Subspace Clustering by Mixture of Gaussian Regression

Baohua Li[1],Ying Zhang[1], Zhouchen Lin[2,3] and Huchuan Lu[1]
[1] Dalian University of Technology
[2] Key Laboratory of Machine Perception (MOE), School of EECS, Peking University
[3] Cooperative Medianet Innovation Center, Shanghai, China

Abstract

Subspace clustering is a problem of finding a multi-subspace representation that best fits sample points drawn from a high-dimensional space. The existing clustering models generally adopt different norms to describe noise, which is equivalent to assuming that the data are corrupted by specific types of noise. In practice, however, noise is much more complex. So it is inappropriate to simply use a certain norm to model noise. Therefore, we propose Mixture of Gaussian Regression (MoG Regression) for subspace clustering by modeling noise as a Mixture of Gaussians (MoG). The MoG Regression provides an effective way to model a much broader range of noise distributions. As a result, the obtained affinity matrix is better at characterizing the structure of data in real applications. Experimental results on multiple datasets demonstrate that MoG Regression significantly outperforms state-of-the-art subspace clustering methods.

1. Introduction

Subspace clustering models high-dimensional data as samples drawn from a union of multiple low-dimensional subspaces. It has been attracting more and more attention in recent years and has found many applications in computer vision and image processing, such as image segmentation [30], motion segmentation [21], face clustering [6], and image representation and compression [24].

1.1. Related Work

A number of approaches to subspace clustering have been proposed in the past two decades. These methods can be roughly divided into four main categories: algebraic methods [14, 22, 4], iterative methods [2, 1], statistical methods [13, 18, 29], and spectral-clustering-based methods [23, 28, 5, 9, 12, 19, 10]. It should be noted that the spectral-clustering-based methods, which are based on the spectral graph theory [3], have shown excellent performance in many real applications.

Generally, the spectral-clustering-based methods consist of two steps. Firstly, an affinity matrix is built to capture the similarity between pairs of sample points. Secondly, graph cut is applied to a graph, whose vertices are the samples and whose weights are prescribed by the affinity matrix, for segmenting the sample points. Building a "good" affinity matrix is key to guarantee a good clustering result. So many subspace clustering methods focus on bringing up a good affinity matrix.

Based on the fact that each data point in a union of subspaces can be represented as a linear or affine combination of other points, the Sparse Subspace Clustering (SSC) algorithm [5] utilizes the ℓ_1-norm to find the sparsest representation of a data point, where points from the same subspace correspond to the nonzero representation coefficients. Low-Rank Representation (LRR) [9] aims to get a low rank representation for robust subspace recovery of the data containing corruptions. Least Squares Regression (LSR) [12] employs the Frobenius norm to speed up the clustering process, while still ensuring the grouping effect of the representation matrix. However, the solution to SSC may be too sparse to encode the data correlation, while both LRR and LSR may result in dense connections between clusters. To achieve a good balance between within-cluster density (which we call *grouping effect* afterwards) and between-cluster sparsity, Correlation Adaptive Subspace Segmentation (CASS) [10] adopts trace Lasso, which is adaptive to the data correlation, to regularize the representation matrix.

As pointed out by Liu et al. [9], noise which always exists in data can perturb the subspace structures, leading to unreliable subspace clustering. To recover the subspaces when the data are corrupted, SSC, LRR, LSR, and CASS employ different norms to describe different types of noise, respectively.

Given data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N) \in \mathbb{R}^{M\times N}$ with N samples in $\mathbb{R}^M$, we denote $\boldsymbol{E} \in \mathbb{R}^{M\times N}$ and $\boldsymbol{Z} \in \mathbb{R}^{N\times N}$ as the noise matrix and the representation matrix, respectively, where the entry Z_{ij} of $\boldsymbol{Z}$ measures the similarity between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. We use $\|\cdot\|_F$, $\|\cdot\|_1$, $\|\cdot\|_2$, $\|\cdot\|_{2,1}$, and $\|\cdot\|_*$ to denote Frobenius norm, the ℓ_1-norm (sum of

absolute values), the ℓ_2-norm, the $\ell_{2,1}$-norm (sum of the ℓ_2-norm of columns of a matrix), and the nuclear norm (sum of singular values), respectively. The mathematical models of existing representative subspace clustering methods are as follows.

Sparse Subspace Clustering (SSC) [5]:

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \parallel \boldsymbol{E} \parallel_1 + \lambda \parallel \boldsymbol{Z} \parallel_1 \\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \mathrm{diag}\left(\boldsymbol{Z}\right) = \boldsymbol{0}. \end{aligned}$$

Low-Rank Representation (LRR) [9]:

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \parallel \boldsymbol{E} \parallel_{2,1} + \lambda \parallel \boldsymbol{Z} \parallel_* \\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}. \end{aligned}$$

Least Squares Regression (LSR) [12]:

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \parallel \boldsymbol{E} \parallel_F^2 + \lambda \parallel \boldsymbol{Z} \parallel_F^2 \\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \mathrm{diag}\left(\boldsymbol{Z}\right) = \boldsymbol{0}. \end{aligned}$$

Correlation Adaptive Subspace Segmentation (CASS) [10]:

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \parallel \boldsymbol{E} \parallel_F^2 + \lambda \sum_{n=1}^{N} \parallel \boldsymbol{X}\mathrm{diag}\left(\boldsymbol{z}_i\right) \parallel_* \\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}. \end{aligned}$$

In the above formulations, $\boldsymbol{z}_i$ is the i-th column of $\boldsymbol{Z}$, $\mathrm{diag}\left(\boldsymbol{z}_i\right)$ is a diagonal matrix with entries of $\boldsymbol{z}_i$ on its diagonal, and $\lambda > 0$ is a parameter to balance the effects of two terms. $\parallel \boldsymbol{E} \parallel_F^2$ is utilized to model Gaussian noise, $\parallel \boldsymbol{E} \parallel_{2,1}$ is for sample-specific corruptions, and $\parallel \boldsymbol{E} \parallel_1$ is for entry-wise corruptions.

All the methods mentioned above rely on specific norms on $\boldsymbol{Z}$ and $\boldsymbol{E}$ to encourage the between-cluster sparsity and grouping effect of the representation matrix. However, they all use a simple norm for the noise term, which has a significant influence on the performance of the subspace clustering model. If the data are contaminated by noise, the subspace structures, grouping effect, and the data similarity are all likely to be corrupted. How the effect will be depends on the distribution of noise. In this situation, if we simply require the representation matrix to be sparse or block diagonal without deeply analyzing the noise, the subspaces may not be accurately recovered, leading to unsatisfactory clustering results. Unfortunately, real noise in applications often exhibits very complex statistical distributions, rather than simply being Gaussian or sparse [31]. So the noise cannot be easily described by a simple norm like the Frobenious norm, ℓ_1-norm, or $\ell_{2,1}$-norm. Therefore, how to properly model the noise is of significant importance for subspace clustering.

To address this issue, we apply a fundamental result of probability theory that almost any distribution can be well approximated by a mixture of a sufficient number of Gaussians. Namely, we utilize the mixture of Gaussian (MoG) model to describe real noise accurately, rather than assuming some specific distribution for noise. The number of Gaussians can be estimated by cross validation. As for the regularization on $\boldsymbol{Z}$, we simply choose the Frobenius norm. The reasons are two-fold. First, we want to demonstrate the effect of noise modeling on subspace clustering. So a simple regularization on $\boldsymbol{Z}$ can better expose such an effect. Second, with the Frobenious norm on $\boldsymbol{Z}$ the computation can be made much easier. For example, we can employ the traditional Expectation Maximization (EM) algorithm to solve our new subspace clustering model.

1.2. Paper Contributions and Organization

We summarize the contributions of this paper as follows:

- We present a novel subspace clustering approach called Mixture of Gaussian Regression (MoG Regression), which employs the MoG model to characterize noise with a complex distribution.
- We prove that MoG Regression has the grouping effect, which is important for subspace clustering.

The remainder of the paper is organized as follows. In Section 2, we motivate and introduce the MoG Regression method for clustering data with complex noise. In Section 3 we prove the grouping effect of the proposed model. Section 4 provides experimental results on motion segmentation, hand-written digits clustering, and complex face clustering to demonstrate the superiority of MoG Regression.

2. Subspace Clustering via MoG Regression

As described in [11], we consider subspace clustering as the following optimization problem:

$$\begin{aligned} &\min_{\boldsymbol{Z},\boldsymbol{E}} \mathcal{L}\left(\boldsymbol{E}\right) + \mathcal{R}\left(\boldsymbol{Z}\right) \\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \end{aligned} \tag{1}$$

where $\mathcal{L}\left(\boldsymbol{E}\right)$ is the loss function to describe noise and $\mathcal{R}\left(\boldsymbol{Z}\right)$ is the regularization term to impose some properties on the representation matrix $\boldsymbol{Z}$.

From (1) we see that how to describe noise has significant importance in subspace clustering. Lu et al. [11] proposed Correntropy Induced L2 (CIL2) graph, which uses correntropy to process non-Gaussian and impulsive noise for robust subspace clustering, and the effectiveness is demonstrated by experiments of face clustering under various types of corruptions and occlusions. In fact, the variation of the width of kernel function makes the behavior of Correntropy Induced Metric changes between ℓ_0, ℓ_1, and ℓ_2 norms, which is effective for many types of noise but not for general noise anyway.

2.1. MoG Regression

In this paper, we propose a novel method called MoG Regression, which employs MoG to characterize general noise for robust subspace clustering.

We assume that each column $\boldsymbol{e}_n\,(n = 1, \ldots, N)$ in $\boldsymbol{E}$ follows an MoG distribution, i.e.,

$$p(\boldsymbol{e}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k),$$

where K is the number of Gaussian components and π_k denotes the mixing weight with $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. $\mathcal{N}(\boldsymbol{e}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k)$ is the zero-mean multivariate Gaussian distribution, with $\boldsymbol{\Sigma}_k (k = 1, 2, \ldots, K)$ denoting the covariance matrix.

Similar to classical regression analysis, all columns in $\boldsymbol{E}$ are assumed to be independently and identically distributed. So we have

$$p(\boldsymbol{E}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k).$$

In a general MoG model, we wish to find $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)^\top$ and $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_K)$ that maximize $p(\boldsymbol{E})$, which is equivalent to minimizing the negative log likelihood function defined as

$$-\ln p(\boldsymbol{E}) = -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right).$$

If we utilize $\mathcal{L}(\boldsymbol{E}) = -\ln p(\boldsymbol{E})$ to replace the Frobenius norm in the LSR model, then the proposed MoG Regression method can be formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{Z}, \boldsymbol{E}, \boldsymbol{\pi}, \boldsymbol{\Sigma}} & -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right) + \lambda \parallel \boldsymbol{Z} \parallel_F^2 \\ & s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \operatorname{diag}(\boldsymbol{Z}) = \mathbf{0}, \\ & \pi_k \geq 0, \boldsymbol{\Sigma}_k \in \mathbb{S}^+, k = 1, ..., K, \sum_{k=1}^{K} \pi_k = 1, \end{aligned} \tag{2}$$

where $\lambda > 0$ is the regularization parameter, $\mathbb{S}^+$ is the set of symmetrical positive definite (SPD) matrices and the constraint $\operatorname{diag}(\boldsymbol{Z}) = \mathbf{0}$ discourages using a sample to represent itself. Here we simply choose the Frobenius norm to regularize $\boldsymbol{Z}$. As stated before, the Frobenious norm on $\boldsymbol{Z}$ can not only reduce the computation cost but also expose the the effect of MoG regression based noise modeling on subspace clustering.

A natural way to solve (2) is the EM algorithm, which finds the maximum-likelihood estimate of the parameters iteratively. It starts from an initial guess and iteratively runs an expectation (E) step, which evaluates the posterior probabilities using currently estimated parameters, and a maximization (M) step, which re-estimates the parameters based on the probabilities calculated in the E step. The iterations stop until some convergence criteria are satisfied [26, 27, 16]. Integrating the traditional steps of the EM algorithm, we can obtain the solution to problem (2) in three main steps.

First, we initialize the representation matrix $\boldsymbol{Z}$, mixing coefficients π_k, and covariance matrices $\boldsymbol{\Sigma}_k, k = 1, ..., K$.

In the E-step, we compute the posterior probabilities based on the current parameters:

$$\gamma_{n,k} = \frac{\pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_j)},$$

where $\widetilde{\boldsymbol{e}}_n = \widetilde{\boldsymbol{X}}_n \boldsymbol{z}_n - \boldsymbol{x}_n$ and $\widetilde{\boldsymbol{X}}_n$ is a copy of $\boldsymbol{X}$ except that the n-th column is $\mathbf{0}$.

In the M-step, we need to minimize our model with respect to the parameters, using the current posterior probabilities.

To find $\boldsymbol{\Sigma}_k,\ k = 1, 2, \ldots, K$, we solve the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\Sigma}_k} & -\sum_{n=1}^{N} \ln \left(\textstyle\sum_{k=1}^{K} \pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right) \\ & s.t.\ \ \boldsymbol{\Sigma}_k \in \mathbb{S}^+. \end{aligned}$$

Setting the derivative of the objective function with respect to $\boldsymbol{\Sigma}_k$ to zero, we obtain

$$\boldsymbol{\Sigma}_k = \frac{1}{\gamma_{n,k}} \left(\sum_{n=1}^{N} \gamma_{n,k} \widetilde{\boldsymbol{e}}_n \cdot \widetilde{\boldsymbol{e}}_n^\top + \epsilon \boldsymbol{I} \right),$$

where $\epsilon > 0$ is a small regularization parameter to ensure that $\boldsymbol{\Sigma}_k$ is invertible.

Each $\pi_k, k = 1, 2, \ldots, K$, is updated by solving

$$\min_{\pi_k \geq 0} -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right) + \beta \left(\sum_{k=1}^{K} \pi_k - 1 \right),$$

where $\beta > 0$ is the Lagrangian multiplier. We find $\beta = N$ and accordingly

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,k}.$$

Each column of $\boldsymbol{Z}$ is updated by solving the following problem:

$$\min_{\boldsymbol{z}_n} -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \right) + \lambda \parallel \boldsymbol{Z} \parallel_F^2.$$

By setting the derivative of object function with respect to $\boldsymbol{z}_n$ to zero, we get

$$\boldsymbol{z}_n = \left(\frac{\sum_{k=1}^{K} \pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k) \widetilde{\boldsymbol{X}}_n^\top \boldsymbol{\Sigma}_k^{-1} \widetilde{\boldsymbol{X}}_n}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_j)} + 2\lambda \boldsymbol{I} \right)^{-1} \boldsymbol{b}_n,$$

where

$$\boldsymbol{b}_n = \frac{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_k\right) \widetilde{\boldsymbol{X}}_n \boldsymbol{\Sigma}_k^{-1}}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n \mid \mathbf{0}, \boldsymbol{\Sigma}_j\right)} \boldsymbol{x}_n.$$

Then we substitute the renewed $\boldsymbol{\Sigma}_k$, π_k, and $\boldsymbol{Z}$ in (2) for the next iteration. The optimization algorithm for solving(2) is summarized in Algorithm 1.

2.2. MoG Regression for Subspace Clustering

Similar to the previous methods [5, 9, 12], our clustering method is also based on the spectral clustering theory [3, 17]. After solving the MoG Regression problem (2) to get the representation matrix $\boldsymbol{Z}$, we define the affinity matrix as

$$\boldsymbol{C} = \mid \boldsymbol{Z} \mid + \mid \boldsymbol{Z}^{\top} \mid,$$

where each entry C_{ij} in $\boldsymbol{C}$ measures the similarity between data points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Figure 1 illustrates the affinity matrices of 10 subjects clustering derived by SSC, LRR, LSR, CASS, CIL2, and the proposed MoG Regression on the AR database, respectively, where the facial variations, illumination variations, and occlusions can be regarded as complex noise added to the original images. We can see that the affinity matrices by SSC and CASS are sparse due to the sparsity regularization, while the correlations within clusters are weak. So they may be less capable of grouping data points in the same cluster. In contrast, the affinity matrices from LRR, LSR, CIL2, and MoG Regression are very dense. The representation coefficients within clusters are large, indicating the good ability to group correlated data together. However, we can see that the contrast between diagonal blocks and non-diagonal blocks of MoG Regression is much higher than those of LRR, LSR, and CIL2, and the differences between coefficients within the same clusters of MoG Regression are also much smaller.

Table 1. The contrast (%) of affinity matrices in Figure 1

SSC	LRR	LSR	CASS	CIL2	Ours
73.51	75.41	52.10	75.18	76.35	**80.32**

To quantitatively evaluate the contrast of the diagonal blocks against the non-diagonal blocks of each method, we define the contrast by $(S_d - S_{nd})/\|\boldsymbol{C}\|_1$, where S_d and S_{nd} are the sums of absolute values of entries in diagonal and non-diagonal blocks, respectively. Table 1 shows the contrast of the affinity matrices from different methods. We see that the contrast value of MoG Regression is much higher than those of other approaches. This demonstrates that, with complex noise aggravating the data, our method is better at describing the distribution of noise, thus showing stronger grouping effect and greater ability to recover the true subspace structures.

In the end, we employ Normalize Cut [19] on the affinity matrix $\boldsymbol{C}$ to produce the final clustering results.

Algorithm 1: Finding the solution of (2) by EM

Initialize: data matrix $\boldsymbol{X}$, covariance matrices $\boldsymbol{\Sigma}_k$, parameter λ, threshold value ε, initial representation matrix $\boldsymbol{Z}$, and the components number K.

Repeat :

1: Compute $\gamma_{n,k}$:

$$\gamma_{n,k} = \frac{\pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n^{old} \mid \mathbf{0}, \boldsymbol{\Sigma}_k^{old}\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n^{old} \mid \mathbf{0}, \boldsymbol{\Sigma}_j^{old}\right)},$$

where $\widetilde{\boldsymbol{e}}_n^{old} = \widetilde{\boldsymbol{X}}_n \boldsymbol{z}_n^{old} - \boldsymbol{x}_n$.

2: Update the $\boldsymbol{\Sigma}_k$, π_k, and $\boldsymbol{Z}$:

$$\boldsymbol{\Sigma}_k^{new} = \frac{1}{\gamma_{n,k}} \left(\sum_{n=1}^{N} \gamma_{n,k} \widetilde{\boldsymbol{e}}_n^{old} \left(\widetilde{\boldsymbol{e}}_n^{old}\right)^{\top} + \epsilon \boldsymbol{I} \right),$$

$$\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} \gamma_{n,k},$$

$$\boldsymbol{z}_n^{new} = \left(\frac{\sum_{k=1}^{K} \xi_k \widetilde{\boldsymbol{X}}_n^{\top} \left(\boldsymbol{\Sigma}_k^{new}\right)^{-1} \widetilde{\boldsymbol{X}}_n}{\sum_{j=1}^{K} \xi_j} + 2\lambda \boldsymbol{I} \right)^{-1} \boldsymbol{b}_n,$$

where

$$\boldsymbol{b}_n = \frac{\sum_{k=1}^{K} \xi_k \widetilde{\boldsymbol{X}}_n \left(\boldsymbol{\Sigma}_k^{new}\right)^{-1}}{\sum_{j=1}^{K} \xi_j} \boldsymbol{x}_n,$$

and

$$\xi_k = \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n^{old} \mid \mathbf{0}, \boldsymbol{\Sigma}_k^{new}\right), \ \xi_j = \pi_j \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n^{old} \mid \mathbf{0}, \boldsymbol{\Sigma}_j^{new}\right).$$

3:

$$\boldsymbol{\Sigma}_k^{old} \longleftarrow \boldsymbol{\Sigma}_k^{new}, \ \pi_k^{old} \longleftarrow \pi_k^{new}, \ \boldsymbol{z}_n^{old} \longleftarrow \boldsymbol{z}_n^{new}.$$

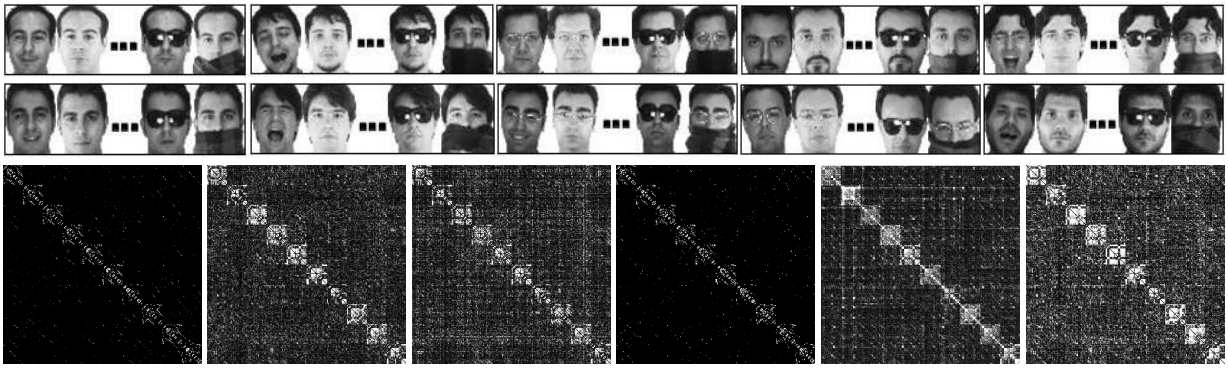
Until :

$$\| \boldsymbol{Z}^{old} - \boldsymbol{Z}^{new} \|_F \leq \varepsilon \text{ and } \| \boldsymbol{\Sigma}^{old} - \boldsymbol{\Sigma}^{new} \|_F \leq \varepsilon.$$

Output: The representation matrix $\boldsymbol{Z}$.

3. The Grouping Effect

In this section we will prove the effectiveness of our proposed model in subspace clustering by investigating its



(a) SSC [5] (b) LRR [9] (c) LSR [12] (d) CASS [10] (e) CIL2 [11] (f) Ours

Figure 1. The affinity matrices of 10 objects obtained by different methods on the AR database.

grouping effect. The grouping effect is an important criterion for measuring the validity of a clustering method, which tends group highly correlated data together [12]. We state the grouping effect of MoG Regression as follows.

Theorem 3.1 *Given a sample point $\boldsymbol{x} \in \mathbb{R}^M$, the normalized data matrix $\boldsymbol{X}$ and the regularization parameter λ, let $\widehat{\boldsymbol{z}}$ be the optimal solution to*

$$\min_{\boldsymbol{z}} -\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{X}\boldsymbol{z}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right)+\lambda\parallel\boldsymbol{z}\parallel^2,$$

then there exists a constant a such that

$$\mid\widehat{z}^i-\widehat{z}^j\mid\leq\frac{a}{\lambda}\sqrt{\frac{1-\rho}{2}},$$

where $\rho=\cos\langle\boldsymbol{x}_i,\boldsymbol{x}_j\rangle$. Here we denote $\widehat{z}^i$ and $\widehat{z}^j$ as the i-th and j-th entries of vector $\widehat{\boldsymbol{z}}$, and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as the i-th and j-th columns of $\boldsymbol{X}$, respectively.

Proof 3.1 *Let*

$$f\left(\boldsymbol{z}\right)=-\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{X}\boldsymbol{z}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right)+\lambda\parallel\boldsymbol{z}\parallel^2.$$

Since $\widehat{\boldsymbol{z}}=\underset{\boldsymbol{z}}{\operatorname{argmin}}\, f\left(\boldsymbol{z}\right)$, we have

$$\left.\frac{\partial f\left(\boldsymbol{z}\right)}{\partial\boldsymbol{z}}\right|_{\boldsymbol{z}=\widehat{\boldsymbol{z}}}=0.$$

This gives

$$\frac{\boldsymbol{x}_i^\top\left(\sum_{k=1}^{K}\xi_k\boldsymbol{\Sigma}_k^{-1}\right)\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)}{\sum_{k=1}^{K}\xi_k}+2\lambda\widehat{z}^i=0,$$

and

$$\frac{\boldsymbol{x}_j^\top\left(\sum_{k=1}^{K}\xi_k\boldsymbol{\Sigma}_k^{-1}\right)\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)}{\sum_{k=1}^{K}\xi_k}+2\lambda\widehat{z}^j=0,$$

where $\xi_k=\pi_k\mathcal{N}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)$.

From the above two equations for $\widehat{z}^i$ and $\widehat{z}^j$ we deduce that

$$\widehat{z}^i-\widehat{z}^j=\frac{\left(\boldsymbol{x}_i^\top-\boldsymbol{x}_j^\top\right)\left(\sum_{k=1}^{K}\xi_k\boldsymbol{\Sigma}_k^{-1}\right)\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)}{2\lambda\sum_{k=1}^{K}\xi_k}.$$

Note that

$$\begin{aligned}\left\|\sum_{k=1}^{K}\xi_k\boldsymbol{\Sigma}_k^{-1}\right\|_2&\leq\sum_{k=1}^{K}\xi_k\parallel\boldsymbol{\Sigma}_k^{-1}\parallel_2\\&\leq\left(\max_k\parallel\boldsymbol{\Sigma}_k^{-1}\parallel_2\right)\sum_{k=1}^{K}\xi_k.\end{aligned}$$

So we get

$$\mid\widehat{z}^i-\widehat{z}^j\mid\leq\frac{\parallel\boldsymbol{x}_i-\boldsymbol{x}_j\parallel_2\cdot\parallel\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\parallel_2\cdot\left(\max_k\parallel\boldsymbol{\Sigma}_k^{-1}\parallel_2\right)}{2\lambda}.$$

Note that $\widehat{\boldsymbol{z}}$ is a minimizer of $f(\boldsymbol{z})$. So we have

$$f\left(\widehat{\boldsymbol{z}}\right)\leq f\left(\boldsymbol{0}\right),$$

which yields

$$\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right)\geq\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right).$$

On the other hand, if we define

$$V=\underset{k}{\operatorname{argmax}}\,\pi_k$$

and

$$S=\underset{k}{\operatorname{argmax}}\,\mathcal{N}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right),$$

then we get

$$\ln\left(K\pi_V\mathcal{N}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_S\right)\right)\geq\ln\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right),$$

which is equivalent to

$$\frac{\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)^{\top}\boldsymbol{\Sigma}_S^{-1}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)}{2}$$

$$\leq\ln\left(\frac{K\pi_V}{\left(2\pi\right)^{\frac{M}{2}}\mid\boldsymbol{\Sigma}_S\mid^{\frac{1}{2}}\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right)}\right).$$

Since $\boldsymbol{\Sigma}_S^{-1}$ is a symmetric positive definite matrix, whose unitary similar matrix is a diagonal matrix, we can list the diagonal entries in descending order. Then We have

$$\mid\boldsymbol{\Sigma}_S\mid^{-1}=U^{\top}\begin{pmatrix}\lambda_1&\cdots&\boldsymbol{0}\\\vdots&\ddots&\vdots\\\boldsymbol{0}&\cdots&\lambda_{\min}\end{pmatrix}U,$$

Where U is the unitary matrix, $\lambda_1\geq\cdots\cdots\geq\lambda_{\min}$ denote the eigenvalues of $\boldsymbol{\Sigma}_S^{-1}$. Thus we get

$$\begin{aligned}&\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)^{\top}\boldsymbol{\Sigma}_S^{-1}\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)\\=&\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)^{\top}U^{\top}\begin{pmatrix}\lambda_1&\cdots&\boldsymbol{0}\\\vdots&\ddots&\vdots\\\boldsymbol{0}&\cdots&\lambda_{\min}\end{pmatrix}U\left(\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\right)\\\geq&\lambda_{\min}\parallel\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\parallel^2.\end{aligned}$$

It yields to

$$\parallel\boldsymbol{X}\widehat{\boldsymbol{z}}-\boldsymbol{x}\parallel^2\leq Q,$$

where

$$Q=\frac{1}{\lambda_{\min}}\ln\left(\frac{K\pi_V}{\left(2\pi\right)^{\frac{M}{2}}\mid\boldsymbol{\Sigma}_S\mid^{\frac{1}{2}}\left(\sum_{k=1}^{K}\pi_k\mathcal{N}\left(\boldsymbol{x}\mid\boldsymbol{0},\boldsymbol{\Sigma}_k\right)\right)}\right)^2.$$

Then we get

$$\mid\widehat{z}^i-\widehat{z}^j\mid\leq\frac{a}{\lambda}\sqrt{\frac{1-\rho}{2}},$$

where

$$a=\left(\max_k\parallel\boldsymbol{\Sigma}_k^{-1}\parallel_2\right)\sqrt{Q}$$

is a constant.

From Theorem 3.1 we can see that, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are highly correlated, then ρ is close to 1 and further the upper bound of the difference between $\widehat{z}^i$ and $\widehat{z}^j$ approaches 0. Therefore, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ would be grouped into the same cluster. An illustration of grouping effect can be seen in Figure 1, where the differences between representation coefficients are small within the same cluster while larger between different clusters.

4. Experiments

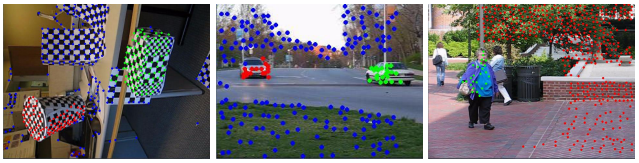
In this section, we show the performance of the proposed MoG Regression method on the Hopkins 155 database [20], the Rotated MNIST Dataset [7], the AR database [15], and the Extended Yale Face Dataset B [25]. Experimental results show that the proposed method is effective and robust to noise in motion segmentation, handwritten digits clustering, and complex face clustering.

We also apply SSC [5], LRR [9], LSR [12], CASS [10], and CIL2 [11] to these datasets. We tune the parameters of each method to achieve the best performance for fair comparison, and the clustering accuracy [5] is employed in quantitative evaluation. The comparison shows that our approach outperforms five state-of-the-art methods.

4.1. Hopkins 155 Database

The Hopkins 155 motion segmentation database [20] contains 155 video sequences, where 120 of the videos contain 2 motions and 35 of the videos have 3 motions. On average, each sequence of the 2 motions has 266 feature trajectories and 30 frames, and each sequence of the 3 motions has 398 feature trajectories and 29 frames. For each sequence, a tracker is used to extract the point trajectories and a subspace clustering task is defined. Thus we have 155 subspace clustering tasks in total.

We first use PCA to reduce the dimensionality of the data. Then we test the MoG Regression method on each video sequence. Some motion segmentation results of our approach are shown in Figure 2, where motions of different objects and background motions can be accurately segmented.



(a) (b) (c)

Figure 2. Exemplar results of motion segmentation on the Hopkins 155 Database. (a) Checkerboard. (b) Cars. (c) People.

Table 2 presents the clustering accuracies of different methods. We can see that MoG Regression achieves significantly higher accuracies than the state-of-the-art methods.

4.2. MNIST-Back-Rand Dataset

The MNIST-back-rand database consists of 50000 images of hand-written digits from 0 to 9. It is first selected from the MNIST dataset [8] and then transformed into more challenging images by inserting random noise into the original images. Figure 3 shows some example images from the dataset.

Table 2. The clustering accuracies (%) on the Hopkins 155 database.

	SSC	LRR	LSR	CASS	CIL2	Ours
2 motions	95.69	96.43	97.48	97.01	97.63	**98.76**
3 motions	91.97	92.35	93.21	94.06	94.34	**95.03**

To reduce memory consumption in experiments, we randomly select 10 images for each digit to build a subset that contains 100 samples. Experimental results are reported in Table 3. We can see the advantage of our method is remarkable. This experiment shows that when the data are contaminated with non-Gaussian or complex noise, the proposed method is more capable of clustering the subspaces with the help of MoG.

Figure 3. Examples of the MNIST-back-rand database, where digits are corrupted with random noise.

Table 3. The clustering accuracies (%) on the MNIST-Back-Rand database

SSC	LRR	LSR	CASS	CIL2	Ours
33.56	22.85	20.55	29.05	36.50	**51.98**

We also conduct experiments to see how the number K of Gaussians affects the clustering accuracy of the proposed model. The accuracies are shown in Figure 4. We can see that when the number of Gaussians increases, the accuracy increases at first and then fluctuates, reaching the maximum value when $K = 5$. This is because when the number of Gaussians is too small, MoG may not characterize the noise accurately. On the other hand, when K is too large the computation cost will increase and the grouping effect will be suppressed (Difference bound demonstrated in Theorem 3.1 will increase). In either case the distribution of noise is not modeled well.

4.3. AR Dataset

The AR database [15] contains over 4,000 facial images corresponding to 126 subjects (70 men and 56 women). For each subject, 26 facial images are taken in two separate sessions. 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.

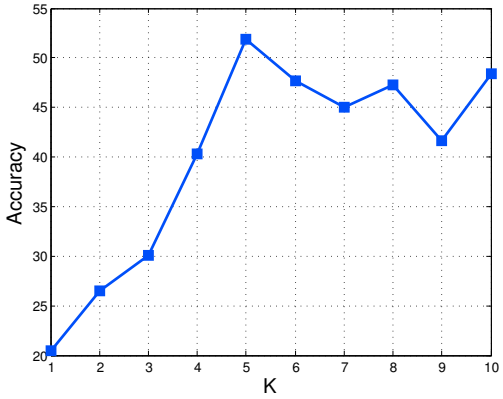


Figure 4. The clustering accuracies (%) of MoG Regression with different values of K on the MNIST-Back-Rand database.

Table 4. The clustering accuracies (%) on the AR database.

	SSC	LRR	LSR	CASS	CIL2	Ours
5 subjects	83.05	84.41	87.69	78.46	85.38	**93.85**
10 subjects	75.06	78.54	63.07	77.69	80.39	**88.85**

We build two subspace clustering tasks by selecting first 5 and 10 subjects from this dataset, respectively. The clustering results on the AR database of different algorithms are shown in Table 4. We can see that MoG Regression performs much better than other state-of-the-art methods in both clustering tasks. This is because MoG Regression has a strong grouping effect on this challenging database, which can be seen in Figure 1.

4.4. Extended Yale Face Dataset B

The Extended Yale Face Dataset B [25] consists of 2,414 frontal face images of 38 subjects, where there are 64 faces for each subject, acquired under various lighting, poses, and illumination conditions. To reduce the computational cost and the memory requirements, we resize the grayscale images to a resolution of 32×32 pixels.

To evaluate the robustness of different methods, we conduct experiments on corrupted Extended Yale Face Dataset B, where each image is corrupted by replacing random image pixels with samples from a uniform distribution on the interval from 0 to 255 [11], and the percentage of corrupted pixels varies from 10% to 100%. Figure 5 shows the clustering accuracies of all methods on the corrupted Extended Yale B database.

From Figure 5 we can see that the proposed approach performs much better when face images are randomly cor-

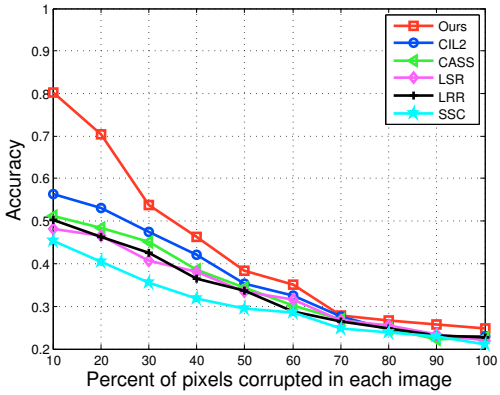


Figure 5. The clustering accuracies (%) with pixel corruption on the Extended Yale B database.

rupted at a level from 10% to 40%, showing better adaptability and greater robustness in noise handling. When the percentage of corrupted pixels is larger than 60%, the discriminative information are badly damaged, thus weakening the performance of all methods.

5. Conclusions

In this paper, we propose a new subspace clustering method by employing the MoG model to describe the distribution of complex noise. Theoretical analysis shows that the proposed MoG Regression method maintains the grouping effect. Experiments on motion segmentation, handwritten digits clustering, and complex face clustering demonstrate the superiority of the proposed method, regarding stability and robustness in handling general noise, over the state-of-the-art subspace clustering methods, SSC, LRR, LSR, and CASS, which assume Gaussian or sparse noise. In the future, we will work on accelerating the solution of MoG Regression.

Acknowledgements. B. Li, Y. Zhang and H. Lu are supported by the Natural Science Foundation of China #61472060 and the Fundamental Research Funds for the Central Universities under Grant DUT14YQ101. Z. Lin is supported by NSF China (grant nos. 61231002 and 61272341), 973 Program of China (grant no. 2015CB352502), and Microsoft Research Asia Collaborative Research Program.

References

[1] P. K. Agarwal and N. H. Mustafa. K-means projective clustering. In *Proceedings of the Twenty-third ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 14-16, 2004, Paris, France*, pages 155–165, 2004.

[2] P. S. Bradley and O. L. Mangasarian. K-plane clustering. *Journal of Global Optimization*, 16(1):23–32, 2000.

[3] F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

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

[5] E. Elhamifar and R. Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2765–2781, 2013.

[6] J. Ho, M. H. Yang, J. Lim, K. C. Lee, and D. Kriegman. Clustering appearances of objects under varying illumination conditions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages I–11, 2003.

[7] H. Larochelle, D. Erhan, A. Courville, J. Bergstra, and Y. Bengio. An empirical evaluation of deep architectures on problems with many factors of variation. In *Proceedings of International Conference on Machine Learning*, pages 473–480, 2007.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

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

[10] C. Lu, J. Feng, Z. Lin, and S. Yan. Correlation adaptive subspace segmentation by trace lasso. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1345–1352, 2013.

[11] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin. Correntropy induced l2 graph for robust subspace clustering. In *Proceedings of IEEE International Conference on Computer Vision*, pages 1801–1808, 2013.

[12] C. Y. Lu, H. Min, Z. Q. Zhao, L. Zhu, D. S. Huang, and S. Yan. Robust and efficient subspace segmentation via least squares regression. In *Proceedings of European Conference on Computer Vision*, pages 347–360. 2012.

[13] Y. Ma, H. Derksen, W. Hong, and J. Wright. Segmentation of multivariate mixed data via lossy data coding and compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(9):1546–1562, 2007.

[14] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum. Estimation of subspace arrangements with applications in modeling and segmenting mixed data. *SIAM review*, 50(3):413–458, 2008.

[15] A. M. Martinez. The AR face database. *CVC Technical Report*, 24, 1998.

[16] D. Nettleton. Convergence properties of the EM algorithm in constrained parameter spaces. *Canadian Journal of Statistics*, 27(3):639–648, 1999.

[17] A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.

[18] S. R. Rao, R. Tron, R. Vidal, and Y. Ma. Motion segmentation via robust subspace separation in the presence of outlying, incomplete, or corrupted trajectories. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2008.

[19] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[20] R. Tron and R. Vidal. A benchmark for the comparison of 3-d motion segmentation algorithms. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[21] R. Vidal and R. Hartley. Motion segmentation with missing data using powerfactorization and GPCA. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–310. IEEE, 2004.

[22] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12):1945–1959, 2005.

[23] U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

[24] K. H. Wei Hong, John Wright and Y. Ma. Multi-scale hybrid linear models for lossy image representation. In *IEEE Transactions on Image Processing*, volume 15, pages 3655–3671, 2006.

[25] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):210–227, 2009.

[26] C. J. Wu. On the convergence properties of the EM algorithm. *The Annals of statistics*, pages 95–103, 1983.

[27] L. Xu and M. I. Jordan. On convergence properties of the EM algorithm for gaussian mixtures. *Neural Computation*, 8(1):129–151, 1996.

[28] J. Yan and M. Pollefeys. A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate. In *Proceedings of European Conference on Computer Vision*, pages 94–106. 2006.

[29] A. Y. Yang, S. R. Rao, and Y. Ma. Robust statistical estimation and segmentation of multiple subspaces. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pages 99–99, 2006.

[30] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry. Unsupervised segmentation of natural images via lossy data compression. *Computer Vision and Image Understanding*, 110(2):212–225, 2008.

[31] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang. Robust principal component analysis with complex noise. In *Proceedings of International Conference on Machine Learning*, 2014.