

Robust subspace clustering via penalized mixture of Gaussians



Jing Yao¹, Xiangyong Cao¹, Qian Zhao*, Deyu Meng, Zongben Xu

School of Mathematics and Statistics and Ministry of Education Key Lab of Intelligent Networks and Network Security, Xi'an Jiaotong University, Xian 710049, PR China

ARTICLE INFO

Article history:

Received 12 October 2016

Revised 5 April 2017

Accepted 21 May 2017

Available online 1 September 2017

Keywords:

Subspace clustering

Low-rank representation

Mixture of Gaussians

Expectation maximization

ABSTRACT

Many problems in computer vision and pattern recognition can be posed as learning low-dimensional subspace structures from high-dimensional data. Subspace clustering represents a commonly utilized subspace learning strategy. The existing subspace clustering models mainly adopt a deterministic loss function to describe a certain noise type between an observed data matrix and its self-expressed form. However, the noises embedded in practical high-dimensional data are generally non-Gaussian and have much more complex structures. To address this issue, this paper proposes a robust subspace clustering model by embedding the Mixture of Gaussians (MoG) noise modeling strategy into the low-rank representation (LRR) subspace clustering model. The proposed MoG-LRR model is capitalized on its adapting to a wider range of noise distributions beyond current methods due to the universal approximation capability of MoG. Additionally, a penalized likelihood method is encoded into this model to facilitate selecting the number of mixture components automatically. A modified Expectation Maximization (EM) algorithm is also designed to infer the parameters involved in the proposed PMoG-LRR model. The superiority of our method is demonstrated by extensive experiments on face clustering and motion segmentation datasets.

© 2017 Elsevier B.V. All rights reserved.

1. Introduction

With dramatic development of techniques on data collection and feature extraction in recent years, the data obtained from real applications are often with high dimensionality. Such high-dimensional characteristic of data not only tends to bring large computation burden to the later data processing implementation, but also might possibly degenerates the performance of the utilized data process technique due to the curse of dimensionality issue.

An effective strategy to alleviate this problem is to find the intrinsic low-dimensional subspace where such high-dimensional data intrinsically reside, and then make implementations on the low-dimensional projections of data. This is always feasible in real scenarios since features of practically collected data are generally with evident correlations. Such a “subspace learning” methodology has attracted much attention in the past decades and various related methods have been proposed for different machine learning, computer vision and pattern recognition tasks [1–3].

In the recent years, a new trend on subspace learning, called subspace clustering, has appeared by simultaneously clustering data and extracting multiple subspaces, each corresponding to one data cluster [4–6]. Compared with traditional methods which assume data lie on a unique low-dimensional subspace, such “subspace clustering” model better complies with many real scenarios where data are located on multiple subspace clusters. Typical applications include face clustering [7], image segmentation [8], metric learning [9], feature grouping [10] and image representation [11]. Accordingly this research has been attracting increasing attention in the recent years.

Now let's introduce the formal definition for the subspace clustering problem as below [4]:

Definition 1.1 (Subspace Clustering). Given a set of sampled data $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_k] = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$ drawn from a union of k subspaces $\{\mathcal{S}_i\}_{i=1}^k$, where $\mathbf{X}_i$ denotes a collection of n_i samples drawn from the subspace $\mathcal{S}_i$, and $n = \sum_{i=1}^k n_i$. The task of subspace clustering is to cluster the samples according to the underlying subspaces they are drawn from.

The main assumption underlying the subspace clustering is that each datum is sampled from one of several low-dimensional subspaces, and hence can be well represented as a linear combination of the other data from the same subspace. The representation matrix composed of all coefficients of such combinations is

* Corresponding author.

E-mail addresses: jasonyao92@gmail.com (J. Yao), caoxiangyong45@gmail.com (X. Cao), timmy.zhaoqian@mail.xjtu.edu.cn, timmy.zhaoqian@gmail.com (Q. Zhao), dymeng@mail.xjtu.edu.cn (D. Meng), zbxu@mail.xjtu.edu.cn (Z. Xu).

¹ 1 indicates equal contribution

firstly taken as the encoding representation for the intrinsic subspace clusters and then is utilized to extract subspace knowledge from data. This subspace learning task is mathematically formulated as follows:

$$\min_{\mathbf{C}} \mathcal{R}(\mathbf{C}) + \lambda \mathcal{L}(\mathbf{X} - \tilde{\mathbf{X}}\mathbf{C}), \quad (1)$$

where $\tilde{\mathbf{X}}$ is a predefined dictionary, the first term regularize the representation matrix $\mathbf{C}$ to encode the subspace clustering prior knowledge in it, the second term is the loss function to fit the additive noise and λ is a positive trade-off parameter.

A natural choice for the loss function $\mathcal{L}(\cdot)$ in Eq. (1) is the $\|\cdot\|_F$ norm, which, from the probabilistic perspective, mainly characterize Gaussian noise. However, real noises in applications are generally non-Gaussian, and thus other types of loss functions were considered, such as $\|\cdot\|_1$ norm and $\|\cdot\|_{2,1}$ norm², corresponding to Laplacian noise and sample specific Gaussian noise, respectively [12]. These noise assumptions, however, still have limitations, since data noise in practical applications often exhibits much more complex statistical structures [13–16]. Therefore, a pre-fixed simple loss term is generally incapable of well fitting practical noise in data. The clustering accuracy of the utilized subspace clustering method [6] thus tends to be negatively influenced by this improper assumption. In this sense, it's crucial to propose a robust subspace clustering model to tackle complex noise.

To address this issue, this paper presents a novel subspace clustering method that is robust against a wider range of noise distributions beyond traditional Gaussian, Laplacian or sample Gaussian noises. Our basic idea is to encode data noise as Mixture of Gaussians (MoG) in the subspace clustering model, and simultaneously extract subspace cluster knowledge and adapt data noise. Here we prefer to employ MoG as the noise model due to its universal approximation property to any continuous distribution [17]. Such idea is inspired by some recent noise modeling methodology designed on some typical machine learning and computer vision tasks, such as low-rank matrix factorization (LRMF) [18], robust principle component analysis (RPCA) [19] and tensor factorization [20], and has been substantiated to be effective in the complicated noise scenarios. There exist multiple subspace clustering approaches recently, such as Sparse Subspace Clustering (SSC) [5], Low-Rank Representative (LRR) [6] and Least Square Regression (LSR) [21]. In this work we readily adopt the LRR model [6,22] due to its utilization of clean data as the dictionary matrix, instead of taking original corrupted ones as most others, which better complies with the insight of subspace clustering. In addition, regarding the automatic selection of mixture Gaussian components, we propose a penalized MoG-LRR model through adopting a penalized likelihood technique inspired by [23,24]. We further design an effective Expectation Maximization (EM) algorithm to infer all parameters involved in this model. The superiority of the proposed method is substantiated on face clustering and motion segmentation problems as compared with the current state-of-the-art methods on subspace clustering.

Specifically, the contribution of this work can be summarized as follows:

- On subspace clustering, we integrated MoG noise modeling methodology into the LRR model, which enhances a robust subspace clustering strategy with capability of adaptively fitting a wide range of data noises beyond current methods.
- On MoG noise modeling methodology, through employing the penalized likelihood technique, the EM algorithm designed on the proposed MoG-LRR model is capable of automatically selecting a proper Gaussian mixture component number as well

Table 1

Some utilized notations in this paper.

Notation	Definition
k	Number of subspaces
N	Data size
D	Data dimensionality
$\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_N]$	Observed data
$\mathbf{A} = [\mathbf{a}_1, \dots, \mathbf{a}_N]$	Clean data
$\mathbf{C} = [\mathbf{c}_1, \dots, \mathbf{c}_N]$	Representation matrix
$\mathbf{E} = [\mathbf{e}_1, \dots, \mathbf{e}_N]$	Sample-wise noise
$\boldsymbol{\pi} = \{\pi_1, \dots, \pi_K\}$	Mixture proportions
$\boldsymbol{\Sigma} = \{\boldsymbol{\Sigma}_1, \dots, \boldsymbol{\Sigma}_K\}$	Covariance matrices

as other involved parameters in this model. This also prompts the frontier of noise modeling and makes it easy for the selection of this important parameter.

Our paper is organized as follows. Related work is introduced in Section 2. Section 3 proposes our model called *Penalized Mixture of Gaussians Low-Rank Representation* (PMoG-LRR) and then presents a modified EM algorithm for solving this model. Section 4 presents experimental results implemented on synthetic and real data sets to substantiate the superiority of our proposed method over other state-of-the-arts. Finally, a brief conclusion is drawn in Section 5. Throughout the paper, we denote scalars, vectors, matrices as the non-bold, bold lower case and bold upper case letters, respectively. Some notations used in this paper are summarized in Table 1.

2. Related work

The past two decades has witnessed a rapid development in the field of subspace clustering. The related methods can be roughly classified into four categories: algebraic methods, iterative methods, statistical methods, and spectral-clustering-based methods [4].

Algebraic methods, typically represented by Matrix Factorization-based methods [25–27], first find a permutation matrix and calculate the multiplication matrix of data and the permutation matrix, and then factorize this multiplication matrix into two rank- r matrices, a base-matrix and a block diagonal matrix, respectively. But these methods are generally sensitive to noise and require the knowledge of the rank r of data matrix. The iterative methods, e.g., K-subspaces [28] use an iterative way to model and segment data. Specifically, such methods first assign data to pre-defined multiple subspaces, and then update the subspaces and reassign each data point to the closest subspace. The drawback of above two methods is that they incline to be sensitive to initialization and outliers. Besides, they need to know the number of subspace and their corresponding dimensions in advance. The statistical methods, e.g., Mixtures of Probabilistic PCA (MPPCA) [29], assume that the sample data are generated from a Mixture of Gaussians (MoG) distribution and then uses the Expectation Maximization (EM) algorithm to update the data segmentation and model parameters alternatively under the Maximum Likelihood Estimation (MLE) framework. One disadvantage of these methods is that the model always cannot fit the cases that the intrinsic distributions of the data inside each subspace are not Gaussian.

Recently, spectral-clustering-based subspace clustering methods has been attracting more attention [5,6,21] due to its rational methodology and successful performance in applications [30]. The fundament of these methods is to assume that each data point can be linearly represented by all the other data points from the same subspace cluster. These methods generally contain two steps. Firstly, an affinity matrix is built to capture the similarity between pairs of data points. Then, the segmentation of data is obtained by applying spectral clustering algorithm [31] to the affinity matrix.

² The three norms are calculated as $\|\cdot\|_F = \sqrt{\sum_{j=1}^N (\sum_{i=1}^D (\cdot)_{ij}^2)}$, $\|\cdot\|_1 = \sum_{j=1}^N (\sum_{i=1}^D |(\cdot)_{ij}|)$ and $\|\cdot\|_{2,1} = \sum_{j=1}^N (\sum_{i=1}^D (\cdot)_{ij}^2)^{\frac{1}{2}}$, respectively.

Since this methodology is the main focus of this work, we give a more detailed introduction to its recent developments as follows.

Elhamifar and Vidal [5] proposed the Sparse Subspace Clustering (SSC) to find the sparsest representation of each data point by utilizing the l_1 norm regularization term. Low-Rank Representation (LRR) [6] aims to get a low rank representation for robust subspace recovery of the data containing corruptions by imposing the nuclear norm on the representative matrix. Besides, based on the l_2 norm, Least Squares Regression (LSR) [21] was proposed for subspace clustering under Enforced Block Diagonal (EBD) conditions. Low-rank subspace clustering (LRSC) [22] extends both methods and proposes an alternative non-convex formulation, which can be solved efficiently by a closed-form solution when there is no outliers.

To improve clustering accuracy, multiple amelioration techniques on subspace learning, including Bayesian method [32], quadratic programming [33], manifold regularization [12] and Markov random walks [34], have also been recently attempted to further ameliorate the capability of subspace clustering. Here, we list some typical instances as follows: Based on block-diagonal structure prior on ideal representation matrix, Feng et al. [35] attempt to explicitly pursue such a block diagonal structure by proposing a graph Laplacian constraint based formulation. In addition, to handle more complex noise, many robust subspace clustering methods have been studied recently. Both Lu et al. [36] and He et al. [37] use correntropy as a robust measure to suppress the influence of large corruptions and make subspace clustering model be robust to non-Gaussian noises. The MoG noise modeling strategy was firstly considered in LSR subspace clustering model in [13]. However, this strategy is inferior in that it needs to empirically while not automatically tune the number of mixture Gaussian components, and besides, in the method the representation matrix between data is constructed on the corrupted data while not clean ones, which tends to make it still sensitive in the presence of noises or outliers, as substantiated by our experiments in Section 4.

3. Robust subspace clustering by penalized mixture of Gaussians

In this section, we first briefly introduce the LRR subspace clustering model [6]. Then we propose our PMoG-LRR model. Finally, we design an efficient EM algorithm to solve this model.

3.1. LRR model revisit

To better illustrate the main idea of our method in the context of LRR, we briefly review the basic idea of LRR in this subsection.

Let's consider the case that data $\mathbf{X}$ are drawn from a union of multiple subspaces given by $\cup_{i=1}^k S_i$, where $S_1, S_2, \dots, S_k$ represent the low-dimensional subspaces of underlying data. The traditional LRR model can be formulated as the following rank minimization problem:

$$\min_{\mathbf{C}, \mathbf{E}} \text{rank}(\mathbf{C}) + \lambda \|\mathbf{E}\|_p \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XC} + \mathbf{E}, \quad (2)$$

where the constraint means a sample can be linearly represented by the ones located in the same cluster with it, $\mathbf{E}$ indicates the noises embedded in data, $\|\cdot\|_p$ denotes a certain loss function, like L_2 -norm, L_1 -norm and $L_{2,1}$ -norm losses, and the parameter $\lambda > 0$ is the compromising parameter between two objective terms. Unfortunately, (2) is both non-smooth and non-convex, and difficult to solve. Generally, a tractable optimization problem is utilized by replacing the rank term with the nuclear norm $\|\mathbf{C}\|_* = \sum_i \sigma_i(\mathbf{C})$, where $\sigma_i(\cdot)$ denotes the i th singular value of a certain matrix.

Then the first term of the LRR model turns to be a convex component as follows:

$$\min_{\mathbf{C}, \mathbf{E}} \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_p \quad \text{s.t.} \quad \mathbf{X} = \mathbf{XC} + \mathbf{E}. \quad (3)$$

Note that in traditional LRR as well as most other subspace clustering models, the observed data $\mathbf{X}$ is directly used as the basis matrix in the model. This is actually improper since the expected basis matrix should be the clean matrix without noises underlying $\mathbf{X}$. An ameliorated version of LRR is thus recently proposed in [22] to deal with this issue. The modified LRR model is

$$\min_{\mathbf{A}, \mathbf{C}, \mathbf{E}} \|\mathbf{C}\|_* + \lambda \|\mathbf{E}\|_p \quad \text{s.t.} \quad \mathbf{A} = \mathbf{AC}, \quad \mathbf{X} = \mathbf{A} + \mathbf{E}, \quad (4)$$

where $\mathbf{A}$ represents the clean data matrix underlying $\mathbf{X}$. It is evident that the new model better formulates the subspace representation correlation, and we thus prefer to employ this model as the baseline in our work.

3.2. Formulation of PMoG-LRR model

In this subsection, we will introduce our PMoG-LRR model. We utilize the MoG distribution to model the real noise, which constructs a universal approximator to any continuous density function in theory [17]. Specifically, we assume that each column $\mathbf{e}_n$ ($n = 1, 2, \dots, N$) of $\mathbf{E}$ in (3) follows an MoG, i.e.,

$$\mathbb{P}(\mathbf{e}_n) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \Sigma_k), \quad (5)$$

where π_k is the mixing proportion with $\pi_k \geq 0$ and $\sum_{k=1}^K \pi_k = 1$, K is the number of the mixture components and $\mathcal{N}(\mathbf{e}_n | \mathbf{0}, \Sigma_k)$ denotes the k th Gaussian component with zero-mean and covariance matrix Σ_k . To reduce the possible over-fitting issue, each covariance matrix Σ_k is assumed to be a diagonal matrix and is represented as $\Sigma_k = \text{diag}(\sigma_{kj}^2)$, $j = 1, 2, \dots, D$. In addition, we denote $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_K)^\top$ and $\boldsymbol{\Sigma} = (\Sigma_1, \Sigma_2, \dots, \Sigma_K)$ and assume each column of $\mathbf{E}$ is independent and identically distributed (i.i.d). Then, the likelihood function can be written as

$$\mathbb{P}(\mathbf{E}; \boldsymbol{\pi}, \boldsymbol{\Sigma}) = \prod_{n=1}^N \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \Sigma_k), \quad (6)$$

and the negative log-likelihood function is

$$-\log \mathbb{P}(\mathbf{E}; \boldsymbol{\pi}, \boldsymbol{\Sigma}) = -\sum_{n=1}^N \log \left[\sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \Sigma_k) \right]. \quad (7)$$

As aforementioned, inspired by [22], given the corrupted data matrix $\mathbf{X}$, instead of utilizing this matrix as the basis matrix, we directly employ the clean data $\mathbf{A}$ underlying $\mathbf{X}$, i.e., $\mathbf{A} = \mathbf{AC}$. The relationship between two matrices is: $\mathbf{X} = \mathbf{A} + \mathbf{E}$.

In our model, we further consider the crucial issue of how to automatically select the number of mixture components K . Various model selection techniques can be readily employed to resolve this issue. Most conventional methods are based on the likelihood function and some information theoretic criteria, such as AIC and BIC. Here we adopt a penalized likelihood method [24], which has been proved to be empirically effective and theoretically sound, to shrink the mixing weights continuously and retain effective components in the mixture distributions. This penalty function is defined as

$$\mathbb{Q}(\boldsymbol{\pi}; \beta) = n\beta \sum_{k=1}^K p_k [\log(\epsilon + \pi_k) - \log(\epsilon)], \quad (8)$$

where n is the number of columns of the data matrix, namely the number of the data, ϵ is a very small positive number, β is a tuning parameter ($\beta \geq 0$), and p_k is the number of free parameters for the k th component.

We can then formulate the PMoG-LRR model as:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{C}, \mathbf{E}, \boldsymbol{\pi}, \boldsymbol{\Sigma}} \quad & \|\mathbf{C}\|_* - \frac{\lambda}{2} [\log \mathbb{P}(\mathbf{E}; \boldsymbol{\pi}, \boldsymbol{\Sigma}) - \mathcal{Q}(\boldsymbol{\pi}; \beta)] \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{AC}, \quad \mathbf{X} = \mathbf{A} + \mathbf{E}, \\ & \pi_k \geq 0, \quad \sum_{k=1}^K \pi_k = 1, \\ & \boldsymbol{\Sigma}_k = \text{diag}(\sigma_{kj}^2) \in \mathcal{S}^+, \quad k = 1, \dots, K, \end{aligned} \quad (9)$$

where $\mathcal{S}^+$ represents the set of positive definite matrices.

3.3. Modified EM Algorithm for solving PMoG-LRR model

An elegant and powerful way to solve (9) is the Expectation Maximization (EM) algorithm, which aims to find the maximum-likelihood estimation of the parameters iteratively. In this subsection, we develop a modified EM algorithm for solving the proposed model.

In the E step, we compute the responsibility parameters γ_{nk} based on the current parameters $\boldsymbol{\Theta}^{(t)} = \{\mathbf{X}^{(t)}, \mathbf{A}^{(t)}, K^{(t)}, \boldsymbol{\pi}^{(t)}, \boldsymbol{\Sigma}^{(t)}\}$:

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \boldsymbol{\Sigma}_k)}{\sum_{l=1}^K \pi_l \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \boldsymbol{\Sigma}_l)}, \quad (10)$$

where $\mathbf{e}_n = \mathbf{x}_n - \mathbf{a}_n$, $\mathbf{x}_n$ and $\mathbf{a}_n$ denote the n th columns of $\mathbf{X}$ and $\mathbf{A}$, respectively, and the superscript denotes the iteration number. Based on the responsibility so obtained, we can compute the Q-function $\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)})$ as

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{(t)}) = & \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} [\log \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \boldsymbol{\Sigma}_k) + \log \pi_k] \\ & - n\beta \sum_{k=1}^K p_k [\log(\epsilon + \pi_k) - \log(\epsilon)]. \end{aligned} \quad (11)$$

In the M step, we need to maximize (11) to achieve the updated MoG parameters $\boldsymbol{\Theta}^{(t+1)}$. We first cope with problem of updating $\boldsymbol{\Sigma}_k$, $k = 1, 2, \dots, K$ by optimizing the following problem:

$$\begin{aligned} \min_{\boldsymbol{\Sigma}_k} \quad & - \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \mathcal{N}(\mathbf{e}_n; \mathbf{0}, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \boldsymbol{\Sigma}_k = \text{diag}(\sigma_{kj}^2) \in \mathcal{S}^+, \quad j = 1, 2, \dots, D. \end{aligned} \quad (12)$$

By taking the first derivative of the objective function with respect to $\boldsymbol{\Sigma}_k$, we obtain

$$\boldsymbol{\Sigma}_k^{(t+1)} = \text{diag} \left(\text{diag} \left(\frac{1}{n_k} \left(\sum_{n=1}^N \gamma_{nk} \mathbf{e}_n \mathbf{e}_n^T + \eta \mathbf{I}_{D \times D} \right) \right) \right), \quad (13)$$

where $n_k = \sum_{n=1}^N \gamma_{nk}$, $\eta > 0$ is a small regularization parameter to make $\boldsymbol{\Sigma}_k^{(t+1)}$ invertible, and $\mathbf{I}_{D \times D}$ is an identity matrix with D dimensions.

To update π_k , $k = 1, \dots, K$, we need to introduce a Lagrange multiplier α to take the constraint $\sum_{k=1}^K \pi_k = 1$ into consideration, and then minimize

$$\begin{aligned} & - \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \pi_k + \alpha \left(\sum_k \pi_k - 1 \right) \\ & - n\beta D \sum_{k=1}^K [\log(\epsilon + \pi_k) - \log(\epsilon)]. \end{aligned} \quad (14)$$

Setting the derivative of (14) with respect to π_k to zero and considering that ϵ is close to zero, we obtain:

$$\pi_k^{(t+1)} = \max \left\{ 0, \frac{1}{1 - \beta KD} \left[\frac{1}{n} \sum_{n=1}^N \gamma_{nk} - \beta D \right] \right\}. \quad (15)$$

To update $\mathbf{A}$ and $\mathbf{C}$, we should solve the following sub-problem:

$$\begin{aligned} \min_{\mathbf{A}, \mathbf{C}} \quad & \|\mathbf{C}\|_* - \frac{\lambda}{2} \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \mathcal{N}(\mathbf{x}_n - \mathbf{a}_n; \mathbf{0}, \boldsymbol{\Sigma}_k) \\ \text{s.t.} \quad & \mathbf{A} = \mathbf{AC}, \end{aligned} \quad (16)$$

which can be mathematically reformulated as:

$$\min_{\tilde{\mathbf{C}}, \tilde{\mathbf{A}}} \|\tilde{\mathbf{C}}\|_* + \frac{\lambda}{2} \|\tilde{\mathbf{X}} - \tilde{\mathbf{A}}\|_F^2 \quad \text{s.t.} \quad \tilde{\mathbf{A}} = \tilde{\mathbf{A}}\tilde{\mathbf{C}}, \quad (17)$$

where $\tilde{\mathbf{X}} = [\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \dots, \tilde{\mathbf{x}}_N]$, $\tilde{\mathbf{x}}_n = \mathbf{H}_n^{\frac{1}{2}} \mathbf{x}_n$, $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1, \tilde{\mathbf{a}}_2, \dots, \tilde{\mathbf{a}}_N]$, $\tilde{\mathbf{a}}_n = \mathbf{H}_n^{\frac{1}{2}} \mathbf{a}_n$ and $\mathbf{H}_n = \sum_{k=1}^K \gamma_{nk} \boldsymbol{\Sigma}_k^{-1}$. From (17), we can obtain a closed-form solution by applying Lemma 2 and 3 in [38]. Specifically, the solution to (17) is

$$\tilde{\mathbf{A}} = \mathbf{U}_r \mathbf{S}_r \mathbf{V}_r^T, \quad \mathbf{C}^{(t+1)} = \mathbf{V}_r \mathbf{V}_r^T, \quad (18)$$

where $\mathbf{S}_r$ represents the top $r = \arg \min(k + \frac{\lambda}{2} \sum_{i>k} \sigma_i^2)$ singular values, $\mathbf{U}_r$ and $\mathbf{V}_r$ are the matrices composed by the corresponding left and right singular vectors of $\tilde{\mathbf{X}}$, respectively. It should be noted that here in the limiting case, r can also be equivalently determined by thresholding the singular value σ_i at $\sqrt{2/\lambda}$. Then, we can update $\mathbf{A}$ by

$$\mathbf{a}_n^{(t+1)} = \mathbf{H}_n^{-\frac{1}{2}} \tilde{\mathbf{a}}_n, \quad n = 1, 2, \dots, N. \quad (19)$$

The procedure of the proposed EM algorithm is summarized in Algorithm 1.

Algorithm 1 EM Algorithm for PMoG-LRR.

Input:

Data matrix $\mathbf{X}$, initial model parameters $\boldsymbol{\Theta}^{(0)} = \{\text{mixture parameters } \boldsymbol{\pi}^{(0)} \text{ and } \boldsymbol{\Sigma}^{(0)}, \text{initial component number } K^{(0)}, \text{clean data matrix } \mathbf{A}^{(0)}, \text{representation matrix } \mathbf{C}^{(0)}\}$ and $t = 1$.

Output:

The representation matrix $\mathbf{C}$ and the clean data matrix $\mathbf{A}$.

- 1: **repeat**
 - 2: (E step): Update $\boldsymbol{\gamma}^{(t)}$ via (10).
 - 3: (M step for $\boldsymbol{\Sigma}$): Update $\boldsymbol{\Sigma}^{(t)}$ via (13).
 - 4: (M step for $\boldsymbol{\pi}$): Update $\boldsymbol{\pi}^{(t)}$ via (15).
 - 5: (M step for $\mathbf{C}, \mathbf{A}$): Update $\mathbf{C}^{(t)}$ and $\mathbf{A}^{(t)}$ via (18) and (19), respectively.
 - 6: $t \leftarrow t + 1$.
 - 7: **until** converge;
-

3.4. Implementation details

In our experiments, the compromising parameter λ between regularization term and penalized likelihood term is determined by cross validation. Additionally, we provide a series of tuning parameter β with alternative values. Then the Bayesian Information Criterion (BIC) or the Schwarz criterion [39] is used to select the β with largest BIC value. We set initialized mixture components $K^{(0)}$ as 5 throughout our experiments. For each component, we sample from a standard normal distribution to initialize the diagonal elements of covariance matrix $\boldsymbol{\Sigma}^{(0)}$. We use $\mathbf{U}_0 \mathbf{D}_0 \mathbf{V}_0$ to initialize $\mathbf{A}^{(0)}$, where $\mathbf{D}_0$ is the non-zero part of singular values of $\mathbf{X}$, $\mathbf{U}_0$ and $\mathbf{V}_0$ are matrices composed by the corresponding left and right singular vectors, respectively. Finally, based on the theory in [26,40], suppose the skinny SVD decomposition of $\mathbf{A}^{(0)}$ is $\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}$, and then $\mathbf{C}^{(0)}$ is initialized with $\mathbf{V}\mathbf{V}^T$.

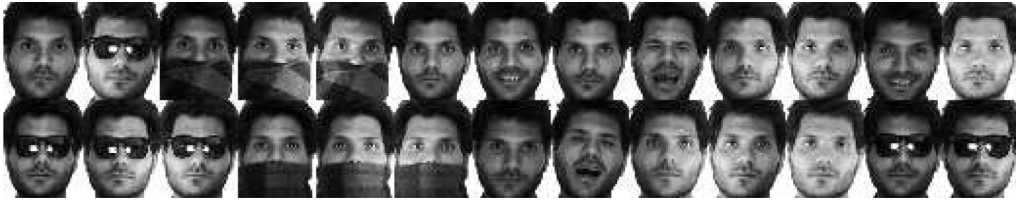


Fig. 1. The face images of the fifth subject from the AR database.

3.5. Convergence

We proposed an EM type algorithm for solving optimization problem (9). During the iterations, each involved subproblem is convex and has global solution. Therefore, by virtue of the general theory of EM algorithm [41], we can conclude that the objective function of (9) monotonically decreases. Consequently, the proposed algorithm converge in the sense of objective value.

4. Experimental results

To evaluate the performance of our proposed PMoG-LRR method, we conducted a series of experiments on two applications of subspace clustering: face clustering and motion segmentation. Seven subspace clustering methods were considered for comparison, including local subspace affinity (LSA) [42], SSC [5], LRR [6], LSR [21], block-diagonal LRR (BD-LRR) [35], Low-Rank Subspace Clustering (LRSC) [22] and MoGR [13]. These methods represent the current state-of-the-art along this research line.

We utilize two existing metrics for quantitative evaluation. Let l_i and g_i be the clustered label and ground truth label of the i th sample, respectively, and then clustering accuracy [5] is defined as $Accuracy = \frac{\sum_{i=1}^N \mathbf{1}(g_i, \text{map}(l_i))}{N} \times 100$, where $\text{map}(l_i)$ denotes the best map from obtained label l_i to equivalent label in ground truth and $\mathbf{1}(a, b)$ is the indicator function that equals to one if $a = b$ and zero otherwise.

For two clusters C and C' , representing ground truth and calculated clustering results, respectively, we can compute their mutual information (MI) as: $MI(C, C') = \sum_{c_i \in C, c'_j \in C'} p(c_i, c'_j) \log_2 \frac{p(c_i, c'_j)}{p(c_i)p(c'_j)}$, where $p(c_i)$ and $p(c'_j)$ denote the probabilities that an arbitrary data point belongs to the clusters c_i and c'_j , respectively, and $p(c_i, c'_j)$ denotes the joint probability that an arbitrary data point belongs to the clusters c_i and c'_j at the same time. $MI(C, C')$ can be used to determine how similar the clusters C and C' are. It takes values between zero and $\max(H(C), H(C'))$, where $H(\cdot)$ denotes the entropy. So we normalize the mutual information to simplify the comparison between different pairs of cluster sets as $NMI(C, C') = \frac{MI(C, C')}{\max(H(C), H(C'))}$, which equals to one when two sets of clusters are identical and zero when the two clusters have not shared even one data point.

All parameters involved in the competing methods were empirically tuned or specified as suggested by the related literatures to guarantee their possibly good performance. Some experimental results of existing methods reported in related papers have been directly utilized. All experiments are carried out on a Windows system with 3.6GHz Intel Core i7 processor using MATLAB2014b.

4.1. Face clustering

Face clustering is to cluster face images collected from multiple individuals based on their identities. The AR Face Database³

[43] and Extended Yale Dataset B⁴ [44] are adopted as the benchmark datasets for this task.

4.1.1. AR face dataset

The AR Face Database contains over 4,000 facial images from 126 subjects (70 men and 56 women). For each subject, 26 facial images are taken in two separate parts for training and testing, respectively. These images suffer different facial variations, including various facial expressions (neutral, smile, anger, and scream), illumination variations (left light on, right light on, and all side lights on), and occlusion by sunglasses or scarf. Fig. 1 shows some typical samples from the dataset for demonstration.

We select images of first $c = \{5, 8, 10\}$ subjects and apply all competing methods to implementing them. We also adopt dimension reduction procedure with standard PCA in this scenario. The clustering results including accuracy and NMI are shown in Table 2. Fig. 2 demonstrates the corresponding affinity matrix constructed from different algorithms.

From Table 2, it can be easily seen that our method performs better than other competing methods in both the clustering accuracy and NMI, and takes evidently less time than most of the competing methods, while only unsubstantially slower than SSC, LSR and LRSC. This shows that our method can perform robust in presence of noise or outliers. The superiority of the proposed method can also be verified visually by observing Fig. 2. It is easy to see that the affinity matrices obtained by LRR, BD-LRR, LRSC, MoGR and our method have more evidently depicted the configuration of subspace clusters as compared with those obtained by LSA and SSC, which lack of within-cluster connectivities. It can also be observed that the contrast between within-cluster and inter-cluster relationships have been significantly magnified by MoGR and PMoG-LRR, which is attributed to the better modeling of noise. Besides, those clusters are not perfectly detected by MoGR, e.g., cluster 2 and 6, have been more clearly recognized by PMoG-LRR, with evidently larger representation coefficients. This better visual result, as compared with those of other methods, indicates that our method has better ability of clustering the correlated data located in the similar subspace.

4.1.2. Extended Yale dataset

The Extended Yale Database B consists of 2,414 frontal face images of 38 subjects, where there are 64 faces for each subject, acquired under different lighting, poses, and illumination conditions. Each image is cropped into 192×168 pixels. Fig. 3 shows some typical samples from this dataset.

Following the experimental setting in [5,35], we build four types of tasks according to the following scheme. First, 38 subjects are separated into four groups: subjects 1 to 10, subjects 11 to 20, subjects 21 to 30, and subjects 31 to 38. Then we consider choice of $c = \{2, 5, 8, 10\}$ subjects for each of the first three groups and $c = \{2, 5, 8\}$ for the last group. We implement all the subspace clustering algorithm on each set of c subjects.

To reduce the computational cost, each face image is resized to a resolution of 48×42 pixels. As shown in [45], the faces with

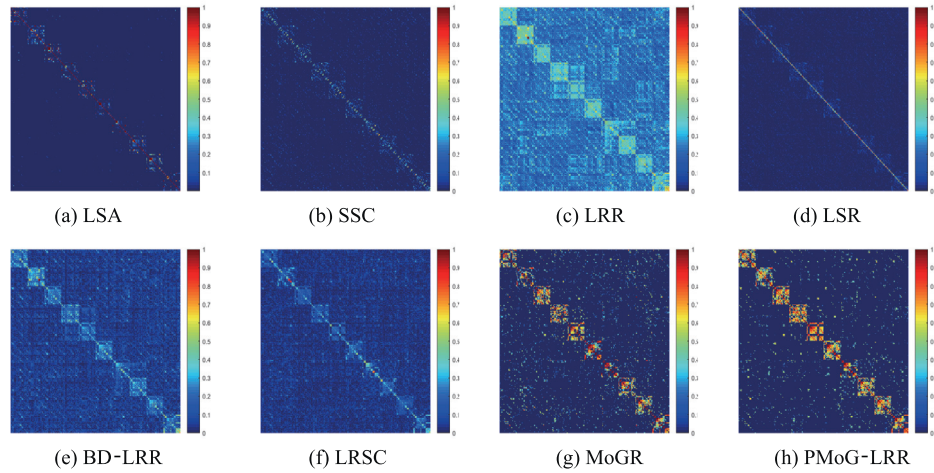
³ <http://www-sipl.technion.ac.il/new/DataBases/Aleix%20Face%20Database.htm>.

⁴ <http://vision.ucsd.edu/leekc/ExtYaleDatabase/ExtYaleB.html>.

Table 2

The average clustering accuracy (%), NMI (%) and running time (s) on the AR database for all competing methods.

Method	LSA	SSC	LRR	LSR	BD-LRR	LRSC	MoGR	PMoG-LRR
5 Subjects								
Accuracy	73.77	83.08	84.23	93.85	86.40	93.08	93.85	95.38
NMI	74.68	82.17	82.97	86.79	83.92	83.60	91.11	92.93
Time	2.22	0.89	2.24	1.33	5.99	1.21	16.04	1.57
8 Subjects								
Accuracy	66.78	75.00	79.81	80.77	83.21	90.38	90.38	94.23
NMI	69.79	75.54	78.37	83.70	80.38	85.78	88.64	92.22
Time	5.62	1.05	4.54	1.67	14.17	1.50	34.70	1.92
10 Subjects								
Accuracy	68.31	77.69	84.23	79.23	85.68	92.69	88.85	93.46
NMI	72.47	78.74	83.38	81.15	85.31	89.81	87.96	91.11
Time	8.90	1.25	6.79	2.19	22.16	2.07	111.13	4.67

**Fig. 2.** The affinity matrices of 10 obtained by all competing methods from the AR database.**Fig. 3.** Some face samples of one subject from the YALE B database.

different types of diffusion or background illumination fall into a 4-dimensional subspace. So before the vectorized image data are inputted into our algorithm, we also resort to standard PCA and project our original vectorized data onto $4 \times c$ dimensional subspace. In order to make a fair comparison, the post-processing procedure is not performed in this scenario. The detailed results are presented in Table 3.

From Table 3, it is easy to observe that PMoG-LRR achieves the best clustering accuracy in almost all cases except the 2 subject case, where it performs the second best, slightly worse than BD-LRR in accuracy Mean and Median. Besides, compared with the traditional LRR, our method obtains at least 5% increase of the clustering accuracy. This substantiates the superiority of the proposed method in such face modeling scenarios.

4.2. Motion segmentation

Motion segmentation aims to segment the trajectories associated with n different moving objects into different groups

according to their motions in a video sequence. The Hopkins 155 dataset⁵[46] is one of the most commonly utilized benchmark databases to evaluate performance of subspace clustering algorithms. It consists of 155 video sequences, where 120 ones contain 2 motions and 35 ones have 3 motions. For each sequence, a tracker is used to extract feature trajectories and outliers are removed manually. Some examples with extracted features are shown in Fig. 4. As traditional pre-processing procedures for this task, we adopt PCA to project the original data into a 12-dimensional subspace according to [5]. The procedure of subspace clustering is performed on the trajectory spatial coordinates of each sequence. The results of the clustering results using different methods are shown in Table 4.

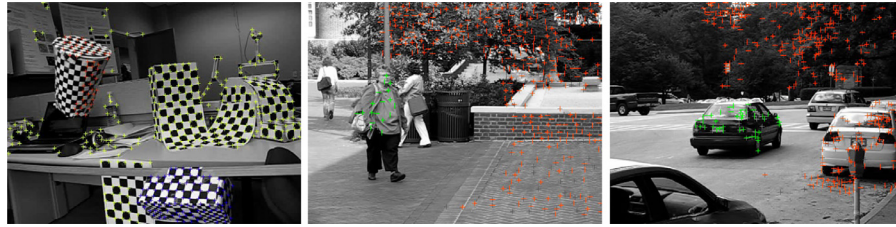
From this table, we can observe that PMoG-LRR achieves the highest clustering accuracy in most cases except slightly worse than LSR and BD-LRR in terms of accuracy Mean. For all 155 video sequences, our method achieves better clustering accuracy in Min, Median and Std., except the second best in Mean. Note that although MoGR has also considered MoG noise modeling strategy in LSR model, it still cannot perform as robust as the proposed method. This is possibly because they compute the coefficient matrix on the basis of the original corrupted data while not the clean one, which possibly degenerates its performance. Also we should note that compared with the baseline LRR method, our method achieves an evident gain in clustering accuracy (about 3.5%) for all 155 video sequences, indicating that MoG noise modeling significantly enhances the capability of LRR to make it perform robust to complex scenarios and fit wider range of noises.

⁵ <http://www.vision.jhu.edu/data/hopkins155/>.

Table 3

The clustering accuracies (%) on the Extended Yale Dataset B for all competing methods.

Method	LSA	SSC	LRR	LSR	BD-LRR	LRSC	MoGR	PMoG-LRR
2 Subjects								
Min	64.09	86.14	85.63	76.73	90.01	86.08	79.48	90.17
Mean	68.92	89.43	88.14	79.66	93.19	90.43	84.71	93.10
Median	70.04	90.10	88.91	80.63	92.48	90.69	85.09	92.24
5 Subjects								
Min	52.18	74.89	80.77	72.08	82.08	79.31	72.93	85.17
Mean	55.67	81.59	85.01	76.91	86.03	84.76	78.24	90.17
Median	54.93	82.19	85.13	76.17	85.73	85.69	80.34	87.93
8 Subjects								
Min	39.87	64.11	62.30	68.09	70.94	64.08	67.71	74.35
Mean	42.05	68.09	68.76	71.03	73.00	70.01	70.38	76.29
Median	43.11	68.62	69.32	70.14	73.34	70.49	69.56	76.08
10 Subjects								
Min	35.87	49.31	57.06	56.14	62.08	60.15	61.66	63.24
Mean	37.69	59.08	61.28	61.04	64.49	62.47	63.08	68.66
Median	37.01	62.96	60.37	59.89	64.77	62.01	63.85	70.00

**Fig. 4.** Some samples with the extracted features from the Hopkins 155 dataset.**Table 4**

The clustering accuracies (%) on the Hopkins 155 database for all competing methods.

Method	LSA	SSC	LRR	LSR	BD-LRR	LRSC	MoGR	PMoG-LRR
2 Motions								
Min	59.10	51.10	59.70	63.64	75.07	69.47	63.40	78.36
Mean	93.20	96.30	96.80	99.60	99.31	91.36	96.76	98.43
Median	97.20	100.00	99.70	100.00	100.00	92.43	98.34	100.00
Std.	8.00	9.70	8.20	6.80	6.47	10.45	7.16	6.02
3 Motions								
Min	53.40	55.40	58.50	65.44	73.39	65.39	60.51	75.28
Mean	83.20	88.60	92.20	93.37	96.31	87.84	95.03	96.55
Median	84.40	96.70	97.20	98.79	99.20	89.55	97.51	99.17
Std.	12.60	15.00	10.30	9.97	7.01	15.44	11.56	6.81
All								
Min	53.40	51.10	58.50	63.64	73.39	65.39	60.51	77.36
Mean	90.94	94.56	95.76	98.19	98.63	90.57	96.40	98.23
Median	95.20	100.00	99.33	99.69	100.00	90.47	98.47	100.00
Std.	10.10	11.60	8.90	8.62	6.56	12.73	10.31	5.96

5. Conclusion

In this paper, we propose a robust subspace clustering method by utilizing the MoG noise modeling strategy and the self-expressed property of clean data basis. Besides, we also adopt a penalized likelihood method to learn the mixture component automatically. Compared with the current subspace clustering methods, our method performs more robust and achieves the state-of-the-art subspace clustering performance in the real complex noise scenarios, including face clustering and motion segmentation.

In future research, we will try to extend the noise modeling methodology to more machine learning and pattern recognition tasks, e.g., tensor subspace clustering [47], and will investigate more theoretical insights under such subspace clustering framework.

References

- [1] D. Meng, Q. Zhao, Z. Xu, Improve robustness of sparse PCA by L1-norm maximization, *Pattern Recognit.* 45 (1) (2012) 487–497.
- [2] F. De La Torre, M.J. Black, A framework for robust subspace learning, *Int. J. Comput. Vision* 54 (1–3) (2003) 117–142.
- [3] Y. Li, On incremental and robust subspace learning, *Pattern Recognit.* 37 (7) (2004) 1509–1518.
- [4] R. Vidal, A tutorial on subspace clustering, *IEEE Signal Process. Mag.* 28 (2) (2010) 52–68.
- [5] E. Elhamifar, R. Vidal, Sparse subspace clustering: algorithm, theory, and applications, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (11) (2013) 2765–2781.
- [6] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, Y. Ma, Robust recovery of subspace structures by low-rank representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 35 (1) (2013) 171–184.
- [7] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, D. Kriegman, Clustering appearances of objects under varying illumination conditions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, 1–11–18.

- [8] A.Y. Yang, J. Wright, Y. Ma, S.S. Sastry, Unsupervised segmentation of natural images via lossy data compression, *Comput. Vision Image Underst.* 110 (2) (2008) 212–225.
- [9] J. Wang, Z. Deng, K.-S. Choi, Y. Jiang, X. Luo, F.-L. Chung, S. Wang, Distance metric learning for soft subspace clustering in composite kernel, *Pattern Recognit.* 52 (1) (2016) 113–134.
- [10] G. Gan, M.K.-P. Ng, Subspace clustering with automatic feature grouping, *Pattern Recognit.* 48 (11) (2015) 3703–3713.
- [11] W. Hong, J. Wright, K. Huang, Y. Ma, Multiscale hybrid linear models for lossy image representation, *IEEE Trans. Image Process.* 15 (12) (2006) 3655–3671.
- [12] J. Liu, Y. Chen, J. Zhang, Z. Xu, Enhancing low-rank subspace clustering by manifold regularization, *IEEE Trans. Image Process.* 23 (9) (2014) 4022–4030.
- [13] B. Li, Y. Zhang, Z. Lin, H. Lu, C.M.I. Center, Subspace clustering by mixture of gaussian regression, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2094–2102.
- [14] Z. Ma, A. Leijon, Bayesian estimation of beta mixture models with variational inference, *IEEE Trans. Pattern Anal. Mach. Intell.* 33 (11) (2011) 2160–2173, doi:10.1109/TPAMI.2011.63.
- [15] Z. Ma, A.E. Teschendorff, A. Leijon, Y. Qiao, H. Zhang, J. Guo, Variational Bayesian matrix factorization for bounded support data, *IEEE Trans. Pattern Anal. Mach. Intell.* 37 (4) (2015) 876–889, doi:10.1109/TPAMI.2014.2353639.
- [16] Z. Ma, J.H. Xue, A. Leijon, Z.H. Tan, Z. Yang, J. Guo, Decorrelation of neutral vector variables: theory and applications, *IEEE Trans. Neural Netw. Learn. Syst.* PP (99) (2016) 1–15, doi:10.1109/TNNLS.2016.2616445.
- [17] V. Maz'ya, G. Schmidt, On approximate approximations using Gaussian kernels, *IMA J. Numer. Anal.* 16 (1) (1996) 13–29.
- [18] D. Meng, F. Torre, Robust matrix factorization with unknown noise, *Proceedings of the International Conference on Computer Vision* (2013) 1337–1344.
- [19] Q. Zhao, D. Meng, Z. Xu, W. Zuo, L. Zhang, Robust principal component analysis with complex noise, *Proceedings of the International Conference on Machine Learning* (2014) 55–63.
- [20] X. Chen, Z. Han, Y. Wang, Q. Zhao, D. Meng, Y. Tang, Robust tensor factorization with unknown noise, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5213–5221.
- [21] C. Lu, H. Min, Z. Zhao, L. Zhu, D. Huang, S. Yan, Robust and efficient subspace segmentation via least squares regression, *Proceedings of the European Conference on Computer Vision* (2012) 347–360.
- [22] R. Vidal, P. Favaro, Low rank subspace clustering (LRSC), *Pattern Recognit. Lett.* 43 (2014) 47–61.
- [23] X. Cao, Y. Chen, Q. Zhao, D. Meng, Y. Wang, D. Wang, Z. Xu, Low-rank matrix factorization under general mixture noise distributions, in: *Proceedings of the International Conference on Computer Vision*, 2015, pp. 1493–1501.
- [24] T. Huang, H. Peng, K. Zhang, Model selection for Gaussian mixture models, *Stat. Sin.* 27 (1) (2017) 147–169.
- [25] T.E. Boult, L.G. Brown, Factorization-based segmentation of motions, *Proceedings of the IEEE Workshop on Visual Motion* (1991) 179–186.
- [26] J.P. Costeira, T. Kanade, A multibody factorization method for independently moving objects, *Int. J. Comput. Vision* 29 (3) (1998) 159–179.
- [27] C.W. Gear, Multibody grouping from motion images, *Int. J. Comput. Vision* 29 (2) (1998) 133–150.
- [28] P. Tseng, Nearest q-flat to m points, *J. Optim. Theory Appl.* 105 (1) (2000) 249–252.
- [29] M.E. Tipping, C.M. Bishop, Mixtures of probabilistic principal component analyzers, *Neural Comput.* 11 (2) (1999) 443–482.
- [30] X. Lu, Y. Wang, Y. Yuan, Graph-regularized low-rank representation for destriping of hyperspectral images, *IEEE Trans. Geosci. Remote Sens.* 51 (7) (2013) 4009–4018.
- [31] H. Lu, Z. Fu, X. Shu, Non-negative and sparse spectral clustering original research article, *Pattern Recognit.* 47 (1) (2014) 418–426.
- [32] S.D. Babacan, S. Nakajima, M. Do, Probabilistic low-rank subspace clustering, *Adv. Neural Inf. Process. Syst.* 25 (2012) 2744–2752.
- [33] S. Wang, X. Yuan, T. Yao, S. Yan, J. Shen, Efficient subspace segmentation via quadratic programming, in: *Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011, pp. 519–524.
- [34] R. Liu, Z. Lin, Z. Su, Learning Markov random walks for robust subspace clustering and estimation, *IEEE Trans. Neural Netw.* 59 (2014) 1–15.
- [35] J. Feng, Z. Lin, H. Xu, S. Yan, Robust subspace segmentation with block-diagonal prior, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 3818–3825.
- [36] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, Z. Lin, Correntropy induced L2 graph for robust subspace clustering, *Proceedings of the International Conference on Computer Vision* (2013) 1801–1808.
- [37] R. He, Y. Zhang, Z. Sun, Q. Yin, Robust subspace clustering with complex noise, *IEEE Trans. Image Process.* 24 (11) (2015) 4001–4013.
- [38] P. Favaro, R. Vidal, A. Ravichandran, A closed form solution to robust subspace estimation and clustering, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 1801–1807.
- [39] G. Schwarz, et al., Estimating the dimension of a model, *The Ann. Stat.* 6 (2) (1978) 461–464.
- [40] S. Wei, Z. Lin, Analysis and improvement of low rank representation for subspace segmentation, *arXiv:1107.1561* (2011).
- [41] C.J. Wu, On the convergence properties of the EM algorithm, *The Ann. Stat.* 11 (1) (1983) 95–103.
- [42] J. Yan, M. Pollefeys, A general framework for motion segmentation: independent, articulated, rigid, non-rigid, degenerate and non-degenerate, *Proceedings of the European Conference on Computer Vision* (2006) 94–106.
- [43] A.M. Martinez, The AR face database, CVC Technical Report #24 (1998).
- [44] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, Robust face recognition via sparse representation, *IEEE Trans. Pattern Anal. Mach. Intell.* 31 (2) (2009) 210–227.
- [45] R. Basri, D.W. Jacobs, Lambertian reflectance and linear subspaces, *IEEE Trans. on Pattern Anal. Mach. Intell.* 25 (2) (2003) 218–233.
- [46] R. Tron, R. Vidal, A benchmark for the comparison of 3-D motion segmentation algorithms, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.
- [47] Y. Fu, J. Gao, D. Tien, Z. Lin, X. Hong, Tensor LRR and sparse coding-based subspace clustering, *IEEE Trans. Neural Netw. Learn. Syst.* 27 (10) (2016) 2120–2133.



Jing Yao received the B.Sc. degree from Northwestern University, Xi'an, China, in 2014. He is currently pursuing the Ph.D. degree in Xi'an Jiaotong University. His current research interests include low-rank matrix factorization and hyperspectral image analysis.



Xiangyong Cao received the B.Sc. degree from Xi'an Jiaotong University, Xi'an, China, in 2012, where he is currently pursuing the Ph.D. degree. His current research interests include low-rank matrix analysis, and Bayesian method for machine learning.



Qian Zhao received the B.Sc. and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2009 and 2015, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2013 to 2014. He is currently a Lecturer with the School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include low-matrix/ tensor analysis, Bayesian modeling, and self-paced learning.



Deyu Meng received the B.Sc., M.Sc., and Ph.D. degrees from Xi'an Jiaotong University, Xi'an, China, in 2001, 2004, and 2008, respectively. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, from 2012 to 2014. He is currently an Associate Professor with the Institute for Information and System Sciences, School of Mathematics and Statistics, Xi'an Jiaotong University. His current research interests include principal component analysis, nonlinear dimensionality reduction, feature extraction and selection, compressed sensing, and sparse machine learning methods.



Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987. He currently serves as the Academician of the Chinese Academy of Sciences, the Chief Scientist of the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences with Xi'an Jiaotong University. His current research interests include nonlinear functional analysis and intelligent information processing. Prof. Xu was a recipient of the National Natural Science Award of China in 2007 and the winner of the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a talk at the International Congress of Mathematicians in 2010.