



Multi-view low-rank sparse subspace clustering



Maria Brbić, Ivica Kopriva*

Laboratory for Machine Learning and Knowledge Representation, Division of Electronics, Rudjer Boskovic Institute, Bijenicka cesta 54, 10000, Zagreb, Croatia

ARTICLE INFO

Article history:

Received 13 March 2017

Revised 10 July 2017

Accepted 23 August 2017

Available online 24 August 2017

Keywords:

Subspace clustering

Multi-view data

Low-rank

Sparsity

Alternating direction method of multipliers

Reproducing kernel Hilbert space

ABSTRACT

Most existing approaches address multi-view subspace clustering problem by constructing the affinity matrix on each view separately and afterwards propose how to extend spectral clustering algorithm to handle multi-view data. This paper presents an approach to multi-view subspace clustering that learns a joint subspace representation by constructing affinity matrix shared among all views. Relying on the importance of both low-rank and sparsity constraints in the construction of the affinity matrix, we introduce the objective that balances between the agreement across different views, while at the same time encourages sparsity and low-rankness of the solution. Related low-rank and sparsity constrained optimization problem is for each view solved using the alternating direction method of multipliers. Furthermore, we extend our approach to cluster data drawn from nonlinear subspaces by solving the corresponding problem in a reproducing kernel Hilbert space. The proposed algorithm outperforms state-of-the-art multi-view subspace clustering algorithms on one synthetic and four real-world datasets.

© 2017 Elsevier Ltd. All rights reserved.

1. Introduction

In many real-world machine learning problems the same data is comprised of several different representations or views. For example, same documents may be available in multiple languages [1] or different descriptors can be constructed from the same images [2]. Although each of these individual views may be sufficient to perform a learning task, integrating complementary information from different views can reduce the complexity of a given task [3]. Multi-view clustering seeks to partition data points based on multiple representations by assuming that the same cluster structure is shared across views. By combining information from different views, multi-view clustering algorithms attempt to achieve more accurate cluster assignments than one can get by simply concatenating features from different views.

In practice, high-dimensional data often reside in a low-dimensional subspace. When all data points lie in a single subspace, the problem can be set as finding a basis of a subspace and a low-dimensional representation of data points. Depending on the constraints imposed on the low-dimensional representation, this problem can be solved using e.g. Principal Component Analysis (PCA) [4], Independent Component Analysis (ICA) [5] or Non-negative Matrix Factorization (NMF) [6–8]. On the other hand, data points can be drawn from different sources and lie in a union

of subspaces. By assigning each subspace to one cluster, one can solve the problem by applying standard clustering algorithms, such as k-means [9]. However, these algorithms are based on the assumption that data points are distributed around centroid and often do not perform well in the cases when data points in a subspace are arbitrarily distributed. For example, two points can have a small distance and lie in different subspaces or can be far and still lie in the same subspace [10]. Therefore, methods that rely on a spatial proximity of data points often fail to provide a satisfactory solution. This has motivated the development of subspace clustering algorithms [10]. The goal of subspace clustering is to identify the low-dimensional subspaces and find the cluster membership of data points. Spectral based methods [11–13] present one approach to subspace clustering problem. They have gained a lot of attention in the recent years due to the competitive results they achieve on arbitrarily shaped clusters and their well defined mathematical principles. These methods are based on the spectral graph theory and represent data points as nodes in a weighted graph. The clustering problem is then solved as a relaxation of the min-cut problem on a graph [14].

One of the main challenges in spectral based methods is the construction of the affinity matrix whose elements define the similarity between data points. Sparse subspace clustering [15] and low-rank subspace clustering [16–19] are among most effective methods that solve this problem. These methods rely on the self-expressiveness property of the data by representing each data point as a linear combination of other data points. Low-Rank Representation (LRR) [16,17] imposes low-rank constraint on the data

* Corresponding author.

E-mail addresses: maria.brbic@irb.hr (M. Brbić), ivica.kopriva@irb.hr, ikopriva@gmail.com, ikopriva@irb.hr (I. Kopriva).

representation matrix and captures global structure of the data. Low-rank implies that data matrix is represented by a sum of small number of outer products of left and right singular vectors weighted by corresponding singular values. Under assumption that subspaces are independent and data sampling is sufficient, LRR guarantees exact clustering. However, for many real-world datasets this assumption is overly restrictive and the assumption that data is drawn from disjoint subspaces would be more appropriate [20,21]. On the other hand, Sparse Subspace Clustering (SSC) [15] represents each data point as a sparse linear combination of other points and captures local structure of the data. Learning representation matrix in SSC can be interpreted as sparse coding [22–27]. However, compared to sparse coding where dictionary is learned such that the representation is sparse [28,29], SSC is based on self-representation property i.e. data matrix stands for a dictionary. SSC also succeeds when data is drawn from independent subspaces and the conditions have been established for clustering data drawn from disjoint subspaces [30]. However, theoretical analysis in [31] shows that it is possible that SSC over-segments subspaces when the dimensionality of data points is higher than three. Experimental results in [32] show that LRR misclassifies different data points than SSC. Therefore, in order to capture global and the local structure of the data, it is necessary to combine low-rank and sparsity constraints [32,33].

Multi-view subspace clustering can be considered as a part of multi-view or multi-modal learning. Multi-view learning method in [34] learns view generation matrices and representation matrix, relying on the assumption that data from all the views share the same representation matrix. The multi-view method in [35] is based on the canonical correlation analysis in extraction of two-view filter-bank-based features for image classification task. Similarly, in [36] the authors rely on tensor-based canonical correlation analysis to perform multi-view dimensionality reduction. This approach can be used as a preprocessing step in multi-view learning in case of high-dimensional data. In [37] low-rank representation matrix is learned on each view separately and learned representation matrices are concatenated to a matrix from which a unified graph affinity matrix is obtained. The method in [38] relies on learning a linear projection matrix for each view separately. High-order distance-based multi-view stochastic learning is proposed in [39], to efficiently explore the complementary characteristics of multi-view features for image classification. The method in [40] is application oriented towards image reranking and assumes that multi-view features are contained in hypergraph Laplacians that define different modalities. In [41] authors propose multi-view matrix completion algorithm for handling multi-view features in semi-supervised multi-label image classification.

Previous multi-view subspace clustering works [42–45] address the problem by constructing affinity matrix on each view separately and then extend algorithm to handle multi-view data. However, since input data may often be corrupted by noise, this approach can lead to the propagation of noise in the affinity matrices and degrade clustering performance. Different from the existing approaches, we propose multi-view spectral clustering framework that jointly learns a subspace representation by constructing single affinity matrix shared by multi-view data, while at the same time encourages low-rank and sparsity of the representation. We propose Multi-view Low-rank Sparse Subspace Clustering (MLRSSC) algorithms that enforce agreement: (i) between affinity matrices of the pairs of views; (ii) between affinity matrices towards a common centroid. Opposed to [35,40,46], the proposed approach can deal with highly heterogeneous multi-view data coming from different modalities. We present optimization procedure to solve the convex dual optimization problems using Alternating Direction Method of Multipliers (ADMM) [47]. Furthermore, we propose the kernel extension of our algorithms by solving the

Table 1

Notations and abbreviations.

Notation	Definition
N	Number of data points
k	Number of clusters
v	View index
n_v	Number of views
$D^{(v)}$	Dimension of data points in a view v
$\mathbf{X}^{(v)} \in \mathbb{R}^{D^{(v)} \times N}$	Data matrix in a view v
$\mathbf{C}^{(v)} \in \mathbb{R}^{N \times N}$	Representation matrix in a view v
$\mathbf{C}^* \in \mathbb{R}^{N \times N}$	Centroid representation matrix
$\mathbf{W} \in \mathbb{R}^{N \times N}$	Affinity matrix
$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$	Singular value decomposition (SVD) of $\mathbf{X}$
$\Phi(\mathbf{X}^{(v)})$	Data points in a view v mapped into high-dimensional feature space
$\mathbf{K}^{(v)} \in \mathbb{R}^{N \times N}$	Gram matrix in a view v

problem in a Reproducing Kernel Hilbert Space (RKHS). Experimental results show that MLRSSC algorithm outperforms state-of-the-art multi-view subspace clustering algorithms on several benchmark datasets. Additionally, we evaluate performance on a novel real-world heterogeneous multi-view dataset from biological domain.

The remainder of the paper is organized as follows. Section 2 gives a brief overview of the low-rank and sparse subspace clustering methods. Section 3 introduces two novel multi-view subspace clustering algorithms. In Section 4 we present the kernelized version of the proposed algorithms by formulating subspace clustering problem in RKHS. The performance of the new algorithms is demonstrated in Section 5. Section 6 concludes the paper.

2. Background and related work

In this section, we give a brief introduction to Sparse Subspace Clustering (SSC) [15], Low-Rank Representation (LRR) [16,17] and Low-rank Sparse Subspace Clustering (LRSSC) [32].

2.1. Main notations

Throughout this paper, matrices are represented with bold capital symbols and vectors with bold lower-case symbols. $\|\cdot\|_F$ denotes the Frobenius norm of a matrix. The ℓ_1 norm, denoted by $\|\cdot\|_1$, is the sum of absolute values of matrix elements; infinity norm $\|\cdot\|_\infty$ is the maximum absolute element value; and the nuclear norm $\|\cdot\|_*$ is the sum of singular values of a matrix. Trace operator of a matrix is denoted by $\text{tr}(\cdot)$ and $\text{diag}(\cdot)$ is the vector of diagonal elements of a matrix. $\mathbf{0}$ denotes null vector. Table 1 summarizes some notations used throughout the paper.

2.2. Related work

Consider the set of N data points $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^D\}_{i=1}^N$ that lie in a union of $k > 1$ linear subspaces of unknown dimensions. Given the set of data points $\mathbf{X}$, the task of subspace clustering is to cluster data points according to the subspaces they belong to. The first step is the construction of the affinity matrix $\mathbf{W} \in \mathbb{R}^{N \times N}$ whose elements define the similarity between data points. Ideally, the affinity matrix is a block diagonal matrix such that a nonzero distance is assigned to the points from the same subspace. LRR, SSC and LRSSC construct the affinity matrix by enforcing low-rank, sparsity and low-rank plus sparsity constraints, respectively.

Low-Rank Representation (LRR) [16,17] seeks to find a low-rank representation matrix $\mathbf{C} \in \mathbb{R}^{N \times N}$ for input data $\mathbf{X}$. The basic model of LRR is the following:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad (1)$$

where the nuclear norm is used to approximate the rank of $\mathbf{C}$ and that results in the convex optimization problem.

Denote the SVD of $\mathbf{X}$ as $\mathbf{U}\Sigma\mathbf{V}^T$. The minimizer of Eq. (1) is uniquely given by [16]:

$$\hat{\mathbf{C}} = \mathbf{V}\mathbf{V}^T. \quad (2)$$

In the cases when data is contaminated by noise, the following problem needs to be solved:

$$\min_{\mathbf{C}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda \|\mathbf{C}\|_*. \quad (3)$$

The optimal solution of Eq. (3) has been derived in [18]:

$$\hat{\mathbf{C}} = \mathbf{V}_1 \left(\mathbf{I} - \frac{1}{\lambda} \Sigma_1^{-2} \right) \mathbf{V}_1^T, \quad (4)$$

where $\mathbf{U} = [\mathbf{U}_1 \ \mathbf{U}_2]$, $\Sigma = \text{diag}(\Sigma_1 \ \Sigma_2)$ and $\mathbf{V} = [\mathbf{V}_1 \ \mathbf{V}_2]$. Matrices are partitioned according to the sets $\mathcal{I}_1 = \{i : \sigma_i > \frac{1}{\sqrt{\lambda}}\}$ and $\mathcal{I}_2 = \{i : \sigma_i \leq \frac{1}{\sqrt{\lambda}}\}$.

Sparse Subspace Clustering (SSC) [15] requires that each data point is represented by a small number of data points from its own subspace and it amounts to solve the following minimization problem:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \text{diag}(\mathbf{C}) = \mathbf{0}. \quad (5)$$

The ℓ_1 norm is used as the tightest convex relaxation of the ℓ_0 quasi-norm that counts the number of nonzero elements of the solution. Constraint $\text{diag}(\mathbf{C}) = \mathbf{0}$ is used to avoid trivial solution of representing a data point as a linear combination of itself.

If data is contaminated by noise, the following minimization problem needs to be solved:

$$\min_{\mathbf{C}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \lambda \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \text{diag}(\mathbf{C}) = \mathbf{0}. \quad (6)$$

This problem can be efficiently solved using ADMM optimization procedure [47].

Low-Rank Sparse Subspace Clustering (LRSSC) [32] combines low-rank and sparsity constraints:

$$\min_{\mathbf{C}} \|\mathbf{C}\|_* + \lambda \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \mathbf{X} = \mathbf{X}\mathbf{C}, \quad \text{diag}(\mathbf{C}) = \mathbf{0}. \quad (7)$$

In the case of the corrupted data the following problem needs to be solved to approximate $\mathbf{C}$:

$$\min_{\mathbf{C}} \frac{1}{2} \|\mathbf{X} - \mathbf{X}\mathbf{C}\|_F^2 + \beta_1 \|\mathbf{C}\|_* + \beta_2 \|\mathbf{C}\|_1 \quad \text{s.t.} \quad \text{diag}(\mathbf{C}) = \mathbf{0}. \quad (8)$$

Once matrix $\mathbf{C}$ is obtained by LRR, SSC or LRSSC approach, the affinity matrix $\mathbf{W}$ is calculated as:

$$\mathbf{W} = |\mathbf{C}| + |\mathbf{C}|^T. \quad (9)$$

Given affinity matrix $\mathbf{W}$, spectral clustering [11,12] finds cluster membership of data points by applying k-means clustering to the eigenvectors of the graph Laplacian matrix $\mathbf{L} \in \mathbb{R}^{N \times N}$ computed from the affinity matrix $\mathbf{W}$.

3. Multi-view low-rank sparse subspace clustering

In this section we present Multi-view Low-rank Sparse Subspace Clustering (MLRSSC) algorithm with two different regularization approaches. We assume that we are given a dataset $\mathbf{X} = \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \dots, \mathbf{X}^{(n_v)}\}$ of n_v views, where each $\mathbf{X}^{(i)} = \{\mathbf{x}_j^{(i)} \in \mathbb{R}^{D^{(i)}}\}_{j=1}^N$ is described with its own set of $D^{(i)}$ features. Our objective is to find a joint representation matrix $\mathbf{C}$ that balances trade-off between the agreement across different views, while at the same time promotes sparsity and low-rankness of the solution.

We formulate joint objective function that enforces representation matrices $\{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}\}$ across different views to be

regularized towards a common consensus. Motivated by [42], we propose two regularization schemes of the MLRSSC algorithm: (i) MLRSSC based on pairwise similarities and (ii) centroid-based MLRSSC. The first regularization encourages similarity between pairs of representation matrices. The centroid-based approach enforces representations across different views towards a common centroid. Standard spectral clustering algorithm can then be applied to the jointly inferred affinity matrix.

3.1. Pairwise multi-view low-rank sparse subspace clustering

We propose to solve the following joint optimization problem over n_v views:

$$\begin{aligned} \min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}} & \sum_{v=1}^{n_v} (\beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1) \\ & + \sum_{1 \leq v, w \leq n_v, v \neq w} \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ \text{s.t.} & \quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{C}^{(v)}, \quad \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v, \end{aligned} \quad (10)$$

where $\mathbf{C}^{(v)} \in \mathbb{R}^{N \times N}$ is the representation matrix for view v . Parameters β_1 , β_2 and $\lambda^{(v)}$ define the trade-off between low-rank, sparsity constraint and the agreement across views, respectively. In the cases where we do not have a prior information that one view is more important than the others, $\lambda^{(v)}$ does not depend on a view v and the same value of $\lambda^{(v)}$ is used across all views $v = 1, \dots, n_v$. The last term in the objective in (10) is introduced to encourage similarities between pairs of representation matrices across views.

With all but one $\mathbf{C}^{(v)}$ fixed, we minimize the function (10) for each $\mathbf{C}^{(v)}$ independently:

$$\begin{aligned} \min_{\mathbf{C}^{(v)}} & \beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1 + \lambda^{(v)} \sum_{1 \leq w \leq n_v, w \neq v} \|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ \text{s.t.} & \quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{C}^{(v)}, \quad \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}. \end{aligned} \quad (11)$$

By introducing auxiliary variables $\mathbf{C}_1^{(v)}$, $\mathbf{C}_2^{(v)}$, $\mathbf{C}_3^{(v)}$ and $\mathbf{A}^{(v)}$, we reformulate the objective:

$$\begin{aligned} \min_{\mathbf{C}_1^{(v)}, \mathbf{C}_2^{(v)}, \mathbf{C}_3^{(v)}, \mathbf{A}^{(v)}} & \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \beta_2 \|\mathbf{C}_2^{(v)}\|_1 + \lambda^{(v)} \sum_{1 \leq w \leq n_v, w \neq v} \|\mathbf{C}_3^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ \text{s.t.} & \quad \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{A}^{(v)}, \quad \mathbf{A}^{(v)} = \mathbf{C}_2^{(v)} - \text{diag}(\mathbf{C}_2^{(v)}), \\ & \quad \mathbf{A}^{(v)} = \mathbf{C}_1^{(v)}, \quad \mathbf{A}^{(v)} = \mathbf{C}_3^{(v)}. \end{aligned} \quad (12)$$

The augmented Lagrangian is:

$$\begin{aligned} \mathcal{L}(\{\mathbf{C}_i^{(v)}\}_{i=1}^3, \mathbf{A}^{(v)}, \{\Lambda_i^{(v)}\}_{i=1}^4) &= \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \beta_2 \|\mathbf{C}_2^{(v)}\|_1 \\ &+ \lambda^{(v)} \sum_{1 \leq w \leq n_v, w \neq v} \|\mathbf{C}_3^{(v)} - \mathbf{C}^{(w)}\|_F^2 + \frac{\mu_1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{A}^{(v)}\|_F^2 \\ &+ \frac{\mu_2}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)} + \text{diag}(\mathbf{C}_2^{(v)})\|_F^2 + \frac{\mu_3}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}\|_F^2 \\ &+ \frac{\mu_4}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)}\|_F^2 + \text{tr}[\Lambda_1^{(v)T} (\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{A}^{(v)})] \\ &+ \text{tr}[\Lambda_2^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)} + \text{diag}(\mathbf{C}_2^{(v)}))] + \text{tr}[\Lambda_3^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)})] \\ &+ \text{tr}[\Lambda_4^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)})], \end{aligned} \quad (13)$$

where $\{\mu_i > 0\}_{i=1}^3$ are penalty parameters that need to be tuned and $\{\Lambda_i^{(v)}\}_{i=1}^4$ are Lagrange dual variables.

To solve the convex optimization problem in (12), we use Alternating Direction Method of Multipliers (ADMM) [47]. ADMM converges for the objective composed of two-block convex separable problems, but here the terms $\mathbf{C}_1^{(v)}$, $\mathbf{C}_2^{(v)}$ and $\mathbf{C}_3^{(v)}$ do not depend on each other and can be observed as one variable block.

Update rule for $\mathbf{A}^{(v)}$ at iteration $k+1$. Given $\{\mathbf{C}_i^{(v)}\}_{i=1}^3, \{\Lambda_i^{(v)}\}_{i=1}^4$ at iteration k , the matrix $\mathbf{A}^{(v)}$ that minimizes the objective in Eq. (13) is updated by the following update rule:

$$\begin{aligned} \mathbf{A}^{(v)} &= [\mu_1 \mathbf{X}^{(v)T} \mathbf{X}^{(v)} + (\mu_2 + \mu_3 + \mu_4) \mathbf{I}]^{-1} \\ &\quad \times (\mu_1 \mathbf{X}^{(v)T} \mathbf{X}^{(v)} + \mu_2 \mathbf{C}_2^{(v)} + \mu_3 \mathbf{C}_1^{(v)} \\ &\quad + \mu_4 \mathbf{C}_3^{(v)} + \mathbf{X}^{(v)T} \Lambda_1^{(v)} - \Lambda_2^{(v)} - \Lambda_3^{(v)} - \Lambda_4^{(v)}). \end{aligned} \quad (14)$$

The update rule follows straightforwardly by setting the partial derivative of $\mathcal{L}$ in Eq. (13) with respect to $\mathbf{A}^{(v)}$ to zero.

Update rule for $\mathbf{C}_1^{(v)}$ at iteration $k+1$. Given $\mathbf{A}^{(v)}$ at iteration $k+1$ and $\Lambda_3^{(v)}$ at iteration k , we minimize the objective in Eq. (13) with respect to $\mathbf{C}_1^{(v)}$:

$$\begin{aligned} \min_{\mathbf{C}_1^{(v)}} \mathcal{L}(\mathbf{C}_1^{(v)}, \mathbf{A}^{(v)}, \Lambda_3^{(v)}) \\ &= \min_{\mathbf{C}_1^{(v)}} \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \frac{\mu_3}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}\|_F^2 + \text{tr}[\Lambda_3^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)})] \\ &= \min_{\mathbf{C}_1^{(v)}} \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \frac{\mu_3}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}\|_F^2 + \text{tr}[\Lambda_3^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)})] \\ &\quad + \frac{\|\Lambda_3^{(v)}\|_F^2}{2\mu_3} \\ &= \min_{\mathbf{C}_1^{(v)}} \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \frac{\mu_3}{2} \left\| \mathbf{A}^{(v)} - \mathbf{C}_1^{(v)} + \frac{\Lambda_3^{(v)}}{\mu_3} \right\|_F^2, \end{aligned} \quad (15)$$

From [48], it follows that the unique minimizer of (15) is:

$$\mathbf{C}_1^{(v)} = \Pi_{\frac{\beta_1}{\mu_3}} \left(\mathbf{A}^{(v)} + \frac{\Lambda_3^{(v)}}{\mu_3} \right), \quad (16)$$

where $\Pi_{\beta}(\mathbf{Y}) = \mathbf{U} \pi_{\beta}(\Sigma) \mathbf{V}^T$ performs soft-thresholding operation on the singular values of $\mathbf{Y}$ and $\mathbf{U} \Sigma \mathbf{V}^T$ is the skinny SVD of $\mathbf{Y}$, here $\mathbf{Y} = \mathbf{A}^{(v)} + \mu_3^{-1} \Lambda_3^{(v)}$. $\pi_{\beta}(\Sigma)$ denotes soft thresholding operator defined as $\pi_{\beta}(\Sigma) = (|\Sigma| - \beta)_+ \text{sgn}(\Sigma)$ and $t_+ = \max(0, t)$.

Update rule for $\mathbf{C}_2^{(v)}$ at iteration $k+1$. Given $\mathbf{A}^{(v)}$ at iteration $k+1$ and $\Lambda_2^{(v)}$ at iteration k , we minimize the $\mathcal{L}$ in Eq. (13) with respect to $\mathbf{C}_2^{(v)}$:

$$\begin{aligned} \min_{\mathbf{C}_2^{(v)}} \mathcal{L}(\mathbf{C}_2^{(v)}, \mathbf{A}^{(v)}, \Lambda_2^{(v)}) \\ &= \min_{\mathbf{C}_2^{(v)}} \beta_2 \|\mathbf{C}_2^{(v)}\|_1 + \frac{\mu_2}{2} \left\| \mathbf{A}^{(v)} - \mathbf{C}_2^{(v)} + \frac{\Lambda_2^{(v)}}{\mu_2} \right\|_F^2 \\ &\quad \mathbf{C}_2^{(v)} = \mathbf{C}_2^{(v)} - \text{diag}(\mathbf{C}_2^{(v)}). \end{aligned} \quad (17)$$

The minimization of (17) gives the following update rules for matrix $\mathbf{C}_2^{(v)}$ [49,50]:

$$\begin{aligned} \mathbf{C}_2^{(v)} &= \pi_{\frac{\beta_2}{\mu_2}} \left(\mathbf{A}^{(v)} + \frac{\Lambda_2^{(v)}}{\mu_2} \right) \\ \mathbf{C}_2^{(v)} &= \mathbf{C}_2^{(v)} - \text{diag}(\mathbf{C}_2^{(v)}), \end{aligned} \quad (18)$$

where π_{β} denotes soft thresholding operator applied entry-wise to $(\mathbf{A}^{(v)} + \mu_2^{-1} \Lambda_2^{(v)})$.

Update rule for $\mathbf{C}_3^{(v)}$ at iteration $k+1$. Given $\mathbf{A}^{(v)}$ at iteration $k+1$ and $\Lambda_4^{(v)}$, $\sum_{1 \leq w \leq n_v, w \neq v} \mathbf{C}^{(w)}$ at iteration k , we minimize the

objective in Eq. (13) with respect to $\mathbf{C}_3^{(v)}$:

$$\begin{aligned} \min_{\mathbf{C}_3^{(v)}} \mathcal{L}(\mathbf{C}_3^{(v)}, \mathbf{A}^{(v)}, \Lambda_4^{(v)}) \\ &= \min_{\mathbf{C}_3^{(v)}} \lambda^{(v)} \sum_{1 \leq w \leq n_v, w \neq v} \|\mathbf{C}_3^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ &\quad + \frac{\mu_4}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)}\|_F^2 + \text{tr}[\Lambda_4^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)})]. \end{aligned} \quad (19)$$

The partial derivative of $\mathcal{L}$ in Eq. (13) with respect to $\mathbf{C}_3^{(v)}$:

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \mathbf{C}_3^{(v)}} &= [2\lambda^{(v)}(n_v - 1) + \mu_4] \mathbf{C}_3^{(v)} - 2\lambda^{(v)} \\ &\quad \times \sum_{1 \leq w \leq n_v, w \neq v} \mathbf{C}^{(w)} - \mu_4 \mathbf{A} - \Lambda_4^{(v)}. \end{aligned} \quad (20)$$

Setting the partial derivative in (20) to zero:

$$\mathbf{C}_3^{(v)} = [2\lambda^{(v)}(n_v - 1) + \mu_4]^{-1} (2\lambda^{(v)} \sum_{1 \leq w \leq n_v, w \neq v} \mathbf{C}^{(w)} + \mu_4 \mathbf{A} + \Lambda_4^{(v)}). \quad (21)$$

Update rules for dual variables $\{\Lambda_i^{(v)}\}_{i=1}^4$ at iteration $k+1$.

Given $\mathbf{A}^{(v)}$, $\{\mathbf{C}_i^{(v)}\}_{i=1}^3$ at iteration $k+1$, dual variables are updated with the following equations:

$$\begin{aligned} \Lambda_1^{(v)} &= \Lambda_1^{(v)} + \mu_1 (\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{A}^{(v)}) \\ \Lambda_2^{(v)} &= \Lambda_2^{(v)} + \mu_2 (\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)}) \\ \Lambda_3^{(v)} &= \Lambda_3^{(v)} + \mu_3 (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}) \\ \Lambda_4^{(v)} &= \Lambda_4^{(v)} + \mu_4 (\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)}). \end{aligned} \quad (22)$$

If data is contaminated by noise and does not perfectly lie in the union of subspaces, we modify the objective function as follows:

$$\begin{aligned} \min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \left(\frac{1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{C}^{(v)}\|_F^2 + \beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1 \right) \\ + \sum_{1 \leq v, w \leq n_v, v \neq w} \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ \text{s.t. } \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v. \end{aligned} \quad (23)$$

Update rule for $\mathbf{A}^{(v)}$ at iteration $k+1$ for corrupted data.

Given $\{\mathbf{C}_i^{(v)}\}_{i=1}^3, \{\Lambda_i^{(v)}\}_{i=1}^4$ at iteration k , the matrix $\mathbf{A}^{(v)}$ is obtained by equating to zero partial derivative of the augmented Lagrangian of problem (23):

$$\begin{aligned} \mathbf{A}^{(v)} &= [\mathbf{X}^{(v)T} \mathbf{X}^{(v)} + (\mu_2 + \mu_3 + \mu_4) \mathbf{I}]^{-1} \\ &\quad \times (\mathbf{X}^{(v)T} \mathbf{X}^{(v)} + \mu_2 \mathbf{C}_2^{(v)} + \mu_3 \mathbf{C}_1^{(v)} + \mu_4 \mathbf{C}_3^{(v)} - \Lambda_2^{(v)} - \Lambda_3^{(v)} - \Lambda_4^{(v)}). \end{aligned} \quad (24)$$

Update rules for $\{\mathbf{C}_i^{(v)}\}_{i=1}^3$ and dual variables $\{\Lambda_i^{(v)}\}_{i=2}^4$ are the same as in (16), (18), (21), (22), respectively.

These update steps are then repeated until the convergence or until the maximum number of iteration is reached. We check the convergence by verifying the following constraints at each iteration k : $\|\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}\|_{\infty} \leq \epsilon$, $\|\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)}\|_{\infty} \leq \epsilon$, $\|\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)}\|_{\infty} \leq \epsilon$ and $\|\Lambda_k^{(v)} - \Lambda_{k-1}^{(v)}\|_{\infty} \leq \epsilon$, for $v = 1, \dots, n_v$. After obtaining representation matrix for each view $\{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}\}$, we combine them by taking the element-wise average across all views. The next step of the algorithm is to find the assignment of the data points to corresponding clusters by applying spectral clustering algorithm to

Algorithm 1 Pairwise MLRSSC.

Input: $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, k , β_1 , β_2 , $\{\lambda^{(v)}\}_{v=1}^{n_v}$, $\{\mu_i\}_{i=1}^4$, μ^{max} , ρ
Output: Assignment of the data points to k clusters

- 1: Initialize: $\{\mathbf{C}_i^{(v)} = \mathbf{0}\}_{i=1}^3$, $\mathbf{A}^{(v)} = \mathbf{0}$, $\{\Lambda_i^{(v)} = \mathbf{0}\}_{i=1}^4$, $i = 1, \dots, n_v$
- 2: **while** not converged **do**
- 3: **for** $v = 1$ **to** n_v **do**
- 4: Fix others and update $\mathbf{A}^{(v)}$ by solving (14) in the case of clean data or (24) in the case of corrupted data
- 5: Fix others and update $\mathbf{C}_1^{(v)}$ by solving (16)
- 6: Fix others and update $\mathbf{C}_2^{(v)}$ by solving (18)
- 7: Fix others and update $\mathbf{C}_3^{(v)}$ by solving (21)
- 8: Fix others and update dual variables $\Lambda_2^{(v)}$, $\Lambda_3^{(v)}$, $\Lambda_4^{(v)}$ by solving (22) and also $\Lambda_1^{(v)}$ in the case of clean data
- 9: **end for**
- 10: Update $\mu_i = \min(\rho\mu_i, \mu^{max})$, $i = 1, \dots, 4$
- 11: **end while**
- 12: Combine $\{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}\}$ by taking the element-wise average
- 13: Apply spectral clustering [12] to the affinity matrix $\mathbf{W} = |\mathbf{C}_{avg}| + |\mathbf{C}_{avg}|^T$

the joint affinity matrix $\mathbf{W} = |\mathbf{C}_{avg}| + |\mathbf{C}_{avg}|^T$. Algorithm 1 summarizes the steps of the pairwise MLRSSC. Due to the practical reasons, we use the same initial values of $\{\mu_i\}_{i=1}^4$, ρ and μ^{max} for different views v and update $\{\mu_i\}_{i=1}^4$ after the optimizations of all views. However, it is possible to have more general approach with different initial values of $\{\mu_i\}_{i=1}^4$, ρ and μ^{max} for each view v , but this significantly increases the number of variables for optimization.

The problem in (10) is convex subject to linear constraints and all its subproblems can be solved exactly. Hence, theoretical results in [51] guarantee the global convergence of ADMM. The computational complexity of Algorithm 1 is $O(Tn_vN^3)$, where T is the number of iterations, $n_v \ll N$ is the number of views and N is the number of data points. In the experiments, we set the maximal T to 100, but the algorithm converged before the maximal number of iterations is exceeded ($T \approx 15 - 20$). Importantly, the computational complexity of spectral clustering step is $O(N^3)$, so the computational cost of the proposed representation learning step is Tn_v times higher.

3.2. Centroid-based multi-view low-rank sparse subspace clustering

In addition to the pairwise MLRSSC, we also introduce objective for the centroid-based MLRSSC which enforces view-specific representations towards a common centroid. We propose to solve the following minimization problem:

$$\min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} (\beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1 + \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^*\|_F^2) \quad (25)$$

$$\text{s.t. } \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{C}^{(v)}, \quad \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v,$$

where $\mathbf{C}^*$ denotes consensus variable.

This objective function can be minimized by the alternating minimization cycling over the views and consensus variable. Specifically, the following two steps are repeated: (1) fix consensus variable $\mathbf{C}^*$ and update each $\mathbf{C}^{(v)}$, $v = 1, \dots, n_v$ while keeping all others fixed and (2) fix $\mathbf{C}^{(v)}$, $v = 1, \dots, n_v$ and update $\mathbf{C}^*$.

By fixing all variables except one $\mathbf{C}^{(v)}$, we solve the following problem:

$$\min_{\mathbf{C}^{(v)}} \beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1 + \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^*\|_F^2 \quad (26)$$

$$\text{s.t. } \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{C}^{(v)}, \quad \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}.$$

Again, we solve the convex optimization problem using ADMM. We introduce auxiliary variables $\mathbf{C}_1^{(v)}$, $\mathbf{C}_2^{(v)}$, $\mathbf{C}_3^{(v)}$ and $\mathbf{A}^{(v)}$ and reformulate the original problem:

$$\min_{\mathbf{C}_1^{(v)}, \mathbf{C}_2^{(v)}, \mathbf{C}_3^{(v)}, \mathbf{A}^{(v)}} \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \beta_2 \|\mathbf{C}_2^{(v)}\|_1 + \lambda^{(v)} \|\mathbf{C}_3^{(v)} - \mathbf{C}^*\|_F^2$$

$$\text{s.t. } \mathbf{X}^{(v)} = \mathbf{X}^{(v)} \mathbf{A}^{(v)}, \quad \mathbf{A}^{(v)} = \mathbf{C}_2^{(v)} - \text{diag}(\mathbf{C}_2^{(v)}),$$

$$\mathbf{A}^{(v)} = \mathbf{C}_1^{(v)}, \quad \mathbf{A}^{(v)} = \mathbf{C}_3^{(v)}. \quad (27)$$

The augmented Lagrangian is:

$$\mathcal{L}(\{\mathbf{C}_i^{(v)}\}_{i=1}^3, \mathbf{A}^{(v)}, \{\Lambda_i^{(v)}\}_{i=1}^4) = \beta_1 \|\mathbf{C}_1^{(v)}\|_* + \beta_2 \|\mathbf{C}_2^{(v)}\|_1$$

$$+ \lambda^{(v)} \|\mathbf{C}_3^{(v)} - \mathbf{C}^*\|_F^2 + \frac{\mu_1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{A}^{(v)}\|_F^2$$

$$+ \frac{\mu_2}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)} + \text{diag}(\mathbf{C}_2^{(v)})\|_F^2 + \frac{\mu_3}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)}\|_F^2$$

$$+ \frac{\mu_4}{2} \|\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)}\|_F^2 + \text{tr}[\Lambda_1^{(v)T} (\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{A}^{(v)})]$$

$$+ \text{tr}[\Lambda_3^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)})] + \text{tr}[\Lambda_2^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_2^{(v)} + \text{diag}(\mathbf{C}_2^{(v)}))]$$

$$+ \text{tr}[\Lambda_4^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_3^{(v)})]. \quad (28)$$

Update rule for $\mathbf{C}_3^{(v)}$ at iteration $k+1$. Given $\mathbf{A}^{(v)}$ at iteration $k+1$ and $\mathbf{C}^*$, $\Lambda_4^{(v)}$ at iteration k , minimization of the objective in Eq. (28) with respect to $\mathbf{C}_3^{(v)}$ leads to the following update rule for $\mathbf{C}_3^{(v)}$:

$$\mathbf{C}_3^{(v)} = (2\lambda^{(v)} + \mu_4)^{-1} (2\lambda^{(v)} \mathbf{C}^* + \mu_4 \mathbf{A}^{(v)} + \Lambda_4^{(v)}). \quad (29)$$

Update rule for $\mathbf{C}^*$. By setting the partial derivative of the objective function in Eq. (25) with respect to $\mathbf{C}^*$ to zero we get the closed-form solution to $\mathbf{C}^*$:

$$\mathbf{C}^* = \frac{\sum_{v=1}^{n_v} \lambda^{(v)} \mathbf{C}^{(v)}}{\sum_{v=1}^{n_v} \lambda^{(v)}}. \quad (30)$$

It is easy to check that update rules for variables $\mathbf{A}^{(v)}$, $\mathbf{C}_1^{(v)}$, $\mathbf{C}_2^{(v)}$ and dual variables $\{\Lambda_i^{(v)}\}_{i=1}^4$ are the same as in the pairwise similarities based multi-view LRSSC (equations (14), (16), (18) and (22)).

In order to extend the model to the data contaminated by additive white Gaussian noise, the objective in (25) is modified as follows:

$$\min_{\mathbf{C}^{(1)}, \dots, \mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \frac{1}{2} \|\mathbf{X}^{(v)} - \mathbf{X}^{(v)} \mathbf{C}^{(v)}\|_F^2 + \beta_1 \|\mathbf{C}^{(v)}\|_* + \beta_2 \|\mathbf{C}^{(v)}\|_1$$

$$+ \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^*\|_F^2 \quad (31)$$

$$\text{s.t. } \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v.$$

Compared to the model for clean data, the only update rule that needs to be modified is for $\mathbf{A}^{(v)}$, which is the same as in pairwise MLRSSC given in Eq. (24).

In centroid-based MLRSSC there is no need to combine affinity matrices across views, since the joint affinity matrix can be directly computed from the centroid matrix i.e. $\mathbf{W} = |\mathbf{C}^*| + |\mathbf{C}^*|^T$. Algorithm 2 summarizes the steps of centroid-based MLRSSC. The computational complexity of Algorithm 2 is the same as the complexity of Algorithm 1.

Algorithm 2 Centroid-based MLRSSC.

Input: $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, k , β_1 , β_2 , $\{\lambda^{(v)}\}_{v=1}^{n_v}$, $\{\mu_i\}_{i=1}^4$, μ^{max} , ρ
Output: Assignment of the data points to k clusters

- 1: Initialize: $\{\mathbf{C}_i^{(v)} = \mathbf{0}\}_{i=1}^3$, $\mathbf{C}^* = \mathbf{0}$, $\mathbf{A}^{(v)} = \mathbf{0}$, $\{\Lambda_i^{(v)} = \mathbf{0}\}_{i=1}^4$, $i = 1, \dots, n_v$
- 2: **while** not converged **do**
- 3: **for** $v = 1$ **to** n_v **do**
- 4: Fix others and update $\mathbf{A}^{(v)}$ by solving (14) in the case of clean data or (24) in the case of corrupted data
- 5: Fix others and update $\mathbf{C}_1^{(v)}$ by solving (16)
- 6: Fix others and update $\mathbf{C}_2^{(v)}$ by solving (18)
- 7: Fix others and update $\mathbf{C}_3^{(v)}$ by solving (29)
- 8: Fix others and update dual variables $\Lambda_2^{(v)}$, $\Lambda_3^{(v)}$, $\Lambda_4^{(v)}$ by solving (22) and also $\Lambda_1^{(v)}$ in the case of clean data
- 9: **end for**
- 10: Update $\mu_i = \min(\rho\mu_i, \mu^{max})$, $i = 1, \dots, 4$
- 11: Fix others and update centroid $\mathbf{C}^*$ by solving (30)
- 12: **end while**
- 13: Apply spectral clustering [12] to the affinity matrix $\mathbf{W} = |\mathbf{C}^*| + |\mathbf{C}^*|^T$

4. Kernel multi-view low-rank sparse subspace clustering

The spectral decomposition of Laplacian enables spectral clustering to separate data points with nonlinear hypersurfaces. However, by representing data points as a linear combination of other data points, the MLRSSC algorithm learns the affinity matrix that models the linear subspace structure of the data. In order to recover nonlinear subspaces, we propose to solve the MLRSSC in RKHS by implicitly mapping data points into a high dimensional feature space.

We define $\Phi: \mathcal{X} \rightarrow \mathcal{F}$ to be a function that maps the original input space $\mathcal{X}$ to a high (possibly infinite) dimensional feature space $\mathcal{F}$. Since the presented update rules for the corrupted data of both pairwise and centroid-based MLRSSC depend only on the dot products $\langle \mathbf{X}^{(v)}, \mathbf{X}^{(w)} \rangle = \mathbf{X}^{(v)T} \mathbf{X}^{(w)}$, $v = 1, \dots, n_v$, both approaches can be solved in RKHS and extended to model nonlinear manifold structure.

Let $\Phi(\mathbf{X}^{(v)}) = \{\Phi(\mathbf{x}_i^{(v)}) \in \mathcal{F}\}_{i=1}^N$ denote the set of data points $\mathbf{X}^{(v)} = \{\mathbf{x}_i^{(v)} \in \mathbb{R}^D\}_{i=1}^N$ mapped into high-dimensional feature space. The objective function of pairwise kernel MLRSSC for data contaminated by noise is the following:

$$\begin{aligned} \min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \left(\frac{1}{2} \|\Phi(\mathbf{X}^{(v)}) - \Phi(\mathbf{X}^{(v)})\mathbf{C}^{(v)}\|_F^2 + \beta_1 \|\mathbf{C}^{(v)}\|_* \right. \\ \left. + \beta_2 \|\mathbf{C}^{(v)}\|_1 \right) + \sum_{1 \leq v, w \leq n_v, v \neq w} \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^{(w)}\|_F^2 \\ \text{s.t. } \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v. \end{aligned} \quad (32)$$

Similarly, the objective function of centroid-based MLRSSC in feature space for corrupted data is:

$$\begin{aligned} \min_{\mathbf{C}^{(1)}, \mathbf{C}^{(2)}, \dots, \mathbf{C}^{(n_v)}} \sum_{v=1}^{n_v} \left(\frac{1}{2} \|\Phi(\mathbf{X}^{(v)}) - \Phi(\mathbf{X}^{(v)})\mathbf{C}^{(v)}\|_F^2 + \beta_1 \|\mathbf{C}^{(v)}\|_* \right. \\ \left. + \beta_2 \|\mathbf{C}^{(v)}\|_1 + \lambda^{(v)} \|\mathbf{C}^{(v)} - \mathbf{C}^*\|_F^2 \right) \\ \text{s.t. } \text{diag}(\mathbf{C}^{(v)}) = \mathbf{0}, \quad v = 1, \dots, n_v. \end{aligned} \quad (33)$$

Since $\mathbf{A}^{(v)}$ is the only variable that depends on $\mathbf{X}^{(v)}$, the update rules for $\{\mathbf{C}_i^{(v)}\}_{i=1}^3$ and dual variables $\{\Lambda_i^{(v)}\}_{i=2}^4$ remain unchanged.

Table 2

Statistics of the multi-view datasets.

Dataset	Samples	Views	Clusters
UCI Digit	2000	3	10
Reuters	600	5	6
3-sources	169	3	6
Prokaryotic	551	3	4
Synthetic	1000	2	2

Update rule for $\mathbf{A}^{(v)}$ at iteration $k+1$. Given $\{\mathbf{C}_i^{(v)}\}_{i=1}^3, \{\Lambda_i^{(v)}\}_{i=2}^4$ at iteration k , the $\mathbf{A}^{(v)}$ is updated by the following update rule:

$$\begin{aligned} \mathbf{A}^{(v)} = [\Phi(\mathbf{X}^{(v)})^T \Phi(\mathbf{X}^{(v)}) + (\mu_2 + \mu_3 + \mu_4)\mathbf{I}]^{-1} \\ \times [\Phi(\mathbf{X}^{(v)})^T \Phi(\mathbf{X}^{(v)}) + \mu_2 \mathbf{C}_2^{(v)} + \mu_3 \mathbf{C}_1^{(v)} \\ + \mu_4 \mathbf{C}_3^{(v)} - \Lambda_2^{(v)} - \Lambda_3^{(v)} - \Lambda_4^{(v)}]. \end{aligned} \quad (34)$$

Substituting the dot product $\langle \Phi(\mathbf{X}^{(v)}), \Phi(\mathbf{X}^{(v)}) \rangle$ with the Gram matrix $\mathbf{K}^{(v)}$, we get the following update rule for $\mathbf{A}^{(v)}$:

$$\begin{aligned} \mathbf{A}^{(v)} = [\mathbf{K}^{(v)} + (\mu_2 + \mu_3 + \mu_4)\mathbf{I}]^{-1} \\ \times [\mathbf{K}^{(v)} + \mu_2 \mathbf{C}_2^{(v)} + \mu_3 \mathbf{C}_1^{(v)} + \mu_4 \mathbf{C}_3^{(v)} - \Lambda_2^{(v)} - \Lambda_3^{(v)} - \Lambda_4^{(v)}]. \end{aligned} \quad (35)$$

Update rule for $\mathbf{A}^{(v)}$ is the same in pairwise and centroid-based versions of the algorithm.

5. Experiments

In this section we present results that demonstrate the effectiveness of the proposed algorithms. The performance is measured on one synthetic and three real-world datasets that are commonly used to evaluate the performance of multi-view algorithms. Moreover, we introduce novel real-world multi-view dataset from molecular biology domain. We compared MLRSSC with the state-of-the-art multi-view subspace clustering algorithms, as well as with two baselines: best single view LRSSC and feature concatenation LRSSC.

5.1. Datasets

We report the experimental results on synthetic and four real-world datasets. We give a brief description of each dataset. Statistics of the datasets are summarized in Table 2.

UCI Digit dataset is available from the UCI repository.¹ This dataset consists of 2000 examples of handwritten digits (0–9) extracted from Dutch utility maps. There are 200 examples in each class, each represented with six feature sets. Following experiments in [45], we used three feature sets: 76 Fourier coefficients of the character shapes, 216 profile correlations and 64 Karhunen-Love coefficients.

Reuters dataset [52] contains features of documents available in five different languages and their translations over a common set of six categories. All documents are in the bag-of-words representation. We use documents originally written in English as one view and their translations to French, German, Spanish and Italian as four other views. We randomly sampled 100 documents from each class, resulting in a dataset of 600 documents.

3-sources dataset² is news articles dataset collected from three online news sources: BBC, Reuters, and The Guardian. All articles are in the bag-of-words representation. Of 948 articles, we

¹ <http://archive.ics.uci.edu/ml/datasets/Multiple+Features>.

² <http://mlg.ucd.ie/datasets/3sources.html>.

used 169 that are available in all three sources. Each article in the dataset is annotated with a dominant topic class.

Prokaryotic phyla dataset contains 551 prokaryotic species described with heterogeneous multi-view data including textual data and different genomic representations [53]. Textual data consists of bag-of-words representation of documents describing prokaryotic species and is considered as one view. In our experiments we use two genomic representations: (i) the proteome composition, encoded as relative frequencies of amino acids (ii) the gene repertoire, encoded as presence/absence indicators of gene families in a genome. In order to reduce the dimensionality of the dataset, we apply principal component analysis (PCA) on each of the three views separately and retain principal components explaining 90% of the variance. Each species in the dataset is labeled with the phylum it belongs to. Unlike previous datasets, this dataset is unbalanced. The most frequently occurring cluster contains 313 species, while the smallest cluster contains 35 species.

Synthetic dataset was generated in a way described in [42,54]. 1000 points are generated from two views, where data points for each view are generated from two-component Gaussian mixture models. Cluster means and covariance matrices for view 1 are: $\mu_1^{(1)} = (1 \ 1)$, $\Sigma_1^{(1)} = (1 \ 0.5; \ 0.5 \ 1.5)$, $\mu_2^{(1)} = (2 \ 2)$, $\Sigma_2^{(1)} = (0.3 \ 0; \ 0 \ 0.6)$, and for view 2 are: $\mu_1^{(2)} = (2 \ 2)$, $\Sigma_1^{(2)} = (0.3 \ 0; \ 0 \ 0.6)$, $\mu_2^{(2)} = (1 \ 1)$, $\Sigma_2^{(2)} = (1 \ 0.5; \ 0.5 \ 1.5)$.

5.2. Compared methods and parameters

We compare pairwise MLRSSC, centroid-based MLRSSC and kernel extensions of both algorithms (KMLRSSC) with the best performing state-of-the-art multi-view subspace clustering algorithms, including Co-regularized Multi-view Spectral Clustering (Co-Reg) [42], Robust Multi-view Spectral Clustering (RMSC) [44] and Convex Sparse Multi-view Spectral Clustering (CSMSC) [45]. Moreover, we also compare MLRSSC algorithms with two LRSSC baselines: (i) best single view Low-rank Sparse Subspace Clustering (LRRSC) [32] that performs single view LRSSC on each view and takes the individual view that achieves the best performance, and (ii) feature concatenation LRRSC that concatenates features of each individual view and performs single-view LRSSC on the joint view representation.

Co-regularized multi-view SC has a parameter α that we vary from 0.01 to 0.05 with step 0.01 [42]. We choose λ in RMSC from the set of the values: {0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100}, as tested in [44]. Parameter α in CSMSC is chosen from $\{10^{-1}, 10^{-2}\}$ and parameter β from $\{10^{-3}, 10^{-4}, 10^{-5}\}$ [45]. For all these algorithms the standard deviation of Gaussian kernel used to build similarity matrix is set to the median of the pairwise Euclidean distances between the data points [42,44,45]. The number of iterations of the Co-Reg SC is set to 100, but it converged within less than 10 iterations. The number of iterations of the CSMSC is set to 200 [45] and of the RMSC to 300, as set in the available source code provided by the authors. All other parameters of these algorithms are set to values based on the respective source codes provided by their authors.

For LRSSC and MLRSSC we first choose penalty parameter μ from the set of values {10, 100, 1000, 10000} with fixed β_1 , β_2 and $\lambda^{(v)}$. We set the same value μ for all constraints ($\mu_i, i = 1, \dots, 4$), but one can also optimize μ for each of the constraints. In each iteration we update μ to be $\rho\mu$ with fixed ρ of 1.5 and till the maximal value of μ (set to 10^6) is reached. For single-view LRSSC ρ is set to 1. Low-rank parameter β_1 is tuned from 0.1 to 0.9 with step 0.2 and sparsity parameter β_2 is set to $(1 - \beta_1)$. Consensus parameter λ is tuned from 0.3 to 0.9 with step 0.2. It is also possible to use different $\lambda^{(v)}$ for each view v , but since we did not have any prior information about the importance of views, we use the

same $\lambda = \lambda^{(v)}$ for each view v . For all datasets we use the variant adjusted for the corrupted data, except for the UCI digit dataset. In the kernel extension of MLRSSC, we use Gaussian kernel and optimize standard deviation for each view separately in range {0.5, 1, 5, 10, 50} times the median of the pairwise Euclidean distances between the data points, while holding other parameters fixed. Best sigma for pairwise MLRSSC was also used for centroid MLRSSC without further optimization. The maximum number of iterations is set to 100 and the convergence error tolerance to $\epsilon = 10^{-3}$ for linear MLRSSC and $\epsilon = 10^{-5}$ for kernel MLRSSC. We tune the parameters of each algorithm and report the best performance.

All compared methods have k-means as the last step of the algorithm. Since k-means depends on the initial cluster centroid positions and can yield different solution with different initializations, we run k-means 20 times and report the means and standard deviations of the performance measures. We evaluate clustering performance using five different metrics: precision, recall, F-score, normalized mutual information (NMI) and adjusted rand index (Adj-RI) [55]. For all these metrics, the higher value indicates better performance.

5.3. Results

Table 3 compares the clustering performance of the MLRSSC with other algorithms on four real-world datasets and one synthetic dataset. Results indicate that MLRSSC consistently outperforms all other methods in terms of all tested measures. On all five datasets, MLRSSC improves performance to a large extent which demonstrates the importance of combined low-rank and sparsity constraints. More specifically, the average NMI of the MLRSSC is higher than the second best method by 7%, 9%, 4%, 12% and 2% on the 3-sources, Reuters, UCI digit, Prokaryotic and synthetic datasets, respectively. Similar improvements can also be observed when using other metrics for measuring clustering performance.

Pairwise and centroid-based MLRSSC perform comparably, except on Prokaryotic dataset where pairwise MLRSSC is significantly better than the centroid-based MLRSSC, except in recall. When comparing linear MLRSSC with the kernel MLRSSC, linear MLRSSC performs better on 3-sources and Reuters datasets. Kernel MLRSSC outperforms linear MLRSSC on UCI Digit, Prokaryotic and synthetic datasets, although the difference on the UCI Digit dataset is not significant. However, this comes with the cost of tuning more parameters for computing the kernel. Better performance of linear MLRSSC on 3-sources and Reuters datasets is not surprising, since these datasets are very sparse (more than 95% values are zeros) and have a large number of features, much higher than the number of data points. On the other hand, UCI Digit, Prokaryotic and especially synthetic datasets have dense lower-dimensional feature vectors and benefit from the projection to a high-dimensional feature space.

5.4. Parameter sensitivity

MLRSSC trades-off low-rank, sparsity and consensus parameters: β_1 , β_2 and $\lambda^{(v)}$, respectively. In this section, we test the effect of these parameters on the performance of the MLRSSC. In all experiments, we set the sparsity parameter β_2 to $1 - \beta_1$, i.e. the higher value of a low-rank parameter leads to the lower value of a sparsity parameter and vice versa. This depends on whether the problem being solved requires exploiting more global or the local structure of the data.

Fig. 1 shows how the NMI metrics changes with different values of low-rank parameter β_1 for both pairwise and centroid-based MLRSSC, while keeping $\lambda^{(v)}$ parameter fixed. On the 3-sources, Reuters and UCI Digit, MLRSSC algorithm outperforms the second best algorithm regardless of the choice of β_1 . On the Prokaryotic

Table 3

Performance of different algorithms on five multi-view datasets. The mean and standard deviation of 20 runs of k-means clustering algorithm with different random initializations are reported.

Dataset	Method	F-score	Precision	Recall	NMI	Adj-RI
3-sources	Best Single View LRSSC	0.569 (0.039)	0.604 (0.015)	0.541 (0.058)	0.496 (0.024)	0.449 (0.042)
	Feat Concat LRSSC	0.579 (0.048)	0.593 (0.031)	0.571 (0.078)	0.521 (0.015)	0.455 (0.054)
	Co-Reg Pairwise	0.463 (0.020)	0.504 (0.049)	0.437 (0.033)	0.519 (0.036)	0.315 (0.033)
	Co-reg Centroid	0.505 (0.032)	0.551 (0.052)	0.467 (0.025)	0.514 (0.026)	0.370 (0.045)
	RMSC	0.477 (0.033)	0.515 (0.034)	0.453 (0.036)	0.517 (0.024)	0.330 (0.045)
	CSMSC	0.482 (0.026)	0.518 (0.056)	0.464 (0.027)	0.518 (0.026)	0.335 (0.039)
	Pairwise MLRSSC	0.659 (0.049)	0.707 (0.051)	0.619 (0.056)	0.594 (0.025)	0.565 (0.060)
	Centroid MLRSSC	0.654 (0.042)	0.696 (0.055)	0.619 (0.052)	0.595 (0.021)	0.557 (0.053)
	Pairwise KMLRSSC	0.541 (0.025)	0.619 (0.032)	0.482 (0.033)	0.529 (0.020)	0.424 (0.029)
Reuters	Centroid KMLRSSC	0.556 (0.045)	0.622 (0.049)	0.503 (0.044)	0.533 (0.031)	0.439 (0.056)
	Best Single View LRSSC	0.333 (0.003)	0.313 (0.007)	0.357 (0.019)	0.245 (0.008)	0.191 (0.005)
	Feat Concat LRSSC	0.347 (0.005)	0.319 (0.010)	0.384 (0.022)	0.283 (0.006)	0.204 (0.008)
	Co-Reg Pairwise	0.371 (0.009)	0.344 (0.016)	0.410 (0.023)	0.300 (0.014)	0.233 (0.017)
	Co-reg Centroid	0.362 (0.017)	0.331 (0.022)	0.409 (0.020)	0.291 (0.014)	0.221 (0.023)
	RMSC	0.361 (0.019)	0.325 (0.012)	0.412 (0.023)	0.297 (0.018)	0.217 (0.015)
	CSMSC	0.365 (0.005)	0.327 (0.010)	0.420 (0.014)	0.295 (0.020)	0.220 (0.008)
	Pairwise MLRSSC	0.428 (0.012)	0.389 (0.024)	0.486 (0.019)	0.390 (0.018)	0.300 (0.021)
	Centroid MLRSSC	0.432 (0.010)	0.395 (0.023)	0.482 (0.025)	0.394 (0.015)	0.306 (0.017)
UCI digit	Pairwise KMLRSSC	0.429 (0.013)	0.415 (0.018)	0.446 (0.016)	0.380 (0.018)	0.311 (0.017)
	Centroid KMLRSSC	0.426 (0.013)	0.410 (0.018)	0.443 (0.015)	0.373 (0.016)	0.307 (0.017)
	Best Single View LRSSC	0.702 (0.033)	0.659 (0.033)	0.755 (0.027)	0.754 (0.020)	0.666 (0.038)
	Feat Concat LRSSC	0.698 (0.038)	0.671 (0.046)	0.728 (0.032)	0.751 (0.021)	0.663 (0.043)
	Co-Reg Pairwise	0.694 (0.057)	0.671 (0.068)	0.718 (0.047)	0.739 (0.036)	0.658 (0.065)
	Co-reg Centroid	0.754 (0.067)	0.735 (0.082)	0.775 (0.050)	0.783 (0.033)	0.726 (0.075)
	RMSC	0.742 (0.070)	0.728 (0.080)	0.757 (0.061)	0.778 (0.040)	0.713 (0.079)
	CSMSC	0.775 (0.045)	0.725 (0.069)	0.836 (0.015)	0.819 (0.019)	0.748 (0.051)
	Pairwise MLRSSC	0.830 (0.048)	0.809 (0.070)	0.854 (0.027)	0.851 (0.023)	0.810 (0.054)
Prokaryotic	Centroid MLRSSC	0.835 (0.047)	0.819 (0.066)	0.854 (0.027)	0.854 (0.023)	0.817 (0.053)
	Pairwise KMLRSSC	0.827 (0.063)	0.800 (0.078)	0.861 (0.022)	0.855 (0.027)	0.807 (0.072)
	Centroid KMLRSSC	0.840 (0.043)	0.820 (0.065)	0.862 (0.019)	0.858 (0.020)	0.822 (0.048)
	Best Single View LRSSC	0.579 (0.057)	0.551 (0.016)	0.634 (0.100)	0.233 (0.026)	0.280 (0.051)
	Feat Concat LRSSC	0.584 (0.054)	0.542 (0.015)	0.644 (0.092)	0.218 (0.029)	0.275 (0.057)
	Co-Reg Pairwise	0.468 (0.023)	0.568 (0.023)	0.398 (0.022)	0.286 (0.021)	0.213 (0.031)
	Co-reg Centroid	0.459 (0.010)	0.567 (0.010)	0.386 (0.012)	0.296 (0.018)	0.206 (0.012)
	RMSC	0.447 (0.027)	0.567 (0.038)	0.369 (0.023)	0.315 (0.041)	0.198 (0.044)
	CSMSC	0.462 (0.026)	0.565 (0.024)	0.391 (0.026)	0.269 (0.022)	0.206 (0.033)
Synthetic	Pairwise MLRSSC	0.591 (0.016)	0.624 (0.003)	0.566 (0.036)	0.322 (0.002)	0.345 (0.016)
	Centroid MLRSSC	0.574 (0.028)	0.530 (0.014)	0.756 (0.124)	0.202 (0.018)	0.258 (0.032)
	Pairwise KMLRSSC	0.591 (0.056)	0.725 (0.068)	0.499 (0.048)	0.437 (0.039)	0.398 (0.082)
	Centroid KMLRSSC	0.582 (0.070)	0.712 (0.079)	0.492 (0.062)	0.424 (0.046)	0.384 (0.100)
	Best Single View LRSSC	0.624 (0.000)	0.560 (0.000)	0.704 (0.000)	0.182 (0.000)	0.152 (0.000)
	Feat Concat LRSSC	0.682 (0.000)	0.682 (0.000)	0.682 (0.000)	0.283 (0.000)	0.364 (0.000)
	Co-Reg Pairwise	0.660 (0.000)	0.637 (0.000)	0.685 (0.000)	0.260 (0.000)	0.295 (0.000)
	Co-reg Centroid	0.646 (0.000)	0.630 (0.000)	0.664 (0.000)	0.229 (0.000)	0.274 (0.000)
	RMSC	0.715 (0.000)	0.715 (0.000)	0.715 (0.000)	0.338 (0.000)	0.430 (0.000)
Synthetic	CSMSC	0.730 (0.000)	0.729 (0.000)	0.731 (0.000)	0.366 (0.000)	0.459 (0.000)
	Pairwise MLRSSC	0.689 (0.000)	0.689 (0.000)	0.689 (0.000)	0.294 (0.000)	0.379 (0.000)
	Centroid MLRSSC	0.690 (0.002)	0.690 (0.002)	0.690 (0.002)	0.296 (0.003)	0.380 (0.004)
	Pairwise KMLRSSC	0.742 (0.000)	0.742 (0.000)	0.742 (0.000)	0.385 (0.000)	0.484 (0.000)
	Centroid KMLRSSC	0.743 (0.000)	0.743 (0.000)	0.805 (0.000)	0.388 (0.002)	0.487 (0.000)

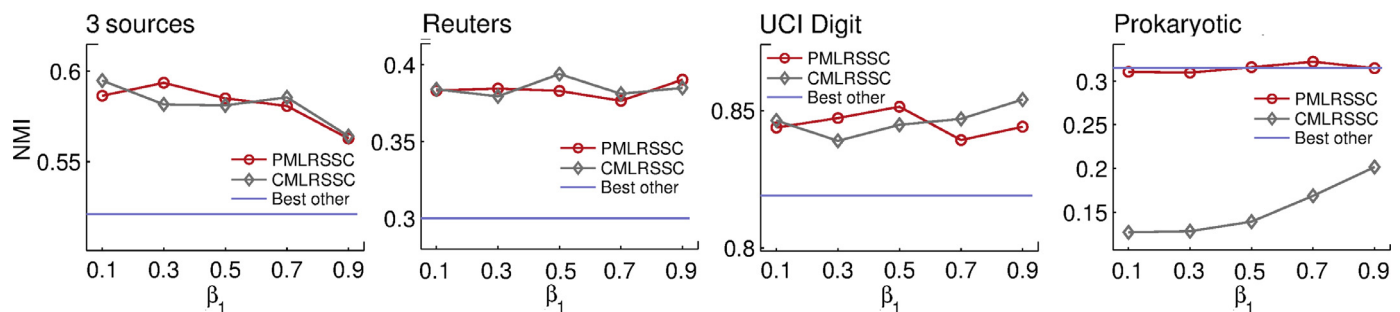


Fig. 1. The performance of the MLRSSC w.r.t. NMI measure when varying low-rank parameter β_1 and keeping consensus parameter $\lambda^{(v)}$ fixed. Sparsity parameter β_2 is set to $1 - \beta_1$. Blue line shows the best performing algorithm besides MLRSSC, among the algorithms listed in Table 3. PMLRSSC stands for pairwise MLRSSC and CMLRSSC for centroid-based MLRSSC.

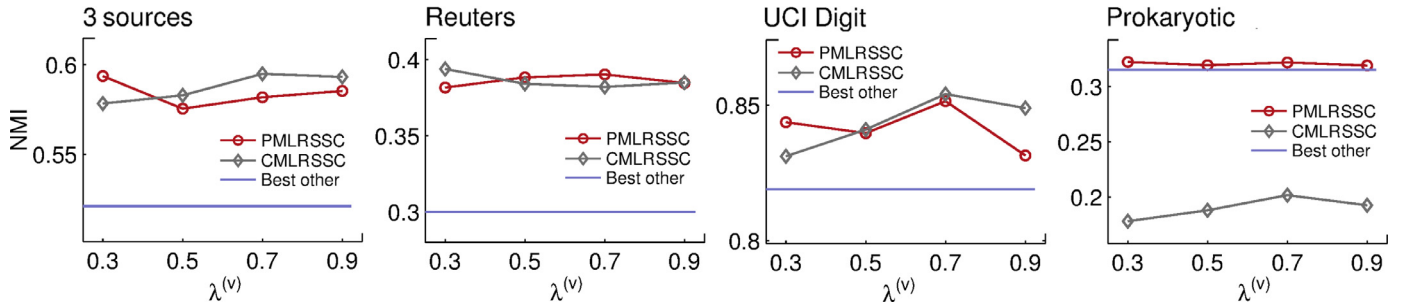


Fig. 2. The performance of the MLRSSC w.r.t. NMI measure when varying consensus parameter $\lambda^{(v)}$ and keeping low-rank parameter β_1 and sparsity parameter β_2 fixed. Blue line shows the best performing algorithm besides MLRSSC, among the algorithms listed in Table 3.

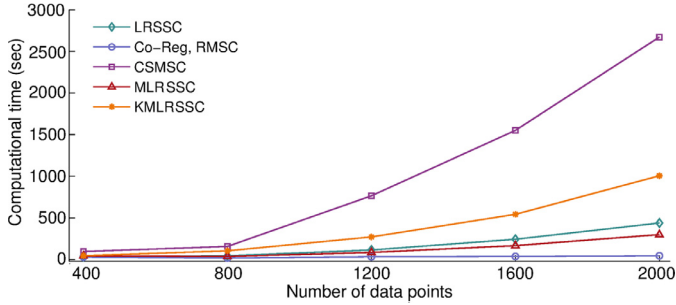


Fig. 3. Average computational time in seconds as a function of the number of data points, measured on the UCI Digit dataset. For the Co-Reg and MLRSSC algorithm times for pairwise regularization are shown, but they are similar for the centroid regularization. The difference between Co-Reg and RMSC can not be seen on this scale, so these two algorithms are shown together.

dataset, pairwise MLRSSC performs comparably to RMSC, but again the algorithm is insensitive to the β_1 parameter. On the other hand, centroid-based MLRSSC lags behind on this dataset with respect to NMI measure, but consistently improves its performance with the higher values of β_1 .

Next, we vary consensus parameter $\lambda^{(v)}$ and keep the low-rank parameter β_1 and sparsity parameter β_2 fixed. Fig. 2 shows the performance of the MLRSSC with respect to NMI measure for different values of $\lambda^{(v)}$. Similarly as when varying β_1 parameter, the MLRSSC performs consistently better than other algorithms regardless of the choice of $\lambda^{(v)}$. Again, the only exception is the centroid-based MLRSSC on the Prokaryotic dataset. These results prove that MLRSSC is pretty stable regardless of the choice of its parameters β_1 , β_2 and $\lambda^{(v)}$, as long as the parameters are chosen in an appropriate range.

5.5. Computational time and convergence

In order to check how computational time of the MLRSSC scales with the increase of the number of data points, we perform experiments on the UCI digit dataset and compare MLRSSC with other al-

gorithms. Computational time depends on the number of iterations and convergence conditions. We use the same number of iterations and error tolerance as when comparing performance of the algorithms. Fig. 3 shows the computational time averaged over 10 runs as a function of the number of data points. Figure 3 demonstrates that MLRSSC is more efficient than CSMSC. Compared to Co-Reg SC and RMSC, the better performance of MLRSSC comes with a higher computational cost.

Fig. 4 demonstrates the behavior of convergence conditions for pairwise MLRSSC. For ease of illustration, the errors are normalized and summed across views. It can be seen that on all four real-world datasets, the algorithm converges within 20 iterations. Centroid MLRSSC exhibits very similar behavior. Fig. 5 shows objective function value for both pairwise and centroid MLRSSC with the respect to number of iterations.

6. Concluding remarks

In this paper we proposed multi-view subspace clustering algorithm, called Multi-view Low-rank Sparse Subspace Clustering (MLRSSC), that learns a joint subspace representation across all views. The main property of the algorithm is to jointly learn an affinity matrix constrained by sparsity and low-rank. We defined optimization problems and derived ADMM-based algorithms for pairwise and centroid-based regularization schemes. In addition, we extended the proposed MLRSSC algorithm to nonlinear subspaces by solving the related optimization problem in reproducing kernel Hilbert space. Experimental results on multi-view datasets from various domains showed that proposed algorithms outperforms state-of-the-art multi-view subspace clustering algorithms.

High computational complexity presents serious drawback of spectral clustering algorithms. In the future work, we plan to explore how to improve the efficiency of the proposed approach to be applicable to large-scale multi-view problems. Moreover, we may consider how to extend the MLRSSC algorithm to handle incomplete data.

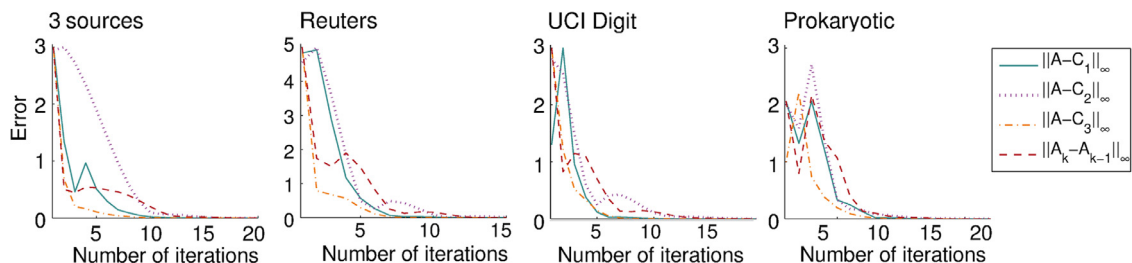


Fig. 4. Sum of normalized errors across views for pairwise MLRSSC. Behavior is very similar for centroid MLRSSC.

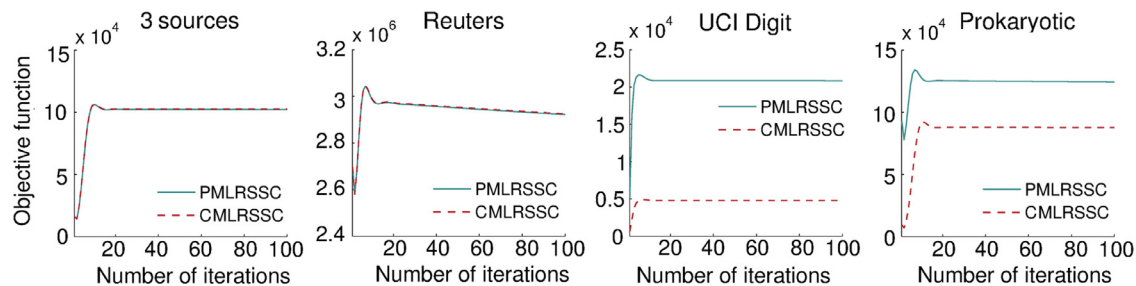


Fig. 5. Objective function value with the respect to number of iterations for pairwise and centroid MLRSSC.

Funding

This work has been supported by the Croatian Science Foundation grant IP-2016-06-5235 (Structured Decompositions of Empirical Data for Computationally-Assisted Diagnoses of Disease) and by the Croatian Science Foundation grant HRZZ-9623 (Descriptive Induction).

References

- [1] M. Amini, N. Usunier, C. Goutte, Learning from multiple partially observed views - an application to multilingual text categorization, in: *Advances in Neural Information Processing Systems* 22, 2009, pp. 28–36.
- [2] A. Vedaldi, V. Gulshan, M. Varma, A. Zisserman, Multiple kernels for object detection, in: *Proceedings of the International Conference on Computer Vision (ICCV)*, 2009.
- [3] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, *arXiv preprint arXiv:1304.5634* (2013).
- [4] I. Jolliffe, *Principal Component Analysis*, Springer Verlag, 1986.
- [5] P. Comon, Independent component analysis, a new concept? *Signal Process.* 36 (3) (1994) 287–314.
- [6] D.D. Lee, H.S. Seung, Learning the parts of objects by nonnegative matrix factorization, *Nature* 401 (1999) 788–791.
- [7] T. Liu, D. Tao, On the performance of Manhattan nonnegative matrix factorization, *IEEE Trans. Neural Netw. Learn. Syst.* 27 (9) (2016) 1851–1863.
- [8] J. Liu, C. Wang, J. Gao, J. Han, Multi-view clustering via joint nonnegative matrix factorization, in: *Proceedings of the 2013 SIAM International Conference on Data Mining, SIAM*, 2013, pp. 252–260.
- [9] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: *5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1967, pp. 281–297.
- [10] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2011) 52–68.
- [11] J. Shi, J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* 22 (8) (2000) 888–905.
- [12] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: *Advances in Neural Information Processing Systems* 14, 2001, pp. 849–856.
- [13] M. Filippone, F. Camastra, F. Masulli, S. Rovetta, A survey of kernel and spectral methods for clustering, *Pattern Recognit.* 41 (1) (2008) 176–190.
- [14] U. von Luxburg, A tutorial on spectral clustering, *Stat. Comput.* 17 (4) (2007) 395–416.
- [15] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (11) (2013) 2765–2781.
- [16] G. Liu, Y. Lin Zhouchen and Yu, Robust subspace segmentation by low-rank representation, in: *26th International Conference on Machine Learning*, 2010, pp. 663–670.
- [17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1) (2013) 171–184.
- [18] P. Favaro, R. Vidal, A. Ravichandran, A closed form solution to robust subspace estimation and clustering, in: *CVPR*, 2011, pp. 1801–1807.
- [19] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognit. Lett.* 43 (2014) 47–61.
- [20] K. Tang, R. Liu, Z. Su, J. Zhang, Structure-constrained low-rank representation, *IEEE Trans. Neural Netw. Learn. Syst.* 25 (12) (2014) 2167–2179.
- [21] K. Tang, D.B. Dunson, Z. Sub, R. Liub, J. Zhanga, J. Dong, Subspace segmentation by dense block and sparse representation, *Neural Netw.* 75 (1) (2016) 66–76.
- [22] B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, *Nature* 381 (6583) (1996) 607.
- [23] M. Aharon, M. Elad, A. Bruckstein, K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation, *IEEE Trans. Signal Process.* 54 (11) (2006) 4311–4322.
- [24] M. Elad, M. Aharon, Image denoising via sparse and redundant representations over learned dictionaries, *IEEE Trans. Image Process.* 15 (12) (2006) 3736–3745.
- [25] J. Mairal, M. Elad, G. Sapiro, Sparse representation for color image restoration, *IEEE Trans. Image Process.* 17 (1) (2008) 53–69.
- [26] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, *J. Mach. Learn. Res.* 11 (2010) 19–60.
- [27] M. Filipović, I. Kopriva, A comparison of dictionary based approaches to inpainting with an emphasis to independent component analysis learned dictionaries, *Inverse problems and imaging* 5 (4) (2011) 815–841.
- [28] W. Liu, Z.J. Zha, Y. Wang, K. Lu, D. Tao, p -Laplacian regularized sparse coding for human activity recognition, *IEEE Trans. Ind. Electron.* 63 (8) (2016) 5120–5129.
- [29] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, *IEEE Trans. Image Process.* 23 (5) (2014) 2019–2032.
- [30] E. Elhamifar, R. Vidal, Clustering disjoint subspaces via sparse representation, in: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 1926–1929.
- [31] B. Nasihatkon, R. Hartley, Graph connectivity in sparse subspace clustering, in: *CVPR*, 2011, pp. 2137–2144.
- [32] Y.-X. Wang, H. Xu, C. Leng, Provable subspace clustering: when LRR meets SSC, in: *Advances in Neural Information Processing Systems* 26, 2013, pp. 64–72.
- [33] Y. Liu, X. Li, C. Liu, H. Liu, Structure-constrained low-rank and partial sparse representation with sample selection for image classification, *Pattern Recognit.* 59 (C) (2016) 5–13.
- [34] C. Xu, D. Tao, C. Xu, Multi-view intact space learning, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (12) (2015) 2531–2544.
- [35] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, *Inf. Sci.* 385 (C) (2017) 338–352.
- [36] Y. Luo, D. Tao, K. Ramamohanarao, C. Xu, Y. Wen, Tensor canonical correlation analysis for multi-view dimension reduction, *IEEE Trans. Knowl. Data Eng.* 27 (11) (2015) 3111–3124.
- [37] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, *IEEE Trans. Image Process.* 24 (12) (2015) 5659–5670.
- [38] J. Yu, X. Yang, F. Gao, D. Tao, Deep multimodal distance metric learning using click constraints for image ranking, *IEEE Trans. Cybern.* (99) (2017) 1–11.
- [39] J. Yu, Y. Rui, Y.Y. Tang, D. Tao, High-order distance-based multiview stochastic learning in image classification, *IEEE Trans. Cybern.* 44 (12) (2014) 2431–2442.
- [40] J. Yu, Y. Rui, B. Chen, Exploiting click constraints and multi-view features for image re-ranking, *IEEE Trans. Multimedia* 16 (1) (2014) 159–168.
- [41] Y. Luo, T. Liu, D. Tao, C. Xu, Multiview matrix completion for multilabel image classification, *IEEE Trans. Image Process.* 24 (8) (2015) 2355–2368.
- [42] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: *Advances in Neural Information Processing Systems* 24, 2011, pp. 1413–1421.
- [43] B. Cheng, G. Liu, J. Wang, Z. Huang, S. Yan, Multi-task low-rank affinity pursuit for image segmentation, in: *2011 International Conference on Computer Vision*, 2011, pp. 2439–2446.
- [44] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition, in: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI Press, 2014, pp. 2149–2155.
- [45] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: single-view to multi-view, *IEEE Trans. Image Process.* 25 (6) (2016) 2833–2843.
- [46] J. Yu, B. Zhang, Z. Kuang, D. Lin, J. Fan, Iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning, *IEEE Trans. Inf. Forensics Secur.* 12 (5) (2017) 1005–1016.
- [47] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (1) (2011) 1–122.
- [48] J.-F. Cai, E.J. Candes, Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optim.* 20 (4) (2010) 1956–1982.
- [49] D.L. Donoho, De-noising by soft-thresholding, *IEEE Trans. Inf. Theory* 41 (3) (1995) 613–627.
- [50] I. Daubechies, M. DeFrise, C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Commun. Pure Appl. Math.* 57 (11) (2004) 1413–1457.
- [51] M. Hong, Z.-Q. Luo, On the linear convergence of the alternating direction method of multipliers, *Math Program.* 162 (1) (2017) 165–199.

- [52] D.D. Lewis, Y. Yang, T.G. Rose, F. Li, RCV1: a new benchmark collection for text categorization research, *J. Mach. Learn. Res.* 5 (2004) 361–397.
- [53] M. Brbić, M. Piškorec, V. Vidulin, A. Kriško, T. Šmuc, F. Supek, The landscape of microbial phenotypic traits and associated genes, *Nucleic Acids Res.* 44 (21) (2016) 10074–10090.
- [54] X. Yi, Y. Xu, C. Zhang, Multi-view EM Algorithm for Finite Mixture Models, Springer Berlin Heidelberg, pp. 420–425.
- [55] C.D. Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, New York, NY, USA, 2008.

Maria Brbic graduated at the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia, in 2013. She is currently employed at the Rudjer Boskovic Institute, Zagreb and she is pursuing Ph.D. degree in Computer Science. Her research interests are focused on unsupervised learning with applications to computational biology.

Ivica Kopriva received PhD degree in electrical engineering from the University of Zagreb, Croatia, in 1998, with the topic on blind source separation. He has been senior research scientist with the ECE Department, The George Washington University, Washington, DC, USA, 2001–2005. Since 2006, he is a senior scientist at the Rudjer Boskovic Institute, Zagreb, Croatia. His research is focused on unsupervised learning with applications in medical imaging and chemometrics. He has co-authored over 40 papers in internationally recognized journals. He holds three US patents.