^{n\times k}$ consists of k eigenvectors associated with the k smallest eigenvalues of $Diag(\mathbf{S1}) - \mathbf{S}$.

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

$$\begin{aligned}\mathbf{E}^* &= \arg\min ||\mathbf{E}||_{2,1} \\ &+ < \mathbf{Y}, \mathbf{X} - \mathbf{XZ} - \mathbf{E} > + \frac{\mu}{2}||\mathbf{X} - \mathbf{XZ} - \mathbf{E}||_F^2 \\ &= \arg\min \frac{1}{\mu}||\mathbf{E}||_{2,1} + \frac{1}{2}||\mathbf{E} - (\mathbf{X} - \mathbf{XZ} + \mathbf{Y}/\mu)||_F^2 \quad (9)\end{aligned}$$

This subproblem can be efficiently solved by Lemma 3.2 in [9].

Updating Multipliers:

$$\begin{cases} \mathbf{Y} = \mathbf{Y} + \mu(\mathbf{X} - \mathbf{XZ} - \mathbf{E}) \\ \mu = \min(\gamma\mu, \mu_{max}) \end{cases} \quad (10)$$

The procedure depicted as in Algorithm 1 solves the problem in Eq. (3).

Algorithm 1 Joint Feature Selection and Self-Representation Learning

Input:
Data matrices $\mathbf{X}$, hyper-parameters ρ, λ, and the number of clusters k.
Initial:
$\mathbf{E} = 0$, $\mathbf{S} = \mathbf{Z} = 0$, $\mathbf{Y} = 0$, $\mu = 10^{-6}$, $\gamma = 1.1$, $\varepsilon = 10^{-4}$, $\mu_{max} = 10^6$.
While not converge do
Update $\mathbf{E}$ according to Eq. (9)
Update $\mathbf{S}$ according to Eq. (7)
Update $\mathbf{W}$ according to Eq. (8)
Update $\mathbf{Z}$ according to Eq. (6)
Update multipliers $\mathbf{Y}$ and μ according to Eq. (10)
Check the convergence conditions:
$||\mathbf{X} - \mathbf{XZ} - \mathbf{E}||_\infty < \varepsilon$
End while
Output: Z, S

IV. Multi-view subspace learning

In this section, we extend JFSSR to multi-view clustering.

A. Formulations

Given multi-view observations $\{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, ..., \mathbf{X}^{(v)}\}$ which consist of v different views, we aim to discover latent clustering structure by learning a consistent self-representation coefficient matrix $\mathbf{Z} \in \mathbb{R}^{n\times n}$ shared by all views. Self-representation property with block diagonal regularizer in each view can be denoted as

$$\begin{aligned}\min_{\mathbf{Z},\mathbf{S},\mathbf{W},\mathbf{E}^{(i)}} \quad & \sum_{i=1}^{v} ||\mathbf{E}^{(i)}||_{2,1} + \lambda < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > + \rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ s.t., \quad & \mathbf{X}^{(i)} = \mathbf{X}^{(i)}\mathbf{Z} + \mathbf{E}^{(i)}, i = 1, 2, ..., v \\ & diag(\mathbf{S}) = \mathbf{0}, \mathbf{S} \geq 0, \mathbf{S} = \mathbf{S}^T \\ & 0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \quad (11)$$

where $\mathbf{X}^{(i)} \in \mathbb{R}^{d_i\times n}$ denotes the feature matrix with the dimension d_i corresponding to the i-th view. $\mathbf{E}^{(i)} \in \mathbb{R}^{d_i\times n}$ denotes the reconstruction error in each view. It is worth mentioning that this $\ell_{2,1}$-norm regularized model can be treated as a self-weighted model, which implicitly assigns an adaptive weight to each view without introducing additional parameters [1].

Besides the consistency term $\mathbf{Z}$ comprised in Eq. (11), the unique part in each view could be considered. Thus, we further relax the restrictions on $\mathbf{Z}$ and reformulate it as a pretty simplified and compact form:

$$\begin{aligned}\min_{\mathbf{Z},\hat{\mathbf{Z}}^i,\mathbf{S},\mathbf{W},\mathbf{E}} \quad & ||\mathbf{E}||_{2,1} + \beta\sum_{i=1}^{v} ||\hat{\mathbf{Z}}^{(i)}||_F^2 + \rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ & + \lambda < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > \\ s.t., \quad & \mathbf{X} = \mathbf{XZ} + \mathbf{P} + \mathbf{E} \\ & diag(\mathbf{S}) = \mathbf{0}, \mathbf{S} \geq 0, \mathbf{S} = \mathbf{S}^T \\ & 0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \quad (12)$$

where $\mathbf{X} = \left[\mathbf{X}^{(1)}; \mathbf{X}^{(2)}; ...; \mathbf{X}^{(v)}\right]$, $\mathbf{E} = \left[\mathbf{E}^{(1)}; \mathbf{E}^{(2)}; ...; \mathbf{E}^{(v)}\right]$, and $\mathbf{P} = \left[\mathbf{X}^{(1)}\hat{\mathbf{Z}}^{(1)}; \mathbf{X}^{(2)}\hat{\mathbf{Z}}^{(2)}; ...; \mathbf{X}^{(v)}\hat{\mathbf{Z}}^{(v)}\right]$.

B. Optimization

The augmented Lagrange function of the problem in Eq. (12) is as follows:

$$\begin{aligned} & L(\mathbf{E}, \hat{\mathbf{Z}}^{(i)}, \mathbf{S}, \mathbf{Z}, \mathbf{W}) \\ =& ||\mathbf{E}||_{2,1} + \beta\sum_{i=1}^{v} ||\hat{\mathbf{Z}}^{(i)}||_F^2 + \rho||\mathbf{Z} - \mathbf{S}||_F^2 \\ & + \lambda < Diag(\mathbf{S1}) - \mathbf{S}, \mathbf{W} > \\ & + \sum_{i=1}^{v} < \mathbf{Y}^{(i)}, \mathbf{X}^{(i)} - \mathbf{X}^{(i)}(\mathbf{Z} + \hat{\mathbf{Z}}^{(i)}) - \mathbf{E}^{(i)} > \\ & + \sum_{i=1}^{v} \frac{\mu}{2}||\mathbf{X}^{(i)} - \mathbf{X}^{(i)}(\mathbf{Z} + \hat{\mathbf{Z}}^{(i)}) - \mathbf{E}^{(i)}||_F^2 \\ s.t., \quad & \mathbf{S} \geq \mathbf{0}, \mathbf{S} = \mathbf{S}^T, diag(\mathbf{S}) = \mathbf{0} \\ & 0 \preceq \mathbf{W} \preceq \mathbf{I}, Tr(\mathbf{W}) = k \end{aligned} \quad (13)$$

We divide the problem in Eq. (12) into several subproblems, and develop Algorithm 2 to solve them.

We give the procedure of clustering as previous works [9]. Given the data matrix $\mathbf{X}$, we obtain the consistent representation coefficient matrix $\mathbf{Z}$ by solving RMSC problem in Eq. (12) using Algorithm 2. Both of them can be used to infer the data clustering. The affinity matrix can be defined as $(|\mathbf{Z}|+|\mathbf{Z}^T|)/2$, followed by the spectral clustering [12] to achieve the final result.

V. Experiments

A. Date sets and evaluation metrics

In this section, we performed extensive experiments to evaluate the effectiveness of the performance of JFSSR (Algorithm 1) and RMSC (Algorithm 2) on some real-world benchmark data sets. ORL[1] contains 10 different face images of each of

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

Algorithm 2 Robust Multi-view Subspace Clustering

Input:

Multi-view matrices $\{\mathbf{X}^{(1)}, ..., \mathbf{X}^{(v)}\}$,
hyper-parameters β, ρ, λ and the number of clustering k.

Initial:

$\mathbf{E} = 0$, $\mathbf{S} = \mathbf{Z} = 0$, $\hat{\mathbf{Z}}^{(i)} = 0$, $\mathbf{Y} = 0$,
$\mu = 10^{-6}$, $\gamma = 1.1$, $\varepsilon = 10^{-4}$, $\mu_{max} = 10^{6}$.

while not converge do

1. Update $\mathbf{E}$ using

$$\mathbf{E}^* = \arg\min \frac{1}{\mu}||\mathbf{E}||_{2,1} + \frac{1}{2}||\mathbf{E} - (\mathbf{X} - \mathbf{XZ} - \mathbf{P} + \mathbf{Y}/\mu)||_F^2$$

2. Update $\mathbf{S}$ using $\mathbf{S}^* = [(\mathbf{A} + \mathbf{A}^T)/2]_+$
where $\mathbf{A} = (\mathbf{Z} - \frac{\lambda}{2\rho}(diag(\mathbf{W})\mathbf{1}^T - \mathbf{W}))$.
3. Update $\mathbf{W}$ using $\mathbf{W}^* = \mathbf{U}\mathbf{U}^T$
where $\mathbf{U} \in \mathbb{R}^{n\times k}$ consist of k eigenvectors associated with the k smallest eigenvalues of $Diag(\mathbf{S1}) - \mathbf{S}$.
4. Update $\mathbf{Z}$ using

$$\mathbf{Z}^* = (\mu\mathbf{X}^T\mathbf{X} + 2\rho\mathbf{I})^{-1} (\mu\mathbf{X}^T\mathbf{X} + \mathbf{X}^T\mathbf{Y} + 2\rho S - \mu\mathbf{X}^T\mathbf{E} - \mathbf{X}^T\mathbf{P})$$

5. Update $\hat{\mathbf{Z}}^{(i)}$ using

$$\hat{\mathbf{Z}}^{(i)*} = (\frac{2\beta}{\mu}\mathbf{I} + \mathbf{X}^{(i)^T}\mathbf{X}^{(i)})\mathbf{X}^{(i)^T}(\mathbf{X}^{(i)} - \mathbf{X}^{(i)}\mathbf{Z} - \mathbf{E}^{(i)} + \frac{\mathbf{Y}^{(i)}}{\mu})$$

6. Update multipliers $\mathbf{Y}$ using $\mathbf{Y} = \mathbf{Y} + \mu(\mathbf{X} - \mathbf{XZ} - \mathbf{P} - \mathbf{E})$
7. Update the parameter μ by $\mu = \min\{\rho\mu, \max_\mu\}$
8. Check the convergence conditions:
$||\mathbf{X} - \mathbf{XZ} - \mathbf{P} - \mathbf{E}_h||_\infty < \varepsilon$

end while

Output: Z, S

40 distinct subjects, which is associated with three views. 3 Sources [2] includes 169 news stories collected from three online news sources. Each story is manually annotated with one of the six topical labels. Notting-Hill [13] video face dataset is derived from the movie Notting-Hill. The faces of five main casts are collected, including 4,660 faces in 76 tracks. We randomly sample 110 images of each cast.

For evaluation metrics, we use Normalized Mutual Information (NMI), Accuracy (ACC), F-measure (F1), and Rand Index (RI) to comprehensively measure the clustering performance in our experiments.

B. BDR vs. JFSSR

To give an intuitive example to illustrate the effectiveness of $\ell_{2,1}$-norm regularized loss minimization function in Eq. (3), we compare JFSSR with BDR [2]. Because the main difference between both lies in the used regularization for the error term $\mathbf{E}$.

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

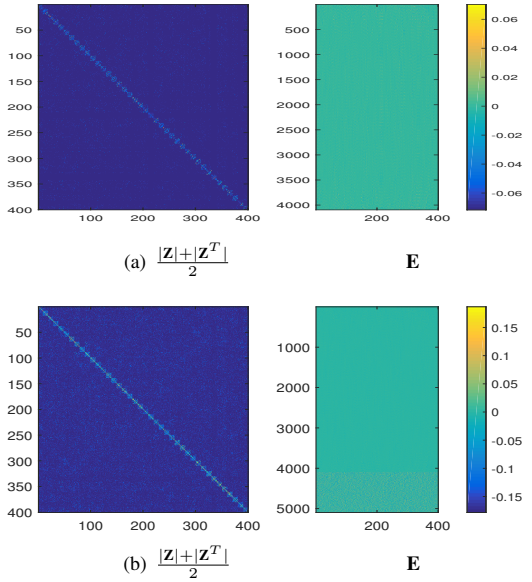


Fig. 2. Affinity matrices and error matrices of JFSSR on (a) View 1 of ORL and (b) View 1 of ORL with additional noisy features ($\lambda = 1$ and $\rho = 10$)

TABLE I
BDR VS. JFSSR ON ORL

Data	Methods	ACC	NMI	F1	RI
View 1	BDR	72.50±0.09	86.69±0.02	64.12±0.10	98.24±0.00
	JFSSR	75.95±0.08	88.78±0.00	68.61±0.04	98.44±0.00
View 1 + Noise	BDR	60.00±0.03	75.29±0.01	44.76±0.02	97.38±0.00
	JFSSR	73.20±0.03	86.17±0.01	64.05±0.06	98.26±0.00

We select View 1 of ORL as $\hat{\mathbf{X}} = [\mathbf{X}_1, \mathbf{X}_2, ..., \mathbf{X}_{40}] \in \mathbb{R}^{4096\times400}$, and randomly generate noisy feature matrix $\mathbf{E}_f \in \mathbb{R}^{1000\times400}$ from $\mathcal{N}(0, 1)$. As shown in Table I, our method achieves much more stable results compared with BDR. And the corresponding affinity matrices $(|\mathbf{Z}| + |\mathbf{Z}^T|)/2$ and error matrices $\mathbf{E}$ obtained by JFSSR are shown in Fig. 2.

C. Experiments to Evaluate RMSC

We compare our method with the baselines such as LMSC [17], GMC [14], SWMC [1], MultiNMF [9], MVKMBD [10], DEKM [11], RDEKM [4], and SPC [3]. For the existing methods, we use the codes released by the authors.

For all the compared methods, we tune the parameters (for some methods, we use the parameters which are given in their codes for some data sets) and use the ones which achieve the best results in most cases for each data set. For our method, we tune all parameters from $\{0.01, 0.1, 1, 10, 100\}$ to report best performances. We run 10 times for each method and report the mean values and standard deviations. The experimental results on the three data sets are presented in Table II, which shows that our approach almost outperforms all the baselines.

To further investigate the improvement of our method, we conduct BDR on each single view and the learned consensus

TABLE II
RESULTS ON THREE DATASETS (MEAN ± STANDARD DEVIATION). THE HIGHEST VALUES ARE IN BOLDFACE, AND THE SECOND BEST ONES ARE IN ITALIC. HIGHER VALUE INDICATES BETTER PERFORMANCE.

Datasets		LMSC	GMC	SWMC	MultiNMF	MVKMBD	DEKM	RDEKM	Co-Reg	RMSC
3 Source	NMI	49.14±4.87	*62.16±0.00*	15.43±0.00	42.36±3.85	40.23±0.00	34.17±0.00	48.09±0.00	52.67±3.14	**62.18±0.02**
	ACC	62.78±4.15	*69.23±0.00*	36.09±0.00	49.92±3.62	54.44±0.00	50.30±0.00	61.54±0.00	56.07±5.48	**70.41±0.07**
	F1	58.74±4.25	*60.47±0.00*	36.91±0.00	42.61±3.35	47.89±0.00	46.40±0.00	60.31±0.00	50.35±5.40	**67.73±0.15**
	RI	81.27±1.92	75.56±0.00	32.23±0.00	72.37±3.49	78.14±0.00	78.16±0.00	*82.94±0.00*	78.25±3.48	**83.54±0.02**
ORL	NMI	*90.31±0.98*	85.71±0.00	84.99±0.00	84.77±1.33	73.90±0.00	70.36±0.00	74.56±0.00	88.62±1.72	**91.93±0.06**
	ACC	*76.48±3.04*	63.25±0.00	74.75±0.00	69.79±2.81	51.25±0.00	44.50±0.00	53.25±0.00	74.45±3.88	**81.50±0.05**
	F1	*70.46±3.04*	35.99±0.00	44.30±0.00	58.45±3.61	38.32±0.00	31.64±0.00	38.93±0.00	68.41±4.57	**74.44±0.15**
	RI	*98.52±0.19*	93.57±0.00	95.16±0.00	97.82±0.26	96.60±0.00	96.06±0.00	96.62±0.00	98.46±0.25	**98.73±0.05**
Notting -Hill	NMI	69.56±5.15	**87.07±0.00**	83.11±000	76.85±1.41	68.28±0.00	68.96±0.00	68.60±0.00	76.08±4.71	*85.17±0.00*
	ACC	81.70±5.28	74.55±0.00	*84.73±0.00*	87.76±0.95	65.64±0.00	68.36±0.00	69.64±0.00	78.10±5.29	**90.36±0.00**
	F1	70.35±7.97	80.12±0.00	*87.78±0.00*	82.71±0.95	65.14±0.00	66.67±0.00	66.62±0.00	77.77±6.66	**88.84±0.00**
	RI	86.78±3.75	91.38±0.00	*94.31±0.00*	92.40±0.40	85.13±0.00	85.74±0.00	85.63±0.00	90.43±2.78	**94.74±0.00**

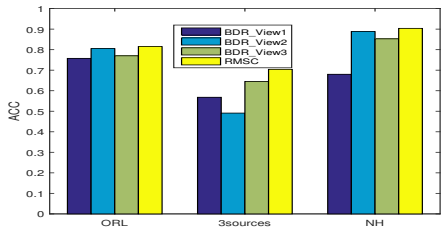


Fig. 3. Performance comparison between BDR with each view and RMSC versus ACC on three datasets.

BDR by RMSC on all three datasets, respectively. According to the results in Fig. 3, the clustering performances with consensus BDR are usually better than those of each single view, which empirically proves that the learnt consensus BDR is more reasonable than each single view.

VI. CONCLUSION

In this paper, we introduce joint feature selection and self-representation learning framework and extend it to multi-view subspace clustering. Multi-view latent cluster structure is encoded by a block diagonal self-representation coefficient matrix. Moreover, irrelevant features and the view without discriminant information could be removed by the proposed $\ell_{2,1}$-norm minimized loss function. Experimental results demonstrate the effectiveness of the proposed methods.

ACKNOWLEDGMENT

This work was supported in part by NSF of China (Grant no. 61773215, no.61772273, no.61703209), National Defense Pre-research Foundation (Grant no. 41412010101), NSF (Grants III-1526499, III-1763325, III-1909323, SaTC-1930941), and CNS-1626432.

REFERENCES

[1] F. Nie, J. Li and X. Li. *Self-weighted multiview clustering with multiple graphs*. International Joint Conference on Artificial Intelligence, pp. 2564-2570, 2017.

[2] C. Lu, J. Feng, Z. Lin, T. Mei and S. Yan. *Subspace clustering by block diagonal representation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 487-501, 2018.

[3] A. Kumar and H. Daume. *A co-training approach for multi-view spectral clustering*. International Conference on Machine Learning, pp. 393-400, 2011.

[4] J. Xu, J. Han, F. Nie and X. Li. *Re-weighted discriminatively embedded k-means for multi-view clustering*. IEEE Transactions on Image Processing, pp. 3016-3027, 2017.

[5] M. Gonen and E. Alpaydin. *Multiple kernel learning algorithms*. Journal of Machine Learning Research. pp. 2211-2268, 2011.

[6] X. Yuan and J. Yang. *Sparse and low-rank matrix decomposition via alternating direction methods*[Online]. Available: http://www.math.hkbu.edu.hk/xmyuan/Publication.html, 2013.

[7] F. Nie, H. Huang, X. Cai, and C. Ding. *Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization*. Conference on Neural Information Processing Systems, 2010.

[8] Y. Cheng and R. Zhao. *Multiview spectral clustering via ensemble*. IEEE International Conference on Granular Computing, pp. 101-106, 2009.

[9] J. Liu, C. Wang, J. Gao, and J. Han. *Multi-view clustering via joint nonnegative matrix factorization*. SIAM International Conference on Data Mining, pp. 252-260, 2013.

[10] X. Cai, F. Nie, and H. Huang. *Multi-view K-means clustering on big data*. International Joint Conference on Artificial Intelligence, pp. 2598-2604, 2013.

[11] J. Xu, J. Han, and F. Nie. *Discriminatively embedded k-means for multi-view clustering*. IEEE Conference on Computer Vision and Pattern Recognition, pp. 5356-5364, 2016.

[12] A. Ng, M. Jordan and Y. Weiss. *On spectral clustering: analysis and an algorithm*. Conference on Neural Information Processing Systems, 2002

[13] Y. Zhang , C. Xu, H. Lu and Y. Huang. *Character identification in feature-length films using global face-name matching*. IEEE Transactions on Multimedia, pp. 1276-1288, 2009.

[14] H. Wang, Y. Yang and B. Liu. *GMC: graph-based multi-view clustering*. IEEE Transactions on Knowledge and Data Engineering, 2019.

[15] 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), pp. 171-184, 2013.

[16] C. Zhang, H. Fu, S. Liu, G. Liu, and X. Cao. *Low-rank tensor constrained multiview subspace clustering*. IEEE International Conference on Computer Vision, pp. 1582-1590, 2015.

[17] C. Zhang, Q. Hu, H. Fu, P. Zhu, and X. Cao. *Latent multi-view subspace clustering*. IEEE Conference on Computer Vision and Pattern Recognition, pp. 4279-4287, 2017.

[18] U. Luxburg. *A tutorial on spectral clustering*. Statistics and Computing, 17(4), pp. 395-416, 2007.

[19] Z. Lin, R. Liu, and Z. Su. *Linearized alternating direction method with adaptive penalty for low-rank representation*. Conference on Neural Information Processing Systems, pp. 612-620, 2011.

[20] J. Dattorro. *Convex optimization euclidean distance geometry*. http://meboo.convexoptimization.com/Meboo.html. 2016.