Neural Networks 59 (2014) 1–15

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

Learning Markov random walks for robust subspace clustering and estimation

Risheng Liu [a,*], Zhouchen Lin [b], Zhixun Su [a]

[a] Dalian University of Technology, Dalian, China

[b] Key Lab. of Machine Perception (MOE), School of EECS, Peking University, Beijing, China

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form 18 April 2014
Accepted 11 June 2014
Available online 25 June 2014

Keywords:
Spectral clustering
Dimensionality reduction
Markov random walks
Transition probability learning
Subspace clustering and estimation

ABSTRACT

Markov Random Walks (MRW) has proven to be an effective way to understand spectral clustering and embedding. However, due to less global structural measure, conventional MRW (e.g., the Gaussian kernel MRW) cannot be applied to handle data points drawn from a mixture of subspaces. In this paper, we introduce a regularized MRW learning model, using a low-rank penalty to constrain the global subspace structure, for subspace clustering and estimation. In our framework, both the local pairwise similarity and the global subspace structure can be learnt from the transition probabilities of MRW. We prove that under some suitable conditions, our proposed local/global criteria can exactly capture the multiple subspace structure and learn a low-dimensional embedding for the data, in which giving the true segmentation of subspaces. To improve robustness in real situations, we also propose an extension of the MRW learning model based on integrating transition matrix learning and error correction in a unified framework. Experimental results on both synthetic data and real applications demonstrate that our proposed MRW learning model and its robust extension outperform the state-of-the-art subspace clustering methods.

1. Introduction

The graph spectral techniques (Chung, 1997; Von Luxburg, 2007) have many applications in machine learning, exploratory data analysis, computer vision and pattern recognition. Normalized Cut (NCut) (Shi & Malik, 2000), as one of the most successful spectral clustering methods, views the data set as a graph, whose nodes represent data points and whose edges are weighted according to the similarity between data samples. The success of such algorithms heavily depends on the choice of the affinity matrix. In addition to the graph cut interpretations, spectral clustering can also be understood in a probabilistic manner. The work in Meila and Shi (2001) views the local pairwise similarities as edge flows in Markov Random Walks (MRW) and studies the properties of the resulting transition matrix. In this view, the NCut criterion can be nicely interpreted in a general MRW framework. Along this direction, the work in Qiu and Hancock (2007) uses the commute time of a random walk for clustering and embedding. MRW can also be considered as a metric or a similarity structure over the data space, which is used by the clustering (Nadler, Lafon, & Coifman, 2005; Ng, Jordan, & Weiss, 2001) and embedding (Lafon & Lee, 2006) algorithms.

The most important problem with conventional MRW methods (e.g., Gaussian kernel MRW) is that they only consider the local similarity of the data set and there is no global structure constraint for the data. Thus these methods might be unsuitable for modeling data sampled from a mixture subspaces. The main reason is that the affinity in typical spectral methods is modeled only based on a characterization of "locality", which may fail to reveal the global subspace structure. However, in real applications several types of visual data, such as motion (Rao, Tron, Vidal, & Ma, 2010), face (Geng, Smith-Miles, Zhou, & Wang, 2011; Huang, Liu, & Metaxas, 2011) and video sequences (Mei & Ling, 2011; Wang, Tieu, & Grimson, 2010), have been known to be well characterized by subspaces. Therefore, there is a need to extend conventional spectral methods to model a mixture of subspaces. Recent advances in low-rank modeling have led to increasingly concise descriptions of the subspace structure. For instance, the work in Candès, Li, Ma, and Wright (2011) showed that the data points sampled from a single subspace can be exactly recovered by the rank minimization model. It is also shown in Liu, Lin, and Yu (2010) that the multiple subspace structure can be revealed by the "lowest rank" representation coefficients of a given dictionary. However, as discussed below, the spectrum properties of such representation matrix cannot be guaranteed. Thus further efforts

* Corresponding author. Tel.: +86 411 84708351.
E-mail addresses: rsliu0705@gmail.com, rsliu@dlut.edu.cn (R. Liu), zlin@pku.edu.cn (Z. Lin), zxsu@dlut.edu.cn (Z. Su).

http://dx.doi.org/10.1016/j.neunet.2014.06.005
0893-6080/

should be made to build upon the connection between the learnt representation and the affinity matrix used in spectral methods.

In this paper, by considering both the local pairwise similarity and the global subspace structure at the same time, we provide a new spectral framework from the MRW viewpoint for subspace clustering and estimation. Specifically, we learn a transition matrix by our local/global criteria and estimate a low-dimensional embedding from this graph. Then data points can be clustered into different subspaces in this feature space. In the following, we will review previous work on subspace clustering and then highlight the contributions of our research.

1.1. Previous work

A number of approaches for subspace clustering have been proposed in the past two decades. According to the mechanisms for data structure modeling, the existing works can be roughly divided into four main categories: algebraic, statistical, factorization, and compressive sensing methods.

Generalized Principal Component Analysis (GPCA) (Vidal, Ma, & Sastry, 2005) is an algebraic method for subspace clustering. The idea behind GPCA is that one can fit the data with polynomials. By representing subspaces with a set of homogeneous polynomials, subspace clustering is reduced to a problem of fitting data points with polynomials. This method does not impose any restriction on the subspaces. But the main drawback of GPCA is that it is difficult to estimate the polynomial coefficients when the data contains large noise. Recently, Robust Algebraic Segmentation (RAS) (Rao, Yang, Sastry, & Ma, 2010) has been proposed to resolve the robustness issue of GPCA. However, due to the computation difficulty in fitting large scale polynomials, RAS can only work for data with low-dimensionality and a small number of subspaces.

Statistical approaches usually model mixed data as a set of independent samples drawn from a mixture of probabilistic distributions (e.g., mixture of Gaussian). Then the problem of clustering is converted to a model estimation problem, which can be tackled by either Expectation–Maximization (EM) (Gruber & Weiss, 2004) or estimating the mixture structure by iteratively finding a min–max estimation (Fischler & Bolles, 1981). The Bayesian Ying-Yang harmony learning technique presented in Xu (0000) and Xu (2002) is a unified statistical framework to model unsupervised learning and recent investigations in Shi, Liu, Tu, and Xu (2014) show that this theory can be successfully applied for cluster number selection and determining the dimension for principal subspace. The main limitation of statistical models is the optimization difficulty. For example, due to the usage of EM algorithm, most statistical methods can only converge to a local minimum, thus are sensitive to initialization. Also, the sensibility to large errors and outliers is also a bottleneck for these methods.

The idea behind factorization methods (Costeira & Kanade, 1998; Gear, 1998) is to seek clustering from the factorization of the data matrix. The factorization can be computed from SVD (Costeira & Kanade, 1998) or the row echelon canonical form Gear (1998). However, all these methods are sensitive to noise. The work in Gruber and Weiss (2004) adds extra regularization terms to the formulation to reduce the effects of noise. Due to the optimization difficulty of the modified non-convex problem, this method may also get stuck at local minimum.

Compressive sensing has proven to be an extremely powerful tool for signal processing. Recently, there has been a surge of methods (Elhamifar & Vidal, 2009; Favaro, Vidal, & Ravichandran, 2011; Liu, Lin, De la Torre, & Su, 2012; Liu et al., 2010; Nasihatkon & Hartley, 2011; Ni, Sun, Yuan, Yan, & Cheong, 2010; Yu & Schuurmans, 2011) exploiting the discriminative nature of compact representation for subspace clustering. One type of methods, such as Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009; Nasihatkon & Hartley, 2011), is based on discovering the sparsest representations (SR) for the data set. According to the theoretical work of Nasihatkon and Hartley (2011), the within-subspace connectivity assumption for SSC holds only for 2- and 3-dimensional subspaces. In this view, it is possible for SSC to over-segment subspaces for dimension higher than 3. Therefore, extra post-processing stage is needed to overcome this intrinsic drawback for high dimensional data set.

Another type of method, such as Low-Rank Representation (LRR) (Liu et al., 2012, 2013, 2010), is based on minimizing the rank of the representation matrix. It has been proven that, under certain conditions, such non-convex problem can be efficiently solved by minimizing the nuclear norm (as a measure of 2D sparsity) of the matrix (Cai, Candès, & Shen, 2010). Theoretical analysis in Wei and Lin (0000) shows that in essence LRR is a kind of factorization method. Several extensions of this work have been developed. In Favaro et al. (2011), Favaro et al. extend the standard LRR to learn both clean dictionary and low-rank representation for subspace clustering. Indeed, a particular case of this method is equivalent to PCA (Jolliffe, 2002). Thus this method can also be utilized for single subspace estimation. A major drawback of this model is that it may be sensitive to sparse outliers due to the Frobenius norm measure for the noise term. The work in Yu and Schuurmans (2011) also proposes some theoretical analysis on LRR related optimization problems and proves that under the Simultaneous Block (SB) and/or Simultaneous Diagonal (SD) assumptions, a class of rank/norm based subspace clustering models can be solved in closed forms. However, due to the strict SB and SD assumptions on the data matrix, it is unclear whether or not their results can be extended to general problems and applied to real applications.

Overall, although compressive sensing based methods (i.e., SSC, LRR and their variations discussed above) all aim to learn an affinity matrix for spectral clustering, the spectrum properties (i.e., symmetric and nonnegative) of the representation matrix has been bypassed. Without consideration in this aspect, the validity of the constructed graph is poorly justified.

1.2. Our contribution

In this paper, we propose a novel method, called Low-Rank Markov Random Walks (LR-MRW), to learn a specific transition matrix (with low-rank property) to transfer the multiple subspaces structure from the observed data space to a low-dimensional discriminant feature space for subspaces clustering and estimation. Our motivations in this work are two-fold: the success of MRW in understanding spectral clustering and the matrix rank viewpoint for measuring the subspace structure.

On one hand, the intuition motivating this study is that since random walks reflect the combined effect of all possible weighted paths between a pair of nodes, the transition matrix can lead to a measure of cluster cohesion that is less sensitive than using edge weight alone, which underpins algorithms such as NCut. Therefore, it is natural to assume transition probabilities as a metric or a similarity measure over the data space for clustering. On the other hand, inspired by recent works on low-rank modeling (Liu et al., 2010; Wright, Ganesh, Rao, & Ma, 2009), we utilize rank as a measure of subspace structure for the transition matrix. In general, by introducing such local/global criteria, our work learns specific transition probabilities from the original data set to characterize both local pairwise relationship and global multiple linear subspaces structure. For noisy and corrupted data, we propose a robust extension of LR-MRW, which integrates transition matrix learning and noise corruption in a unified framework. Moreover, as a nontrivial byproduct, we propose closed-form solutions for a general class of nuclear norm regularized lease square problems. In the following, we highlight main contributions of the proposed approach:

1. Conventional MRW methods (Lafon & Lee, 2006; Meila & Shi, 2001) use a Gaussian kernel to define the transition matrix for random walks, which may fail to reveal the subspace structure. In contrast, LR-MRW aims at directly learning transition probabilities by incorporating local/global prior knowledge into random walks to model the multiple subspaces structure.
2. Another major shortcoming of conventional MRW is its brittleness to grossly corrupted observations. By integrating transition matrix learning and noise correction in a unified framework, the Robust Low-Rank MRW (RLR-MRW) can successfully recover the corrupted data and reveal the multiple subspaces structure at the same time.
3. Compared to compressive sensing based methods, in which the properties of the affinity matrix cannot be gauged and thus need some extra post-processing, our model advocates to enforce the symmetric and nonnegative constraints explicitly in the optimization model. In this way, LR-MRW can directly learn a valid transition matrix to capture the multiple subspace structure.
4. We obtain closed form solutions to a general class of nuclear norm regularized least square problems. Compared to the work in Ni et al. (2010) and Toh and Yun (2010), our result is more general and we present an entirely different proof, which can be extended in a relatively straightforward way to other nuclear norm minimization problems.

1.3. Paper organization

The outline of the paper is as follows. In Section 2, we review NCut and its link to MRW on the graph. Section 3 introduces our proposed Low-Rank Markov Random Walks (LR-MRW) framework for subspace clustering and estimation. We discuss the computational issues related to LR-MRW in Section 4. The robust extension for LR-MRW is proposed in Section 5. The experimental results are shown in Section 6. Finally, we provide some concluding remarks and suggestions for future work in Section 7.

2. Understanding spectral clustering by Markov random walk

In this section, we review how to understand spectral clustering in the viewpoint of NCut and describe its relationship to MRW on a graph. The material presented here provides the prerequisites for our study and is a summary of results in Belkin and Niyogi (2003), Goh and Vidal (2007), Meila and Shi (2001), Shi and Malik (2000) and Von Luxburg (2007).

2.1. Notations

Hereafter, bold capital letter denotes a matrix (e.g., $\mathbf{X}$), bold lower-case letter denotes a column vector (e.g., $\mathbf{x}$). $\mathbf{x}_i$ represents the ith column of $\mathbf{X}$. x_{ij} denotes the scalar in the row i and the column j of $\mathbf{X}$. $\mathbf{X} \geq 0$ denotes that all $x_{ij} \geq 0$. $\mathbf{1}_n$ is the all-one column vector of length n. $\mathbf{I}$ denotes the identity matrix. $\mathrm{tr}(\mathbf{X}) = \sum_i x_{ii}$ is the trace of $\mathbf{X}$. $\mathrm{rank}(\mathbf{X})$ is the rank of $\mathbf{X}$. $\mathrm{span}(\mathbf{X})$ is the subspace spanned by the columns of $\mathbf{X}$. $\mathrm{diag}(\mathbf{x})$ is a diagonal matrix whose diagonal entries are $\mathbf{x}$. A variety of norms on vectors and matrices will be used. $\|\mathbf{x}\|_2$ and $\|\mathbf{x}\|_1$ denote the l_2 and l_1 norm of $\mathbf{x}$, respectively. $\|\mathbf{X}\|_F$ designates the Frobenius norm of $\mathbf{X}$. $\|\mathbf{X}\|_*$ is the nuclear norm of $\mathbf{X}$ (the sum of singular values of $\mathbf{X}$). $\|\mathbf{X}\|_{2,1} = \sum_j \|\mathbf{x}_j\|_2$ is the $l_{2,1}$ norm of $\mathbf{X}$. $\|\mathbf{X}\|_\infty = \max_{ij}(|x_{ij}|)$. The space of $n \times n$ symmetric matrices is denoted by $\mathcal{S}^n$.

2.2. Spectral clustering via normalized cut

Given a set of data points $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{m\times n}$ (each column is a sample), let G be an undirected weighted graph with vertex set $V = \{v_1, \ldots, v_n\}$ and weighted adjacency matrix $\mathbf{W} = [w_{ij}]_{n\times n}$, where vertex v_i represents the data point $\mathbf{x}_i$ and w_{ij} is the weight of edge between v_i and v_j. The degree of a vertex $v_i \in V$ is defined as $d_i = \sum_j w_{ij}$. The degree matrix $\mathbf{D}$ is defined as the diagonal matrix with degree $d_1, \ldots, d_n$ on the diagonal. For $A \subseteq V$, the set of edges between A and its complement $\bar{A}$ is an edge cut. The NCut criterion in Shi and Malik (2000) is to find the cut that minimizes the following cost function:

$$\mathrm{NCut}(A, \bar{A}) = \left(\frac{1}{\mathrm{vol}(A)} + \frac{1}{\mathrm{vol}(\bar{A})}\right) \sum_{i\in A, j\in \bar{A}} w_{ij}, \tag{1}$$

over all cuts between A and $\bar{A}$, where $\mathrm{vol}(A) := \sum_{i\in A} d_i$. The algorithm in Shi and Malik (2000) is a continuous approximation for solving (1) by computing a generalized eigenvalues problem

$$\mathbf{Lh} = \lambda \mathbf{Dh}, \tag{2}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the Laplacian matrix.

2.3. A Markov random walk view of NCut

Based on MRW on a graph, we present a simple probabilistic interpretation that can offer insights and serve as an analysis tool for NCut. A MRW on a graph is a stochastic process which randomly jumps from vertex to vertex. Therefore, data clustering can be interpreted as trying to find a partition of the graph such that the random walks stay long within the same cluster and seldom jump between clusters. Formally, the transition probability of jumping in one step of the random walk is proportional to the edge weight $\mathbf{W}$ and is given by $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$.

If the graph is connected and non-bipartite, then we can define $\pi = [\pi_1, \ldots, \pi_n]^T$ by $\pi_i = d_i/\mathrm{vol}(V)$. It is easy to verify that $\mathbf{P}^T\pi = \pi$ and thus π is a unique stationary distribution of the Markov chain on the graph. As shown in Chung (1997), many properties of a graph can be expressed in terms of the corresponding transition matrix $\mathbf{P}$.

For any disjoint subsets $A, B \subseteq V$, we define $P(B|A) := Pr(A \to B|A)$ as the probability of the random walks transiting from set A to set B in one step if the current state is in A and the random walk is started in its stationary distribution π. Then the following proposition (Von Luxburg, 2007) demonstrates the equivalence between NCut and MRW:

Proposition 1. *Let G be connected and non-bipartite, then we have:*

$$\mathrm{NCut}(A, \bar{A}) = P(\bar{A}|A) + P(A|\bar{A}). \tag{3}$$

Following this proposition, it is easy to understand that when minimizing NCut, we actually look for a cut on the graph such that a random walk seldom transitions from A to $\bar{A}$ and vice versa. This probabilistic interpretation of NCut as MRW not only sheds new lights on why and how spectral methods work in clustering, but also offers a principled way of learning the similarity function for clustering.

3. Modeling multiple subspace via Markov random walks

One of the main challenges in applying MRW for subspace clustering is how to define a proper transition matrix $\mathbf{P}$ to model the multiple subspace structure. Conventional MRW methods (Azran & Ghahramani, 2006; Lafon & Lee, 2006; Meila & Shi, 2001) usually calculate the transition matrix by the distance-based local similarity (e.g., Gaussian kernel). A deficiency in these methods is that the intrinsic data structure is *only* weakly modeled in local viewpoint, which may fail to reveal the global multiple subspace structure (Vidal, 2010). To avoid this unnatural bias for subspace clustering, we introduce a new model that respects both the local pairwise similarity and the global multiple subspaces structure in the data.

3.1. The basic formulation

Given a graph G, we define the following set of probability matrices

$$\mathbb{P} := \{\mathbf{P} = [p_{ij}]_{n\times n} | 0 \le p_{ij} \le 1\}, \tag{4}$$

where p_{ij} gives the probability of jumping in one step from v_i to v_j. To remove meaningless elements in $\mathbb{P}$ (e.g., matrix with all zero or zero within-subspace probabilities), we assume that for each point there exists at least one nonzero within-subspace jumping probability (i.e., $\forall v_i$, $\exists$ at least one point v_j such that v_i and v_j belong to the same subspace and $p_{ij} > 0$).

Now we consider the problem of subspace clustering and estimation from the viewpoint of MRW. Specifically, we aim to learn a transition matrix from the probability matrix set $\mathbb{P}$ to reveal the multiple subspace structure. Our criteria consist of two parts: one is derived from the local similarity of the original data space, and the other is contributed from the global structure of multiple subspaces.

3.1.1. Local pairwise measure

Recall that, in a general clustering problem, in order to preserve the geometric structure of adjacent samples, we need to define an affinity matrix that encodes the pairwise affinities between data samples. In MRW view, this means that the adjacent samples should have higher transition probabilities to jump from one to another. Consequently, the local pairwise similarity of the data samples can be revealed by minimizing

$$\mathcal{J}(\mathbf{P}) = \sum_{ij} p_{ij}\|\mathbf{x}_i - \mathbf{x}_j\|_2^2. \tag{5}$$

3.1.2. Global structural measure

Now we turn to model the global subspace structure of the data set. In compressive sensing theory, we always have that each sample in a union of subspaces has a linear representation with respect to a dictionary formed by all other data samples (Vidal, 2010). Different from previous works (e.g., Liu et al., 2012, 2010), which directly utilize the linear combination coefficients themselves to reveal subspace relationship, this paper would like to consider the probabilities behind this linear reconstruction. That is, in global structure viewpoint, we consider jth column in $\mathbf{P}$ as the probabilities of samples appearing in the linear reconstruction for jth sample (i.e., reconstruction probabilities). Thus based on the observation that *samples should only be written as a linear combination of other samples from the same subspace*,[1] we desire that the reconstruction probabilities to the jth sample should be high when they are in the same subspace while low when they do not. So in the ideal case we would like to have reconstruction probabilities $p_{ij} \approx 1$ on pairs of samples belonging to the same subspace and $p_{ij} \approx 0$ otherwise and the "prefect" $\mathbf{P}$ for subspace clustering should be an approximation to the following rank k block matrix

$$\mathbf{P} := \begin{bmatrix} \mathbf{1}_{n_1}\mathbf{1}_{n_1}^T & 0 & 0 & 0 \\ 0 & \mathbf{1}_{n_2}\mathbf{1}_{n_2}^T & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{1}_{n_k}\mathbf{1}_{n_k}^T \end{bmatrix}, \tag{6}$$

where k is the number of subspaces. In practice, however, it is challenging to achieve such "prefect" probability matrix. Fortunately, when the sampling of the data set is sufficient, we can observe that $\mathbf{P}$ in (6) is always a low-rank matrix ($k \ll n$). So we may adopt a rank regularizer as a relaxed global structure constraint to $\mathbf{P}$ in (5), namely optimizing the following minimization problem

$$\min_{\mathbf{P}\in\mathbb{P}} \mathcal{J}(\mathbf{P}) + \mu \cdot \mathrm{rank}(\mathbf{P}), \tag{7}$$

where μ is a parameter to balance the loss function and the regularizer. Following the common strategy in low-rank methods (Candès & Recht, 2009; Recht, Fazel, & Parrilo, 2010), we also use the nuclear norm as the convex surrogate for matrix rank. Therefore, our criteria for modeling multiple subspace in the viewpoint of MRW can be formulated as

$$\min_{\mathbf{P}\in\mathbb{P}} \mathcal{J}(\mathbf{P}) + \mu\|\mathbf{P}\|_*. \tag{8}$$

The theoretical analysis in next section shows that though we cannot obtain the ideal transition matrix in (6), the optimization model (8) can achieve an approximate block-diagonal matrix, which is powerful for subspace clustering and estimation.

3.2. Analysis on the proposed criteria

Let $\{\mathcal{C}_j\}_{j=1}^k$ be a collection of k subspaces each of which has a dimension of $d_{\mathcal{C}_j} > 0$. Without loss of generality, we suppose that each $\mathbf{X}_j$ is a collection of n_j samples from $\mathcal{C}_j$ and $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_k]$ (i.e., the indices have been rearranged to satisfy the label of the data). By putting following assumption on the observed data and the set of probability matrices, we establish some theoretical analysis regarding the proposed local and global criteria.

Assumption. The sampling of $\mathbf{X}$ is sufficient but without repetition, i.e., $\|\mathbf{x}_i - \mathbf{x}_j\|_{l_2} > 0$ for $\forall i \ne j$.

Theorem 2. [2] *Let $\mathbb{P}$ be a convex feasible solution set. Then there exists an optimal solution $\mathbf{P}^* \in \mathbb{P}$ to problem (8) with the following block-diagonal structure:*

$$\mathbf{P}^* = \begin{bmatrix} \mathbf{P}_1^* & 0 & 0 & 0 \\ 0 & \mathbf{P}_2^* & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \mathbf{P}_k^* \end{bmatrix}_{n\times n}, \tag{9}$$

where $\mathbf{P}_j^*$ is an $n_j \times n_j$ nonnegative matrix.

Theorem 2 actually establishes a theoretical guarantee for the proposed local and global criteria. Namely, under some suitable conditions, the minimizer to problem (8) has the nature of high within-subspace homogeneity and large between-subspace margin.

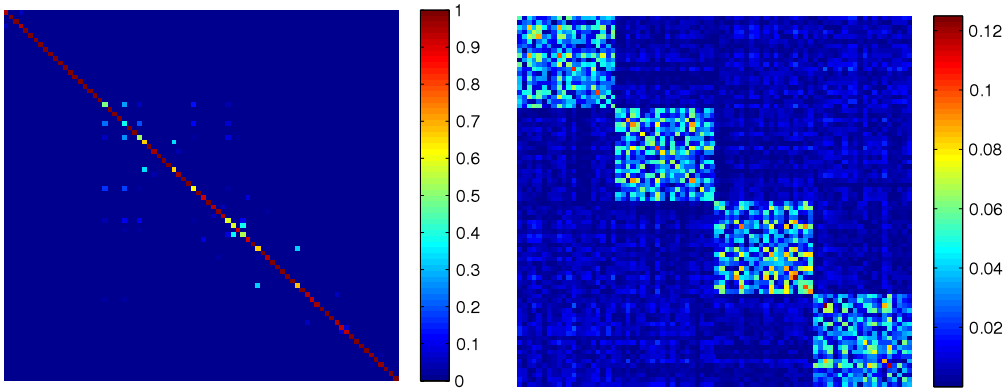
The goal of enforcing convex constraint to the feasible set $\mathbb{P}$ is to guarantee that we can obtain a global optimal solution to the problem. However, in practical implementation the form of the feasible set should be specified for computation. Therefore, we exploit some necessary convex constraints to the MRW learning model in the following subsection.

3.3. The completed MRW learning model

For a given graph, it is natural to consider the transition matrix as a metric or a similarity structure over the space of vertices. In this view, we should assume $\mathbf{P} \in \mathcal{S}^n$. This implies that the relationship between v_i and v_j is symmetric. Furthermore, due to the

[1] This assumption has been verified by previous works (Liu et al., 2012, 2010; Nasihatkon & Hartley, 2011).

[2] Please see Appendix A for the proof of Theorem 2.



(a) G-MRW. (b) LR-MRW.

Fig. 1. The transition probabilities of G-MRW (a) and LR-MRW (b) for data sampled from a mixture of subspaces.

properties of random walks, the row sum of probabilities matrix should be 1. Putting the row normalization, symmetric and nonnegative constraints together, we have the following optimization model

$$\min_{\mathbf{P}} \mathcal{J}(\mathbf{P}) + \mu\|\mathbf{P}\|_*, \quad \text{s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{P} \in \mathcal{S}^n,\ \mathbf{P} \geq 0. \tag{10}$$

Although it is generally challenging to exactly specify $\mathbb{P}$ for computation, we will show in Section 6 that the convex feasible set $\{\mathbf{P} \in \mathbb{R}^{n\times n} | \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \in \mathcal{S}^n, \mathbf{P} \geq 0\}$ yields good performance with respect to our proposed criteria (e.g., Fig. 1(b) illustrates that the transition matrix obtained by LR-MRW model (10) is near block diagonal in reality). The numerical issues of (10) will be discussed in Section 4.

3.4. Comparisons with LRR related works

The LRR related works (Liu et al., 2013, 2010) aim to find a low-rank representation to capture the structure of the data set. These works are inspired by compressive sensing and proposed by solving the following sparse-coding-like model

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_*, \quad \text{s.t. } \mathbf{X} = \mathbf{X}\mathbf{Z}. \tag{11}$$

Now we would like to show that LRR can be considered as a special case of LR-MRW. Specifically, in the case of data contaminated by noise, LRR cannot write a data point as an exact linear combination of other points. Instead, a penalty in the Frobenius norm of the error should be added to the objective function. Thus the representation matrix should be found by solving the problem

$$\min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{1}{\mu}\|\mathbf{X} - \mathbf{X}\mathbf{Z}\|_F^2 = \min_{\mathbf{Z}} \|\mathbf{Z}\|_* + \frac{1}{\mu}\mathrm{tr}(\mathbf{X}(\mathbf{I} - \mathbf{Z})^2\mathbf{X}^T), \tag{12}$$

where μ is a penalty parameter and $\mathbf{M} = (\mathbf{I} - \mathbf{Z})^2$ is known as iterated Laplacian matrix (Belkin & Niyogi, 2003). So although LRR is built from the viewpoint of compressive sensing (i.e., learning a representation from the given dictionary), it actually can be reformulated as a special case of the graph-based model (8) with iterated Laplacian matrix.

However, the most important drawback of LRR is that it does not consider the spectrum properties of $\mathbf{Z}$ such that the validity of the graph constructed by the representation matrix is poorly justified.[3] In contrast, the spectrum properties (i.e., symmetric and nonnegative) of the transition matrix are explicitly enforced in LR-MRW. Therefore, LR-MRW has more transparent connections to spectral graph clustering and embedding. Moreover, the normalization constraint corresponding to the properties of Markov random walks can also further improve the clustering performance for multiple subspaces (Liu et al., 2012).

3.5. Subspace clustering and estimation

Now we show how to extract discriminant multiple subspaces structure from the newly built MRW for subspace clustering and estimation.

It has been shown in Belkin and Niyogi (2003) that the spectral clustering approaches (e.g., NCut), which utilize the eigenvectors of the graph Laplacian, can be interpreted in the framework of nonlinear dimensionality reduction. In this sense, the clustering task on complex data set can be performed by first (nonlinearly) embedding high-dimensional data into a low-dimensional discriminant feature space and then achieving clustering by some standard central clustering techniques. For example, the work in Souvenir and Pless (2005) combines ISOMAP (Tenenbaum, De Silva, & Langford, 2000) with EM and (Goh & Vidal, 2007) combines LLE (Roweis & Saul, 2000) with K-means (David, 2003).

Inspired by the above idea, here we would like to develop a MRW based nonlinear embedding model to extract the membership of multiple subspaces. That is, based on the learnt transition matrix, we transfer the separable structure of mixture of subspaces into a low-dimensional feature space and then cluster data into different subspaces using K-means. Our theoretical analysis in Proposition 3 and Corollary 4 will show that the proposed embedding based model can exactly identify the subspace memberships for the data set.

In the following, we first assume that the number of subspaces is known beforehand in Section 3.5.1 and then provide a spectrum based strategy for estimating the subspace number in Section 3.5.2.

3.5.1. Clustering and estimation with known subspace number

Let the $d \times n$ matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_n]$ be the discriminant features of $\mathbf{X}$, where the ith column provides the representation of the ith sample. As we aim at deriving the intrinsic multiple subspaces structure by the learnt transition matrix $\mathbf{P}$, the embedding $\mathbf{H}$ can be learnt by minimizing the cost function

$$\mathcal{J}(\mathbf{H}) = \sum_{ij} p_{ij}\|\mathbf{h}_i - \mathbf{h}_j\|_2^2 = \mathrm{tr}(\mathbf{H}(\mathbf{D} - \mathbf{P})\mathbf{H}^T), \tag{13}$$

where $\mathbf{D} = \mathrm{diag}(\mathbf{P}\mathbf{1}_n)$. To remove an arbitrary scaling factor in the embedding, we impose constraint $\mathbf{H}\mathbf{H}^T = \mathbf{I}$ to the problem and the

[3] Please note that the SSC related works (Elhamifar & Vidal, 2009; Nasihatkon & Hartley, 2011) also suffer this issue. Although the work in Ni et al. (2010) introduces the positive semi-definite (PSD) constraint for $\mathbf{Z}$, PSD cannot gauge the spectral properties of the affinity matrix (in general, affinity matrices are symmetric and nonnegative, but not necessarily PSD) and the post-processing is still needed.

optimization model reduces to

$$\min_{\mathbf{H}} \operatorname{tr}(\mathbf{H}(\mathbf{D}-\mathbf{P})\mathbf{H}^T), \quad \text{s.t. } \mathbf{H}\mathbf{H}^T = \mathbf{I}. \tag{14}$$

With the normalization property of $\mathbf{P}$ (i.e., $\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$, see problem (10)) and the constraint $\mathbf{H}\mathbf{H}^T = \mathbf{I}$, it is easy to see that the above problem has the following equivalent variation

$$\max_{\mathbf{H}} \operatorname{tr}(\mathbf{H}\mathbf{P}\mathbf{H}^T), \quad \text{s.t. } \mathbf{H}\mathbf{H}^T = \mathbf{I}. \tag{15}$$

It is easy to check that this problem can be solved by the eigen-decomposition of $\mathbf{P}$: the d-column matrix $\mathbf{H}^T$ corresponds to the d eigenvectors associated with the d largest eigenvalues of $\mathbf{P}$.

Now we present the following proposition to infer the connection between multiple subspaces clustering and our random walk based embedding.

Proposition 3. *Assume that the transition matrix $\mathbf{P}$ has the block-diagonal structure (9) and each subspace is connected. Then there exists a k-dimensional eigenspace of $\mathbf{P}$ with the largest eigenvalue which admits a basis $\{\mathbf{v}_j\}_{j=1}^k$ such that $\mathbf{v}_j$ corresponds to the jth subspace:*

$$\mathbf{v}_j(i) = \begin{cases} 1, & \mathbf{x}_i \in \mathcal{C}_j, \\ 0, & \textit{otherwise}. \end{cases} \tag{16}$$

Proof. In view of (9) and the connected assumption for each subspace, we have that $\mathbf{P} = \operatorname{diag}(\mathbf{P}_1, \ldots, \mathbf{P}_k)$, where $\mathbf{P}_j$ is an $n_j \times n_j$ nonnegative matrix for subspace $\mathcal{C}_j$. As a direct consequence of the row normalization of the transition matrix, we have that $\mathbf{1}_{n_j}$ is an eigenvector of $\mathbf{P}_j$ with eigenvalue 1. Therefore, there exists a basis $\{\mathbf{v}_j\}_{j=1}^k$, each vector of which is eigenvector of $\mathbf{P}$ with the eigenvalue 1, which indicates the membership for $\mathcal{C}_j$, as claimed. Finally, from the spectral properties of the transition matrix (Vempala, 2005), we know that 1 is the largest eigenvalue of $\mathbf{P}$, which finishes our proof. □

Corollary 4. *Let $\mathbf{H}$ be the optimal solution to (15). With the same assumptions in Proposition 3, we have that $\mathbf{h}_i = \frac{n}{\sqrt{n_j}}\mathbf{v}_j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, k$, respectively.*

Proof. We define matrix $\mathbf{M}_{\mathcal{C}_j} = \mathbf{v}_j\mathbf{v}_j^T/n_j$. Let $\mathcal{P}_e$ be the projection matrix for the k-dimensional eigenspace of $\mathbf{P}$ with respect to the largest eigenvalue can be written as

$$\mathcal{P}_e = \sum_{j=1}^{k} \frac{1}{n_j}\mathbf{M}_{\mathcal{C}_j} \quad \text{and} \quad \mathcal{P}_e = \frac{1}{n}\mathbf{H}\mathbf{H}^T.$$

Thus for any $i_1, i_2 \in \{1, \ldots, n\}$, $\frac{1}{n}\mathbf{h}_{i_1}^T\mathbf{h}_{i_2} = (\mathcal{P}_e)_{i_1,i_2}$. In particular, if there exists $j \in \{1, \ldots, k\}$ such that $i_1, i_2 \in \mathcal{C}_j$, then $\frac{1}{n}\mathbf{h}_{i_1}^T\mathbf{h}_{i_2} = \frac{1}{n_j}$. When i_1 and i_2 belong to separate subspaces, then $\mathbf{h}_{i_1}^T\mathbf{h}_{i_2} = 0$. For $i_1, i_2 \in \mathcal{C}_j$,

$$\frac{\mathbf{h}_{i_1}^T\mathbf{h}_{i_2}}{\|\mathbf{h}_{i_1}\|_2\|\mathbf{h}_{i_2}\|_2} = \frac{1/n_j}{(1/n_j)(1/n_j)} = 1$$

giving that $\mathbf{h}_{i_1}$ and $\mathbf{h}_{i_2}$ are in the same direction. As they have the same magnitude as well. $\mathbf{h}_{i_1}$ and $\mathbf{h}_{i_2}$ coincide for any two indices i_1 and i_2 belonging to the same subspace. Thus letting $w_j = n_j/n$ for $j = 1, \ldots, k$, there are k perpendicular vectors $\{\mathbf{v}_1, \ldots \mathbf{v}_k\}$ corresponding to the k subspace such that $\mathbf{h}_i = \frac{1}{\sqrt{w_j}}\mathbf{v}_j^T$. □

Proposition 3 implies that the k dimensional eigenspace of $\mathbf{P}$ gives the subspace membership of each samples. However, there are many possible choices for an orthogonal basis of this eigenspace. Thus we cannot assume that any particular basis of the eigenspace will directly provide indicators of the various subspaces. Fortunately, in practice, it is not necessary to compute the set of basis $\{\mathbf{v}_j\}_{j=1}^k$ themselves. This is because Corollary 4 indeed demonstrates that feature points in the embedded eigenspace aggregate to k distinct centroids located on k corners of a simplex (also known as the simplex spectral embedding theory). Therefore, the clustering can be obtained by using K-means on columns of $\mathbf{H}$.

It should be emphasized that the embedding $\mathbf{H}$ used in this work is related to the clustering of the multiple subspaces, but not the low-dimensional approximation of the data set in general nonlinear dimensionality reduction methods (Roweis & Saul, 2000; Tenenbaum et al., 2000). Also, based on our analysis in Proposition 3 and Corollary 4, the dimension of this embedding is theoretically determined by the number of subspace, i.e., $d = k-1$, where k is the number of subspaces.[4] In contrast, conventional nonlinear dimensionality reduction methods (Roweis & Saul, 2000; Tenenbaum et al., 2000), which focus on reconstructing the geometric structures of the data set in a low-dimensional space, often needs to estimate a particular feature dimension for the data set.

For samples belonging to one cluster, we can further estimate the intrinsic dimension and find basis and/or low-dimensional approximation for the subspace using standard subspace learning methods. For example, one may perform PCA (Jolliffe, 2002) on each cluster to find the subspace basis or run Robust PCA (RPCA) (Candès et al., 2011) to directly achieve the intrinsic low-rank approximations to the data samples.[5] Algorithm 1 summarizes the whole clustering and estimation algorithm of LR-MRW.

Algorithm 1 Subspace Clustering and Estimation via LR-MRW Framework

Input: Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ be a set of n data points sampled from k subspaces.
Step 1: Solve (10) to obtain the transition matrix $\mathbf{P}$.
Step 2: Compute the largest eigenvectors of $\mathbf{P}$ to obtain the estimation of data samples $\mathbf{H}$.
Step 3: Apply K-means to the columns of $\mathbf{H}$ to cluster the original data points into k different subspaces $\{\mathcal{C}_j\}_{j=1}^k$.
Step 4: Apply single subspace learning method (e.g., PCA or RPCA) to each group to obtain the basis and/or the low-dimensional approximation for each subspace.

3.5.2. Estimating the subspace number

Now we consider the problem of estimating the number of the subspaces (i.e., clusters). Actually this model selection problem can be solved by considering the spectrum of our learnt MRW. Specifically, the proof of Proposition 3 reveals that the top k eigenvectors have a corresponding eigenvalue of magnitude 1 and others do not. So in principle, the number of subspaces (i.e., clusters) can be found by simply looking at the eigenvalues: the subspace number is equal to the number of eigenvalues of magnitude 1. In practice, we use the following approach to estimate the subspace number $\bar{k}$

$$\bar{k} = n - \operatorname{int}\left(\sum_{i=1}^{n} f_\tau(\lambda_i)\right), \tag{17}$$

where $\{\lambda_i\}_{i=1}^n$ are the eigenvalues of $\mathbf{P}$, $\operatorname{int}(\cdot)$ is the function output the nearest integer of a real number and $f_\tau(\cdot)$ is a soft thresholding operator defined as

$$f_\tau(\lambda) = \begin{cases} 1 & \lambda \le \tau, \\ \log_2\left(1 + \dfrac{\tau^2}{\lambda^2}\right) & \text{otherwise}. \end{cases}$$

[4] This is because there always exists a constant eigenvector $\mathbf{1}_n$ of $\mathbf{P}$ with eigenvalue 1 and we leave out this eigenvector from $\mathbf{H}$ to obtain the embedding.

[5] As single subspace estimation has been well studied in many papers, we omit analysis and experimental verification for this step.

Here $0 < \tau < 1$ is the thresholding parameter. The whole subspaces number estimation process is summarized in Algorithm 2.

Algorithm 2 Subspace Number Estimation via the Spectrum

Input: Let $\mathbf{X} \in \mathbb{R}^{m\times n}$ be a set of n data points.
Step 1: Solve (10) to obtain the transition matrix $\mathbf{P}$.
Step 2: Compute the subspace number $\bar{k}$ by (17).

4. Numerical solution

In this section, we present a practical solution to the LR-MRW problem (10) by applying the alternating direction method (ADM) (Lin, Chen, Wu, & Ma, 0000). We begin with the definition of a key building block, namely the eigenvalue thresholding operator, which can be considered as the extension of the singular value thresholding operator (Cai et al., 2010) for nuclear norm regularized least square problem with the symmetry constraint.

4.1. Eigenvalue thresholding operator

Consider the eigenvalue decomposition (EVD) of a matrix $\mathbf{G} \in \mathcal{S}^n$ of rank r:

$$\mathbf{G} = \mathbf{U}\Lambda(\mathbf{G})\mathbf{U}^T, \tag{18}$$

where $\mathbf{U}$ is an $n \times r$ matrix with orthonormal columns, $\lambda(\mathbf{G}) = [\lambda_1(\mathbf{G}), \ldots, \lambda_r(\mathbf{G})]^T$ are eigenvalues arranged in nonincreasing order and $\Lambda(\mathbf{G}) = \operatorname{diag}(\lambda(\mathbf{G}))$. For each $\mu \geq 0$, we introduce the *eigenvalue thresholding operator* $\mathcal{E}_\mu$ defined as follows:

$$\mathcal{E}_\mu(\mathbf{G}) := \mathbf{U}\operatorname{diag}(\mathcal{T}_\mu(\lambda(\mathbf{G})))\mathbf{U}^T, \tag{19}$$

where

$$\mathcal{T}_\mu(\lambda_i(\mathbf{G})) = \begin{cases} \lambda_i(\mathbf{G}) - \mu, & \lambda_i(\mathbf{G}) > \mu, \\ \lambda_i(\mathbf{G}) + \mu, & \lambda_i(\mathbf{G}) < -\mu, \\ 0, & \text{otherwise}, \end{cases} \tag{20}$$

is the component-wise shrinkage operator. In other words, this operator simply applies a shrinkage rule to the eigenvalues of $\mathbf{G}$, effectively shrinking them towards zero.

Theorem 5. [6] *For each $\mu \geq 0$, and each square matrix $\mathbf{G} \in \mathbb{R}^{n\times n}$. Let $\bar{\mathbf{G}} = (\mathbf{G} + \mathbf{G}^T)/2$. Suppose that $\mathbf{U}\Lambda(\bar{\mathbf{G}})\mathbf{U}^T$ is the eigenvalue decomposition of $\bar{\mathbf{G}}$. Then $\mathbf{K}^* = \mathbf{U}\operatorname{diag}(\mathbf{k}^*)\mathbf{U}^T$ is an optimal solution of the problem*

$$\min_{\mathbf{K}} \frac{1}{2}\|\mathbf{K} - \mathbf{G}\|_F^2 + \mu\|\mathbf{K}\|_*, \quad s.t.\ \mathbf{K} \in \mathcal{S}^n, \tag{21}$$

where $\mathbf{k}^*$ is an optimal solution of the problem

$$\min_{\mathbf{k}} \frac{1}{2}\|\mathbf{k} - \lambda(\bar{\mathbf{G}})\|_2^2 + \mu\|\mathbf{k}\|_1. \tag{22}$$

Remark. Although the work in Ni et al. (2010) and Toh and Yun (2010) also consider a similar problem, it can be seen in the following corollary that their conclusion is actually a special case of Theorem 5. Moreover, the proof of this theorem can also be utilized to understand the connection between nuclear norm regularized and l_1 regularized least square problems and extended in a relatively straightforward way to other problems.

Corollary 6. *Under the same assumption as in Theorem 5, $\mathcal{E}_\mu(\bar{\mathbf{G}})$ is an optimal solution of the problem*

$$\min_{\mathbf{K}} \frac{1}{2}\|\mathbf{K} - \mathbf{G}\|_F^2 + \mu\|\mathbf{K}\|_*, \quad s.t.\ \mathbf{K} \in \mathcal{S}^n. \tag{23}$$

[6] Please see Appendix B for the proof of Theorem 5.

Proof. It is not hard to observe that problem (22) has closed form solution $\mathcal{T}_\mu(\lambda(\bar{\mathbf{G}}))$ (Hale, Yin, & Zhang, 0000). It thus follows from Theorem 5 that the conclusion holds. □

4.2. Solving LR-MRW model by ADM

Now we show how to apply the method of ADM to solve problem (10). With auxiliary variables $\mathbf{Z}$ and $\mathbf{Y}$, problem (10) is equivalent to

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{Z},\mathbf{Y}} \mu\|\mathbf{Z}\|_* - \operatorname{tr}(\mathbf{XPX}^T), \\ &\text{s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{P} = \mathbf{Z},\ \mathbf{P} = \mathbf{Y}, \\ &\mathbf{Z} \in \mathcal{S}^n,\ \mathbf{Y} \geq 0. \end{aligned} \tag{24}$$

By introducing Lagrange multipliers $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$ to remove the equality constraints, one has the augmented Lagrangian function of (24):

$$\begin{aligned} \mathcal{L}_A(\mathbf{P}, \mathbf{Z}, \mathbf{Y}, \{\mathbf{L}_i\}_{i=1}^3) = {} & \mu\|\mathbf{Z}\|_* - \operatorname{tr}(\mathbf{XPX}^T) + \langle \mathbf{L}_1, \mathbf{P}\mathbf{1}_n - \mathbf{1}_n\rangle \\ & + \langle \mathbf{L}_2, \mathbf{P} - \mathbf{Z}\rangle + \langle \mathbf{L}_3, \mathbf{P} - \mathbf{Y}\rangle \\ & + \frac{\beta_1}{2}\|\mathbf{P}\mathbf{1}_n - \mathbf{1}_n\|_2^2 + \frac{\beta_2}{2}\|\mathbf{P} - \mathbf{Z}\|_F^2 \\ & + \frac{\beta_3}{2}\|\mathbf{P} - \mathbf{Y}\|_F^2, \end{aligned} \tag{25}$$

where β_1, β_2 and β_3 are penalty parameters. Then the ADM approach updates $\mathbf{P}$, $\mathbf{Z}$, $\mathbf{Y}$, $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$ iteratively. It respectively updates $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{Y}$ by minimizing $\mathcal{L}_A$ with respect to $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{Y}$, with $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$ fixed. Then the amount of violation of the constraints $\mathbf{P}\mathbf{1}_n = \mathbf{1}_n$, $\mathbf{P} = \mathbf{Z}$ and $\mathbf{P} = \mathbf{Y}$ are used to update $\mathbf{L}_1$, $\mathbf{L}_2$ and $\mathbf{L}_3$, respectively. More specifically, the updating schemes can be found to be:

$$\begin{cases} \mathbf{P}^+ = (\mathbf{X}^T\mathbf{X} - \mathbf{L}_1\mathbf{1}_n^T - \mathbf{L}_2 - \mathbf{L}_3 + \beta_1\mathbf{1}_n\mathbf{1}_n^T + \beta_2\mathbf{Z} + \beta_3\mathbf{Y})\mathbf{A}, \\ \mathbf{Z}^+ = \arg\min_{\mathbf{Z}\in\mathcal{S}^n} \dfrac{1}{2}\left\|\mathbf{Z} - \left(\mathbf{P}^+ + \dfrac{\mathbf{L}_2}{\beta_2}\right)\right\|_F^2 + \dfrac{\mu}{\beta_2}\|\mathbf{Z}\|_*, \\ \mathbf{Y}^+ = \mathcal{P}_+(\mathbf{P}^+ + \mathbf{L}_3/\beta_3), \\ \mathbf{L}_1^+ = \mathbf{L}_1 + \beta_1(\mathbf{P}^+\mathbf{1}_n - \mathbf{1}_n), \\ \mathbf{L}_2^+ = \mathbf{L}_2 + \beta_2(\mathbf{P}^+ - \mathbf{Z}^+), \\ \mathbf{L}_3^+ = \mathbf{L}_3 + \beta_3(\mathbf{P}^+ - \mathbf{Y}^+), \end{cases} \tag{26}$$

where superscripts "+" denote that the values are updated and $\mathbf{A} = (\beta_1\mathbf{1}_n\mathbf{1}_n^T + (\beta_2 + \beta_3)\mathbf{I})^{-1}$. By applying Sherman–Morrison formula (Hager, 1989) on $(\beta_1\mathbf{1}_n\mathbf{1}_n^T + (\beta_2 + \beta_3)\mathbf{I})^{-1}$, we can have a numerically easy way to compute $\mathbf{A}$:

$$\mathbf{A} = \frac{\mathbf{I}}{\beta_2 + \beta_3} - \frac{\beta_1\mathbf{1}_n\mathbf{1}_n^T}{(\beta_2 + \beta_3)(n\beta_1 + \beta_2 + \beta_3)}.$$

$\mathcal{P}_+(\mathbf{M})$ is a matrix

$$\mathcal{P}_+(\mathbf{M})_{ij} := \begin{cases} \mathbf{M}_{ij}, & \mathbf{M}_{ij} \geq 0, \\ 0, & \mathbf{M}_{ij} < 0, \end{cases}$$

which is a projection of the matrix $\mathbf{M}$ onto the set of nonnegative matrices of appropriate size. The solution of $\mathbf{Z}^+$ is by Theorem 5 and Corollary 6. The entire procedure is summarized in Algorithm 3.

5. Robust extension for real situations

In real applications, our observations are often noisy, or even grossly corrupted. In this section, we show how to extend LR-MRW to this real situations.

Algorithm 3 Solving Problem (10) via ADM

Input: Observation matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$, parameter μ.
Initialize: Set $\mathbf{P}$, $\mathbf{Z}$, $\mathbf{Y}$ and $\{\mathbf{L}_i\}_{i=1}^3$ to zero matrices of appropriate sizes, $\{\beta_i > 0\}_{i=1}^3$ and $\varepsilon = 10^{-6}$.
while not converged **do**
  Step 1: Update ($\mathbf{P}$, $\mathbf{Z}$, $\mathbf{Y}$, $\{\mathbf{L}_i\}_{i=1}^3$) by (26).
  Step 2: Check the convergence condition:
  $\max\{\|\mathbf{P}^+\mathbf{1}_n - \mathbf{1}_n\|_\infty, \|\mathbf{P}^+ - \mathbf{Z}\|_\infty, \|\mathbf{P}^+ - \mathbf{Y}\|_\infty\} \le \varepsilon$.
end while
Output: Transition matrix $\mathbf{P}$.

5.1. Robust low-rank MRW

Our approach to robust MRW learning is motivated by the recent work on matrix recovery (Wright et al., 2009). Specifically, we assume that the observed data $\mathbf{X}_0$ can be decomposed as

$$\mathbf{X}_0 = \mathbf{X} + \mathbf{E}, \tag{27}$$

where $\mathbf{X}$ is a clean data matrix with the column vectors drawn from a union of subspaces and $\mathbf{E}$ is an unknown matrix of outliers which can be arbitrary in magnitude, but affecting only a fraction of the entries.

For data points sampled from multiple subspaces, the rank of clean data must be less than or equal to the sum of all the intrinsic subspace dimensions. In most applications, this is usually much less than the observed dimension. Therefore, we assume that the clean data $\mathbf{X}$ should be of low rank (or small nuclear norm Wright et al., 2009). For $\mathbf{E}$, as the outliers can be arbitrary in magnitude, but affecting only a fraction of the entries, the $l_{2,1}$ norm (Liu, Ji, & Ye, 2009; Liu et al., 2010) can be employed for this term. Adding the above two constraints to LR-MRW, we have the following robust model for MRW learning, named Robust Low-Rank MRW (RLR-MRW):

$$\begin{aligned}&\min_{\mathbf{X},\mathbf{E},\mathbf{P}} \mathcal{J}_R(\mathbf{X},\mathbf{E},\mathbf{P}) + \mu\|\mathbf{P}\|_*,\\ &\text{s.t. } \mathbf{X}_0 = \mathbf{X} + \mathbf{E},\ \mathbf{P}\mathbf{1}_n = \mathbf{1}_n,\ \mathbf{P} \in \mathcal{S}^n,\ \mathbf{P} \ge 0,\\ &\text{where } \mathcal{J}_R(\mathbf{X},\mathbf{E},\mathbf{P}) = \mathcal{J}(\mathbf{P}) + \eta\|\mathbf{X}\|_* + \gamma\|\mathbf{E}\|_{2,1}.\end{aligned} \tag{28}$$

5.2. Analysis on RLR-MRW model

Now we propose a brief analysis on RLR-MRW model. The optimization problem (28) can be split into the following two subproblems:

$$\begin{aligned}&\text{(P.1): } \min_{\mathbf{X},\mathbf{E}} \operatorname{tr}(\mathbf{X}(\mathbf{I}-\mathbf{P})\mathbf{X}^T) + \eta\|\mathbf{X}\|_* + \gamma\|\mathbf{E}\|_{2,1},\\ &\qquad \text{s.t. } \mathbf{X}_0 = \mathbf{X} + \mathbf{E}.\\ &\text{(P.2): } \min_{\mathbf{P}} \mu\|\mathbf{P}\|_* - \operatorname{tr}(\mathbf{X}\mathbf{P}\mathbf{X}^T),\\ &\qquad \text{s.t. } \mathbf{P}\mathbf{1}_n = \mathbf{1}_n, \mathbf{P} \in \mathcal{S}^n,\ \mathbf{P} \ge 0.\end{aligned} \tag{29}$$

Given a transition matrix $\mathbf{P}$, problem (P.1) can be considered as a random walk regularized extension for RPCA. As discussed in Liu et al. (2013), RPCA cannot handle well the mixture data, since it hinges on the assumption that the underlying data structure is a single low-rank subspace. With respect to this method, subproblem (P.1) is a general one that can leverage the power of both RPCA (outliers detection) and random walk regularization (preserving multiple subspace structure) for data sampled from a mixture of subspaces.

5.3. Solving RLR-MRW model

Based on the analysis in above subsection, we now propose a decomposition-based strategy for solving problem (28), in which each subproblem can be solved by ADM method. Specifically, we consider the equivalent model (29) and iteratively solve (P.1) and (P.2) to update $\mathbf{X}$, $\mathbf{E}$, and $\mathbf{P}$. Clearly, subproblem (P.1) can be reformulated as

$$\begin{aligned}&\min_{\mathbf{X},\mathbf{R},\mathbf{E}} \operatorname{tr}(\mathbf{X}(\mathbf{I}-\mathbf{P})\mathbf{X}^T) + \eta\|\mathbf{R}\|_* + \gamma\|\mathbf{E}\|_{2,1},\\ &\text{s.t. } \mathbf{X} = \mathbf{R},\ \mathbf{X}_0 = \mathbf{X} + \mathbf{E}.\end{aligned} \tag{30}$$

By introducing the augmented Lagrangian function

$$\begin{aligned}\mathcal{L}_A(\mathbf{X},\mathbf{R},\mathbf{E}) =\ & \operatorname{tr}(\mathbf{X}(\mathbf{I}-\mathbf{P})\mathbf{X}^T) + \eta\|\mathbf{R}\|_*\\ &+ \gamma\|\mathbf{E}\|_{2,1} + \langle\mathbf{L}_1, \mathbf{X}-\mathbf{R}\rangle + \langle\mathbf{L}_2, \mathbf{X}_0-\mathbf{X}-\mathbf{E}\rangle\\ &+ \frac{\beta_1}{2}\|\mathbf{X}-\mathbf{R}\|_F^2 + \frac{\beta_2}{2}\|\mathbf{X}_0-\mathbf{X}-\mathbf{E}\|_F^2,\end{aligned}$$

it is not hard to see that problem (P.1) can be solved by the following updating schemes:

$$\begin{cases}\mathbf{X}^+ = (\beta_1\mathbf{R} + \beta_2(\mathbf{X}_0 - \mathbf{E}) - \mathbf{L}_1 + \mathbf{L}_2)\mathbf{B},\\ \mathbf{R}^+ = \mathbf{U}\mathcal{T}_{\frac{\eta}{\beta_1}}(\Sigma)\mathbf{V}^T,\\ \mathbf{E}^+ = \arg\min_{\mathbf{E}} \frac{1}{2}\left\|\mathbf{E} - \left(\mathbf{X}_0 - \mathbf{X}^+ + \frac{\mathbf{L}_2}{\beta_2}\right)\right\|_F^2 + \frac{\gamma}{\beta_2}\|\mathbf{E}\|_{2,1},\\ \mathbf{L}_1^+ = \mathbf{L}_1 + \beta_1(\mathbf{X}^+ - \mathbf{R}^+),\\ \mathbf{L}_2^+ = \mathbf{L}_2 + \beta_2(\mathbf{X}_0 - \mathbf{X}^+ - \mathbf{E}^+),\end{cases} \tag{31}$$

where $\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{X}^+ + \frac{\mathbf{L}_1}{\beta_1}$ is the singular value decomposition (SVD) and $\mathbf{B} = (2(\mathbf{I}-\mathbf{P}) + (\beta_1+\beta_2)\mathbf{I})^{-1}$. The solution of $\mathbf{E}^+$ is by the following lemma (Liu et al., 2010).

Lemma 7. *Let $\mathbf{Q} = [\mathbf{q}_1, \ldots, \mathbf{q}_n]$ be a given matrix. Then $\mathbf{E}^* = [\mathbf{e}_1^*, \ldots, \mathbf{e}_n^*]$ is the optimal solution to the problem*

$$\min_{\mathbf{E}} \frac{1}{2}\|\mathbf{E}-\mathbf{Q}\|_F^2 + \gamma\|\mathbf{E}\|_{2,1}, \tag{32}$$

where the ith column of $\mathbf{E}^*$ is

$$\mathbf{e}_i^* = \begin{cases}\dfrac{\|\mathbf{q}_i\|_2 - \gamma}{\|\mathbf{q}_i\|_2}\mathbf{q}_i, & \gamma < \|\mathbf{q}_i\|_2,\\ 0, & \text{otherwise}.\end{cases} \tag{33}$$

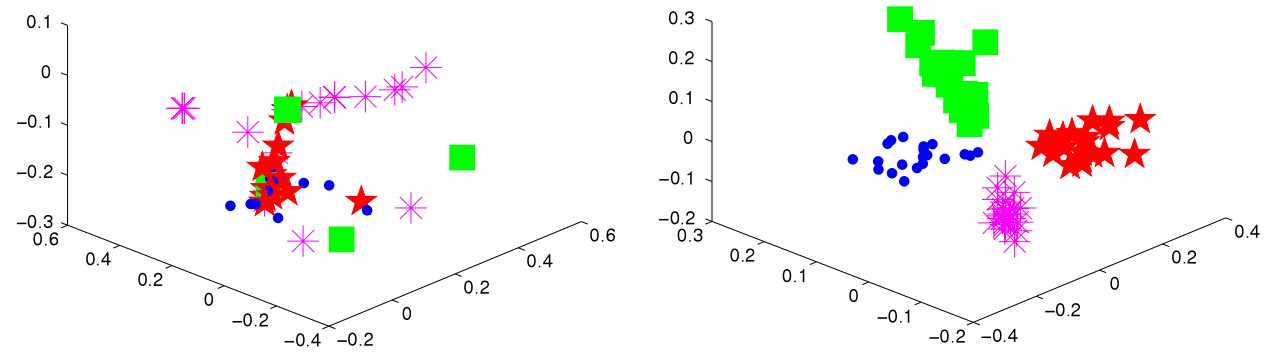
The optimization problem (P.2) is the original LR-MRW model and thus can be efficiently solved by (26). Algorithm 4 summarizes the whole solution method for RLR-MRW.

Algorithm 4 Iteratively Solving (28) via ADM

Input: Observation matrix $\mathbf{X}_0 \in \mathbb{R}^{m\times n}$, number k of subspaces, the dimension d of embedding and $\varepsilon = 10^{-3}$.
Initialization: Initialize $\mathbf{P}$ by solving (10) with $\mathbf{X}_0$.
while not converged **do**
  Step 1: Solve (P.1) in (29) to get solution ($\mathbf{X}^+$, $\mathbf{E}^+$).
  Step 2: Solve (P.2) in (29) to get solution $\mathbf{P}^+$.
  Step 3: Check the convergence condition:
  $\max\{\|\mathbf{X}^+ - \mathbf{X}\|_F/\|\mathbf{X}\|_F, \|\mathbf{P}^+ - \mathbf{P}\|_F/\|\mathbf{P}\|_F\} \le \varepsilon$.
  Step 4: Update $\mathbf{X} = \mathbf{X}^+$, $\mathbf{P} = \mathbf{P}^+$.
end while
Output: Clean data $\mathbf{X}$ and transition matrix $\mathbf{P}$.

6. Experimental results

In this section, we evaluate the performance of our proposed algorithms on both synthetic data and real vision tasks. Some previous state-of-the-art methods are also included. For LR-MRW and RLR-MRW, the transition probabilities are obtained by (10) and (28), respectively.



(a) G-MRW. (b) LR-MRW.

Fig. 2. Comparing the embedding and clustering performance of LR-MRW and G-MRW. The color (and the marker symbol) of the points are the true labels. The clustering accuracy are 48.8% (G-MRW) and 100% (LR-MRW), respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6.1. Synthetic data

In this subsection, we perform subspace clustering on synthetic data to compare the mechanism of LR-MRW (also RLR-MRW) with that of Gaussian kernel based MRW (G-MRW) and demonstrate the advantages of learning low-rank MRW by our proposed local and global criteria for multiple subspace data set. Our theoretical analysis in previous sections can also be verified by these experiments.

For G-MRW, we define the Gaussian kernel $\mathbf{W} = [w_{ij}]_{n\times n}$ as follows:

$$w_{ij} = \exp\left(\frac{-\|\mathbf{x}_i - \mathbf{x}_j\|_2^2}{2\sigma^2}\right), \tag{34}$$

where σ determines the width of the Gaussian kernel. Then the transition matrix for G-MRW is defined as $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$.

The synthetic data, parameterized as (k, p, m, d), is constructed as follows: k independent subspaces $\{\mathcal{C}_i\}_{i=1}^k$ whose basis $\{\mathbf{U}\}_{i=1}^k$ are computed by $\mathbf{U}_{i+1} = \mathbf{T}\mathbf{U}_i$, $1 \le i \le k-1$, where $\mathbf{T}$ is a random rotation and $\mathbf{U}_1$ is a random column orthogonal matrix of dimension $m \times d$. So each subspace has a rank of d and the data has an ambient dimension of m. Then we construct a $m \times kp$ data matrix $\mathbf{X} = [\mathbf{X}_1, \ldots, \mathbf{X}_k]$ by sampling p data vectors from each subspace by $\mathbf{X}_i = \mathbf{U}_i\mathbf{C}_i$, $1 \le i \le k$, with $\mathbf{C}_i$ being a $d \times p$ matrix i.i.d. $\mathcal{N}(0, 1)$.

6.1.1. Exactly clustering and estimating clean data

We first generate data set $(4, 30, 100, 3)$ (without noise) to understand the mechanism of LR-MRW and G-MRW for multiple subspaces data modeling. Fig. 1 presents the transition probabilities learnt from (10) and determined by heat kernel (34), respectively. One can see that the transition probabilities obtained by LR-MRW are much higher when the points belong to the same subspace and much lower when the points belong to different subspaces (Fig. 1(b)), whereas G-MRW only achieves high transition probabilities between local neighbors (Fig. 1(a)). This confirms our theoretical analysis in Section 3.2.

Fig. 2 compares the embedding and clustering results of G-MRW and LR-MRW on our generated mixture of subspaces. We can see from Fig. 2 that LR-MRW embedding (Fig. 2(b)) yields different groups of points that are easily identifiable while G-MRW embedding (Fig. 2(a)) results in points that do not have clear clusters. Therefore, the clustering accuracy of LR-MRW is dramatically higher than that of G-MRW. This is because the transition probabilities learnt from LR-MRW can successfully recover the intrinsic structure of the data set. This confirms the effectiveness of our proposed local and global criteria for modeling multiple subspace structures.

Now we discuss the influence of the embedding dimension to our problem. Specifically, we generate a data set with parameter $(10, 20, 100, 50)$ and compute transition matrix $\mathbf{P}$ by (10) for this data. Then we learn the embedding $\mathbf{H}$ from (15) with the dimension $m \in [1, 10]$. Fig. 3 compares the clustering performances for different embedding dimensions. It can be seen that the k-dimensional embedding (here $k = 10$) with first k eigenvectors ("Emb1" at $d = k$) and the $(k-1)$-dimensional embedding with the second to the kth eigenvectors ("Emb2" at $d = k-1$) all achieved the best clustering performance (i.e., 100%). This evaluation result confirms our theoretical analysis of the optimal dimensionality for embedding in Proposition 3.

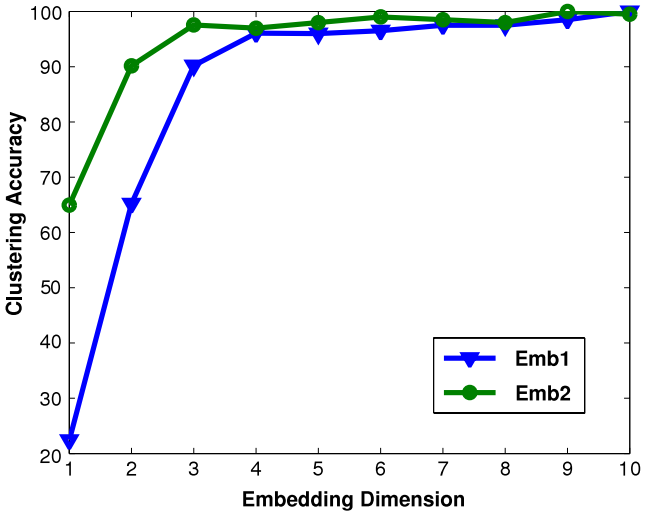


Fig. 3. Comparing the clustering accuracies (% averaged over 20 runs) of LR-MRW with different embedding dimensions. "Emb1" and "Emb2" denote utilizing the first d and the second to the $(d+1)$-th eigenvectors to obtain the embedding, respectively. The x-axis represents the embedding dimension and the y-axis represents the clustering accuracy. "Emb1" with dimension 1 only achieves 20% accuracy because the first eigenvector of $\mathbf{P}$ is $\mathbf{1}_n$.

6.1.2. Robustness to data corruptions

To further test the performance of G-MRW, LR-MRW and RLR-MRW on data with noises and outliers, we generate another data set with quadruple $(6, 20, 100, 5)$ in the same way as in Section 6.1.1.

Firstly, we test the robustness of our algorithms for gross corruptions (i.e., outliers). To do so, some data vectors $\mathbf{x}$ are randomly chosen to be corrupted by Gaussian noise with zero means and standard deviation $0.2\|\mathbf{x}\|_2$. The results in Fig. 4 show that both LR-MRW and RLR-MRW can achieve perfect clustering results when there is no corruption. G-MRW also performs fairly well for this clean data. However, G-MRW is very sensitive to corruptions and

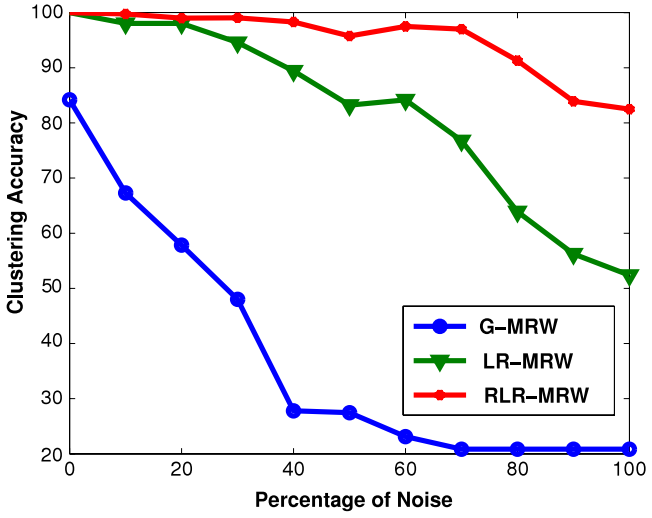


Fig. 4. Comparing the clustering accuracies (% averaged over 20 runs) of GMRW, LR-MRW and RLR-MRW for various percentage of noise.

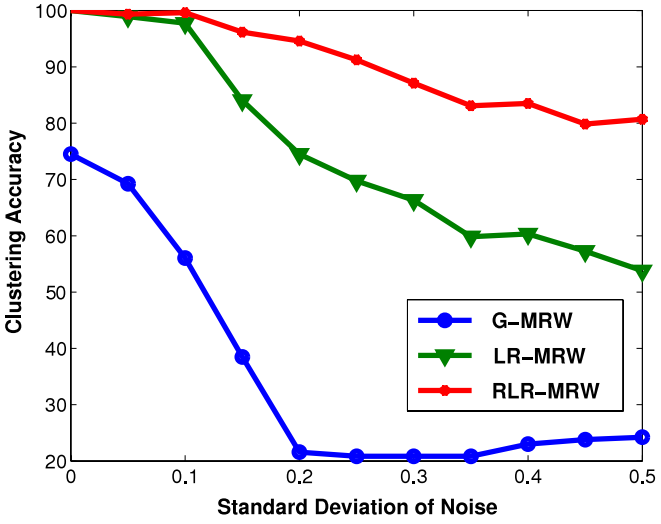


Fig. 5. Comparing the clustering accuracies (% averaged over 20 runs) of GMRW, LR-MRW and RLR-MRW at various noise intensities.

the performances of G-MRW will reduce to only 20% when most of the data points are corrupted (more than 70%). LR-MRW also cannot work well when most of the data points are corrupted. In this case, RLR-MRW successfully removes the outliers and still performs well.

Then we test the performances of these algorithms on data with noises at various intensities. To do so, 50% data points are randomly chosen to be corrupted by adding Gaussian noise with zero mean and standard deviation $\sigma \in [0, 0.5]$. Fig. 5 shows that both LR-MRW and RLR-MRW perform much better than G-MRW. RLR-MRW is more robust to large noises and thus achieves higher clustering accuracy than LR-MRW.

6.2. Motion segmentation

To verify the clustering performance of our proposed models for real world problems, now we utilize LR-MRW and RLR-MRW to address the motion segmentation problem, which refers to the task of separating a video sequence into multiple spatiotemporal regions corresponding to different rigid-body motions. As shown in Rao, Tron et al. (2010), the motion segmentation problem can be preceded by first extracting a set of point trajectories from the video sequences using standard tracking methods. Then the problem reduces to clustering these point trajectories according to different rigid-body motions in the scene.

6.2.1. Segmentation performance on sequences with small noises

We first evaluate our proposed model on the Hopkins155 motion database (Tron & Vidal, 2007). This database consists of 156 sequences of two or three motions (see Fig. 6). As the motion sequences were obtained using an automatic tracker, and errors in tracking were manually corrected for each sequence, it could be regarded that these sequences only contain small noises and there is no attempt to deal with heavily corrupted trajectories. Therefore, we test LR-MRW on this data set. In order to compare LR-MRW with the state-of-the-art approaches, we also list the results of GPCA, Random Sample Consensus (RANSAC) (Fischler & Bolles, 1981), Local Subspace Analysis (LSA) (Yan & Pollefeys, 2006), SR[7] and LRR.[8]

As some conventional methods (e.g., GPCA and RANSAC) may fail to return any results on the raw sequences of Hopkins155 database within a reasonable response time (i.e., 1 day), it is necessary to perform a PCA preprocessing step for these methods (Liu et al., 2010; Vidal, 2010; Vidal et al., 2005) to reduce the dimensionality of the problem for computational efficiency. Moreover, PCA can also reduce some small dense noises in the raw sequences. In order to make a fair comparison for all the methods in the experiments, we project the trajectories into a subspace of lower dimensionality (i.e., 5D or 12D), in which the choices of PCA dimension are respectively suggested by Liu et al. (2010) and Vidal (2010).

For each algorithm and each sequence, we record the classification error defined as

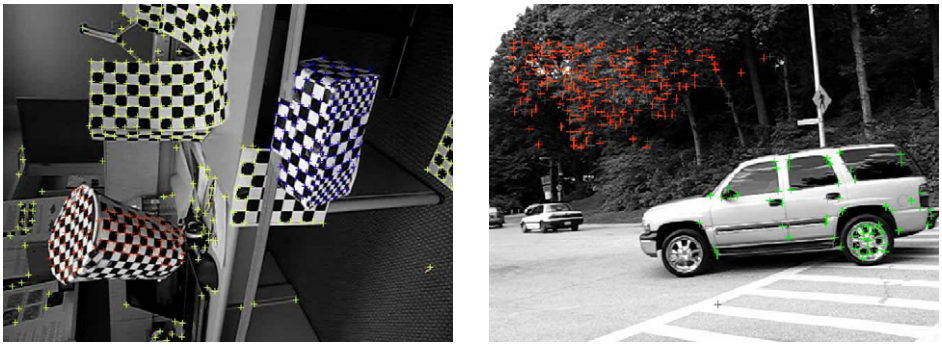
$$\text{classification error} := \frac{\#\text{ of misclassified points}}{\text{total }\#\text{ of points}}\%. \tag{35}$$

The detailed statistics of the classification errors (i.e., the average, standard deviation (std.) and maximum (max.) of the results) are shown in Table 1. As there exists one degenerate sequence in this database, the results are reported for both 155 (discarding the degenerate data) and 156 (all) sequences. It can be seen that all the methods are sensitive to the dimension of the projection. The performances of GPCA, RANSAC, and LR-MRW in 5D space are better than that in 12D space. While LSA, SR and LRR give better results in the 12D space. Although LRR is better than other compared methods when the dimension of the projection is 12D, our LR-MRW with 5D PCA projection achieves the best performance among all the algorithms with all the projections. Figs. 7 and 8 further show the percentage of sequences for which the classification error is below a given percentage of misclassification (for 5D data set). All these results demonstrate that LR-MRW significantly outperforms other methods.

The only parameter in LR-MRW model (10) is $\mu \geq 0$. This parameter is used to balance the effects of the local and global measures in the cost function of LR-MRW. In general, the choice of this parameter depends on the prior knowledge of the data structure. That is, when data samples from the same cluster have strict linear relationships (i.e., can be exactly modeled by subspace), we should use relatively large μ, while when the data samples tend to have high local neighborhood similarity, we should set μ to be relatively small. In the extreme case that the low-rank assumption is invalid on the data set, we can simply set $\mu = 0$ in our model. As shown in Fig. 9, when μ ranges from 0 to 0.4, the classification error varies slightly from 4.6% to 5.8% (for 5D data set).

[7] The SR approach solves the representation matrix $\mathbf{Z}$ and outliers $\mathbf{E}$ by $\min \|\mathbf{Z}\|_1 + lambda\|\mathbf{E}\|_1$, s.t. $\mathbf{X} = \mathbf{XZ} + \mathbf{E}$, $\text{diag}(\mathbf{Z}) = 0$.

[8] The Matlab code of GPCA, RANSAC and LSA is available at http://www.vision.jhu.edu/data/hopkins155/. The Matlab code of SR and LRR can be downloaded from https://sites.google.com/site/guangcanliu/.



(a) 1R2RC. (b) cars1.

Fig. 6. Example image frames from two motion sequences from the Hopkins155 database.

Table 1
Classification errors (%) on Hopkins155. In the right four columns, we report results on both 155 and 156 sequences.

Dim.	Method	2 motions			3 motions			All (155/156)		
		mean	std.	max.	mean	std.	max.	mean	std.	max.
12D	GPCA	20.6	16.6	50.0	20.0	15.4	62.5	22.7/22.8	16.8/16.8	62.5/62.5
	RANSAC	35.0	11.9	49.7	46.9	**11.5**	64.2	37.7/37.9	12.8/13.0	64.2/66.0
	LSA	6.4	13.0	50.0	**9.8**	15.0	53.8	7.2/7.2	13.5/13.5	53.8/53.8
	SR	5.1	9.8	46.3	14.7	15.2	59.7	7.2/7.4	11.9/12.2	59.7/59.7
	LRR	3.4	8.6	**40.3**	**9.8**	12.0	**41.4**	4.8/5.0	9.8/**9.9**	**41.4**/**41.4**
	LR-MRW	6.3	12.3	48.2	11.7	13.0	44.7	7.6/7.6	12.7/12.7	48.2/48.2
5D	GPCA	9.5	13.6	49.0	22.8	15.6	48.2	12.5/12.7	15.1/15.3	49.0/49.0
	RANSAC	5.3	9.4	44.5	17.7	12.8	48.5	8.1/8.3	11.5/11.8	48.5/48.5
	LSA	5.0	9.2	49.4	19.5	16.7	55.0	8.3/8.5	12.6/12.6	55.0/55.0
	SR	10.4	16.0	49.7	18.9	16.6	49.3	12.3/12.5	15.0/15.0	49.7/49.7
	LRR	5.7	10.1	48.6	16.5	14.9	43.8	8.2/8.3	12.7/12.9	48.6/48.6
	LR-MRW	**2.5**	**7.2**	44.6	11.8	12.9	**41.4**	**4.6**/**4.7**	**9.7**/**9.9**	44.6/44.6

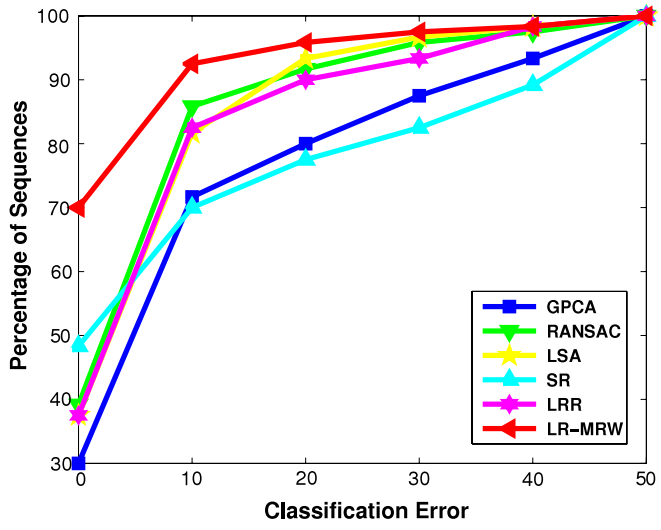


Fig. 7. Percentage of sequences (2 motions) for which the classification error is less than or equal to a given percentage of misclassification.

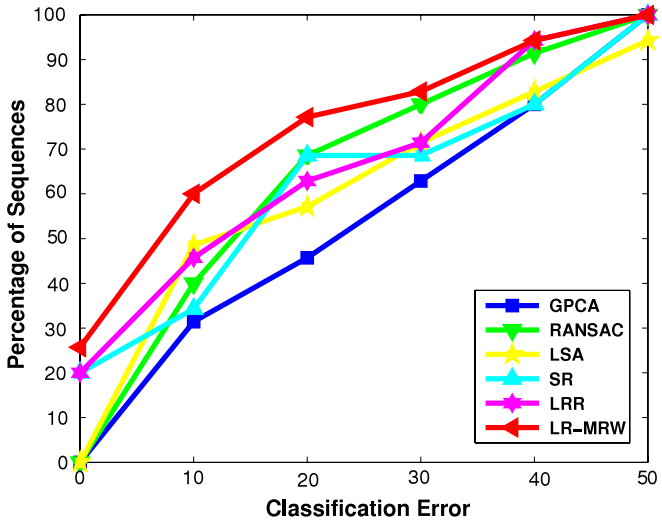


Fig. 8. Percentage of sequences (3 motions) for which the classification error is less than or equal to a given percentage of misclassification.

6.2.2. Segmentation performance on sequences with gross corruptions

We further test the robustness of RLR-MRW for motion sequences with outliers and noise. We first choose two sequences ("1R2RC" and "cars1") from the Hopkins155 database and add 30% outlying trajectories to the data set of a given motion sequence. Outlying trajectories were generated by choosing a random initial point in the first frame and then selecting a random increment between successive frames.[9] We then run 20 trails with different randomly generated outlying trajectories and report the average classification errors for each sequence in Table 2. We also test RLR-MRW on 16 additional sequences of Hopkins155 database that contain real corruption (Rao, Tron et al., 2010). Table 2 shows that RLR-MRW can outperform other state-of-the-art algorithms for motion segmentation with both simulated and real corruptions.

6.2.3. Estimating the number of subspaces

Now we consider the problem of subspace number estimation for motion segmentation. As each sequence in Hopkins155 consists of data vectors drawn from two or three motions, this database can also be used to evaluate the effectiveness of Algorithm 2. In the

[9] The Matlab code for generating the outlying trajectories can be found at http://www.vision.jhu.edu/code/.

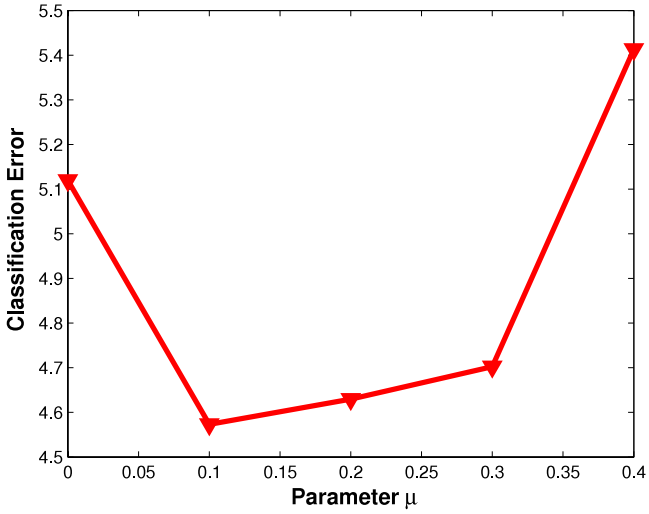


Fig. 9. The influences of the parameter μ in LR-MRW for motion segmentation.

Table 2
Classification errors (%) on corrupted motion sequences, containing simulated corruption (the top two rows) and real corruption (bottom row).

Data	GPCA	RANSAC	LSA	SR	LRR	RLR-MRW
1R2RC	59.5	61.7	60.1	50	50.2	**45.5**
cars1	38.6	44.8	38.8	30.2	40.1	**29.6**
Real	28.1	27.9	35.1	33.1	33.3	**27.1**

Table 3
Subspaces number estimation on Hopkins155.

τ	# predicted sequences	Prediction rate (%)	Absolute error
0.70	86	55.13	0.46
0.725	113	72.44	0.29
0.75	**120**	**76.92**	**0.24**
0.775	116	74.36	0.27
0.80	98	62.82	0.42

results presented in Table 3, the prediction rate and the absolute error averaged over $m = 156$ sequences are respectively defined as

$$\text{prediction rate} := \frac{\#\text{ predicted sequences}}{m}\%,$$

and

$$\text{absolute error} := \left(\sum_{i=1}^{m} |k_i - \bar{k}_i|\right) \Big/ m,$$

where the subscript i denotes the ith sequence. It can be seen that when $\tau = 0.75$ our spectrum based strategy in Algorithm 2 can correctly predict the true subspace number $\bar{k}$ of 120 sequences and the average absolute error of all 156 sequences is only 0.24.

6.3. Temporal segmentation of video sequence

We consider the problem of partitioning a long video sequence into multiple short segments corresponding to different scenes. We assume that all the image frames having the same scene lie in a low-dimensional subspace, and that different scenes correspond to different subspaces. We show that both LR-MRW and RLR-MRW can be applied to solve this problem.

6.3.1. Fox TV video sequence

We first borrow the video sequence from Vidal et al. (2005), which is about an interview at Fox TV (Fig. 10). It consists of 135 images of size 294×413, each containing either the interviewer alone, or the interviewee along, or both. We would like to segment these images into three scenes. We also apply PCA to reduce the dimension of images and then apply all the algorithms to these image sequence.

Table 4
Segmentation accuracies (%) on Fox TV video sequences.

Method	5D	6D	7D	8D	9D	10D
GPCA	70.4	96.3	**100**	63.0	70.4	63.0
RANSAC	55.6	63.0	44.4	66.7	66.7	66.7
LSA	55.6	55.6	85.2	63.0	85.2	74.1
SR	**100**	**100**	66.7	85.2	63.0	48.2
LRR	81.5	85.2	85.2	92.6	92.6	88.9
LR-MRW	**100**	**100**	**100**	**100**	**100**	**100**

Table 4 shows the segmentation accuracy of all algorithms at different PCA dimensions. One can see that GPCA and SR can handle such problem when choosing proper PCA dimensions. However, they are sensitive to the dimension. On the contrary, LR-MRW achieves 100% segmentation accuracy on features at all PCA dimensions.

6.3.2. Variety show video sequence

To further testify the performance of our proposed algorithms in real situations, in this experiment we test on a more complex real-world video sequence. We download a variety show video sequence[10] with length 380 frames (288×352 pixels) from YouTube. This video sequence can be divided into 10 different scenes (8 scenes containing different single person, 1 scene containing multiple persons and 1 scene containing no person, all these scenes have very similar background). To further test the robustness of these algorithms, 50% frames of the sequence are randomly chosen to be corrupted by using large Gaussian noise with zero mean and standard deviation $3.5\|\mathbf{x}_i\|_{l_2}$, where $\mathbf{x}_i$ is the ith frame.

Table 5 shows the performance of various methods by using the 5D data produced by PCA. One can see that both LR-MRW and RLR-MRW perform well on the original video sequences. On the corrupted data, LR-MRW outperforms other state-of-the-art clustering methods. As expected, RLR-MRW performs better than LR-MRW for data with large corruptions.

7. Conclusions and future work

This paper proposes a new way to learn Markov random walks for multiple subspaces clustering and estimation. Unlike conventional spectral clustering algorithms which use the Gaussian kernel to generate the transition matrix, LR-MRW aims at directly learning the transition probabilities under some intuitive criteria to capture both within-subspace homogeneity and between-subspace discrimination. Theoretical analysis shows that under some suitable conditions, our proposed mechanism can successfully reveal the multiple subspaces structure. We also propose a robust extension of LR-MRW (RLR-MRW) to integrate subspace clustering and estimation and error correction in a unified framework. As a non-trivial byproduct, we prove a result that establishes the intrinsic connection between nuclear norm regularized and l_1 regularized least square problems. Our experiments show that LR-MRW and its robust extension are promising for both symmetric data and real applications. However, there still remain several problems for future work.

First, it is interesting to consider whether the criteria presented in LR-MRW can be extended to the general graph learning problems (e.g., spectral clustering and graph embedding) other than

[10] The video can be found on http://sites.google.com/site/rsliu0705/fcwr.zip.

Fig. 10. Three scenes of the Fox TV video sequence. The integers in the brackets are the number of frames for the scene.

Table 5
Segmentation accuracies (%) on variety show video sequence. The top row are results on original data and the bottom row are results on data with simulated corruption.

Data	GPCA	RANSAC	LSA	SR	LRR	LR-MRW	RLR-MRW
Original	51.32	45.26	38.16	82.63	93.42	97.11	**97.51**
Corrupted	45.79	39.47	36.32	68.42	80.79	90.17	**96.92**

MRW. Second, the most expensive computational task required by ADM based algorithm is to perform EVD or SVD inherent with the nuclear norm optimization at each iteration, which becomes increasingly costly when the data size grows. Recently, by combining a linearized version of ADM with an acceleration technique for SVD computation, the work in Lin, Liu, and Su (2011) proposed a fast solver for nuclear norm minimization. It is attractive to apply similar strategy to solve LR-MRW related models for large scale data set. Third, LR-MRW only considers the unsupervised clustering and embedding tasks. We would like to apply our model for more problems, e.g., semi-supervised and supervised learning.

Acknowledgments

Risheng Liu is supported by the National Natural Science Foundation of China (No. 61300086), the China Postdoctoral Science Foundation (No. 2013M530917), the Fundamental Research Funds for the Central Universities (No. DUT12RC(3)67) and the Open Project Program of the State Key Laboratory of CAD&CG, Zhejiang University, Zhejiang, China (No. A1404). Zhouchen Lin is supported by National Natural Science Foundation of China (Nos. 61272341, 61231002, 61121002). Zhixun Su is supported by National Natural Science Foundation of China (Nos. 61173103 and U0935004).

Appendix A. Proof of Theorem 2

The proof of Theorem 2 is based on the following lemma.

Lemma 8. *For any four matrix* $\mathbf{A}, \mathbf{B}, \mathbf{C}$ *and* $\mathbf{F}$ *of compatible dimensions, the following statements hold:*

$$\left\|\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{C} & \mathbf{F}\end{bmatrix}\right\|_* \geq \left\|\begin{bmatrix}\mathbf{A} & 0\\ 0 & \mathbf{F}\end{bmatrix}\right\|_* = \|\mathbf{A}\|_* + \|\mathbf{F}\|_*, \tag{A.1a}$$

$$\operatorname{rank}\left(\begin{bmatrix}\mathbf{A} & 0\\ 0 & \mathbf{F}\end{bmatrix}\right) = \operatorname{rank}(\mathbf{A}) + \operatorname{rank}(\mathbf{F}). \tag{A.1b}$$

Proof of Lemma 8. The proofs of Lemma 3.1 in Liu et al. (2010) and Lemma 3 in Wang and Lu (2009) can directly lead to the above conclusions. □

Proof of Theorem 2. Let $\mathbf{P}^* \in \mathbb{P}$ be an optimal solution to (8). Form a block-diagonal matrix $\tilde{\mathbf{P}}$ by setting

$$\tilde{\mathbf{P}}_{ij} = \begin{cases}\mathbf{P}^*_{ij}, & \text{if } \mathbf{x}_i \text{ and } \mathbf{x}_j \text{ belong to the same subspace,}\\ 0, & \text{otherwise,}\end{cases} \tag{A.2}$$

and denote $\mathbf{C} = \mathbf{P}^* - \tilde{\mathbf{P}}$. By the definition and assumption of $\mathbb{P}$, we have that $\tilde{\mathbf{P}} \in \mathbb{P}$. Furthermore, we have that

$$\mathcal{J}(\tilde{\mathbf{P}}) \leq \mathcal{J}(\mathbf{P}^*) = \mathcal{J}(\tilde{\mathbf{P}}) + \mathcal{J}(\mathbf{C}). \tag{A.3}$$

By (A.1a) in Lemma 8, $\|\tilde{\mathbf{P}}\|_* \leq \|\mathbf{P}^*\|_*$. Therefore, $\tilde{\mathbf{P}}$ is also optimal for (8). As the problem (8) is convex, we have $\mathcal{J}(\tilde{\mathbf{P}}) + \mu\|\tilde{\mathbf{P}}\|_* = \mathcal{J}(\mathbf{P}^*) + \mu\|\mathbf{P}^*\|_*$ and thus $\mathcal{J}(\tilde{\mathbf{P}}) \geq \mathcal{J}(\mathbf{P}^*)$. This together with (A.3) concludes that $\mathcal{J}(\tilde{\mathbf{P}}) = \mathcal{J}(\mathbf{P}^*)$ and $\mathcal{J}(\mathbf{C}) = 0$. By $\|\mathbf{x}_i - \mathbf{x}_j\|_2 > 0$ for any $i \neq j$, we have that $\mathbf{C} = \mathbf{0}$, and so we conclude that $\mathbf{P}^* = \tilde{\mathbf{P}}$ has the block-diagonal structure (9). □

Appendix B. Proof of Theorem 5

The proof of Theorem 5 is based on the following lemma.

Lemma 9 (*Horn & Johnson, 2008*)**.** *For* $\forall \mathbf{A}, \mathbf{B} \in \mathcal{S}^n$*, the following inequality holds:*

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \|\Lambda(\mathbf{A}) - \Lambda(\mathbf{B})\|_F. \tag{B.1}$$

Proof of Theorem 5. Since $\mathbf{K} \in \mathcal{S}^n$, we have $\|\mathbf{K} - \mathbf{G}\|_F^2 = \|\mathbf{K} - \mathbf{G}^T\|_F^2$, which suggests the following relation $\frac{1}{2}\|\mathbf{K} - \mathbf{G}\|_F^2 = \frac{1}{4}\|\mathbf{K} - \mathbf{G}\|_F^2 + \frac{1}{4}\|\mathbf{K} - \mathbf{G}^T\|_F^2 = \frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + C$, where $\bar{\mathbf{G}} = (\mathbf{G} + \mathbf{G}^T)/2 \in \mathcal{S}^n$ and C is a constant. Therefore, optimization problem (21) is equal to the following model:

$$\min_{\mathbf{K}} \frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + \mu\|\mathbf{K}\|_* \quad \text{s.t. } \mathbf{K} \in \mathcal{S}^n. \tag{B.2}$$

Thus we only need to prove that $\mathbf{K}^*$ is the optimal solution to the problem (B.6). As $\mathbf{K}, \bar{\mathbf{G}} \in \mathcal{S}^n$, we know from Lemma 10 that

$$\|\mathbf{K} - \bar{\mathbf{G}}\|_F \geq \|\Lambda(\mathbf{K}) - \Lambda(\bar{\mathbf{G}})\|_F. \tag{B.3}$$

For $\forall \mathbf{K} \in \mathcal{S}^n$, we also have

$$\|\mathbf{K}\|_* = \|\Lambda(\mathbf{K})\|_* = \|\lambda(\mathbf{K})\|_1. \tag{B.4}$$

Using the relations (B.3) and (B.4), we immediately obtain that

$$\begin{aligned}&\frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + \mu\|\mathbf{K}\|_*\\ &\geq \frac{1}{2}\|\Lambda(\mathbf{K}) - \Lambda(\bar{\mathbf{G}})\|_F^2 + \mu\|\Lambda(\mathbf{K})\|_*\\ &= \frac{1}{2}\|\lambda(\mathbf{K}) - \lambda(\bar{\mathbf{G}})\|_2^2 + \mu\|\lambda(\mathbf{K})\|_1, \quad \forall \mathbf{K} \in \mathcal{S}^n,\end{aligned}$$

which together with $\text{diag}(\lambda(\mathbf{K})) \in \mathcal{S}^n$ implies that the optimal value of problem (B.6) is minorized by problem (22). Further, by the definition of $\mathbf{k}^*$, we know that $\text{diag}(\mathbf{k}^*) \in \mathcal{S}^n$, implies that $\mathbf{K}^* \in \mathcal{S}^n$, that is, $\mathbf{K}^*$ is a feasible solution of problem (B.6). Moreover, we observe that $\|\mathbf{K}^*\|_* = \|\text{diag}(\mathbf{k}^*)\|_* = \|\mathbf{k}^*\|_1$ and $\|\mathbf{K}^* - \bar{\mathbf{G}}\|_F = \|\mathbf{U}\,\text{diag}(\mathbf{k}^*)\mathbf{U}^T - \bar{\mathbf{G}}\|_F = \|\mathbf{k}^* - \lambda(\bar{\mathbf{G}})\|_{l_2}$. Thus, the objective function (B.6) reaches the optimal value of problem (22) at $\mathbf{K}^*$. It then immediately follows that problem (B.6) and (22) share the same optimal value, and hence $\mathbf{K}^*$ is an optimal solution of (B.6), which concludes the proof. □

The proof of Theorem 5 is based on the following lemma (Horn & Johnson, 2008).

Lemma 10. *For $\forall \mathbf{A}, \mathbf{B} \in \mathcal{S}^n$, the following inequality holds:*

$$\|\mathbf{A} - \mathbf{B}\|_F \geq \|\Lambda(\mathbf{A}) - \Lambda(\mathbf{B})\|_F. \tag{B.5}$$

Proof of Theorem 5. Since $\mathbf{K} \in \mathcal{S}^n$, we have

$$\|\mathbf{K} - \mathbf{G}\|_F^2 = \|\mathbf{K} - \mathbf{G}^T\|_F^2,$$

which suggests the following relation

$$\begin{aligned}\frac{1}{2}\|\mathbf{K} - \mathbf{G}\|_F^2 &= \frac{1}{4}\|\mathbf{K} - \mathbf{G}\|_F^2 + \frac{1}{4}\|\mathbf{K} - \mathbf{G}^T\|_F^2 \\ &= \frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + C,\end{aligned}$$

where $\bar{\mathbf{G}} = (\mathbf{G} + \mathbf{G}^T)/2 \in \mathcal{S}^n$ and C is a constant. Therefore, optimization problem (21) is equal to the following model:

$$\min_{\mathbf{K}} \frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + \mu\|\mathbf{K}\|_* \quad \text{s.t. } \mathbf{K} \in \mathcal{S}^n. \tag{B.6}$$

Thus we only need to prove that $\mathbf{K}^*$ is the optimal solution to the problem (B.6). As $\mathbf{K}, \bar{\mathbf{G}} \in \mathcal{S}^n$, we know from Lemma 10 that

$$\|\mathbf{K} - \bar{\mathbf{G}}\|_F \geq \|\Lambda(\mathbf{K}) - \Lambda(\bar{\mathbf{G}})\|_F.$$

For $\forall \mathbf{K} \in \mathcal{S}^n$, we also have

$$\|\mathbf{K}\|_* = \|\Lambda(\mathbf{K})\|_* = \|\lambda(\mathbf{K})\|_1.$$

Using the above relations, we immediately obtain that

$$\begin{aligned}&\frac{1}{2}\|\mathbf{K} - \bar{\mathbf{G}}\|_F^2 + \mu\|\mathbf{K}\|_* \\ &\geq \frac{1}{2}\|\Lambda(\mathbf{K}) - \Lambda(\bar{\mathbf{G}})\|_F^2 + \mu\|\Lambda(\mathbf{K})\|_* \\ &= \frac{1}{2}\|\lambda(\mathbf{K}) - \lambda(\bar{\mathbf{G}})\|_2^2 + \mu\|\lambda(\mathbf{K})\|_1, \quad \forall \mathbf{K} \in \mathcal{S}^n,\end{aligned}$$

which together with $\text{diag}(\lambda(\mathbf{K})) \in \mathcal{S}^n$ implies that the optimal value of problem (B.6) is minorized by problem (22). Further, by the definition of $\mathbf{k}^*$, we know that $\text{diag}(\mathbf{k}^*) \in \mathcal{S}^n$, implies that $\mathbf{K}^* \in \mathcal{S}^n$, that is, $\mathbf{K}^*$ is a feasible solution of problem (B.6). Moreover, we observe that

$$\|\mathbf{K}^*\|_* = \|\text{diag}(\mathbf{k}^*)\|_* = \|\mathbf{k}^*\|_1$$

and

$$\begin{aligned}\|\mathbf{K}^* - \bar{\mathbf{G}}\|_F &= \|\mathbf{U}\,\text{diag}(\mathbf{k}^*)\mathbf{U}^T - \bar{\mathbf{G}}\|_F \\ &= \|\mathbf{k}^* - \lambda(\bar{\mathbf{G}})\|_2.\end{aligned}$$

Thus, the objective function (B.6) reaches the optimal value of problem (22) at $\mathbf{K}^*$. It then immediately follows that problem (B.6) and (22) share the same optimal value, and hence $\mathbf{K}^*$ is an optimal solution of (B.6), which concludes the proof. □

References

Azran, A., & Ghahramani, Z. (2006). A new approach to data driven clustering. In *ICML*.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computing*, *15*(6), 1373–1396.

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

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, *58*(3), 11.

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*(6), 717–772.

Chung, F. R. K. (1997). *Spectral graph theory*. American Mathematical Society.

Costeira, J. P., & Kanade, T. (1998). A multibody factorization method for independently moving objects. *International Journal of Computer Vision*, *29*(3), 159–179.

David, M. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Elhamifar, E., & Vidal, R. (2009). Sparse subspace clustering. In *CVPR*.

Favaro, P., Vidal, R., & Ravichandran, A. (2011). A closed form solution to robust subspace estimation and clustering. In *CVPR*.

Fischler, M. A., & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, *24*(6), 381–395.

Gear, C. W. (1998). Multibody grouping from motion images. *International Journal of Computer Vision*, *29*(2), 133–150.

Geng, X., Smith-Miles, K., Zhou, Z.-H., & Wang, L. (2011). Face image modeling by multilinear subspace analysis with missing values. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *41*(3), 881–892.

Goh, A., & Vidal, R. (2007). Segmenting motions of different types by unsupervised manifold clustering. In *CVPR*.

Gruber, A., & Weiss, Y. (2004). Multibody factorization with uncertainty and missing data using the EM algorithm. In *CVPR*.

Hager, W. W. (1989). Updating the inverse of a matrix. *SIAM Review*, *31*(2), 221–239.

Hale, E.T., Yin, W., & Zhang, Y. A fixed-point continuation method for l1-regularized minimization with applications to compressed sensing. CAAM TR07-07. Rice University.

Horn, R., & Johnson, C. (2008). *Topics in matrix analysis*. Cambridge University Press.

Huang, Y., Liu, Q., & Metaxas, D. N. (2011). A component-based framework for generalized face alignment. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, *41*(1), 287–298.

Jolliffe, I. (2002). *Principal component analysis* (2nd ed.). Springer.

Lafon, S., & Lee, A. B. (2006). Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(9), 1393–1403.

Lin, Z., Chen, M., Wu, L., & Ma, Y. The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices. *UIUC technical report UILU-ENG-09-2215*.

Lin, Z., Liu, R., & Su, Z. (2011). Linearized alternating direction method with adaptive penalty for low rank representation. In *NIPS*.

Liu, J., Ji, S., & Ye, J. (2009). Multi-task feature learning via efficient $l_{2,1}$-norm minimization. In *UAI*.

Liu, R., Lin, Z., De la Torre, F., & Su, Z. (2012). Fixed-rank representation for unsupervised visual learning. In *CVPR*.

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

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *ICML*.

Mei, X., & Ling, H. (2011). Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(11), 2259–2272.

Meila, M., & Shi, J. (2001). Learning segmentation with random walk. In *NIPS*.

Nadler, B., Lafon, S., & Coifman, R. (2005). Diffusion maps, spectral clustering and eigenfunctions of Fokker–Planck operators. In *NIPS*.

Nasihatkon, B., & Hartley, R. (2011). Graph connectivity in sparse subspace clustering. In *CVPR*.

Ng, A., Jordan, M., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *NIPS*.

Ni, Y., Sun, J., Yuan, X., Yan, S., & Cheong, L. (2010). Robust low-rank subspace segmentation with semidefinite guarantees. In *ICDM workshop*.

Qiu, H., & Hancock, E. R. (2007). Clustering and embedding using commute times. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(11), 1873–1890.

Rao, S., Tron, R., Vidal, R., & Ma, Y. (2010). Motion segmentation in the presence of outlying, incomplete, or corrupted trajectories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(10), 1832–1845.

Rao, S. R., Yang, A. Y., Sastry, S. S., & Ma, Y. (2010). Robust algebraic segmentation of mixed rigid-body and planar motions from two views. *International Journal of Computer Vision*, *88*(3), 425–446.

Recht, B., Fazel, M., & Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, *52*(3), 471–501.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.
Shi, L., Liu, Z.-Y., Tu, S., & Xu, L. (2014). Learning local factor analysis versus mixture of factor analyzers with automatic model selection. *Neurocomputing*, *139*, 3–14.
Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.
Souvenir, R., & Pless, R. (2005). Manifold clustering. In *ICCV*.
Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.
Toh, K.-C., & Yun, S. (2010). An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, *6*(615–640), 15.
Tron, R., & Vidal, R. (2007). A benchmark for the comparison of 3D montion segmentation algorithms. In *CVPR*.
Vempala, S. (2005). Geometric random walks: a survey. *SRI Volume on Combinatorial and Computational Geometry*.
Vidal, R. (2010). A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, *28*(2), 52–68.
Vidal, R., Ma, Y., & Sastry, S. (2005). Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1945–1959.
Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and Computing*, *17*(4), 395–416.
Wang, J., & Lu, J. (2009). Proof of inequality of rank of matrix on skew field by constructing block matrix. *International Mathematical Forum*, *36*(4), 1803–1808.
Wang, X., Tieu, K., & Grimson, W. E. L. (2010). Correspondence-free activity analysis and scene modeling in multiple camera views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(1), 56–71.
Wei, S., & Lin, Z. Analysis and improvement of low rank representation for subspace segmentation. arXiv Preprint arXiv:1107.1561.
Wright, J., Ganesh, A., Rao, S., & Ma, Y. (2009). Robust principal component analysis: exact recovery of corrupted low-rank matrices via convex optimization. In *NIPS*.
Xu, L. Beyond PCA learnings: from linear to nonlinear and from global representation to local representation (invited talk). In *International conference on neural information processing*.
Xu, L. (2002). BYY harmony learning, structural RPCL, and topological self-organizing on mixture models. *Neural Networks*, *15*(8), 1125–1151.
Yan, J., & Pollefeys, M. (2006). A general framework for montion segmentation: independent, articulated, rigid, non-rigid, degenerate and nondegenerate. In *ECCV*.
Yu, Y., & Schuurmans, D. (2011). Rank/norm regularisation with closed-form solution: application to subspace clustering. In *UAI*.