Subspace Clustering Under Complex Noise

Baohua Li, Huchuan Lu, *Senior Member, IEEE*, Ying Zhang, Zhouchen Lin, *Fellow, IEEE*, and Wei Wu

***Abstract*— In this paper, we study the subspace clustering problem under complex noise. A wide class of reconstruction-based methods model the subspace clustering problem by combining a quadratic data-fidelity term and a regularization term. In a statistical framework, the data-fidelity term assumes to be contaminated by a unimodal Gaussian noise, which is a popular setting in most current subspace clustering models. However, the realistic noise is much more complex than our assumptions. Besides, the coarse representation of the data-fidelity term may depress the clustering accuracy, which is often used to evaluate the models. To address this issue, we propose the mixture of Gaussian regression (MoG Regression) for subspace clustering. The MoG Regression seeks a valid way to model the unknown noise distribution, which approaches the real one as far as possible, so that the desired affinity matrix is better at characterizing the structure of data in the real world, and furthermore, improving the performance. Theoretically, the proposed model enjoys the grouping effect, which encourages the coefficients of highly correlated points are nearly equal. Drawing upon the ideal of the minimum message length, a model selection strategy is proposed to estimate the numbers of the Gaussian components that shows a way how to seek the number of Gaussian components besides determining it by empirical value. In addition, the asymptotic property of our model is investigated. The proposed model is evaluated on the challenging datasets. The experimental results show that the proposed MoG Regression model significantly outperforms several state-of-the-art subspace clustering methods.**

***Index Terms*— Subspace clustering, mixture of Gaussian regression, expectation maximization.**

Manuscript received January 16, 2017; revised August 16, 2017; accepted December 29, 2017. Date of publication January 15, 2018; date of current version April 3, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61725202, Grant 61528101, Grant 61472060, Grant 61625301, and Grant 61731018, in part by the National Basic Research Program of China (973 Program) under Grant 2015CB352502, in part by Qualcomm, and in part by Microsoft Research Asia. This paper was presented at the IEEE Conference on Computer Vision and Pattern Recognition, June 2015 [1]. This paper was recommended by Associate Editor Y. Keller. *(Corresponding author: Huchuan Lu.)*

B. Li is with the School of Electric and Automatic Engineering, Changshu Institute of Technology, Changshu 215500, China (e-mail: libaoh@mail.dlut.edu.cn).

H. Lu and Y. Zhang are with the School of Information and Communication Engineering, Faculty of Electronic Information and Electrical Engineering, Dalian University of Technology, Dalian 116024, China (e-mail: lhchuan@dlut.edu.cn; zydl0907@mail.dlut.edu.cn).

Z. Lin is with the Key Laboratory of Machine Perception (MOE), School of EECS, Peking University, Beijing 100871, China, and also with the Cooperative Medianet Innovation Center, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: zlin@pku.edu.cn).

W. Wu is with the School of Mathematical Sciences, Dalian University of Technology, Dalian 116024, China (e-mail: wuweiw@dlut.edu.cn).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. This supplementary file provides detailed algorithm steps of the concerned optimization problem. The total size of the file is 161.7859 kb.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSVT.2018.2793359

I. Introduction

THE goal of subspace clustering is to gather the given data points into disparate group which contains the data points that come from the same underlying subspace. It has been attracting more and more attentions in recent years and has found many applications in computer vision and image processing, such as image segmentation [2], motion segmentation [3], face clustering [4], and image representation and compression [5].

There exists numbers of major subspace clustering approaches which have been proposed in the past two decades. These methods may be roughly divided into four main categories: algebraic methods [6]–[8], iterative methods [9], [10], statistical methods [11]–[13], and spectral-clustering-based methods [14]–[20]. It should be noted specially that the subspace clustering self-reconstruction based methods [16]–[20], which take root in the elegant spectral graph theory [21], have shown excellent performance in many real applications.

Generally, the subspace clustering self-reconstruction based methods consist of two steps. Firstly, building an affinity matrix which is used to capture the similarity between pairs of sample points. Secondly, graph cut is applied to a undirected 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 which leads to some subspace clustering methods focus on how to build a good affinity matrix.

Based on the ideal that each data point in a union of several subspaces can be represented as a linear or affine combination of other points, the Sparse Subspace Clustering (SSC) algorithm [16] utilizes the ℓ_1-norm regularization to find the sparsest representation of a data point, where points come from the same subspace correspond to the nonzero representation coefficients. Low-Rank Representation (LRR) [17] aims to get a low rank reconstruction coefficient for robust subspace recovery of the data containing corruptions, the $\ell_{2,1}$ is used to make the algorithm more robust to outliers. Least Squares Regression (LSR) [18] employs the Frobenius norm regularization to speed up the clustering process, while still ensuring the grouping effect of the representation matrix. However, the reconstruction coefficient of SSC may be too sparse to encode the data correlation, and the reconstruction coefficient derived by both LRR and LSR may result in dense connections between-clusters besides the within-clusters. In order to achieve a good balance between within-cluster density (which we call *grouping effect* afterwards) and between-cluster sparsity, Correlation Adaptive Subspace Segmentation (CASS) [20] adopts trace Lasso norm regularization, which is adaptive to the data correlation, to trade-off the representation matrix.

As pointed out by Liu *et al.* [17], the noises that always exist in data can perturb the subspace structures, which leads to unreliable subspace clustering result. To cluster the real subspaces when the data are corrupted by noises, SSC, LRR, LSR, and CASS employ different norms to select the solution of various properties, respectively.

Given a data matrix $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N) \in \mathbb{R}^{M\times N}$ with N samples in $\mathbb{R}^M$, here, 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 component Z_{ij} of $\boldsymbol{Z}$ measures the similarity between points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the data matrix. In this paper, 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 mentioned subspace clustering methods are listed as follows.

Sparse Subspace Clustering (SSC) [16]:

$$\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{E}} \ \|\boldsymbol{E}\|_1 + \lambda\|\boldsymbol{Z}\|_1\\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \quad \mathrm{diag}\,(\boldsymbol{Z}) = \boldsymbol{0}.\end{aligned}$$

Low-Rank Representation (LRR) [17]:

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

Least Squares Regression (LSR) [18]:

$$\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{E}} \ \|\boldsymbol{E}\|_F^2 + \lambda\|\boldsymbol{Z}\|_F^2\\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \quad \mathrm{diag}\,(\boldsymbol{Z}) = \boldsymbol{0}.\end{aligned}$$

Correlation Adaptive Subspace Segmentation (CASS) [20]:

$$\begin{aligned}&\min_{\boldsymbol{Z},\boldsymbol{E}} \ \|\boldsymbol{E}\|_F^2 + \lambda\sum_{n=1}^{N}\|\boldsymbol{X}\mathrm{diag}\,(z_i)\|_*\\ &s.t.\ \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}.\end{aligned}$$

In the above mentioned formulations, z_i is represented the i-th column of $\boldsymbol{Z}$, $\mathrm{diag}\,(\boldsymbol{Z})$ is a diagonal matrix with entries of z_{ii} on its diagonal, and $\lambda > 0$ is a regularization parameter to balance the effects of two terms. $\boldsymbol{E}$ denotes the reconstruction error. $\|\boldsymbol{E}\|_F^2$ is utilized to model Gaussian noise, $\|\boldsymbol{E}\|_{2,1}$ is for sample-specific corruptions, and $\|\boldsymbol{E}\|_1$ is for entry-wise corruptions.

All the clustering algorithms mentioned above rely on specific norms on $\boldsymbol{Z}$ and $\boldsymbol{E}$ to encourage either the between-cluster sparsity and within-cluster density or grouping effect of the representation matrix which makes the model be valid. However, they all use a relatively simple norm to describe the data fidelity term that coincides with the noises.

In fact, the real noise scenario in practice often exhibits very complex statistical distributions, rather than simply being a unimodal Gaussian or Laplace [22]. Therefore, to describe the noise by a simple norm like the Frobenious norm, ℓ_1-norm, or $\ell_{2,1}$-norm, may lead to the obtained affinity matrix depress the clustering accuracies.

To alleviate this issue, we employ a fundamental result from the probability theory that almost any distribution can be well approximated by a mixture of a suitable number of Gaussian type distributions. Namely, we employ the mixture of Gaussian (MoG) model to describe the real noise accurately, rather than assuming a specific distribution for the noise. As for the regularization the term, we simply choose the Frobenius norm which means that we select the minimal Frobenius norm solution among the candidates. 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 exhibit such an effect. Second, it makes the computational procedure much easier with the Frobenious norm on $\boldsymbol{Z}$. For example, the traditional Expectation Maximization (EM) algorithm can be used to find the solution of our new subspace clustering model. We prove that the prosed model holds the grouping effect [23] for correlated data points, which encourages the coefficients of correlated data pints are approximately equal. How to determine the number of Gaussians K is another crux problem. Aside from empirically fixing K, we proposed a model selection strategy to estimate K inspired by [24] and [25]. Besides, we prove the asymptotic properties of our model for fixed M and K in the spirit of [26]. In summary, we list the outline of the contributions as follows:

- A Mixture of Gaussian Regression (MoG Regression) based subspace clustering method was proposed.
- We prove that MoG Regression has the grouping effect, which is important for subspace clustering.
- We provide a model selection method based on the minimum message length (MML) criterion to estimate the numbers of Gaussian components.
- To investigate the property of our proposed model under Expectation Maximization(EM) Algorithm, we provide the asymptotic properties of solution.

The remainder of the paper is organized as follows. In Section II, we motivate and introduce the MoG Regression method in detail for clustering data. In Section III we prove that the proposed model possesses the grouping effect. The asymptotic property is shown in section IV. Based on MML, we show how to estimate K in section V. Section VII reports the experimental results. We relegate the main steps of proof to section VIII. Partial results of this paper appear in our conference version [1].

II. Subspace Clustering via MoG Regression

As described in [27], we model the subspace clustering issue as the following optimization problem:

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

where $\mathcal{L}(\boldsymbol{E})$ is to be described the noise in the loss function, and $\mathcal{R}(\boldsymbol{Z})$ is the regularization term to impose some desired properties on the representation matrix $\boldsymbol{Z}$.

In fact, the noise is a nuisance, that may spoil the ability of (1) to cluster the real subspaces. So, it becomes significant importance to describe the unknown noise in subspace clustering problems. Lu *et al.* [27] 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. However, as Liu *et al.* [28] pointed that the correntropy strategy is more suitable for the impulsive noise environment, which may cripple the performance of clustering by correntropy induced metric [27] if the noise is not in the listed specified types.

Inspired by the probability theory that almost any continuous density can be approximated by using s sufficient number of Gaussians to arbitrary accuracy. We propose a novel clustering method called MoG Regression based on the reconstruction, which employs MoG to characterize the general noise, and uses the regression coefficient to carry out the subspace clustering. The previous MoG strategies [29], [30], etc. whose mechanism is to assign the data points that come from the same group to the corresponding component, which is the major difference from the proposed method in this work. We assume that each column $\boldsymbol{e}_n$ $(n = 1, \ldots, N)$ of $\boldsymbol{E}$ follows the MoG distribution, i.e.,

$$p(\boldsymbol{e}_n) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k), \tag{2}$$

where K is the number of Gaussian components and π_k denotes the mixing weight which is satisfy with the constrain $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$. $\mathcal{N}(\boldsymbol{e}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k)$ is denoted the zero-mean multivariate Gaussian distribution, with $\boldsymbol{\Sigma}_k (k = 1, 2, \ldots, K)$ representing the invertible and symmetrical covariance matrix. Note that, we have 0 replaced the unknown means in (2) lies in two aspects: lighting the computation burden of the (5), on the other hand, the clustering accuracies of both estimating means and no estimating means in our frame (5) are about the same.

It is analogous to the classical regression analysis that the columns of $\boldsymbol{E}$ are assumed to be independently and identically distributed in the MoG Regression setting. Thus we have

$$p(\boldsymbol{E}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k). \tag{3}$$

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

$$-\ln p(\boldsymbol{E}) = -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k) \right). \tag{4}$$

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

$$\begin{aligned} \min_{\boldsymbol{Z}, \boldsymbol{\pi}, \boldsymbol{\Sigma}} \; & -\sum_{n=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\boldsymbol{e}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k) \right) + \lambda \|\boldsymbol{Z}\|_F^2 \\ s.t. \; & \boldsymbol{X} = \boldsymbol{X}\boldsymbol{Z} + \boldsymbol{E}, \quad \operatorname{diag}(\boldsymbol{Z}) = \boldsymbol{0}, \\ & \pi_k \geq 0, \quad \boldsymbol{\Sigma}_k \in \mathbb{S}^+, \; k = 1, \ldots, K, \; \sum_{k=1}^{K} \pi_k = 1, \end{aligned} \tag{5}$$

where $\lambda > 0$ is the regularization parameter, $\mathbb{S}^+$ is denoted the set of symmetrical positive definite (SPD) matrices and the constraint $\operatorname{diag}(\boldsymbol{Z}) = \boldsymbol{0}$ discourages using a sample to represent itself. Here we simply choose the Frobenius norm of $\boldsymbol{Z}$ as the regularization term. As declared before, we chose the Frobenious norm on $\boldsymbol{Z}$ that can not only reduce the computation cost but also expose the effect of MoG regression based noise modeling on subspace clustering. In model (5), we will bear some limitations of Frobenius norm, for instance, it is sensitive to outliers, and may obtain the dense coefficient matrix, ect.. On the other hand, using the MoG to describe the unknown noise is very common [31]–[33]ect., however, the mentioned works do not use the coefficient matrix to carry out the subspace clustering.

A natural way to capture the solution of (5) may be the powerful EM algorithm [34], [35], which finds the maximum-likelihood estimate of the parameters iteratively. Its procedure starts from an initial guess and iteratively runs an expectation (E) step, which evaluates the posterior probabilities using currently known parameters, and a maximization (M) step, which will re-estimate the parameters based on the probabilities calculated in the E step. The iterations will stop until some convergence criteria are satisfied [36]–[38]. Integrating the traditional processes of the EM algorithm, we can obtain the solution of problem (5) in the following three main steps.

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

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

$$\gamma_{n,k} = \frac{\pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_j)}, \tag{6}$$

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 $\boldsymbol{0}$.

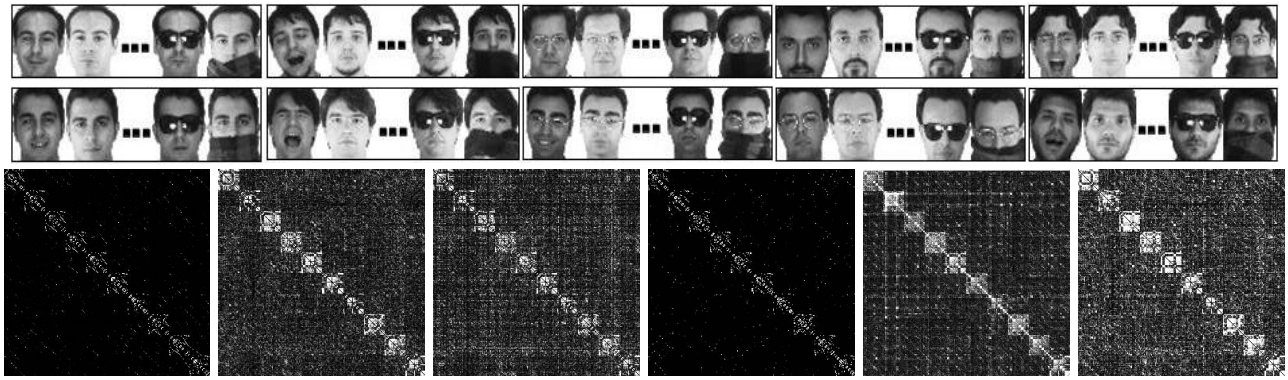
In the M-step, we want to minimize the log likelihood with respect to the parameters, using the current posterior probabilities.

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

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

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

$$\boldsymbol{\Sigma}_k = \frac{1}{\gamma_{n,k}} \left(\sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_j)} \widetilde{\boldsymbol{e}}_n \cdot \widetilde{\boldsymbol{e}}_n^\top + \epsilon \boldsymbol{I} \right), \tag{7}$$



(a) (b) (c) (d) (e) (f)

Fig. 1. The affinity matrices of 10 objects obtained by different methods on the AR database. (a) SSC [16]. (b) LRR [17]. (c) LSR [18]. (d) CASS [20]. (e) CIL2 [27]. (f) Ours.

where $\epsilon > 0$ is a small regularization parameter to avoid that the determinant of $\mathbf{\Sigma}_k$ equals to zero.

Each mixing weighting π_k, $k = 1, 2, \ldots, K$, is updated by solving

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

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

$$\pi_k = \frac{1}{N} \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \mathbf{\Sigma}_k\right)}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \mathbf{\Sigma}_j\right)}. \tag{8}$$

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

$$\min_{\boldsymbol{z}_n} -\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \mathbf{\Sigma}_k\right)\right) + \lambda \parallel \boldsymbol{z}_n \parallel_F^2 . \tag{9}$$

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

$$\boldsymbol{z}_n = \left(\frac{\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \mathbf{\Sigma}_k\right) \widetilde{\boldsymbol{X}}_n^{\top} \mathbf{\Sigma}_k^{-1} \widetilde{\boldsymbol{X}}_n}{\sum_{j=1}^{K} \pi_j \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \mathbf{\Sigma}_j\right)} + 2\lambda \boldsymbol{I}\right)^{-1} \boldsymbol{b}_n, \tag{10}$$

where

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

Then we plug the renewed $\mathbf{\Sigma}_k$, π_k $(k = 1, 2, \cdots, K)$, and $\boldsymbol{Z}$ in (5) for the next round iteration. The optimization procedure for solving (5) is the standard EM iteration, which is shown in our conference version [1].

A. MoG Regression for Subspace Clustering

The proposed method falls into the category of spectral based clustering [21], [39], which is the analogy to the previous methods [16]–[18]. After solving the MoG Regression problem (5), the desired representation matrix $\boldsymbol{Z}$ is found. We define the affinity matrix as

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

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

We illustrate the affinity matrices of 10 subjects clustering derived by SSC, LRR, LSR, CASS, CIL2, and the proposed MoG Regression, respectively with Figure 1 on the AR database, where the facial variations, illumination variations, and occlusions can be regarded as complex noise added to the original images. Because of the function of variable selection, the affinity matrix derived by SSC is sparse, which depresses the correlations within clusters. Although the trace lasso norm enjoys correlation adaptive, the complex noises breaks this property, and the obtained affinity matrix shows the unseemly correlations. So they may be less ability of grouping data points in the same cluster. On contrast, the affinity matrices derived by LRR, LSR, CIL2, and MoG Regression are very dense. The value of representation coefficients within clusters are large, which indicates the relevant clustering method to be good ability to group correlated data together. Meanwhile, we can see that the contrast between diagonal blocks and non-diagonal parts of MoG Regression is much higher than those of LRR, LSR, and CIL2.

TABLE I

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**

In order to quantitatively evaluate the contrast of the diagonal blocks against the non-diagonal parts of affinity matrices derived by 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 parts, respectively. Table I lists the contrast of the affinity matrices from different methods. We notice that the contrast value of MoG Regression precedes other approaches. This demonstrates that, with complex noise corruption the data, our method is suitable for describing the distribution of noise, thus presenting

stronger grouping effect and greater ability to recover the true subspace structures.

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

III. The Grouping Effect

In this section we will theoretically expound the validity of the proposed MOG regression model for subspace clustering. A regression method shows the grouping effect if the coefficients of a group of correlated data tend to be equal. In [23] and [40] the grouping effect is detailed studied. The validity of clustering comes from the grouping effect for the models in [18], [20], and [27] has been proved. In this section we will show that our proposed MoG Regression model also possesses the grouping effect for correlated data. Now, we declare the grouping effect of MoG Regression as follows.

Theorem 1: Given a sample point $\boldsymbol{x} \in \mathbb{R}^M$, the normalized data matrix $\boldsymbol{X}$ and the regularization parameter λ, let $\widehat{z}$ be the optimal solution to

$$\min_{z} \; -\ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\boldsymbol{X}\boldsymbol{z}-\boldsymbol{x}|\boldsymbol{0}, \boldsymbol{\Sigma}_k\right)\right)+\lambda\|\boldsymbol{z}\|^2, \qquad (11)$$

then there exists a constant a such that

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

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 $\widehat{z}$, and $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ as the i-th and j-th columns of $\boldsymbol{X}$, respectively.

From the above **Theorem** 1 we can see that, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are highly correlated, i.e. ρ is close to 1, then the upper bound of the difference between $\widehat{z}^i$ and $\widehat{z}^j$ approaches 0. In this case, $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ would be grouped into the same cluster due to the grouping effect, which encourages the clustering performance.

IV. Asymptotic Property

In fact, our method is similar to [16]–[18], [20], and [27] belonging to the self reconstruction category which needs enough data points. On the other hand, due to the non-convex of our model, it is necessary to analysis the asymptotic property of $\boldsymbol{Z}$ in (5). We assume that the number of mixture Gaussian components K and the dimension of each data point M are fixed. Let $\theta_{\boldsymbol{Z}}^{ini}$ and $\theta_{\boldsymbol{Z}}^{t}$ denote the initial value and the true parameter value of $\boldsymbol{Z}$, where $\theta_{\boldsymbol{Z}}^{ini} = \left(z_1^{ini}, \ldots, z_N^{ini}\right)$ and $\theta_{\boldsymbol{Z}}^{t} = \left(z_1^{t}, \ldots, z_N^{t}\right)$. In the spirit of [26] and [41], we obtain the following result.

Theorem 2: Let the columns of data matrix $\boldsymbol{X} \in \mathbb{R}^{M\times N}$ with independent and identically distributed $(i.i.d)$, If $\frac{\lambda}{\sqrt{n}} = o(1)$, $\theta_{\boldsymbol{Z}}^{ini} - \theta_{\boldsymbol{Z}}^{t} = \mathcal{O}_p\left(n^{\frac{-1}{2}}\right)$, then, under the regularity conditions $(A)-(C)$ [26] on (9), and keep K invariably. We obtain that the local minimizer $\theta_{\boldsymbol{Z}}^{lm}$ of model (9) holds

$$\sqrt{n}\left(z_i^{lm}-z_i^{t}\right) \xrightarrow{d} \mathcal{N}\left(0, I_s\left(z_i^{t}\right)^{-1}\right), \quad i=1, \ldots, n \qquad (12)$$

Where the notions $\mathcal{O}_p$ and $\xrightarrow{d}$ denote the order to be equal in probability, and convergence in distribution [42] respectively, and $I_s(\cdot)$ denotes the information matrix [42]. Equipped with the regularity conditions$(A)-(C)$ [26] we derive the theorem 2 and postpone the proof in appendix

V. A Strategy for Finding the Number of Mixture Components

So far, we assume the number of the components K is fixed by empirical value. Inspired by [24], [25], and [30] we provide a strategy to estimation the number of components K based on the minimum message length (MML) criterion [43], [44]. For the sake of self-contained, we give a glance of MML criterion.

To formalize the MML ideal, we notice that the equation (3) can be viewed as $p(\boldsymbol{E}|\Theta)$ which agrees with parameter

$$\Theta = (\pi_1, \cdots \pi_K, \Sigma_1, \cdots, \Sigma_K) \qquad (13)$$

In the spirit of [43] and [44], the parameter estimation issue boils down a transmission encoding problem. If a short code can be found for the provide data, we will obtain a good data generation mode [43], [45], [46]. This leads to

$$length(\boldsymbol{E}, \Theta) = length(\boldsymbol{E}|\Theta) + length(\Theta) \qquad (14)$$

where $length(\boldsymbol{E}, \Theta) = -\ln p(\boldsymbol{E}, \Theta)$. In this context, the finite code length can only be obtained by quantizing the parameter Θ to finite precision after we undergo a loop of finding $\boldsymbol{Z}$. In fact, a fine precision is truncated, $length(\Theta)$ may be large, but $length(\boldsymbol{E}|\Theta)$ will be small because Θ closes to the optimal value. Conversely, a coarse precision is used, $length(\Theta)$ may be small, but $length(\boldsymbol{E}|\Theta)$ will be large because Θ departs from the optimal value [25]. In [25], the Taylor approximation method is used to balance the optimal quantization. In this case the optimal Θ is found by

$$\widehat{\Theta} = \arg\min_{\Theta}\{-\ln p(\Theta) - \ln p(\boldsymbol{E}|\Theta) + \frac{1}{2}\ln|\mathrm{F}(\Theta)| + \frac{D(\Theta)}{2}\left(1+\ln\frac{1}{12}\right)\} \qquad (15)$$

where $|\mathrm{F}(\Theta)|$ denotes the determinant of the expected Fisher information matrix and $D(\Theta)$ denotes the dimension of the parameter Θ.

We adopt the approach in [25], which allows equation (15) to be rewritten as the following equivalent problem:

$$\widehat{\Theta} = \arg\min_{\Theta}\{-\ln p(\Theta) - \ln p(\boldsymbol{E}|\Theta) + \frac{1}{2}\ln|\mathrm{F}_c(\Theta)| + \frac{D(\Theta)}{2}\left(1+\ln\frac{1}{12}\right)\} \qquad (16)$$

where $\mathrm{F}_c(\Theta)$ denotes the expected complete Fisher information matrix that is shown

$$\mathrm{F}_c(\Theta) = diag\left(\pi_1\sum_{i=1}^{n}\mathrm{F}_i(\Sigma_1), \cdots, \pi_K\sum_{i=1}^{n}\mathrm{F}_i(\Sigma_K), n\mathrm{M}\right)$$

$\mathrm{F}_i(\Sigma_k)$ is the Fisher information matrix of the ith observation for the kth component, and M is the Fisher information matrix of the multinomial distribution, that is $\mathrm{M} = diag(\pi_1, \cdots, \pi_K)^{-1}$ [47]. We adopt the same setting

with [25] for the distributions $p(\Sigma_k)$ $(k = 1, \cdots, K)$ and $p(\pi_1, \cdots, \pi_K)$. Furthermore, let K_{no} denote the number of non-zero components and D denote the dimensionality of covariance matrix Σ_k $(k = 1, \cdots, K)$. We have

$$\begin{aligned}&\min_{\Theta, \boldsymbol{Z}} \mathcal{L}(\Theta, \boldsymbol{Z}) \\ &= \min_{\Theta, \boldsymbol{Z}} \{ \frac{D}{2} \sum_{k:\pi_k>0} \ln\left(\frac{n\pi_k}{12}\right) + \frac{\mathrm{K}_{no}}{2} \ln\left(\frac{n}{12}\right) + \lambda \parallel \boldsymbol{Z} \parallel_F^2 \\ &\quad + \frac{\mathrm{K}_{no}(D+1)}{2} - \sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k\right)\right)\} \\ &\quad s.t.\ \pi_k \geq 0, \quad \boldsymbol{\Sigma}_k \in \mathbb{S}^+,\ k = 1, \ldots, K,\ \sum_{k=1}^{K} \pi_k = 1\end{aligned} \tag{17}$$

in our particular case. Θ is the same as (13).

Based on this preparatory work, we will discuss how to determine the numbers of the Gaussian components. Notice that the optimal problem (17) is equal to use a symmetric improper Dirichlet type prior which is conjugate to multinomial likehoods [48]. Therefore, the modified estimation mixing weight is

$$\pi_k^{mnew} = \frac{\max\{0, \sum_{n=1}^{N} \gamma_{n,k} - \frac{D}{2}\}}{\sum_{j=1}^{K} \max\{0, \sum_{n=1}^{N} \gamma_{n,j} - \frac{D}{2}\}} \tag{18}$$

From (18), it shows whether the kth component is annihilated. Once the required support $\frac{D}{2}$ can not be reached from the provide data, the mixing weight π_k equals to zero. In this scenario, the corresponding component is removed. Thus we can estimate the number of the components. As discussed in [25] and [30], we should provide both the maximum K and the minimum K. If we begin with a large K, which leads to several empty components. To avoid this singular case, we use the component-wise EM procedure [25], [49]. The crucial difference between the modified EM algorithm and the older version in that the formulation (8) is replaced by formulation (18). We list the modified EM in algorithms 2. Each iterative t runs the component-wise E and M step. If one of the components is removed, the parameters are updated accordingly, until $|\triangle Lenth|$ is below a given threshold. In this case, if the current length is not more than the $Lenth_{\min}$, the current parameters, coefficient matrix, and length are assigned to $\Theta_{\min}$, $\boldsymbol{Z}_{\min}$ and $\mathcal{L}(\Theta^{mne}, \boldsymbol{Z}^{new})$ respectively. For sake of exploring the full range of K^+, the less populated component is artificially removed, run the component-wise EM procedure again in sprit of [25].

We evaluate this tentative EM strategy on the contaminated Extended Yale Face Dataset B, and compare it with Algorithm 1 in the Supplementary Material. Although the modified EM algorithm 2 provides a way for automatically selecting the number of Gaussian components rather than by empirical value. The price it pays is more heavier computational burden than the common EM algorithm [25].

VI. Computational Complexity

In this section, we provide a concise computational complexity analysis of our proposed MoG method in a round.

Algorithm 2 Finding the Solution of (5) by Modified EM Based on MML

Initialize: input data matrix $\boldsymbol{X}$, covariance matrices $\boldsymbol{\Sigma}_k$, π_k, parameter λ, threshold value ε, initial representation matrix $\boldsymbol{Z}^{old}$, and the components number $K_{\min}$ and $K_{\max}$
output: The minimum length mixture model: The coefficient matrix $\boldsymbol{Z}$, K^+
Set $t \longleftarrow 0$, $K^+ \longleftarrow K_{\max}$, $Lenth_{\min} \longleftarrow +\infty$
while $K^+ \geq K_{\min}$ do
repeat
t = t + 1;
for k = 1 to $K_{\max}$ do
E–step: Compute $\gamma_{n,k}$:

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

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

$$\pi_k^{mnew} = \frac{\max\{0, \sum_{n=1}^{N} \gamma_{n,k} - \frac{D}{2}\}}{\sum_{j=1}^{K} \max\{0, \sum_{n=1}^{N} \gamma_{n,j} - \frac{D}{2}\}}$$

if $\pi_k^{mnew} > 0$

$$\boldsymbol{\Sigma}_k^{mnew} = \frac{1}{\gamma_{n,k}} \left(\sum_{n=1}^{N} \frac{\pi_k^{mnew} \mathcal{N}(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_k) \widetilde{\boldsymbol{e}}_n^{old} \left(\widetilde{\boldsymbol{e}}_n^{old}\right)^\top}{\sum_{j=1}^{K_{\max}} \pi_j^{mnew} \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \boldsymbol{0}, \boldsymbol{\Sigma}_j\right)} + \epsilon \boldsymbol{I} \right),$$

else
$K^+ = K^+ - 1$
end if
end for
updata $\boldsymbol{Z}$

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

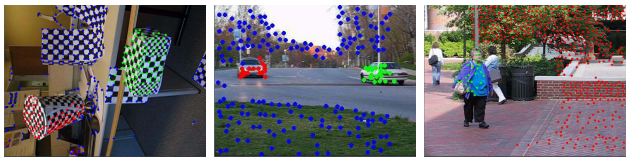
where

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

and

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

$\boldsymbol{Z}^{old} \leftarrow \boldsymbol{Z}^{new}$
Compute the optimal length using the new parameters and $\boldsymbol{Z}$
Compute the length $\mathcal{L}\left(\Theta^{mne}, \boldsymbol{Z}^{new}\right)$ via (17)
until
$|\triangle Lenth| \leq \varepsilon$
if $\mathcal{L}\left(\Theta^{mne}, \boldsymbol{Z}^{new}\right) < Lenth_{\min}$
$\mathcal{L}\left(\Theta^{mne}, \boldsymbol{Z}^{new}\right) \leftarrow Lenth_{\min}$
$\Theta_{\min} \leftarrow \Theta^{mne}$
$\boldsymbol{Z}_{\min} \leftarrow \boldsymbol{Z}^{new}$
end if
k = $\arg\min_k \{\pi_k^{mnew}\}_+$
$\pi_k^{mnew} \leftarrow 0$
$K^+ \leftarrow K^+ - 1$
end while
Using the representation matrix $\boldsymbol{Z}$ to cluster



(a) (b) (c)

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

Assume that we have collected M samples with dimension $\widehat{N}$, which contained the unknown noise that can be approached by mixture Gaussian with K components. Here, $\widehat{N}$ is the dimension after dimension reduction. To compute the determinant needs $\mathcal{O}\left(\widehat{N}^3\right)$ operations, which leads to the E step (6) costs $\mathcal{O}\left(KM\widehat{N}^3\right)$. Using the obtained result from E step, we proceed to run the M step. Thus, the step (7)and (8) need $\mathcal{O}(KM)$ and $\mathcal{O}\left(KM\widehat{N}\right)$ operations respectively. Especially, the step (10) costs due to the matrix inversion. Now, it is easy to see that the modified model (17) needs heavier computational burden than the prime model (5) because the additional mission that how to annihilate the Gaussian components.

VII. Experiments

In this section, we evaluate the proposed MoG Regression model for clustering on the Hopkins 155 database [50], the Rotated MNIST Dataset [51], the AR database [52], and the Extended Yale Face Dataset B [53]. Experimental results demonstrate that the proposed method is valid and robust to noise in motion segmentation, handwritten digits clustering, and complex face clustering.

We also run SSC [16], LRR [17], LSR [18], CASS [20], and CIL2 [27] on these datasets. Meanwhile, we tune the parameters of each method so that every model achieves its best performance. We use PCA method to project data matrix to a low dimension space by maintaining at least 90% energy for all concerned databases. Especially, we determine the regularization parameter and K by empirical value in our model (5). The clustering accuracy [16] is employed in quantitative evaluation. The comparison results show that our approach outperforms the mentioned five state-of-the-art methods.

A. Hopkins 155 Database

In this situation, we will recorder the average accuracy over 2 motions and 3 motions videos on Hopkins 155 motion segmentation database respectively.

After resorting to PCA on the data matrix, we test the proposed MoG Regression method on each video sequence. Some motion segmentation visually results of our approach are shown in Figure 2, where motions of different objects and background trajectories can be accurately segmented.

Table II lists the average clustering accuracies of different methods. We can see that MoG Regression achieves significantly higher accuracies than the other methods.

TABLE II
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**

TABLE III
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**

TABLE IV
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**

B. MNIST-Back-Rand Dataset

In this subsection we randomly select 10 images for each digit of MNIST-back-rand database to build a subset, thus the candidate dataset contains 100 samples. The experiment results are reported in Table III, which declares that the advantage of our method is notable. This experiment also shows that when the data are corrupted with non-Gaussian or complex noise, the proposed method is more capable of clustering the underlying subspaces with the help of MoG.

C. AR Dataset

The AR database is another challenging database for subspace clustering mission. We design two subspace clustering tasks by selecting first 5 and 10 subjects based on this dataset, respectively. The clustering results on the AR database of different algorithms are recorded in Table IV. We can see that the performance of MoG Regression method for subspace clustering is superior to the other methods in both clustering tasks. This is because MoG Regression has both a strong grouping effect on this challenging database and reasonably model the noise, which can be seen in Figure 1.

D. Extended Yale Face Dataset B

In order to further show the ability of MoG model for describing the noises, we add the noise on each image of Extended Yale Database B database by replacing randomly its pixels with samples from a uniform distribution on the interval from 0 to 255 [27], and the percentage of corrupted pixels range from 10% to 100%. In order to reduce the computational cost and memory requirements, we tailor the grayscale images to a resolution of 32×32 pixels. The clustering accuracies of all methods on the corrupted Extended Yale B database

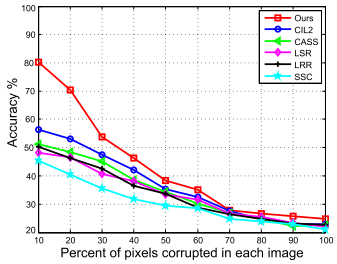


Fig. 3. The clustering accuracies (%) with pixel corruption on the Extended Yale B database.

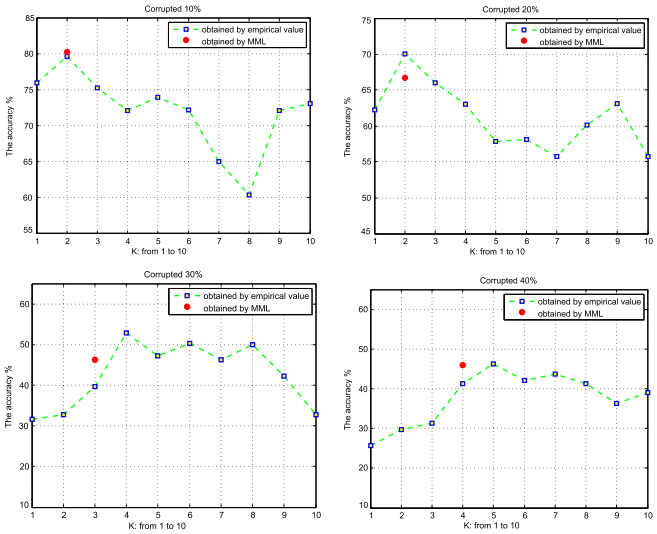


Fig. 4. Comparison of the empirical value K with the estimated K based on MML, and their clustering accuracies

are reported by Figure 3. From Figure 3 we can see that the proposed method performs much better when face images are randomly contaminated at a level from 10% to 40%, exhibiting better adaptability and greater robustness in noise situation. When the percentage of corrupted pixels is over than 60%, the discriminative information are destructive damaged, thus will weaken the performance of all methods.

When the pixels are corrupted over 40%, the accuracies are too low. So, we evaluate K, accuracies obtained by Algorithm 1 and Algorithm 2 respectively on the data set that the pixels are corrupted not more than 40%. Fig 4 reveals the experimental results. We take $K_{\min} = 1$ and $K_{\max} = 25$ for iteration. It can be seen that, the number of Gaussian components of both two algorithms are increase when the level of the pixels corruption is high. While the accuracies obtained by these two algorithms go down when the pixels corruption has high level. In the case of 10% and 20% pixels are corrupted, the estimated K is 2 that equals to the empirical value. In case of 30% and 40% pixels are corrupted, the estimated values of K are 3 and 4 respectively. In this case the empirical values of K are 4 and 5. The accuracies that obtained by empirical values K have relatively small deviations from the accuracies that obtained by MML based. If the data contains simple unknown noise we advocate

TABLE V
THE CLUSTERING ACCURACIES (%) OF UNIMODAL GAUSSIAN SCENARIO

SSC	LRR	LSR	CASS	CIL2	OursI	OursII
78.1	72	70.05	81.05	72.32	**86**	**81.76**

TABLE VI
THE CLUSTERING ACCURACIES (%) OF MIXTURE GAUSSIAN SCENARIO

SSC	LRR	LSR	CASS	CIL2	OursI	OursII
69.41	65.77	60.52	71.94	66.73	**80.87**	**78.02**

to use the empirical value K. When the data suffers from serious corruption, we need more Gaussian components to approximate the unknown noises. In this scenario using the explorative method is seems blind. Notice that the accuracies between MML based EM algorithm and the common EM algorithm are different under the optimal K estimated by MML based EM algorithm. The reason is that the update mechanism of mixing weights and covariance matrices MML based is different from the common EM algorithm.

E. Estimation K Under the Known Noise

In this subsection, we intend to investigate whether the model (17) can find the proper the number of Gaussian components. We use the Gaussian noise $\mathcal{N}(0.01,\ 0.02)$ corrupt the Extended Yale Face Dataset B, and evaluate the related algorithms. In another experiment, we use a mixture noise which is a superposition of $\mathcal{N}(0.01,\ 0.02)$ and $\mathcal{N}(0.03,\ 0.01)$ to corrupt the Extended Yale Face Dataset B, and proceed to evaluate the concerned models. We select the first 5 subjects of mentioned dataset in these two experiments. The experiment results are reported in different tables. We use OursI and OursII to denote the model (5) and (17) respectively in the following tables. In the unimodal Gaussian scenario, the estimated K is 2 by OursII which has a deviation from the ground truth 1. In the mixture Gaussian scenario, the estimated K equals to 2 by OursII which is just the ground truth. Although, the modified model (17) may yield to the imprecise K, the model (17) provides an illuminating strategy to find the proper K.

From the two tables we can see that the proposed model (5) is superior to the other models. And its modified version (17) provides a satisfied performance.

VIII. CONCLUSIONS

In this paper, we propose a new subspace clustering method by employing the MoG model to describe the distribution of complex noise. In fact, the SSC, LRR, LSR, CASS, CIL2, and MoG are all reconstruction based methods for subspace clustering by computing a reconstruction matrix which is also called coefficient matrix. Using the model (1) can be written in a unified form. The models SSC, LRR, LSR, and CASS describe the noise as the unimodal Gaussian or sparse type, while the CIL2 borrows the idea of [28] which deals with the non-Gaussian noise especially for impulsive noise. In real scenario, the noise goes beyond the Gaussian or impulsive

types. Inspired by the property of mix Gaussian distribution, we use MoG model to group the subspaces. On one hand, our proposed model can character the complex noise by the property mix Gaussian model, on the other hand we give the theoretical analysis shows that the MoG Regression model maintains the grouping effect. The experiments on motion segmentation, handwritten digits clustering, and complex face clustering demonstrate the superiority of our proposed method, regarding stability and robustness in handling general noise, over the state-of-the-art subspace clustering methods, SSC, LRR, LSR, CASS, and CIL2 which assume Gaussian or sparse noise or impulsive type. In the future, we will deal with the accelerating aspect of the solution for MoG Regression.

APPENDIX

The detail proof of **theorem** 1 can be found in [1]. We thus omit it.

Proof of **theorem**2: Let

$$Q\left(z_i\right) = -\mathcal{L}\left(z_i\right) + \lambda \|z_i\|_F^2.$$

where $-\mathcal{L}\left(z_i\right)$ replaces $-\sum_{n=1}^{N} \ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\widetilde{\boldsymbol{e}}_n | \mathbf{0}, \boldsymbol{\Sigma}_k\right)\right)$ for simplification.

Since z_i^{lm} is the local minimizer of $Q\left(z_i\right)$, which reads

$$\mathbf{0} = \frac{\partial Q\left(z_i\right)}{\partial z_i}|_{z_i = z_i^{lm}} = -\frac{\partial \mathcal{L}\left(z_i\right)}{\partial z_i}|_{z_i = z_i^{lm}} + 2\lambda z_i^{lm} \tag{19}$$

We let $\frac{\partial \mathcal{L}\left(z_i^{lm}\right)}{\partial z_i}$ denote $\frac{\partial \mathcal{L}\left(z_i\right)}{\partial z_i}|_{z_i = z_i^{lm}}$. Thanks for the Taylor expansion, we obtain

$$\begin{aligned}
&\frac{\partial \mathcal{L}\left(z_i^{lm}\right)}{\partial z_i} \\
&= \frac{\partial \mathcal{L}\left(z_i^t\right)}{\partial z_i} + \left(\frac{\partial^2 \mathcal{L}\left(z_i^t\right)}{\partial z_i^2}\right)^{\top}\left(z_i^{lm} - z_i^t\right) \\
&\quad + \frac{1}{2}\left(I \otimes \left(z_i^{lm} - z_i^t\right)\right)^{\top}\left(\frac{\partial^3 \mathcal{L}\left(z_i^{nt}\right)}{\partial z_i^3}\right)^{\top}\left(z_i^{lm} - z_i^t\right)
\end{aligned} \tag{20}$$

where z_i^{nt} belongs to the ball neighborhood of z_i^t and I is an identity matrix of size $M \times M$. Using (20), we rearrange the equation (19) that yields to

$$\begin{aligned}
\mathbf{0} = &-\frac{1}{n}\frac{\partial \mathcal{L}\left(z_i^t\right)}{\partial z_i} - \frac{1}{n}\left(\frac{\partial^2 \mathcal{L}\left(z_i^t\right)}{\partial z_i^2}\right)^{\top}\left(z_i^{lm} - z_i^t\right) \\
&- \frac{1}{2n}\left(I \otimes \left(z_i^{lm} - z_i^t\right)\right)^{\top}\left(\frac{\partial^3 \mathcal{L}\left(z_i^{nt}\right)}{\partial z_i^3}\right)^{\top}\left(z_i^{lm} - z_i^t\right) \\
&+ \frac{2\lambda}{n} z_i^{lm}
\end{aligned} \tag{21}$$

Now employing the regularity conditions $(A) - (C)$ [26], which reads

$$\frac{1}{n}\left(\frac{\partial^2 \mathcal{L}\left(z_i^t\right)}{\partial z_i^2}\right)^{\top} = -I\left(z_i^t\right) + o_p\left(1\right)$$

where $I\left(z_i^t\right)$ is the information matrix at z_i^t.

$$\frac{1}{n}\left(\frac{\partial^3 \mathcal{L}\left(z_i^t\right)}{\partial z_i^3}\right)^{\top} = \mathcal{O}_p\left(1\right)$$

Notice the consistency of $\theta_{\mathbf{Z}}^{ini} - \theta_{\mathbf{Z}}^t = \mathcal{O}_p\left(N^{\frac{-1}{2}}\right)$, which means that $\theta_{\mathbf{Z}}^{ini} - \theta_{\mathbf{Z}}^t = o_p\left(1\right)$. Thus

$$\frac{1}{2n}\left(I \otimes \left(z_i^{lm} - z_i^t\right)\right)^{\top}\left(\frac{\partial^3 \mathcal{L}\left(z_i^{nt}\right)}{\partial z_i^3}\right)^{\top} = o_p\left(1\right)$$

After rearrange (21), we get

$$\begin{aligned}
-\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}\left(z_i^t\right)}{\partial z_i} = &\left(-I\left(z_i^t\right) + o_p\left(1\right)\right)\sqrt{n}\left(z_i^{lm} - z_i^t\right) \\
&+ \frac{2\lambda}{\sqrt{n}} z_i^{lm}
\end{aligned} \tag{22}$$

Note that $\frac{\lambda}{\sqrt{N}} = o\left(1\right)$ by assumed condition, $I\left(z_i^t\right)$ is the mentioned information matrix. Using the center limit theory, we get $-\frac{1}{\sqrt{n}}\frac{\partial \mathcal{L}\left(z_i^t\right)}{\partial z_i} \xrightarrow{d} \mathcal{N}\left(0, I\left(z_i^t\right)\right)$, which leads to the conclusion.

REFERENCES

[1] B. Li, Y. Zhang, Z. Lin, and H. Lu, "Subspace clustering by mixture of Gaussian regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2094–2102.

[2] A. Y. Yang, J. Wright, Y. Ma, and S. S. Sastry, "Unsupervised segmentation of natural images via lossy data compression," *Comput. Vis. Image Understand.*, vol. 110, no. 2, pp. 212–225, 2008.

[3] R. Vidal and R. Hartley, "Motion segmentation with missing data using powerfactorization and GPCA," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2. Jun./Jul. 2004, pp. II-310–II-316.

[4] J. Ho, M.-H. Yang, J. Lim, K.-C. Lee, and D. Kriegman, "Clustering appearances of objects under varying illumination conditions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, Jun. 2003, pp. I-11–I-18.

[5] W. Hong, J. Wright, K. Huang, and Y. Ma, "Multiscale hybrid linear models for lossy image representation," *IEEE Trans. Image Process.*, vol. 15, no. 12, pp. 3655–3671, Dec. 2006.

[6] Y. Ma, A. Y. Yang, H. Derksen, and R. Fossum, "Estimation of subspace arrangements with applications in modeling and segmenting mixed data," *SIAM Rev.*, vol. 50, no. 3, pp. 413–458, 2008.

[7] R. Vidal, Y. Ma, and S. Sastry, "Generalized principal component analysis (GPCA)," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 12, pp. 1945–1959, Dec. 2005.

[8] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *Int. J. Comput. Vis.*, vol. 29, no. 3, pp. 159–179, 1998.

[9] P. S. Bradley and O. L. Mangasarian, "K-plane clustering," *J. Global Optim.*, vol. 16, no. 1, pp. 23–32, 2000.

[10] P. K. Agarwal and N. H. Mustafa, "K-means projective clustering," in *Proc. 23rd ACM SIGACT-SIGMOD-SIGART Symp. Principles Database Syst.*, Paris, France, 2004, pp. 155–165.

[11] Y. Ma, H. Derksen, W. Hong, and J. Wright, "Segmentation of multivariate mixed data via lossy data coding and compression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 9, pp. 1546–1562, Sep. 2007.

[12] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.

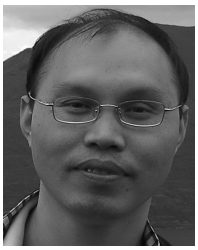
[13] A. Y. Yang, S. R. Rao, and Y. Ma, "Robust statistical estimation and segmentation of multiple subspaces," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop (CVPRW)*, Jun. 2006, p. 99.

[14] U. von Luxburg, "A tutorial on spectral clustering," *Statist. Comput.*, vol. 17, no. 4, pp. 395–416, 2007.

[15] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 94–106.

[16] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[17] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[18] 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 *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 347–360.
[19] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, Aug. 2000.
[20] C. Lu, J. Feng, Z. Lin, and S. Yan, "Correlation adaptive subspace segmentation by trace lasso," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1345–1352.
[21] F. R. K. Chung, *Spectral Graph Theory*, vol. 92. Providence, RI, USA: AMS, 1997.
[22] Q. Zhao, D. Meng, Z. Xu, W. Zuo, and L. Zhang, "Robust principal component analysis with complex noise," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 1–9.
[23] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *J. Roy. Statist. Soc., B Statist. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005.
[24] J. J. Oliver, R. A. Baxter, and C. S. Wallace, "Unsupervised learning using MML," in *Proc. ICML*, 1996, pp. 1–9.
[25] M. A. T. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 381–396, Mar. 2002.
[26] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *J. Amer. Statist. Assoc.*, vol. 96, no. 456, pp. 1348–1360, 2001.
[27] C. Lu, J. Tang, M. Lin, L. Lin, S. Yan, and Z. Lin, "Correntropy induced L2 graph for robust subspace clustering," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1801–1808.
[28] W. Liu, P. P. Pokharel, and J. C. Principe, "Correntropy: Properties and applications in non-Gaussian signal processing," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5286–5298, Nov. 2007.
[29] X. He, D. Cai, Y. Shao, H. Bao, and J. Han, "Laplacian regularized Gaussian mixture model for data clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 9, pp. 1406–1418, Sep. 2011.
[30] I. D. Gebru, X. Alameda-Pineda, F. Forbes, and R. Horaud, "EM algorithms for weighted-data clustering with application to audio-visual scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 12, pp. 2402–2415, Dec. 2016.
[31] J. Xia, F. Liang, and Y. M. Wang, "On clustering fMRI using potts and mixture regression models," in *Proc. IEEE Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Sep. 2009, pp. 4795–4798.
[32] G. K. Befekadu, M. G. Tadesse, T.-H. Tsai, and H. W. Ressom, "Probabilistic mixture regression models for alignment of LC-MS data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 8, no. 5, pp. 1417–1424, Sep./Oct. 2011.
[33] V. P. Oikonomou and K. Blekas, "An adaptive regression mixture model for fMRI cluster analysis," *IEEE Trans. Med. Imag.*, vol. 32, no. 4, pp. 649–659, Apr. 2013.
[34] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Roy. Statist. Soc. B Methodol.*, vol. 39, no. 1, pp. 1–38, 1977.
[35] G. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, vol. 382. Hoboken, NJ, USA: Wiley, 2007.
[36] C. F. J. Wu, "On the convergence properties of the EM algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, 1983.
[37] L. Xu and M. I. Jordan, "On convergence properties of the EM algorithm for Gaussian mixtures," *Neural Comput.*, vol. 8, no. 1, pp. 129–151, 1996.
[38] D. Nettleton, "Convergence properties of the EM algorithm in constrained parameter spaces," *Can. J. Statist.*, vol. 27, no. 3, pp. 639–648, 1999.
[39] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 849–856.
[40] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *J. Comput. Biol.*, vol. 10, no. 6, pp. 961–980, 2003.
[41] N. Städler, P. Bühlmann, and S. van de Geer, "ℓ_1-penalization for mixture regression models," *Test*, vol. 19, no. 2, pp. 209–285, 2010.
[42] E. L. Lehmann, *Elements of Large-Sample Theory*. New York, NY, USA: Springer, 1999.
[43] C. S. Wallace and P. R. Freeman, "Estimation and inference by compact coding," *J. Roy. Statist. Soc. B Methodol.*, vol. 49, no. 3, pp. 240–265, 1987.
[44] C. S. Wallace and D. L. Dowe, "MML clustering of multi-state, poisson, von Mises circular and Gaussian distributions," *Statist. Comput.*, vol. 10, no. 1, pp. 73–83, 2000.
[45] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*, vol. 15. Singapore: World scientific, 1998.
[46] C. S. Wallace and D. L. Dowe, "Minimum message length and kolmogorov complexity," *Comput. J.*, vol. 42, no. 4, pp. 270–283, 1999.
[47] D. M. Titterington, A. F. M. Smith, and U. E. Makov, *Statistical Analysis of Finite Mixture Distributions*. Hoboken, NJ, USA: Wiley, 1985.
[48] J. Bernardo, *Bayesian Theory* (Wiley Series in Probability and Statistics). Hoboken, NJ, USA: Wiley, 2000, p. 586.
[49] G. Celeux, S. Chretien, F. Forbes, and A. Mkhadri, "A component-wise EM algorithm for mixtures," *J. Comput. Graph. Statist.*, vol. 10, no. 4, pp. 697–712, 2012.
[50] R. Tron and R. Vidal, "A benchmark for the comparison of 3-D motion segmentation algorithms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[51] 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 *Proc. Int. Conf. Mach. Learn.*, 2007, pp. 473–480.
[52] A. M. Martínez and R. Benavente, "The AR face database," CVC, New Delhi, India, Tech. Rep. 24, 1998.
[53] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.
[54] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.
[55] S. Puntanen and G. P. H. Styan, "Schur complements in statistics and probability," in *The Schur Complement and Its Applications*. New York, NY, USA: Springer, 2005, pp. 163–226.
[56] Y. Tian and Y. Takane, "The inverse of any two-by-two nonsingular partitioned matrix and three matrix inverse completion problems," *Comput. Math. Appl.*, vol. 57, no. 8, pp. 1294–1304, 2009.

Baohua Li received the master's degree in computational mathematics from Lanzhou University, Lanzhou, China, in 2009. He is currently pursuing the Ph.D. degree with the School of Mathematical Sciences, Dalian University of Technology. His research interests include optimization, clustering, and high-dimensional data processing.



Huchuan Lu (SM'12) received the M.Sc. degree in signal and information processing and the Ph.D. degree in system engineering from the Dalian University of Technology (DUT), China, in 1998 and 2008, respectively. He has been with the School of Information and Communication Engineering, DUT, as a Faculty Member since 1998 and as a Professor since 2012. His research interests include computer vision and pattern recognition. In recent years, he focuses on visual tracking and segmentation. He currently serves as an Associate Editor of the IEEE TRANSACTIONS ON CYBERNETICS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY.

Ying Zhang received the M.S. degree in information and communication engineering from the Dalian University of Technology in 2015, where she is currently pursuing the Ph.D. degree with the School of Signal and Information Processing. Her research interests include person re-identification, anomaly detection, and saliency detection.

Zhouchen Lin (M'00–SM'08–F'18) received the Ph.D. degree in applied mathematics from Peking University in 2000. He is currently a Professor with the Key Laboratory of Machine Perception, School of Electronics Engineering and Computer Science, Peking University. His research interests include computer vision, image processing, machine learning, pattern recognition, and numerical optimization. He is a fellow of IAPR. He was the Area Chair of CVPR 2014/2016, ICCV 2015, and NIPS 2015 and a Senior Program Committee Member of AAAI 2016/2017/2018 and IJCAI 2016/2018. He is an Associate Editor of the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and the *International Journal of Computer Vision*.

Wei Wu received the bachelor's and master's degrees from Jilin University, Changchun, China, in 1974 and 1981, respectively, and the Ph.D. degree from Oxford University, Oxford, U.K., in 1987. He is currently with the School of Mathematical Sciences, Dalian University of Technology, Dalian, China. He has published four books and 90 research papers. His current research interests include learning methods of neural networks.