

Robust low-rank kernel multi-view subspace clustering based on the Schatten p -norm and correntropy

Xiaoqian Zhang^{a,b}, Huaijiang Sun^{a,*}, Zhigui Liu^b, Zhenwen Ren^b, Qiongjie Cui^a, Yanmeng Li^a

^aSchool of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, PR China

^bSchool of Information Engineering, Southwest University of Science and Technology, Mianyang 621010, PR China

ARTICLE INFO

Article history:

Received 19 September 2018

Revised 26 October 2018

Accepted 27 October 2018

Available online 28 October 2018

Keywords:

Subspace clustering

Multi-view data

Low-rank kernel

Schatten p -norm

Correntropy

ABSTRACT

Multi-view subspace clustering, which studies the similarities and differences among data in multiple views, is an efficient clustering problem. However, to deal with problems that have non-linear structures and non-Gaussian noise in multi-view data, existing clustering methods relax the original problem convexly. The solutions that are generated by these convex relaxations are not the optimal solutions to the original problem. To overcome this deficiency, this paper presents a robust low-rank kernel multi-view subspace clustering approach that combines the non-convex Schatten p -norm ($0 < p \leq 1$) regularizer with the “kernel trick”, which can efficiently deal with problems that have non-linear structures in multi-view data via non-convex methods. In addition, the correntropy is introduced into our model, which is a robust measure of the corruptions that are caused by non-Gaussian noise. Moreover, our method can learn the joint subspace representation of all views. Because it learns a low-rank kernel mapping, the data in the feature space are both low-rank and self-expressed. The optimization problems are solved efficiently via an iterative algorithm (HQ-ADMM). This algorithm can ensure that each iteration has closed-form solutions, which simplifies the optimization problems substantially. Experimental comparisons on five real-world datasets demonstrate that the proposed algorithm outperforms state-of-the-art multi-view subspace clustering algorithms.

© 2018 Elsevier Inc. All rights reserved.

1. Introduction

Subspace clustering, which refers to the clustering of samples that belong to the same subspace into the same cluster, plays an important role in tasks such as data processing and pattern recognition [9,31,48]. Many methods are summarized in the literature [33], which can be divided into four main groups: algebraic methods, iterative methods, statistical methods, and methods that are based on spectral clustering. For dealing with noise and damage in data points, various approaches that are based on sparse representation theory, rank minimization and the Frobenius norm are proposed. A typical example is the sparse subspace clustering method (SSC), which is proposed in [10] and constructs the affinity matrix according to the sparsest representation coefficient matrix that is produced in the low-dimensional subspace of the data. Analogously, low-rank-representation-based (LRR) and low-rank subspace clustering (LRSC) are proposed [22,34], which construct the

* Corresponding author.

E-mail addresses: zhxq0528@163.com (X. Zhang), sunhuaijiang@njust.edu.cn (H. Sun), liuzhigui@swust.edu.cn (Z. Liu), rzwn@njust.edu.cn (Z. Ren), cuiqiongjie@126.com (Q. Cui), yanmengli6@126.com (Y. Li).

affinity matrix with the low-rank coefficients that are obtained via their frameworks. To ensure a lower computational cost, the Frobenius norm is used to construct a sparse similarity graph (L2-Graph) [27]. Combining the advantages of an LRR (globality) and an auto-encoder (self-supervision-based locality), the low-rank constrained auto-encoder (LRAE) method is proposed [8]. However, these methods are self-representation methods, which can only handle linear (or affine) subspaces. Unfortunately, the data points may not fit exactly a linear subspace model in practice.

To deal with non-linearity in data points, the “kernel trick” is used for subspace clustering: kernel sparse subspace clustering on the SPD Riemannian manifold (KSSCR) [40] applies the log-Euclidean kernel on SPD matrices. To represent the data points in a high-dimensional Hilbert space with a linear subspace structure, a low-rank kernel subspace clustering method is proposed for learning a low-rank kernel mapping, which can better handle non-linear subspaces [17]. However, all of these methods relax the original problem convexly via a nuclear norm. The solutions that are generated by these convex relaxations differ substantially from the solution of the original problem. This enables us to use non-convex agents as much as possible, while ensuring that the computational complexity does not increase significantly.

Many studies on non-convex agents have been proposed recently. For example, [38,50] demonstrate that using a ℓ_p -norm to handle image noise is more effective. In addition, research results on non-convex rank minimization functions were reported recently. The Schatten p -norm is introduced into the mathematical model to recover noisy data [18]. The matrix rank is approximated via a truncated nuclear norm [16]. The Schatten p -norm is also used to regularize the LRR model (SPM) [46].

Although they have been demonstrated to be effective in many applications, these methods principally focus on single-view data. In practical applications, visual data are mostly multi-view [45]. For example, color histograms or shape features can be used to describe an image. In the same learning task, multi-view data can yield better performance than single-view data because different views can provide complementary information [15,37,43]. For investigating the complementary information that is involved in multi-view data, many multi-view clustering methods have been proposed [7]. By ignoring the local structure of each view, some of the earlier methods simply splice all the features in the multi-view data and perform subspace clustering [2]. These methods can't improve the subspace clustering performance on multi-view data. A co-regularization framework is proposed for performing multi-view subspace clustering [19]. Diversity among the representations is explicitly enforced in the model of [6]. The main objective of many researchers is to learn a consistent clustering structure and the subspace representation of each view simultaneously [12]. Furthermore, other methods extend a single-view subspace clustering algorithm to multi-view scenarios, in which the affinity matrix is constructed on each view [24]. However, most of the data are not pure (for example, corrupted by noise). The process of extending a single-view algorithm to process multi-view data may cause noise to propagate in the affinity matrix, thereby degrading the clustering performance. To improve the robustness of the model to single-view data, many scholars add correntropy to the clustering model when the data contain non-Gaussian noise [14,35]. For multi-view data, a robust multi-view spectral clustering method is proposed that is based on a low-rank and sparse decomposition [36]. Unfortunately, these methods still relax the original problem convexly.

Recently, a deep learning (DL)-related method, namely, a canonical correlation analysis network (CCANet), was proposed, in which two-view features are used for image classification [39]. For dimension reduction, an unsupervised deep-learning framework is proposed that utilizes local deep-feature alignment (LDFA) for dimension reduction; this method demonstrates satisfactory clustering performance on a single-view dataset [44]. In [47], view-specific samples are modeled as a non-linear mapping of uniform but latent intact samples for all the views, and the view-specific similarity matrices are estimated based on the uniform latent one. Motivated by a co-regularization framework, a multi-view spectral clustering framework is proposed, which performs clustering tasks by constructing a multi-view data sharing affinity matrix [5]. By extending classical SSC and LRR methods, multi-modal subspace clustering is proposed in [1], which enforces the common representation across the modalities. In [42], an active three-way clustering method that is based on a low-rank matrix is proposed, which can improve the clustering accuracy when clustering high-dimensional multi-view data.

In practice, many multi-view datasets are better modeled by non-linear manifolds. Hence, many of the above methods use the “kernel trick” to extend their models to better handle nonlinear structures in the data. However, these kernels are pre-specified and the original problem is relaxed convexly. With the predefined kernels that are used in these methods, after (implicit) mapping to the feature space, the data points are not guaranteed to be low-rank and, thus, are very unlikely to form multiple low-dimensional subspace structures. In addition, the solutions that are generated by these convex relaxations are not optimal solutions to the original problems. Therefore, this paper proposes a novel robust multi-view subspace clustering model for solving these problems. Since high-dimensional data may contain non-Gaussian noise (such as missing data noise and outliers) and nonlinear structures, we propose a robust low-rank kernel multi-view subspace clustering framework that is based on the Schatten p -norm and correntropy (RLKMSC) and has two regularization schemes for constructing an affinity matrix that is shared among all views: (i) the pairs of views and (ii) toward a common centroid. To evaluate the effectiveness of the proposed method, we conduct extensive experiments on four benchmarks. The main contributions of this study can be summarized below.

- To the best of our knowledge, this is the first work on applying a non-convex low-rank kernel into multi-view data for subspace clustering.
- To better handle non-linear subspaces, a low-rank kernel mapping is learned such that the data in the resulting feature space are both low-rank and self-expressive.

Table 1
Notations and abbreviations.

Notation	Definition
I	Identity matrix
$\mathbf{1}$	All-ones column vector
N	Number of data points
v	v -th view
n_v	Number of views
$D^{(v)}$	Dimension of the data points in a view v
$\mathbf{X}^{(v)} \in \mathbb{R}^{D^{(v)} \times N}$	Data matrix in view v
$\mathbf{Z}^{(v)} \in \mathbb{R}^{N \times N}$	Representation matrix in view v
$\mathbf{Z}^* \in \mathbb{R}^{N \times N}$	Centroid representation matrix
$\mathbf{W} \in \mathbb{R}^{N \times N}$	Affinity matrix
$\phi(\mathbf{X}^{(v)})$	Feature mapping of the data matrix in view v
$\mathbf{K}^{(v)} \in \mathbb{R}^{N \times N}$	Unknown kernel Gram matrix in view v
$\mathbf{K}_G^{(v)} \in \mathbb{R}^{N \times N}$	Specified positive-definite kernel matrix in view v
$\ \cdot\ _q$	Arbitrary matrix norm
$\text{tr}(\cdot)$	Trace operator of a matrix
$\text{diag}(\cdot)$	Vector of the diagonal elements of a matrix

- To approximate the rank of the data in the feature space effectively, the Schatten p -norm regularizer is used and the closed-form solution for solving this sub-problem is presented.
- To handle non-Gaussian noise in the data, the correntropy is used. Moreover, considering the characteristic half-quadratic theories (HQ) [26] and the alternating-direction method of multipliers (ADMM) [4], an iterative algorithm (HQ-ADMM) is developed for implementing the proposed methods.

The remainder of this paper is organized as follows: Section 2 reviews related works. Section 3 introduces two novel models for robust low-rank kernel multi-view subspace clustering that are based on the Schatten p -norm and correntropy. Section 4 presents the experimental results on five datasets, which are used to evaluate the effectiveness and robustness of RLKMSC. Section 5 presents the conclusions of this paper.

2. Background and related work

In this section, the main symbols that are used in this article are listed first, followed by a brief introduction to subspace self-expression, the Schatten p -norm, the kernel strategy and correntropy.

2.1. Main notations

For convenience, we list important mathematic notations that are used throughout the paper in Table 1. Matrices are represented in bold uppercase, while vectors are represented in bold lowercase.

2.2. Related work

2.2.1. Subspace self-expressiveness

Recent self-expressiveness-based methods resort to the so-called subspace self-expressiveness property, namely, a point from a subspace can be expressed as a linear combination of other points in the same subspace, and utilize the self-expression coefficient matrix as the affinity matrix for spectral clustering. Given a set of N data points $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$, $\mathbf{X} = \mathbf{XZ}$ express the self-expressiveness of a subspace, where $\mathbf{Z}$ is the self-expression coefficient matrix. Assuming the subspaces are independent, the optimal solution for $\mathbf{Z}$, which is obtained by minimizing various norms of $\mathbf{Z}$, has a block-diagonal structure: $z_{ij} = 0$ when points i and j are from the same subspace. Namely, we have the following optimization problem:

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_q \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XZ}, \mathbf{Z}_{ii} = 0 \quad (1)$$

where the constraint $\mathbf{Z}_{ii} = 0$ is used to prevent the trivial identity solution for sparse norms of $\mathbf{Z}$. In the literature, various norms on $\mathbf{Z}$ have been used to regularize subspace clustering, such as the ℓ_1 -norm in [10], the nuclear norm in [22,34], the Schatten p -norm in [46], and a mixture of the ℓ_1 - and ℓ_2 -norms in [41].

In another series of local-based higher-order models, the affinity matrix is constructed from the residuals that are fitted by the local subspace model. In contrast, the self-expressiveness-based methods build holistic connections (or affinities) for all points in a single and principled optimization problem.

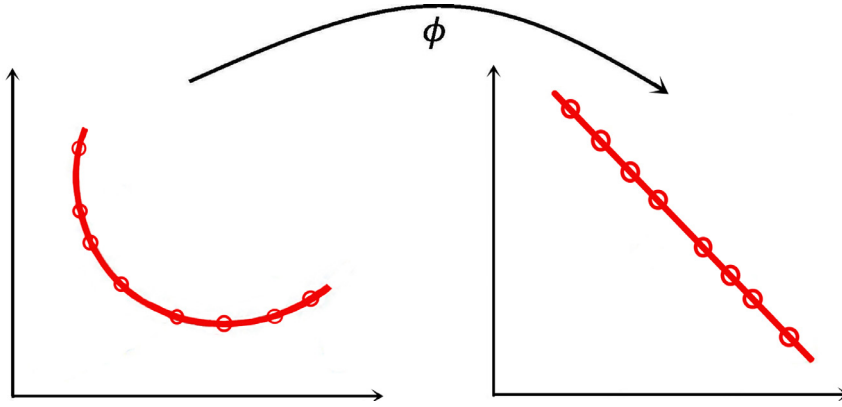


Fig. 1. Subspace-structure-preserving feature mapping: non-linear subspaces are mapped to linear ones in high-dimensional feature space.

2.2.2. Schatten p -norm

The Schatten p -norm of a matrix $\mathbf{X} \in \mathbb{R}^{D \times N}$ is expressed by

$$\|\mathbf{X}\|_{S_p} = \left(\sum_{i=1}^{\min(D,N)} \delta_i^p \right)^{1/p} \quad (2)$$

where $0 < p \leq 1$ and δ_i is the i -th largest singular value. We obtain the following formula:

$$\|\mathbf{X}\|_{S_p}^p = \sum_{i=1}^{\min(D,N)} \delta_i^p \quad (3)$$

This equation is equivalent to nuclear norm $\|\mathbf{X}\|_*$ when $p = 1$, while the Schatten 0-norm is the rank of $\mathbf{X}$. Therefore, $\|\mathbf{X}\|_{S_p}$ is a better approximation of the rank of $\mathbf{X}$ than $\|\mathbf{X}\|_*$.

2.2.3. Kernel trick

With kernel methods, data points can be mapped to higher-dimensional feature spaces. Then, the nonlinear analysis in the input data space can be transformed into a linear analysis in the feature space [29]. The “kernel trick” is a common use of kernel methods, in which feature mapping is implicit and the inner products between pairs of data points in the feature space are computed as kernel values. Fig. 1 illustrates the subspace-structure-preserving feature mapping in 2-D. The ambient dimension is very high in the feature space. For commonly used kernels such as the Gaussian RBF, although the kernel trick is relatively computationally cheap, we can’t accurately determine how the data points are mapped into the feature space. Therefore, in the context of subspace clustering, it is very likely that there will be no expected low-dimensional linear subspace structure in the feature space after implicit feature mapping.

To map the data into a high-dimensional Hilbert space with a linear subspace structure, the feature mapping, which is denoted as $\phi(\mathbf{X})$, should be low-rank and we also expect it to be self-expressive. By considering these features comprehensively and implementing them within the SSC framework [10], we can obtain the following optimization problem:

$$\min_{\mathbf{K}, \mathbf{Z}} \|\phi(\mathbf{X})\|_{S_p}^p + \lambda \|\mathbf{Z}\|_1 \quad \text{s.t.} \quad \phi(\mathbf{X}) = \phi(\mathbf{X})\mathbf{Z}, \mathbf{Z}_{ii} = 0 \quad (4)$$

where $\mathbf{K} = \phi(\mathbf{X})^T \phi(\mathbf{X})$ is the unknown kernel Gram matrix and λ is a tradeoff parameter.

2.2.4. Correntropy

In the literature, [23], the correntropy of two arbitrary variables, namely, x and y , is defined as

$$V(x, y) = E[k_\sigma(x, y)] \quad (5)$$

where $k_\sigma(\cdot, \cdot)$ is a kernel function and $E[\cdot]$ denotes the mathematical expectation. In this paper, we define $k_\sigma(x_1, x_2) = \exp(-\frac{\|x_1 - x_2\|_2^2}{2\sigma^2})$ (Gaussian kernel), where parameter σ denotes the width of the kernel. A general similarity measurement is extended as the correntropy-induced metric (CIM) [23] between any two variables. It is formally defined as

$$\text{CIM}(x, y) = (1 - V(x, y))^{\frac{1}{2}} = \left(1 - \frac{1}{N} \sum_{i=1}^N k_\sigma(x_i, y_i) \right)^{\frac{1}{2}} \quad (6)$$

For an optimization problem, maximizing the correntropy is equivalent to minimizing CIM, which is also called the maximum correntropy criterion (MCC) [23]. CIM is a local metric. When the errors are comparatively small, CIM can approximate

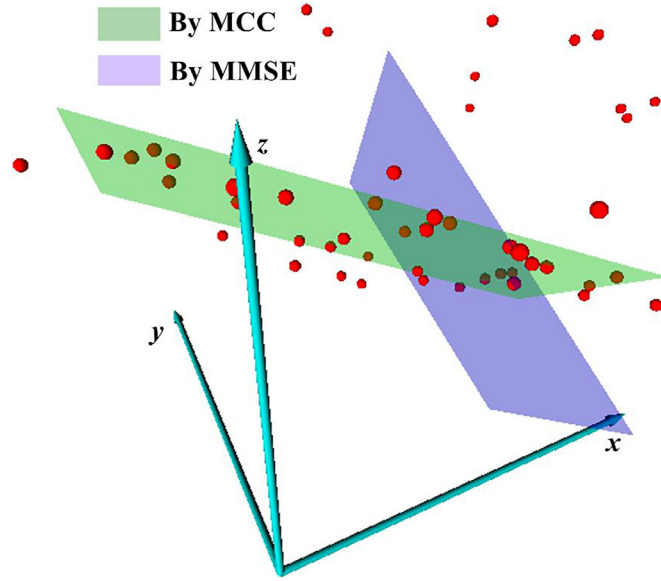


Fig. 2. Regression results with the MCC and MMSE criteria.

the absolute error. For large errors, which are typically caused by outliers, the value of CIM is close to 1; hence, their effect on CIM is limited. This is why CIM is more robust to non-Gaussian noises.

A simple regression experiment is conducted by [35] to help readers better understand the principle of the suppression of non-Gaussian noise by correntropy. As shown in Fig. 2, the objective is to fit a plane to sample points in a 3-D space; however, some of the sample points contain impulse noise. The minimum mean square error (MMSE) is a global measure and is used as a comparison algorithm. MCC substantially outperforms MMSE, which demonstrates the robustness of MCC.

3. Robust low-rank kernel multi-view subspace clustering

In this section, considering the non-linear structure and non-Gaussian noises of data points, we present the robust low-rank kernel multi-view subspace clustering (RLKMSC) framework and implement it with two regularization methods. For a dataset of n_v views, which is defined as $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, where $\mathbf{X}^{(v)} = [\mathbf{x}_1^{(v)}, \mathbf{x}_2^{(v)}, \dots, \mathbf{x}_N^{(v)}] \in \mathbb{R}^{D^{(v)} \times N}$, our objective is to construct the affinity matrix that is shared among all views by learning a joint representation matrix $\mathbf{Z}$, while ensuring that the data (implicit) mapping to the feature space is low-rank.

The joint objective function is formulated, in which the representation matrices $\{\mathbf{Z}^{(v)}\}_{v=1}^{n_v}$ of all views are constrained and regularized toward a consensus. Motivated by [5,35,46], two regularization methods are proposed in the RLKMSC framework: (i) centroid-based RLKMSC and (ii) RLKMSC that is based on pairwise similarities. The former forces $\mathbf{Z}^{(v)}$ toward a common centroid $\mathbf{Z}^*$. The latter encourages similarity between $\mathbf{Z}^{(v)}$ and $\mathbf{Z}^{(w)}$. Fig. 3 presents an overview of the proposed framework. First, with the multi-view data, our methods dependently learn the coefficient matrix $\mathbf{Z}^{(m)}$ of each view. Second, the global coefficient matrix $\mathbf{Z}^*$, which is obtained via the above two regularization methods, is used to explore the globally consistent information from all single views. When the global coefficient matrix satisfies the convergence condition, we can use it to build the affinity matrix. Finally, the spectral clustering algorithm is applied to the affinity matrix to obtain the clustering result. In these models, the Schatten p -norm is combined with the “kernel trick” and correntropy for clustering.

3.1. Centroid-based RLKMSC

According to Eq. (4), we propose a centroid-based regularization method that forces the representations of all views toward a common centroid. The objective function is expressed as follows:

$$\begin{aligned} \min_{\{\mathbf{K}^{(v)}\}_{v=1}^{n_v}, \{\mathbf{Z}^{(v)}\}_{v=1}^{n_v}} & \sum_{v=1}^{n_v} (\|\phi(\mathbf{X}^{(v)})\|_{S_p}^p + \lambda \|\mathbf{Z}^{(v)}\|_1 + \beta^{(v)} \|\mathbf{Z}^{(v)} - \mathbf{Z}^*\|_F^2) \\ \text{s.t. } & \phi(\mathbf{X}^{(v)}) = \phi(\mathbf{X}^{(v)})\mathbf{Z}^{(v)}, \text{diag}(\mathbf{Z}^{(v)}) = \mathbf{0}, v = 1, \dots, n_v \end{aligned} \quad (7)$$

where $\mathbf{K}^{(v)} = \phi(\mathbf{X}^{(v)})^T \phi(\mathbf{X}^{(v)})$ is an unknown kernel Gram matrix, λ and $\beta^{(v)}$ are trade-off parameters for view v and $\mathbf{Z}^*$ denotes the consensus variable. This problem can be transformed into the following problem:

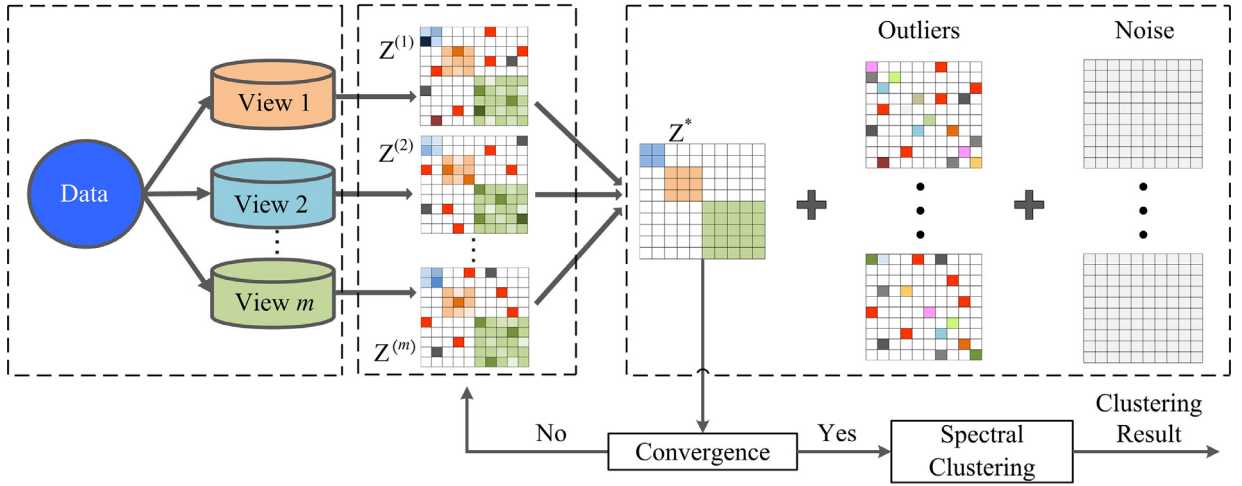


Fig. 3. Overview of the proposed robust low-rank kernel multi-view subspace clustering framework.

$$\begin{aligned} \min_{\mathbf{K}^{(v)}, \mathbf{Z}^{(v)}} \quad & \|\phi(\mathbf{X}^{(v)})\|_{S_p}^p + \lambda \|\mathbf{Z}^{(v)}\|_1 + \beta^{(v)} \|\mathbf{Z}^{(v)} - \mathbf{Z}^*\|_F^2 \\ \text{s.t.} \quad & \phi(\mathbf{X}^{(v)}) = \phi(\mathbf{X}^{(v)})\mathbf{Z}^{(v)}, \text{diag}(\mathbf{Z}^{(v)}) = \mathbf{0} \end{aligned} \quad (8)$$

The main obstacle in the process of optimizing (8) is that $\|\phi(\mathbf{X}^{(v)})\|_{S_p}^p$ depends on $\phi(\mathbf{X}^{(v)})$. This can be circumvented via re-parametrization [13], which leads to a closed-form solution for robust rank minimization in the feature space. Since the kernel matrix $\mathbf{K}^{(v)}$ is symmetric positive semi-definite, we can factorize it as $\mathbf{K}^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)}$ and obtain $\|\phi(\mathbf{X}^{(v)})\|_{S_p}^p = \|\mathbf{B}^{(v)}\|_{S_p}^p$, where $\mathbf{B}^{(v)}$ is a square matrix. Typically, the data points are polluted by noise. Hence, we must add a regularization term to the objective by relaxing the equality constraint in (8): $\|\phi(\mathbf{X}^{(v)}) - \phi(\mathbf{X}^{(v)})\mathbf{Z}^{(v)}\|_F^2 = \text{tr}[(\mathbf{I} - 2\mathbf{Z}^{(v)} + \mathbf{Z}^{(v)}\mathbf{Z}^{(v)T})\mathbf{B}^{(v)T}\mathbf{B}^{(v)}]$. To make the model more robust to corruption (caused by non-Gaussian noise), the correntropy is introduced into model (8) in the form of $\sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)})$ to measure the similarity between $\mathbf{K}_G^{(v)}$ and $\mathbf{B}^{(v)T}\mathbf{B}^{(v)}$. With an additional affine constraint for affine subspaces, problem (8) can be expressed as follows:

$$\begin{aligned} \min_{\mathbf{B}^{(v)}, \mathbf{Z}^{(v)}, \mathbf{E}^{(v)}} \quad & \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr}\left((\mathbf{I} - 2\mathbf{Z}^{(v)} + \mathbf{Z}^{(v)}\mathbf{Z}^{(v)T})\mathbf{B}^{(v)T}\mathbf{B}^{(v)}\right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) + \beta^{(v)} \|\mathbf{Z}^{(v)} - \mathbf{Z}^*\|_F^2 \\ \text{s.t.} \quad & \text{diag}(\mathbf{Z}^{(v)}) = \mathbf{0}, \mathbf{1}^T \mathbf{Z}^{(v)} = \mathbf{1}^T, \mathbf{K}_G^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \mathbf{E}^{(v)} \end{aligned} \quad (9)$$

where $\mathbf{K}_G^{(v)}$ is a predefined kernel matrix, $\mathbf{E}^{(v)}$ models the errors in the data, $\varphi(\mathbf{E}_{ij}^{(v)}) = 1 - \exp(-\frac{\mathbf{E}_{ij}^{(v)2}}{2\sigma^2})$, $\mathbf{E}_{ij}^{(v)}$ denotes the i -th row and j -th column element of matrix $\mathbf{E}^{(v)}$, and σ is the size of the Gaussian kernel.

For convenience, auxiliary variables $\mathbf{Z}_1^{(v)}$, $\mathbf{Z}_2^{(v)}$ and $\mathbf{A}^{(v)}$ are introduced in Eq. (9). Then, the original problem can be reformulated as follows:

$$\begin{aligned} \min_{\mathbf{A}^{(v)}, \mathbf{B}^{(v)}, \mathbf{Z}_1^{(v)}, \mathbf{Z}_2^{(v)}, \mathbf{E}^{(v)}} \quad & \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}_1^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr}\left((\mathbf{I} - 2\mathbf{A}^{(v)} + \mathbf{A}^{(v)}\mathbf{A}^{(v)T})\mathbf{B}^{(v)T}\mathbf{B}^{(v)}\right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) + \beta^{(v)} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^*\|_F^2 \\ \text{s.t.} \quad & \mathbf{A}^{(v)} = \mathbf{Z}_1^{(v)} - \text{diag}(\mathbf{Z}_1^{(v)}), \mathbf{A}^{(v)} = \mathbf{Z}_2^{(v)}, \mathbf{K}_G^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \mathbf{E}^{(v)}, \mathbf{1}^T \mathbf{Z}_1^{(v)} = \mathbf{1}^T \end{aligned} \quad (10)$$

Because of the bilinear terms in the objective, problem (10) is non-convex (or bi-convex). Recently, the ADMM has become popular for solving non-convex problems [20], especially bilinear ones [30]. A convergence analysis of the ADMM for non-convex problems is provided in [21]. Then, we use the ADMM to solve Eq. (10). The augmented Lagrangian function is

$$\begin{aligned} \mathcal{L}(\mathbf{A}^{(v)}, \mathbf{B}^{(v)}, \{\mathbf{Z}_i^{(v)}\}_{i=1}^2, \mathbf{E}^{(v)}, \{\mathbf{Y}_i^{(v)}\}_{i=1}^3, \mathbf{J}_4^{(v)}) = & \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}_1^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr}\left((\mathbf{I} - 2\mathbf{A}^{(v)} + \mathbf{A}^{(v)}\mathbf{A}^{(v)T})\mathbf{B}^{(v)T}\mathbf{B}^{(v)}\right) \\ & + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) + \beta^{(v)} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^*\|_F^2 + \text{tr}(\mathbf{Y}_1^{(v)T}(\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)}))) \\ & + \text{tr}(\mathbf{Y}_2^{(v)T}(\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)})) + \text{tr}(\mathbf{Y}_3^{(v)T}(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T}\mathbf{B}^{(v)} - \mathbf{E}^{(v)})) \end{aligned}$$

$$\begin{aligned}
& + \text{tr}(\mathbf{y}_4^{(v)T} (\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T)) + \frac{\mu}{2} (\|\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})\|_F^2 + \|\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}\|_F^2 \\
& + \|\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}\|_F^2 + \|\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T\|_2^2)
\end{aligned} \quad (11)$$

where $\{\mathbf{Y}_i^{(v)}\}_{i=1}^3 \in \mathbb{R}^{N \times N}$, $\mathbf{y}_4^{(v)} \in \mathbb{R}^{1 \times N}$ are the Lagrange multipliers, and $\mu \geq 0$ denotes the penalty parameter. We update each of the above variables alternately by minimizing Eq. (11) while fixing the other variables. In the following, the details of the process will be provided.

(1) Updating $\mathbf{Z}_1^{(v)}$

$\mathbf{Z}_1^{(v)}$ can be updated by solving the following sub-problem:

$$\min_{\mathbf{Z}_1^{(v)}} \frac{\lambda_1}{\mu} \|\mathbf{Z}_1^{(v)}\|_1 + \frac{1}{2} \|\mathbf{Z}_1^{(v)} - \text{diag}(\mathbf{Z}_1^{(v)}) - \left(\mathbf{A}^{(v)} + \frac{\mathbf{Y}_1^{(v)}}{\mu} \right)\|_F^2 \quad (12)$$

This sub-problem also has a closed-form solution:

$$\mathbf{Z}_1^{(v)} = \mathbf{\Upsilon} - \text{diag}(\mathbf{\Upsilon}) \quad (13)$$

where $\mathbf{\Upsilon} = T_{\frac{\lambda_1}{\mu}}(\mathbf{A}^{(v)} + \frac{\mathbf{Y}_1^{(v)}}{\mu})$ and T is an element-wise soft-thresholding operator that is defined as $T_\tau(x) = \text{sign}(x) \cdot \max(|x| - \tau, 0)$.

(2) Updating $\mathbf{Z}_2^{(v)}$

$\mathbf{Z}_2^{(v)}$ can be updated by solving the following sub-problem:

$$\min_{\mathbf{Z}_2^{(v)}} \beta^{(v)} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^*\|_F^2 + \text{tr}(\mathbf{Y}_2^{(v)T} (\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)})) + \frac{\mu}{2} \|\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}\|_F^2 \quad (14)$$

This sub-problem also has a closed-form solution:

$$\mathbf{Z}_2^{(v)} = ((2\beta^{(v)} + \mu)\mathbf{I})^{-1} (2\beta^{(v)}\mathbf{Z}^* + \mu\mathbf{A}^{(v)} + \mathbf{Y}_2^{(v)}) \quad (15)$$

(3) Updating $\mathbf{Z}^*$

The closed-form solution of $\mathbf{Z}^*$ can be obtained by setting the partial derivative of the objective function with respect to $\mathbf{Z}^*$ to zero and solving for $\mathbf{Z}^*$:

$$\mathbf{Z}^* = \frac{\sum_{v=1}^{n_v} \beta^{(v)} \mathbf{Z}^{(v)}}{\sum_{v=1}^{n_v} \beta^{(v)}} \quad (16)$$

(4) Updating $\mathbf{A}^{(v)}$

The matrix $\mathbf{A}^{(v)}$ can be updated by solving the following sub-problem:

$$\begin{aligned}
\min_{\mathbf{A}^{(v)}} & \frac{\lambda_2}{2} \text{tr}((\mathbf{I} - 2\mathbf{A}^{(v)} + \mathbf{A}^{(v)} \mathbf{A}^{(v)T}) \mathbf{B}^{(v)T} \mathbf{B}^{(v)}) + \text{tr}(\mathbf{Y}_1^{(v)T} \mathbf{A}^{(v)}) + \text{tr}(\mathbf{Y}_2^{(v)T} \mathbf{A}^{(v)}) + \text{tr}(\mathbf{y}_4^{(v)T} \mathbf{1}^T \mathbf{Z}_1^{(v)}) \\
& + \frac{\mu}{2} (\|\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})\|_F^2 + \|\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T\|_2^2)
\end{aligned} \quad (17)$$

This can be achieved by taking the derivative with respect to $\mathbf{A}^{(v)}$ and setting it to zero. This sub-problem has a closed-form solution:

$$\mathbf{A}^{(v)} = (\lambda_2 \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + 2\mu(\mathbf{I} + \mathbf{1}\mathbf{1}^T)^{-1} (\lambda_2 \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{Y}_1^{(v)} - \mathbf{Y}_2^{(v)} - \mathbf{1}\mathbf{y}_4^{(v)} + \mu(\mathbf{Z}_1^{(v)} - \text{diag}(\mathbf{Z}_1^{(v)}))) + \mathbf{Z}_2^{(v)} + \mathbf{1}\mathbf{1}^T) \quad (18)$$

(5) Updating $\mathbf{B}^{(v)}$

To update $\mathbf{B}^{(v)}$, we can solve the following sub-problem,

$$\min_{\mathbf{B}^{(v)}} \|\mathbf{B}^{(v)}\|_{S_p}^p + \frac{\lambda_3}{2} \|\tilde{\mathbf{K}}_{G1}^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)}\|_F^2 \quad (19)$$

where $\tilde{\mathbf{K}}_{G1}^{(v)} = \mathbf{K}_G^{(v)} - \frac{1}{\mu} (\frac{\lambda_2}{2} (\mathbf{I} - 2\mathbf{A}^{(v)T} + \mathbf{A}^{(v)} \mathbf{A}^{(v)T}) - \mathbf{Y}_3^{(v)}) - \mathbf{E}^{(v)}$. At iteration $k+1$, this sub-problem also has a closed-form solution:

$$\mathbf{B}^{(v)} = \mathbf{\Gamma}^* \mathbf{V}^T \quad (20)$$

where $\mathbf{\Gamma}^*$ and $\mathbf{V}$ are both related to the SVD of $\tilde{\mathbf{K}}_{G1}^{(v)}$. This problem can be solved by Lemma 1.

Lemma 1. Let $\mathbf{G} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$ denote the singular value decomposition (SVD) of $\mathbf{G}$, where $\mathbf{\Sigma} = \text{diag}(\delta_1, \dots, \delta_N)$. The following problem is obtained:

$$\min_{\mathbf{L}} \|\mathbf{L}\|_{S_p}^p + \frac{\rho}{2} \|\mathbf{G} - \mathbf{L}^T \mathbf{L}\|_F^2 \quad (21)$$

The closed-form solution of Eq. (21) is given by

$$\mathbf{L}^* = \mathbf{\Gamma}^* \mathbf{V}^T \quad (22)$$

where $\mathbf{\Gamma}^*$ is related to the SVD of $\mathbf{G}$ and $\mathbf{\Gamma}^* = \text{diag}(\gamma_1^*, \dots, \gamma_N^*)$, with $\gamma_i^* = \arg \min_{\gamma_i} \frac{\rho}{2} (\delta_i - \gamma_i^2)^2 + \gamma_i^p$ and $\gamma_i \in \{x \in \mathbb{R}_+ \mid x^3 - \delta_i x + \frac{\rho}{2} x^{p-1} = 0\} \cup \{0\}$, where δ_i is the i -th singular value of $\mathbf{G}$.

As the closed-form solution to (22) is a key contribution of this work, an extended proof of the above lemma is included in the Appendix A.

(6) Updating $\mathbf{E}^{(v)}$

The sub-problem of updating $\mathbf{E}^{(v)}$ can be expressed as:

$$\min_{\mathbf{E}^{(v)}} \frac{\lambda_3}{\mu} \sum_{ij} \varphi(\mathbf{E}_{ij}^{(v)}) + \frac{1}{2} \|\mathbf{E}^{(v)} - \left(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \frac{\mathbf{Y}_3^{(v)}}{\mu} \right)\|_F^2 \quad (23)$$

It is difficult to directly optimize Eq. (23). As shown in [26], the computational speed of the HQ minimization method is substantially higher compared to the gradient-based method. Hence, the HQ technique is used to optimize this function. The optimization procedure of Eq. (23) is expressed as follows:

Fixing $\mathbf{E}_{ij}^{(v)}$, according to HQ theory [26], we can obtain

$$\varphi(\mathbf{E}_{ij}^{(v)}) = \min_{\mathbf{M}_{ij}^{(v)} \in \mathbb{R}} \frac{1}{2} \mathbf{M}_{ij}^{(v)} \mathbf{E}_{ij}^{(v)2} + \psi(\mathbf{M}_{ij}^{(v)}) \quad (24)$$

where $\psi(\cdot)$ is the dual potential function of $\varphi(\cdot)$, and $\mathbf{W}_{ij}^{(v)}$ is the auxiliary variable. Substituting (24) into (23) yields the following augmented function of (23):

$$\min_{\mathbf{E}^{(v)}, \mathbf{M}^{(v)}} \frac{1}{2} \|\mathbf{E}^{(v)} - \left(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \frac{\mathbf{Y}_3^{(v)}}{\mu} \right)\|_F^2 + \frac{\lambda_3}{2\mu} \|\mathbf{M}^{(v)}\|_F^2 \otimes \mathbf{E}^{(v)} + \sum_{ij} \psi(\mathbf{M}_{ij}^{(v)}) \quad (25)$$

where $\otimes$ denotes the dot product. Using alternating minimization, the minimizer $(\mathbf{E}^{(v)*}, \mathbf{M}^{(v)*})$ of (25) can be calculated. Assuming that the result of the h -th iteration is known and denoted by $(\mathbf{E}^{(v)h}, \mathbf{M}^{(v)h})$, the result of the $(h+1)$ -th iteration can be calculated by

a. fixing $\mathbf{E}^{(v)}$: $\mathbf{M}_{ij}^{(v)}$ is confirmed by the minimizer function $\xi(\cdot)$ of $\varphi(\cdot)$:

$$\mathbf{M}_{ij}^{(v)h+1} = \xi(\mathbf{E}_{ij}^{(v)h}) \quad (26)$$

where $\xi(x) = \exp(-\frac{x^2}{2\sigma^2})$.

b. fixing $\mathbf{M}^{(v)}$: the last term of Eq. (25) is negligible and optimization problem (25) can be solved by

$$\begin{aligned} \mathbf{E}^{(v)h+1} &= \arg \min_{\mathbf{E}^{(v)}} \|\mathbf{E}^{(v)} - \left(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \frac{\mathbf{Y}_3^{(v)}}{\mu} \right)\|_F^2 + \frac{\lambda_3}{\mu} \|\mathbf{M}^{(v)h+1}\|_F^2 \otimes \mathbf{E}^{(v)} \\ &= \left(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \frac{\mathbf{Y}_3^{(v)}}{\mu} \right) ./ \left(\frac{\lambda_3}{\mu} \mathbf{M}^{(v)h+1} + \mathbf{1} \right) \end{aligned} \quad (27)$$

where $./$ denotes element-by-element division.

Prior to convergence or reaching the maximum number of iterations, these update steps will be repeated. At each iteration, convergence is checked by evaluating the following constraints: $\|\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})\|_\infty \leq \epsilon$, $\|\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T\|_\infty \leq \epsilon$, $\|\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}\|_\infty \leq \epsilon$ and $\|\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}\|_\infty \leq \epsilon$, for $v = 1, \dots, n_v$.

3.2. Pairwise RLKMSC

We introduce another regularization method of RLKMSC, namely, pairwise RLKMSC, which encourages similarity between pairs of matrices. We obtain the following optimization problem:

$$\begin{aligned} & \min_{\{\mathbf{B}^{(v)}\}_{v=1}^{n_v}, \{\mathbf{Z}^{(v)}\}_{v=1}^{n_v}, \{\mathbf{E}^{(v)}\}_{v=1}^{n_v}} \sum_{v=1}^{n_v} \left(\|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr} \left(\left(\mathbf{I} - 2\mathbf{Z}^{(v)} + \mathbf{Z}^{(v)} \mathbf{Z}^{(v)T} \right) \mathbf{B}^{(v)T} \mathbf{B}^{(v)} \right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) \right) \\ & + \sum_{1 \leq v, w \leq n_v, v \neq w} \beta^{(v)} \|\mathbf{Z}^{(v)} - \mathbf{Z}^{(w)}\|_F^2 \\ & \text{s.t. } \text{diag}(\mathbf{Z}^{(v)}) = \mathbf{0}, \mathbf{1}^T \mathbf{Z}^{(v)} = \mathbf{1}^T, \mathbf{K}_G^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \mathbf{E}^{(v)}, v = 1, \dots, n_v \end{aligned} \quad (28)$$

The introduction of the last term in objective function (28), unlike in the centroid-based RLKMSC, can encourage similarity between pairs of representation matrices of views. This problem can also be separated into separate sub-problems:

$$\begin{aligned} \min_{\mathbf{B}^{(v)}, \mathbf{Z}^{(v)}, \mathbf{E}^{(v)}} & \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr} \left(\left(\mathbf{I} - 2\mathbf{Z}^{(v)} + \mathbf{Z}^{(v)} \mathbf{Z}^{(v)T} \right) \mathbf{B}^{(v)T} \mathbf{B}^{(v)} \right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) \\ & + \beta^{(v)} \sum_{1 \leq w \leq n_v, v \neq w} \|\mathbf{Z}^{(v)} - \mathbf{Z}^{(w)}\|_F^2 \end{aligned} \quad (29)$$

$$\text{s.t. } \text{diag}(\mathbf{Z}^{(v)}) = \mathbf{0}, \mathbf{1}^T \mathbf{Z}^{(v)} = \mathbf{1}^T, \mathbf{K}_G^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \mathbf{E}^{(v)} \quad (29)$$

By introducing auxiliary variables $\mathbf{Z}_1^{(v)}$, $\mathbf{Z}_2^{(v)}$ and $\mathbf{A}^{(v)}$, the original problem can be reformulated as follows:

$$\begin{aligned} \min_{\mathbf{A}^{(v)}, \mathbf{B}^{(v)}, \mathbf{Z}_1^{(v)}, \mathbf{Z}_2^{(v)}, \mathbf{E}^{(v)}} & \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}_1^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr} \left(\left(\mathbf{I} - 2\mathbf{A}^{(v)} + \mathbf{A}^{(v)} \mathbf{A}^{(v)T} \right) \mathbf{B}^{(v)T} \mathbf{B}^{(v)} \right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) \\ & + \beta^{(v)} \sum_{1 \leq w \leq n_v, v \neq w} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^{(w)}\|_F^2 \\ \text{s.t. } & \mathbf{A}^{(v)} = \mathbf{Z}_1^{(v)} - \text{diag}(\mathbf{Z}_1^{(v)}), \mathbf{A}^{(v)} = \mathbf{Z}_2^{(v)}, \mathbf{K}_G^{(v)} = \mathbf{B}^{(v)T} \mathbf{B}^{(v)} + \mathbf{E}^{(v)}, \mathbf{1}^T \mathbf{Z}_1^{(v)} = \mathbf{1}^T \end{aligned} \quad (30)$$

The augmented Lagrangian is

$$\begin{aligned} \mathcal{L}(\mathbf{A}^{(v)}, \mathbf{B}^{(v)}, \{\mathbf{Z}_i^{(v)}\}_{i=1}^2, \mathbf{E}^{(v)}, \{\mathbf{Y}_i^{(v)}\}_{i=1}^3, \mathbf{Y}_4^{(v)}) \\ = \|\mathbf{B}^{(v)}\|_{S_p}^p + \lambda_1 \|\mathbf{Z}_1^{(v)}\|_1 + \frac{\lambda_2}{2} \text{tr} \left(\left(\mathbf{I} - 2\mathbf{A}^{(v)} + \mathbf{A}^{(v)} \mathbf{A}^{(v)T} \right) \mathbf{B}^{(v)T} \mathbf{B}^{(v)} \right) + \lambda_3 \sum_{i,j} \varphi(\mathbf{E}_{ij}^{(v)}) + \beta^{(v)} \sum_{1 \leq w \leq n_v, v \neq w} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^{(w)}\|_F^2 \\ + \text{tr} \left(\mathbf{Y}_1^{(v)T} (\mathbf{A}^{(v)} - \mathbf{C}_1^{(v)} + \text{diag}(\mathbf{C}_1^{(v)})) \right) + \text{tr} \left(\mathbf{Y}_2^{(v)T} (\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}) \right) + \text{tr} \left(\mathbf{Y}_3^{(v)T} (\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}) \right) \\ + \text{tr} \left(\mathbf{Y}_4^{(v)T} (\mathbf{1}^T \mathbf{C}_1^{(v)} - \mathbf{1}^T) \right) + \frac{\mu}{2} \left(\|\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})\|_F^2 + \|\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}\|_F^2 + \|\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}\|_F^2 \right. \\ \left. + \|\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T\|_2^2 \right) \end{aligned} \quad (31)$$

(1) Updating $\mathbf{Z}_1^{(v)}$, $\mathbf{A}^{(v)}$, $\mathbf{B}^{(v)}$ and $\mathbf{E}^{(v)}$

Comparing Eq. (31) and Eq. (11), the updating rules for variables $\mathbf{Z}_1^{(v)}$, $\mathbf{A}^{(v)}$, $\mathbf{B}^{(v)}$ and $\mathbf{E}^{(v)}$ in the pairwise RLKMSC are the same as those in the centroid-based RLKMSC (Eqs. (13), (18), (20) and (27)).

(2) Updating $\mathbf{Z}_2^{(v)}$

$\mathbf{Z}_2^{(v)}$ can be updated by solving the following sub-problem:

$$\min_{\mathbf{Z}_2^{(v)}} \beta^{(v)} \sum_{1 \leq w \leq n_v, v \neq w} \|\mathbf{Z}_2^{(v)} - \mathbf{Z}^{(w)}\|_F^2 + \text{tr} \left(\mathbf{Y}_2^{(v)T} (\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}) \right) + \frac{\mu}{2} \|\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}\|_F^2 \quad (32)$$

The closed-form solution of $\mathbf{Z}_2^{(v)}$ can be obtained by setting the partial derivative of Eq. (32) with respect to $\mathbf{Z}_2^{(v)}$ to zero and solving for $\mathbf{Z}_2^{(v)}$:

$$\mathbf{Z}_2^{(v)} = \left[(2\beta^{(v)}(n_v - 1) + \mu) \mathbf{I} \right]^{-1} \left(2\beta^{(v)} \sum_{1 \leq w \leq n_v, v \neq w} \mathbf{Z}^{(w)} + \mu \mathbf{A}^{(v)} + \mathbf{Y}_2^{(v)} \right) \quad (33)$$

3.3. Complete algorithm

Given a data matrix $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, the steps of the centroid-based RLKMSC and the pairwise RLKMSC are summarized in Algorithms 1 and 2, respectively. After the coefficient matrix $\mathbf{Z}^*$ has been obtained, we can obtain the affinity matrix $\mathbf{W} = \frac{1}{2}(|\mathbf{Z}^*| + |\mathbf{Z}^*|^T)$. Then, the spectral clustering algorithm [25] is applied to $\mathbf{W}$ to obtain the final clustering results. The complete algorithm of RLKMSC is summarized in Algorithm 3.

4. Experiments

In this section, we evaluate the performance of the proposed algorithms on five datasets that are widely used to measure the performances of multi-view algorithms. We select several state-of-the-art multi-view subspace clustering methods as baselines. We also report the running times of all algorithms and evaluate the sensitivity to parameters and the convergence of our algorithms. All experiments are conducted using MATLAB 2015b on a laptop with an Intel Core i5-3230M CPU and 4G RAM.

Algorithm 1 Centroid-based RLKMSC.

Input: Data matrix $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, kernel matrix $\{\mathbf{K}_G^{(v)}\}_{v=1}^{n_v}$, weight parameters $\lambda_1, \lambda_2, \lambda_3, \{\beta^{(v)}\}_{v=1}^{n_v}$.

```

1: Initialize:  $\mathbf{A}^{(v)} = \mathbf{0}, \mathbf{B}^{(v)} = \sqrt{\mathbf{K}_G^{(v)}}, \{\mathbf{Z}_i^{(v)}\}_{i=1}^2 = \mathbf{0}, \mathbf{Z}^* = \mathbf{0}, \{\mathbf{Y}_i^{(v)}\}_{i=1}^3 = \mathbf{0}, \mathbf{y}_4^{(v)} = \mathbf{0}, \mu_0 = 10^{-8}, \mu_m = 10^8, \eta = 20, \epsilon_1 = 10^{-2},$ 
    $i = 1, \dots, n_v.$ 
2: while not converged do
3:   for  $v = 1$  to  $n_v$  do
4:     Sequentially update  $\mathbf{Z}_1^{(v)}, \mathbf{Z}_2^{(v)}, \mathbf{A}^{(v)}$ , and  $\mathbf{B}^{(v)}$  by solving (13), (15), (18), and (20), respectively;
5:     Alternatingly update  $\mathbf{E}^{(v)}$  and  $\mathbf{M}^{(v)}$  by solving (26) and (27), respectively, until:
6:        $\mu \|\mathbf{E}^{(v)h} - \mathbf{E}^{(v)h-1}\|_F / \|\mathbf{K}_G^{(v)}\|_F \leq \epsilon_1$ ;
7:     Update the Lagrange multiplier variables as follows:
8:        $\mathbf{Y}_1^{(v)} := \mathbf{Y}_1^{(v)} + \mu(\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})),$ 
9:        $\mathbf{Y}_2^{(v)} := \mathbf{Y}_2^{(v)} + \mu(\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}),$ 
10:       $\mathbf{Y}_3^{(v)} := \mathbf{Y}_3^{(v)} + \mu(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}),$ 
11:       $\mathbf{y}_4^{(v)} := \mathbf{y}_4^{(v)} + \mu(\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T);$ 
12:   end for
13:   Update the penalty variable  $\mu := \min(\eta\mu, \mu_m)$ ;
14:   Update  $\mathbf{Z}^*$  by solving (16);
15: end while
Output: Coefficient matrix  $\mathbf{Z}^*$ .
```

Algorithm 2 Pairwise RLKMSC.

Input: Data matrix $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, kernel matrix $\{\mathbf{K}_G^{(v)}\}_{v=1}^{n_v}$, weight parameters $\lambda_1, \lambda_2, \lambda_3, \{\beta^{(v)}\}_{v=1}^{n_v}$.

```

1: Initialize:  $\mathbf{A}^{(v)} = \mathbf{0}, \mathbf{B}^{(v)} = \sqrt{\mathbf{K}_G^{(v)}}, \{\mathbf{Z}_i^{(v)}\}_{i=1}^2 = \mathbf{0}, \{\mathbf{Y}_i^{(v)}\}_{i=1}^3 = \mathbf{0}, \mathbf{y}_4^{(v)} = \mathbf{0}, \mu_0 = 10^{-8}, \mu_m = 10^8, \eta = 20, \epsilon_1 = 10^{-2}, i = 1, \dots,$ 
    $n_v.$ 
2: while not converged do
3:   for  $v = 1$  to  $n_v$  do
4:     Sequentially update  $\mathbf{Z}_1^{(v)}, \mathbf{Z}_2^{(v)}, \mathbf{A}^{(v)}$ , and  $\mathbf{B}^{(v)}$  by solving (13), (33), (18), and (20), respectively;
5:     Alternatingly update  $\mathbf{E}^{(v)}$  and  $\mathbf{M}^{(v)}$  by solving (26) and (27), respectively, until:
6:        $\mu \|\mathbf{E}^{(v)h} - \mathbf{E}^{(v)h-1}\|_F / \|\mathbf{K}_G^{(v)}\|_F \leq \epsilon_1$ ;
7:     Update the Lagrange multiplier variables as follows:
8:        $\mathbf{Y}_1^{(v)} := \mathbf{Y}_1^{(v)} + \mu(\mathbf{A}^{(v)} - \mathbf{Z}_1^{(v)} + \text{diag}(\mathbf{Z}_1^{(v)})),$ 
9:        $\mathbf{Y}_2^{(v)} := \mathbf{Y}_2^{(v)} + \mu(\mathbf{A}^{(v)} - \mathbf{Z}_2^{(v)}),$ 
10:       $\mathbf{Y}_3^{(v)} := \mathbf{Y}_3^{(v)} + \mu(\mathbf{K}_G^{(v)} - \mathbf{B}^{(v)T} \mathbf{B}^{(v)} - \mathbf{E}^{(v)}),$ 
11:       $\mathbf{y}_4^{(v)} := \mathbf{y}_4^{(v)} + \mu(\mathbf{1}^T \mathbf{Z}_1^{(v)} - \mathbf{1}^T);$ 
12:   end for
13:   Update the penalty variable  $\mu := \min(\eta\mu, \mu_m)$ ;
14: end while
15: Combine  $\{\mathbf{Z}^{(1)}, \dots, \mathbf{Z}^{(n_v)}\}$  into  $\mathbf{Z}^*$  by taking the element-wise average;
Output: Coefficient matrix  $\mathbf{Z}^*$ .
```

Algorithm 3 RLKMSC algorithm.

Input: Data matrix $\mathbf{X} = \{\mathbf{X}^{(v)}\}_{v=1}^{n_v}$, number of subspaces k .

```

1: Obtain the coefficient matrix  $\mathbf{Z}^*$  via Algorithm 1 or 2;
2: Define the affinity matrix  $\mathbf{W} = \frac{1}{2}(|\mathbf{Z}^*| + |\mathbf{Z}^*|^T)$ ;
3: Apply the spectral clustering algorithm [25] to the affinity matrix;
Output: The clustering results.
```

4.1. Datasets

Five commonly used datasets are utilized in the experiments; their statistical information is summarized in Table 2. We present a short description of each dataset here.

Table 2
Statistical information on the databases.

Dataset	#Instances	#Views	#Clusters
UCI digit	2000	3	10
Caltech 101	75	3	5
Reuters	600	5	6
Prokaryotic	551	3	4
Flower17	1360	7	17

- **UCI Digit dataset¹**: This dataset is available from the UCI repository. It consists of handwritten numerals ('0'-'9') and each numeral has 200 instances, for a total of 2000 samples. These samples are represented with six feature sets, three of which are used in this experiment: 76 Fourier coefficients of the character shapes as view-1, 216 profile correlations as view-2 and 64 Karhunen-Love coefficients as view-3.
- **Caltech-101 dataset²**: This dataset is composed of the Caltech-101 data from the Multiple Kernel Learning repository, which contains more than 8000 pictures of 101 objects [11]. A subset of this dataset is used in this experiment that contains 75 samples with 5 underlying clusters. Following experiments in [24], our three views consist of the "pixel features", "pyramid histogram of gradients", and "sparse localized features".
- **Reuters dataset³**: The documents that are contained in this dataset have features that belong to 6 categories and are written in five different languages: the original documents are written in English and translated to French, German, Spanish and Italian [3]. The original English-language documents are used as the first view and their four translations are used as other four views. We randomly select 100 documents from each category as samples and form a dataset that contains 600 documents.
- **Prokaryotic phyla dataset**: This dataset contains three views: one text data and two genomic representations, each of which contains 551 types of prokaryotic species [28]. These three views are described as follows: (a) the text data are composed of the word bag-of-words representation of the documents that describe the prokaryotic species; (b) the proteome data are composed of the relative frequencies of amino acids; and (c) the gene sequence data are composed of the presence/absence indicators of gene families in a genome. Similar to the data processing method in [5], in the three views, principal component analysis (PCA) is used for each view and the principal component of the 90% variance is preserved, which can reduce the dimension of the dataset. In contrast to the previous dataset, in this dataset the clusters differ in terms of the number of species: there are 313 species in common, while the smallest cluster contains only 35 species.
- **Flower17 dataset⁴**: This dataset consists of 17 flowers that are common in the UK, each with 80 images. There are species that have unique visual appearances and species that are highly similar. These images have large view, scale and photometric variations. Large variability within classes and sometimes small variability between classes make this dataset very challenging. Similar to [36], the seven kernel matrices that are contained in the dataset are used directly as cluster inputs. Hence, we do not report the results of feature concatenation.

4.2. Baseline algorithms and settings

To better evaluate the performances of the proposed methods, we compare them with several state-of-the-art algorithms: (1) **Best Single View**: using the individual view that achieves the best subspace clustering performance with a single view of data; (2) **Feature Concatenation**: concatenating the features of each view and running standard subspace clustering directly on the joint view representation; (3) **Co-Reg**: the co-regularization method for spectral clustering [19]; (4) **RMSC**: the robust multi-view spectral clustering method, which is based on the low-rank and sparse decomposition [36]; (5) **CSMSC**: convex sparse multi-view spectral clustering [24]; and (6) **MLRSSC (KMLRSSC)**: multi-view low-rank sparse subspace clustering (kernel multi-view low-rank sparse subspace clustering) [5].

For the baseline algorithms, when possible, we use the parameters that are suggested in the original papers (if the parameters are specified in it) or adjust them to the optimal values. The ranges of values for various parameters are listed in Table 3. For RLKMSC, parameters p and λ_i ($i = 1, 3$) in our method are determined by searching the grid from 0.1 to 1 in steps of 0.1 and $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$. Parameter λ_2 is set to $0.1 \times \lambda_1$. For the consensus parameter $\beta^{(v)}$, which is tuned from 0.1 to 1 with step of 0.1, we used the same value for each view because we didn't know what information of the views was important beforehand. Of course, we can also try to use a different value of $\beta^{(v)}$ for each view to evaluate the performance of the proposed algorithm; however, this will substantially increase the workload. We use two types of kernels $K_G = (\mathbf{x}_1, \mathbf{x}_2)$: (1) polynomial kernels: $(\mathbf{x}_1^T \mathbf{x}_2 + a)^b$ of degree $b = 2$ and constant a from 5 to 40 in steps of 5; and

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

² <http://www.vision.caltech.edu/archive.html>

³ <http://multilingreuters.iit.nrc.ca>

⁴ <http://www.robots.ox.ac.uk/~vgg/data/flowers/17/index.html>

Table 3
Parameter settings of various algorithms.

Method	Parameter settings	Iterations
Co-Reg	$\lambda \in \{0.01, 0.02, 0.03, 0.04, 0.05\}$	100
RMSC	$\lambda \in \{0.005, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100\}$	300
CSMSC	$\alpha \in \{10^{-1}, 10^{-2}\}$, $\beta \in \{10^{-3}, 10^{-4}, 10^{-5}\}$	200
KMLRSSC	$\beta_1 \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$, $\beta_2 = 1 - \beta_1$, $\lambda \in \{0.3, 0.5, 0.7, 0.9\}$	100

(2) Gaussian kernels: The standard deviation of the kernel is taken to be equal to the median of the pair-wise Euclidean distances between the data points.

We use five metrics to evaluate clustering performances: precision, recall, F-score, normalized mutual information (NMI) and adjusted rand index (Adj-RI) [28]. For these metrics, higher values indicate better performance. Next, we briefly introduce the calculation methods of these five metrics. Suppose there are four parameters (which are obtained from the confusion matrix of the clustering results): true positive (TP), true negative (TN), false negative (FN) and false positive (FP). We obtain $Precision = \frac{TP}{TP+FP}$; $Recall = \frac{TP}{TP+FN}$; $F\text{-score} = 2 \times \frac{Recall \times Precision}{Recall + Precision}$; $NMI(A, B) = 2 \times \frac{I(X, Y)}{H(X) + H(Y)}$, where $I(X, Y)$ is the mutual information of vector X and vector Y , and $H(X)$ is the information entropy of vector X ; and $Adj\text{-}RI = \frac{RI - E(RI)}{\max(RI) - E(RI)}$, where $RI = \frac{TP+TN}{TP+FP+FN+TN}$ denotes the Rand index.

4.3. Results and discussion

The clustering results of the methods on five real-world datasets are listed in Table 4. The best results are shown in bold. For the Caltech-101 dataset, we perform feature concatenation because only the affinity matrices of all views are known. In all the test results, we only keep the first three digits after the decimal point.

As shown in Table 4, the clustering performance of Co-Reg is the worst among the multi-view algorithms. This is probably because the main goal of Co-Reg is to propose a way to combine multiple kernels (or similarity matrices) for the clustering problem, while ignoring the sparseness and noise characteristics of each view. RLKMSC and MLRSSC outperform Co-Reg, RMSC and CSMSC on clustering. This is because the former two methods make full use of the advantages of Co-Reg, and take into account the characteristics of sparse and low-rank of the data while learning the coefficient matrix of each view. This phenomenon demonstrates that the combination of low-rank constraints and sparse constraints can yield a better result.

In addition, two of the best-performing algorithms, namely, RLKMSC and KMLRSSC, adopt “kernel tricks” that can realize the non-linear subspace clustering, which is one of the reasons why they can significantly improve performance. However, the “kernel tricks” that are used by these two methods are different: The predefined kernel that is used in KMLRSSC does not guarantee that the data that are mapped (implicitly) to the feature space are low-rank; hence, it is unlikely to form multiple low-dimensional subspace structures. Our method can learn a low-rank kernel mapping such that the data in resulting feature space are low-rank. Moreover, in the process of learning the representation coefficients of each view, the rank can be estimated effectively via the Schatten p -norm in our model. Meanwhile, the influence of the complex noise in the data are restrained by the correntropy, which improves the robustness of our model. The clustering performance of a single view, which is listed in Table 4, is superior to those of some multi-view algorithms; this demonstrates the contributions of the Schatten p -norm and the correntropy to improving clustering performance.

All the above improvements in our model are the reasons why the accuracy of our method is higher compared to other comparison algorithms. For example, the average NMIs of RLKMSC exceed those of KMLRSSC (the second-best method) by 6%, 10%, 3%, 6% and 6% on the UCI digit, Caltech-101, Reuters, Prokaryotic and Flower17 datasets, respectively. Similar results are obtained by measuring other indices for clustering performance. Compared to other datasets, the clustering performance of RLKMSC on the Reuters dataset is not substantially improved because over 95% of the values in this dataset are zero (very sparse).

According to Table 4, some of the multi-view clustering methods, such as Feature Concatenation and Co-Reg, perform more poorly than the single-view method on various datasets. This is because multi-view clustering mainly takes advantage of the discrimination and diversity among data views. Therefore, the multi-view clustering performance is superior to that of single view only if the views of the data satisfy this requirement. However, in practice, this is not guaranteed. Hence, determining how to make use of the multi-view information of data is highly important.

To further evaluate the performance of our algorithm, we measure the computational time of all the compared algorithms on the UCI Digit dataset⁵. During the test, the convergence conditions of all the algorithms are the same. The test results are shown in Fig. 4. RLKMSC is more efficient than CSMSC and KMLRSSC. Although the calculation speed of RLKMSC is not as fast as those of Co-Reg and RMSC, the clustering accuracy of RLKMSC is significantly higher compared to these two algorithms.

⁵ The computation time of each algorithm in the figure is the average over 10 test runs. For RLKMSC, only the computation time of its centroid regularization is given, which is very close to that of its pairwise regularization.

Table 4

Comparison results on five datasets. On each dataset, the average performance and standard deviation (numbers in parentheses) of 20 test runs of the k-means clustering algorithm with random initializations are reported.

Dataset	Method	F-score	Precision	Recall	NMI	Adj-RI
UCI digit	Best single view	0.732 (0.034)	0.689 (0.034)	0.785 (0.030)	0.796 (0.021)	0.706 (0.035)
	Feat concat	0.718 (0.040)	0.701 (0.046)	0.768 (0.018)	0.792 (0.023)	0.701 (0.041)
	Co-Reg	0.754 (0.067)	0.735 (0.082)	0.775 (0.050)	0.783 (0.033)	0.726 (0.075)
	RMSC	0.762 (0.051)	0.768 (0.080)	0.827 (0.031)	0.818 (0.040)	0.713 (0.048)
	CSMSC	0.795 (0.045)	0.775 (0.069)	0.856 (0.015)	0.839 (0.019)	0.788 (0.056)
	MLRSSC	0.828 (0.049)	0.813 (0.068)	0.846 (0.028)	0.848 (0.025)	0.801 (0.054)
	KMLRSSC	0.835 (0.045)	0.820 (0.066)	0.851 (0.016)	0.849 (0.020)	0.815 (0.049)
	Centroid RLKMSC	0.912 (0.042)	0.911 (0.068)	0.914 (0.016)	0.909 (0.024)	0.903 (0.050)
	Pairwise RLKMSC	0.894 (0.061)	0.890 (0.074)	0.906 (0.020)	0.905 (0.022)	0.885 (0.055)
	Best single view	0.587(0.029)	0.558(0.028)	0.603(0.032)	0.575(0.032)	0.476(0.024)
Caltech-101	Feat concat	–	–	–	–	–
	Co-Reg	0.562(0.027)	0.525(0.025)	0.594(0.029)	0.556(0.048)	0.438(0.034)
	RMSC	0.536(0.007)	0.521(0.003)	0.551(0.014)	0.519(0.007)	0.430(0.007)
	CSMSC	0.558(0.020)	0.532(0.021)	0.583(0.028)	0.559(0.025)	0.448(0.024)
	MLRSSC	0.614(0.032)	0.597(0.041)	0.658(0.031)	0.621(0.038)	0.539(0.038)
	KMLRSSC	0.656(0.025)	0.644(0.035)	0.693(0.027)	0.678(0.034)	0.601(0.027)
	Centroid RLKMSC	0.779(0.018)	0.772(0.020)	0.787(0.025)	0.779(0.025)	0.727(0.022)
	Pairwise RLKMSC	0.775(0.014)	0.759(0.019)	0.781(0.028)	0.768(0.027)	0.722(0.031)
	Best single view	0.374 (0.004)	0.355 (0.006)	0.404 (0.016)	0.317 (0.011)	0.235 (0.006)
	Feat concat	0.388 (0.005)	0.361 (0.012)	0.418 (0.020)	0.333 (0.008)	0.248 (0.009)
Reuters	Co-Reg	0.374 (0.010)	0.352 (0.019)	0.416 (0.020)	0.308 (0.015)	0.231 (0.016)
	RMSC	0.368 (0.013)	0.339 (0.014)	0.407 (0.021)	0.327 (0.018)	0.237 (0.019)
	CSMSC	0.376 (0.008)	0.339 (0.011)	0.430 (0.018)	0.335 (0.023)	0.238 (0.018)
	MLRSSC	0.414 (0.012)	0.374 (0.022)	0.461 (0.026)	0.374 (0.013)	0.284 (0.016)
	KMLRSSC	0.405 (0.011)	0.389 (0.018)	0.436 (0.018)	0.382 (0.020)	0.296 (0.017)
	Centroid RLKMSC	0.436 (0.009)	0.384 (0.020)	0.505 (0.022)	0.413 (0.017)	0.311 (0.019)
	Pairwise RLKMSC	0.428 (0.017)	0.392 (0.026)	0.497 (0.027)	0.408 (0.025)	0.322 (0.017)
	Best single view	0.448 (0.056)	0.574 (0.018)	0.367 (0.101)	0.350 (0.032)	0.202 (0.053)
	Feat concat	0.464 (0.054)	0.582 (0.015)	0.384 (0.092)	0.348 (0.027)	0.205 (0.056)
	Co-Reg	0.468 (0.023)	0.568 (0.023)	0.398 (0.022)	0.286 (0.021)	0.213 (0.031)
Prokaryotic	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)
	MLRSSC	0.591 (0.016)	0.624 (0.003)	0.566 (0.036)	0.322 (0.002)	0.345 (0.016)
	KMLRSSC	0.591 (0.056)	0.725 (0.068)	0.499 (0.048)	0.437 (0.039)	0.398 (0.082)
	Centroid RLKMSC	0.691 (0.044)	0.794 (0.064)	0.611 (0.052)	0.508 (0.040)	0.529 (0.078)
	Pairwise RLKMSC	0.725 (0.045)	0.726 (0.072)	0.727 (0.056)	0.494 (0.042)	0.544 (0.074)
	Best Single View	0.328(0.006)	0.306(0.007)	0.353(0.009)	0.496 (0.006)	0.283(0.007)
	Feat Concat	–	–	–	–	–
	Co-Reg	0.098(0.004)	0.090(0.003)	0.108(0.006)	0.13290.004)	0.039(0.002)
	RMSC	0.412(0.018)	0.403(0.016)	0.415(0.012)	0.516(0.009)	0.368(0.015)
Flower17	CSMSC	0.406(0.023)	0.408(0.018)	0.424(0.011)	0.505(0.012)	0.353(0.017)
	MLRSSC	0.431(0.020)	0.419(0.017)	0.433(0.014)	0.534(0.011)	0.379(0.019)
	KMLRSSC	0.426(0.021)	0.413(0.017)	0.424(0.017)	0.522(0.013)	0.367(0.018)
	Centroid RLKMSC	0.454 (0.020)	0.444 (0.016)	0.464 (0.016)	0.585 (0.012)	0.4190 (0.021)
	Pairwise RLKMSC	0.436 (0.022)	0.431 (0.017)	0.441 (0.019)	0.571 (0.010)	0.400 (0.018)

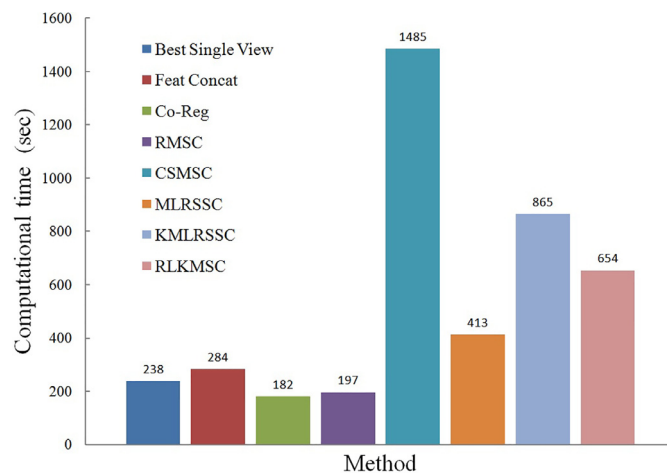


Fig. 4. Computational time (in seconds) of the compared methods on the UCI Digit dataset.

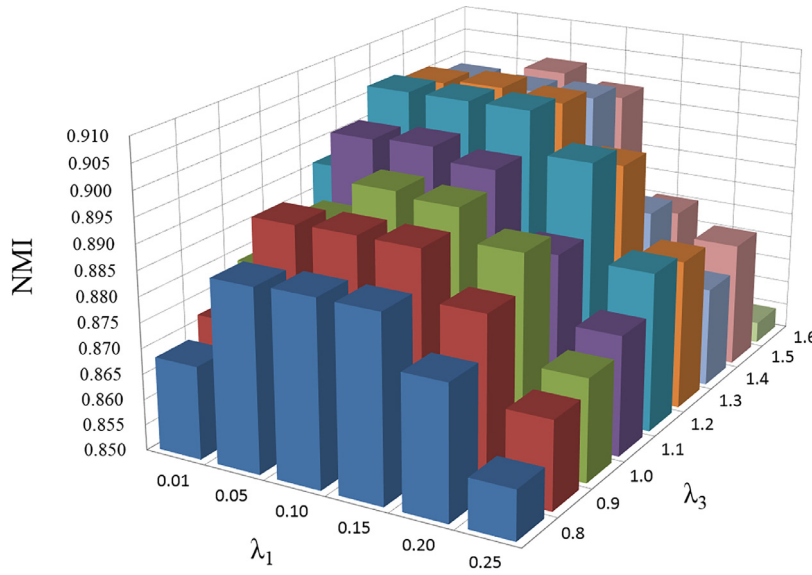


Fig. 5. Clustering performance of centroid-based RLKMSC with respect to the NMI measure when varying parameter λ_i ($i = 1, 3$) and fixing parameter $\beta^{(v)}$.

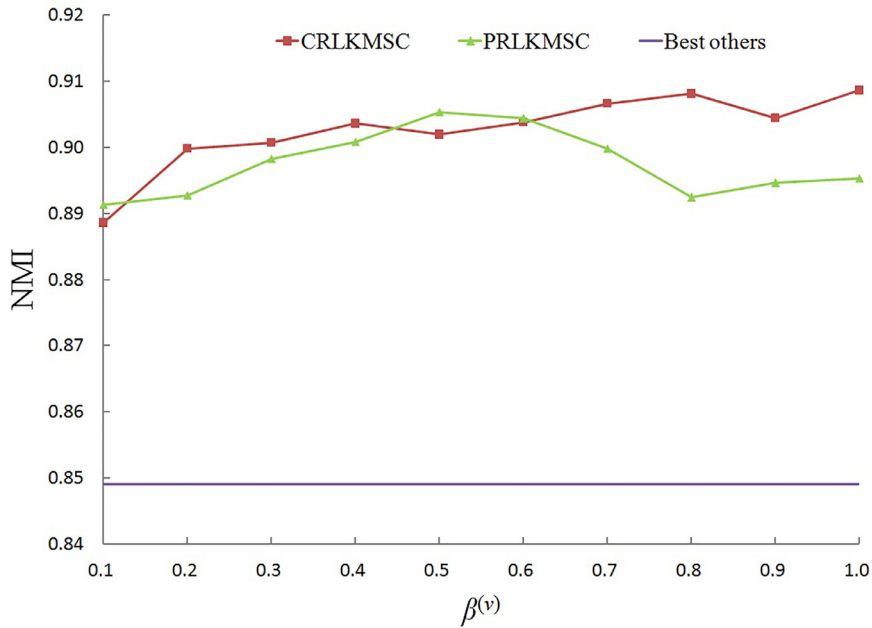


Fig. 6. Clustering performance of RLKMSC with respect to the NMI measure when varying parameter $\beta^{(v)}$ and fixing parameter λ_i ($i = 1, 3$).

4.4. Parameter selection and convergence verification

There are five important parameters in our model: λ_i ($i = 1, 2, 3$), β , and p , where λ_i ($i = 1, 2, 3$) and β are used to balance the effects of data items and regular items. Next, we analyze the effectiveness of the parameters on the UCI Digit dataset.

We select parameter values by varying λ_i ($i = 1, 3$) synchronously. The experimental results are shown in Fig. 5. Our method achieves the desired effect when $\lambda_1 \in [0.05, 0.15]$ and $\lambda_3 \in [0.9, 1.5]$. Fig. 6 shows the clustering performance of RLKMSC's NMI measure when $\beta^{(v)}$ takes various values. If the parameters are within the appropriate ranges, our algorithm can outperform other comparison algorithms. These results demonstrate that RLKMSC is stable.

In Section 2.2.2, we analyzed the characteristics of the Schatten p -norm, which becomes closer to the rank function as p decreases. However, in practice, the data structure may be destroyed by noise, in which case a smaller value of p may cause

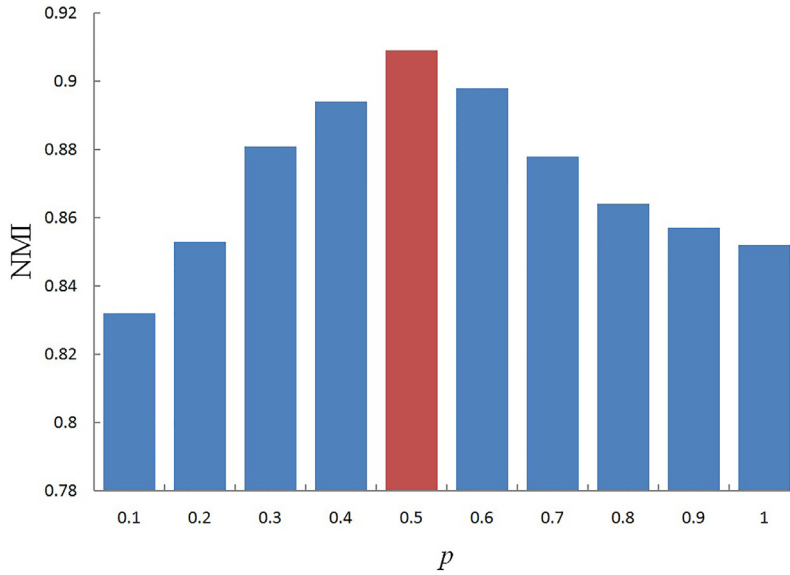


Fig. 7. Clustering performance of centroid-based RLKMSC with the varied parameter p .

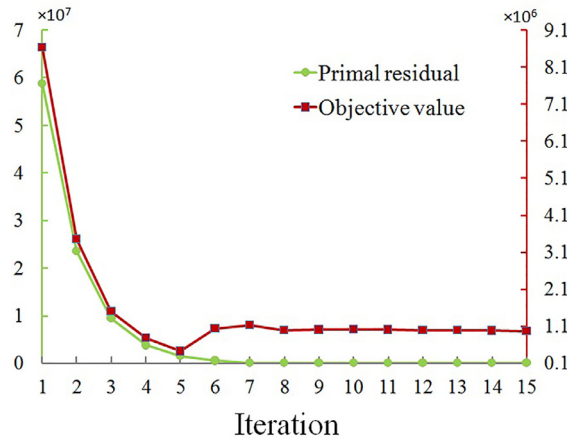


Fig. 8. Convergence curves for centroid-based RLKMSC. Its behavior is highly similar to that of pairwise MLRSSC.

the rank of the data to deviate from the actual rank. The effects of various p -values on clustering results are evaluated and shown in Fig. 7. The best performance is obtained when $p = 0.5$.

In addition, we examine the convergence of our algorithm. The centroid-based RLKMSC algorithm is used as the representative. At each iteration, we compute the primal residual and the objective function value of (10). The primal residual is computed as $\max(\|\mathbf{A} - \mathbf{Z}_1 + \text{diag}(\mathbf{Z}_1)\|_F, \|\mathbf{A} - \mathbf{Z}_2\|_F, \|\mathbf{K}_G - \mathbf{B}^T \mathbf{B} - \mathbf{E}\|_F, \|\mathbf{1}^T \mathbf{Z}_1 - \mathbf{1}^T\|_2)$. As shown in Fig. 8, our algorithm converges within 13 iterations, with the primal residuals quickly approaching zero.

5. Conclusions

This paper proposed a robust low-rank kernel multi-view subspace clustering method (RLKMSC) that integrated the Schatten p -norm, the “kernel trick” and the correntropy elegantly. To make full use of the properties of multi-view data, two regularization methods are used to learn the joint subspace representation of all views: centroid-based and pairwise regularization schemes. In addition, an alternate minimization algorithm (HQ-ADMM) was designed for solving the optimization problems. Our experimental results on five datasets demonstrated that our method achieved the state-of-the-art performance.

However, there remain several issues that are worthy of further study: First, the proposed model contains many parameters, which makes it more difficult to adjust the parameter values. Therefore, the relationship among the parameters should be explored to simplify the adjustment of the parameters. Second, as the research value of incomplete multi-view

data processing is increasing day by day [32,49], we plan to investigate how to handle missing data in a more principled and effective way.

Acknowledgments

This work was supported by the [National Natural Science Foundation of China](#) under Grant 61772272 and the Project of Science and Technology of Jiangsu Province of China under Grant BE2017031.

Appendix A

Proof of Lemma 1. We being proving [Lemma 1](#) by stating the following theorem:

Theorem 1. (Von-Neumann) For any $m \times n$ matrices $\mathbf{F}$ and $\mathbf{H}$, let $r = \min(m, n)$, $\delta(\mathbf{F}) = [\delta_1(\mathbf{F}), \dots, \delta_r(\mathbf{F})]^T$ and $\delta(\mathbf{H}) = [\delta_1(\mathbf{H}), \dots, \delta_r(\mathbf{H})]^T$ be the singular values of $\mathbf{F}$ and $\mathbf{H}$, respectively. Then, $\text{tr}(\mathbf{F}^T \mathbf{H}) \leq \text{tr}(\delta(\mathbf{F})^T \delta(\mathbf{H}))$. Equality occurs if and only if it is possible to find unitary matrices $\mathbf{U}$ and $\mathbf{V}$ that simultaneously singular-value decompose $\mathbf{F}$ and $\mathbf{H}$ in the sense that

$$\mathbf{F} = \mathbf{U} \boldsymbol{\Sigma}_F \mathbf{V}^T \text{ and } \mathbf{H} = \mathbf{U} \boldsymbol{\Sigma}_H \mathbf{V}^T \quad (34)$$

where $\boldsymbol{\Sigma}_F$ and $\boldsymbol{\Sigma}_H$ denote ordered eigenvalue matrices with the singular values $\delta(\mathbf{F})$ and $\delta(\mathbf{H})$, respectively, along the diagonal that have the same order.

Next, suppose matrices $\mathbf{L}$ and $\mathbf{G}$ have SVDs $\mathbf{L} = \mathbf{P} \boldsymbol{\Gamma} \mathbf{Q}^T$ and $\mathbf{G} = \mathbf{U} \boldsymbol{\Sigma} \mathbf{V}^T$, respectively, where both $\boldsymbol{\Gamma} = \text{diag}(\delta_1, \dots, \delta_N)$ and $\boldsymbol{\Sigma} = (\gamma_1, \dots, \gamma_N)$ are diagonal matrices that have the same order (here, non-ascending). Then, $\mathbf{L}^T \mathbf{L} = \mathbf{Q} \boldsymbol{\Gamma}^2 \mathbf{Q}^T$. According to [Theorem 1](#), we obtain

$$\begin{aligned} & \frac{\rho}{2} \|\mathbf{G} - \mathbf{L}^T \mathbf{L}\|_F^2 + \|\mathbf{L}\|_{S_p}^p \\ &= \text{tr}(\boldsymbol{\Sigma}^T \boldsymbol{\Sigma}) + \text{tr}((\boldsymbol{\Gamma}^2)^T \boldsymbol{\Gamma}^2) - 2\text{tr}(\mathbf{G}^T \mathbf{L}^T \mathbf{L}) + \text{tr}(\boldsymbol{\Gamma}^p) \\ &\geq \text{tr}(\boldsymbol{\Sigma}^T \boldsymbol{\Sigma}) + \text{tr}((\boldsymbol{\Gamma}^2)^T \boldsymbol{\Gamma}^2) - 2\text{tr}(\boldsymbol{\Sigma}^T \boldsymbol{\Gamma}^2) + \text{tr}(\boldsymbol{\Gamma}^p) \\ &= \|\boldsymbol{\Sigma} - \boldsymbol{\Gamma}^2\|_F^2 + \text{tr}(\boldsymbol{\Gamma}^p) \end{aligned} \quad (35)$$

Equality holds if and only if $\mathbf{Q} = \mathbf{V}$ according to [Theorem 1](#). Therefore, problem (21) can be reduced to the following problem:

$$\min_{\gamma_i \geq 0} \sum_{i=1}^N \left[\frac{\rho}{2} (\delta_i - \gamma_i^2)^2 + \gamma_i^p \right] \quad (36)$$

The subproblem in (36) can be separated into the following problem:

$$\min_{\gamma_i \geq 0} \left[\frac{\rho}{2} (\delta_i - \gamma_i^2)^2 + \gamma_i^p \right] \quad (37)$$

Hence, we generalize the above problem as

$$\min_x \left[\frac{1}{2} (x^2 - \delta)^2 + x^p \right] \quad (38)$$

Let $\tau^{GTS}(\delta)$ denote the critical value of δ and its corresponding local minimum x_p^* . Thus, to generalize soft-thresholding, we must solve the following nonlinear equation system to determine a correct thresholding value:

$$\frac{1}{2} ((x_p^*)^2 - \tau^{GTS}(\delta))^2 + (x_p^*)^p = \frac{1}{2} (\tau^{GTS}(\delta))^2 \quad (39)$$

$$2(x_p^*)^2 - \tau^{GTS}(\delta)x_p^* + p(x_p^*)^{p-1} = 0 \quad (40)$$

Based on [Eq. \(40\)](#), we can substitute $\tau^{GTS}(\delta)$ in [Eq. \(39\)](#) by $(x_p^*)^2 + \frac{p(x_p^*)^{p-2}}{2}$ and obtain the following equation:

$$\begin{aligned} & \frac{1}{2} \left((x_p^*)^2 - (x_p^*)^2 + \frac{p(x_p^*)^{p-2}}{2} \right)^2 + (x_p^*)^p \\ &= \frac{1}{2} \left((x_p^*)^2 + \frac{p(x_p^*)^{p-2}}{2} \right)^2 \end{aligned} \quad (41)$$

We simplify the above formula to obtain

$$(x_p^*)^p (2 - p - (x_p^*)^{4-p}) = 0 \quad (42)$$

Thus, the only solution of x_p^* that is in the range $(0, +\infty)$ is

$$x_p^* = (2 - p)^{\frac{1}{4-p}} \quad (43)$$

and the thresholding value $\tau^{GTS}(\delta)$ is

$$\tau^{GTS}(\delta) = (2 - p)^{\frac{2}{4-p}} + \frac{p(2 - p)^{\frac{p-2}{4-p}}}{2} \quad (44)$$

In the following, we let

$$f_{\delta,p}(x) = \frac{1}{2}(x^2 - \delta)^2 + x^p \quad (45)$$

We simplify this notation as $f(x)$ for convenience of expression. For any $\delta \in (\tau^{GTS}(\delta), +\infty)$, $f(x)$ has a unique local minimum, which is denoted as $\Theta_p(\delta)$, in the range of $x \in (0, +\infty)$. Corresponding to the larger intersection point between $C_1(x) = px^{p-1}$ and $C_2 = 2(\delta x - x^3)$, the smaller and larger intersection points between $C_1(x)$ and $C_2(x)$ are denoted as $x = p_1$ and $x = p_2$, respectively. Thus, the following hold:

- (1) For $x \in (0, p_1)$, $\frac{d}{dx}f(x) = px^{p-1} - 2(\tau^{GTS}(\delta)x - x^3) > 0$ and, thus, $f(x)$ is monotonically increasing.
 - (2) For $x \in (p_1, p_2)$, $\frac{d}{dx}f(x) = px^{p-1} - 2(\tau^{GTS}(\delta)x - x^3) < 0$ and, thus, $f(x)$ is monotonically decreasing.
 - (3) For $x \in (p_2, +\infty)$, $\frac{d}{dx}f(x) = px^{p-1} - 2(\tau^{GTS}(\delta)x - x^3) > 0$ and, thus, $f(x)$ is monotonically increasing.
- Therefore the larger intersection point p_2 is the unique local minimum of $f(x)$ in $x \in (0, +\infty)$, which satisfies

$$2(\Theta_p(\delta)^2 - \delta)\Theta_p(\delta) + p\Theta_p(\delta)^{p-1} = 0 \quad (46)$$

It follows that there exists γ_i^* such that

$$\begin{aligned} & \frac{\rho}{2} \left(\delta_i^2 - 2\gamma_i^2\delta_i + \gamma_i^4 + \frac{2}{\mu}\gamma_i^p \right) \\ & \geq \frac{\rho}{2} \left(\delta_i - (\gamma_i^*)^2 \right)^2 + (\gamma_i^*)^p, \quad \forall \gamma_i \geq 0 \end{aligned} \quad (47)$$

The proof is complete. $\square$

References

- [1] M. Abavisani, V.M. Patel, Multimodal sparse and low-rank subspace clustering, *Inf. Fusion* 39 (2018) 168–177.
- [2] Z. Akata, C. Thurau, C. Bauckhage, Non-negative matrix factorization in multimodality data for segmentation and label prediction, 16th Computer vision winter workshop, 2011.
- [3] 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*, 2009, pp. 28–36.
- [4] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al., Distributed optimization and statistical learning via the alternating direction method of multipliers, *Found. Trends Mach. Learn.* 3 (1) (2011) 1–122.
- [5] M. Brbić, I. Kopriva, Multi-view low-rank sparse subspace clustering, *Pattern Recognit.* 73 (2018) 247–258.
- [6] X. Cao, C. Zhang, H. Fu, S. Liu, H. Zhang, Diversity-induced multi-view subspace clustering, in: *Computer Vision and Pattern Recognition (CVPR)*, 2015 IEEE Conference on, IEEE, 2015, pp. 586–594.
- [7] G. Chao, S. Sun, Consensus and complementarity based maximum entropy discrimination for multi-view classification, *Inf. Sci.* 367 (2016) 296–310.
- [8] Y. Chen, L. Zhang, Z. Yi, Subspace clustering using a low-rank constrained autoencoder, *Inf. Sci.* 424 (2018) 27–38.
- [9] Z. Deng, K.-S. Choi, Y. Jiang, J. Wang, S. Wang, A survey on soft subspace clustering, *Inf. Sci.* 348 (2016) 84–106.
- [10] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (11) (2013) 2765–2781.
- [11] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories, *Comput. Vision Image Understanding* 106 (1) (2007) 59–70.
- [12] H. Gao, F. Nie, X. Li, H. Huang, Multi-view subspace clustering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4238–4246.
- [13] R. Garg, A. Eriksson, I. Reid, Non-linear dimensionality regularizer for solving inverse problems, *CoRR* (2016). [abs/1603.05015](https://arxiv.org/abs/1603.05015).
- [14] R. He, Y. Zhang, Z. Sun, Q. Yin, Robust subspace clustering with complex noise, *IEEE Trans. Image Process.* 24 (11) (2015) 4001–4013.
- [15] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, *IEEE Trans. Ind. Electron.* 62 (6) (2015) 3742–3751.
- [16] Y. Hu, D. Zhang, J. Ye, X. Li, X. He, Fast and accurate matrix completion via truncated nuclear norm regularization, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (9) (2013) 2117–2130.
- [17] P. Ji, I. Reid, R. Garg, H. Li, M. Salzmann, Low-rank kernel subspace clustering, *CoRR* (2017). [abs/1707.04974](https://arxiv.org/abs/1707.04974).
- [18] D. Kong, M. Zhang, C. Ding, Minimal shrinkage for noisy data recovery using Schatten-p norm objective, in: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2013, pp. 177–193.
- [19] A. Kumar, P. Rai, H. Daume, Co-regularized multi-view spectral clustering, in: *Advances in neural information processing systems*, 2011, pp. 1413–1421.
- [20] G. Li, T.K. Pong, Global convergence of splitting methods for nonconvex composite optimization, *SIAM J. Optim.* 25 (4) (2015) 2434–2460.
- [21] Z. Lin, R. Liu, Z. Su, Linearized alternating direction method with adaptive penalty for low-rank representation, in: *Advances in neural information processing systems*, 2011, pp. 612–620.
- [22] 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.
- [23] W. Liu, P.P. Pokharel, J.C. Principe, Correntropy: properties and applications in non-gaussian signal processing, *IEEE Trans. Signal Process.* 55 (11) (2007) 5286–5298.

- [24] C. Lu, S. Yan, Z. Lin, Convex sparse spectral clustering: single-view to multi-view, *IEEE Trans. Image Process.* 25 (6) (2016) 2833–2843.
- [25] A.Y. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: Analysis and an algorithm, in: *Advances in neural information processing systems*, 2002, pp. 849–856.
- [26] M. Nikolova, M.K. Ng, Analysis of half-quadratic minimization methods for signal and image recovery, *SIAM J. Sci. Comput.* 27 (3) (2005) 937–966.
- [27] X. Peng, Z. Yu, Z. Yi, H. Tang, Constructing the l_2 -graph for robust subspace learning and subspace clustering, *IEEE Trans. Cybern.* 47 (4) (2017) 1053–1066.
- [28] 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.
- [29] J. Shawe-Taylor, N. Cristianini, *Kernel methods for pattern analysis*, Cambridge university press, 2004.
- [30] Y. Shen, Z. Wen, Y. Zhang, Augmented lagrangian alternating direction method for matrix separation based on low-rank factorization, *Optim. Methods Software* 29 (2) (2014) 239–263.
- [31] Ł. Struski, J. Tabor, P. Spurek, Lossy compression approach to subspace clustering, *Inf. Sci.* 435 (2018) 161–183.
- [32] Q. Tan, G. Yu, C. Domeniconi, J. Wang, Z. Zhang, Incomplete multi-view weak-label learning., in: *IJCAI*, 2018, pp. 2703–2709.
- [33] R. Vidal, Subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2011) 52–68.
- [34] R. Vidal, P. Favaro, Low rank subspace clustering (lsrc), *Pattern Recognit. Lett.* 43 (2014) 47–61.
- [35] G. Xia, H. Sun, L. Feng, G. Zhang, Y. Liu, Human motion segmentation via robust kernel sparse subspace clustering, *IEEE Trans. Image Process.* 27 (1) (2018) 135–150.
- [36] R. Xia, Y. Pan, L. Du, J. Yin, Robust multi-view spectral clustering via low-rank and sparse decomposition., in: *AAAI*, 2014, pp. 2149–2155.
- [37] C. Xu, D. Tao, C. Xu, A survey on multi-view learning, *CoRR* (2013). [abs/1304.5634](https://arxiv.org/abs/1304.5634).
- [38] Z. Xu, X. Chang, F. Xu, H. Zhang, $l_{1/2}$ Regularization: a thresholding representation theory and a fast solver, *IEEE Trans. Neural Netw. Learn. Syst.* 23 (7) (2012) 1013–1027.
- [39] X. Yang, W. Liu, D. Tao, J. Cheng, Canonical correlation analysis networks for two-view image recognition, *Inf. Sci.* 385 (2017) 338–352.
- [40] M. Yin, Y. Guo, J. Gao, Z. He, S. Xie, Kernel sparse subspace clustering on symmetric positive definite manifolds, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5157–5164.
- [41] C. You, C.-G. Li, D.P. Robinson, R. Vidal, Oracle based active set algorithm for scalable elastic net subspace clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3928–3937.
- [42] H. Yu, X. Wang, G. Wang, X. Zeng, An active three-way clustering method via low-rank matrices for multi-view data, *Inf. Sci.* (2018).
- [43] J. Yu, Y. Rui, D. Tao, et al., Click prediction for web image reranking using multimodal sparse coding., *IEEE Trans. Image Process.* 23 (5) (2014) 2019–2032.
- [44] J. Zhang, J. Yu, D. Tao, Local deep-feature alignment for unsupervised dimension reduction, *IEEE Trans. Image Process.* 27 (5) (2018) 2420–2432.
- [45] X. Zhang, H. Gao, G. Li, J. Zhao, J. Huo, J. Yin, Y. Liu, L. Zheng, Multi-view clustering based on graph-regularized nonnegative matrix factorization for object recognition, *Inf. Sci.* 432 (2018) 463–478.
- [46] X. Zhang, C. Xu, X. Sun, G. Baci, Schatten-q regularizer constrained low rank subspace clustering model, *Neurocomputing* 182 (2016) 36–47.
- [47] Z. Zhang, Z. Zhai, L. Li, Uniform projection for multi-view learning, *IEEE Trans Pattern Anal. Mach. Intell.* 39 (8) (2017) 1675–1689.
- [48] C. Zhao, W.-L. Hwang, C.-L. Lin, W. Chen, Greedy orthogonal matching pursuit for subspace clustering to improve graph connectivity, *Inf. Sci.* 459 (2018) 135–148.
- [49] H. Zhao, H. Liu, Y. Fu, Incomplete multi-modal visual data grouping., in: *IJCAI*, 2016, pp. 2392–2398.
- [50] W. Zuo, D. Meng, L. Zhang, X. Feng, D. Zhang, A generalized iterated shrinkage algorithm for non-convex sparse coding, in: *Computer Vision (ICCV), 2013 IEEE International Conference on*, IEEE, 2013, pp. 217–224.