Journal of Computational Mathematics
Vol.36, No.3, 2018, 426–440.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1708-m2016-0830

REDUCED-RANK MODELING FOR HIGH-DIMENSIONAL MODEL-BASED CLUSTERING*

Lei Yang
Department of Environmental Medicine, New York University, New York, NY, USA
Email: ly888@nyu.edu
Junhui Wang
Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong
Email: j.h.wang@cityu.edu.hk
Shiqian Ma
Department of Mathematics, University of California, Davis, CA 95616, USA
Email: sqma@math.ucdavis.edu

Abstract

Model-based clustering is popularly used in statistical literature, which often models the data with a Gaussian mixture model. As a consequence, it requires estimation of a large amount of parameters, especially when the data dimension is relatively large. In this paper, reduced-rank model and group-sparsity regularization are proposed to equip with the model-based clustering, which substantially reduce the number of parameters and thus facilitate the high-dimensional clustering and variable selection simultaneously. We propose an EM algorithm for this task, in which the M-step is solved using alternating minimization. One of the alternating steps involves both nonsmooth function and nonconvex constraint, and thus we propose a linearized alternating direction method of multipliers (ADMM) for solving it. This leads to an efficient algorithm whose subproblems are all easy to solve. In addition, a model selection criterion based on the concept of clustering stability is developed for tuning the clustering model. The effectiveness of the proposed method is supported in a variety of simulated and real examples, as well as its asymptotic estimation and selection consistencies.

Mathematics subject classification: 62-07, 90C30
Key words: Clustering, Gaussian mixture model, Group Lasso, ADMM, Reduced-rank model.

1. Introduction

Cluster analysis is to assign observations into a number of clusters so that observations within the same cluster are more similar compared with those in different clusters. In literature, the similarity is often measured by certain pre-specified forms of distance, leading to various clustering algorithms such as K-means, hierarchical clustering, and spectral clustering [17]. On the contrary, model-based clustering [28] approaches cluster analysis from a probabilistic perspective. It models the data as a finite mixture of parametric distributions, and converts cluster analysis into estimation of the unknown parameters.

The model-based clustering is popular in literature, especially when the data dimension is small [13]. However, challenges arise when the data dimension becomes relatively large,

* Received January 8, 2017 / Revised version received May 3, 2017 / Accepted August 9, 2017 / Published online March 28, 2018 /

due to the large number of unknown parameters [8]. To circumvent the difficulty, several remedial treatments have been proposed. For instances, [14] and [16] propose to first conduct principal component analysis to reduce the dimension and then conduct model-based clustering on the low-dimensional space. [20] and [11] propose to conduct a forward stepwise or pre-screening variable selection for the model-based clustering based on a BIC criterion. [18] restricts the covariance matrix of each Gaussian mixture component to be diagonal. Such treatments can substantially alleviate the computation cost, yet may be too restrictive to capture all the important information. More recently, [7] developed a sparse Fisher-EM algorithm, which combines the model-based clustering and Fisher's linear discriminant analysis and can conduct clustering in high-dimensional space.

In this paper, reduced-rank model and group-sparsity regularization are proposed to equip with the model-based clustering to tackle the high-dimensional cluster analysis and variable selection simultaneously. The key motivation is to project the original data onto a low-dimensional space so that the projected data follows a finite mixture of Gaussian distributions. A group Lasso penalty is enforced on each row of the projection matrix to assure that the non-informative variables can be automatically identified if they do not contribute to any of the reduced dimension. We propose an EM algorithm for this task. One of the steps in the EM algorithm involves both nonsmooth function and nonconvex constraint, and thus we propose a linearized ADMM for solving it. This leads to an efficient algorithm whose subproblems are all easy to solve. To optimally determine the tuning parameters, a stability-based selection criterion [24, 26] is employed, which searches for the tuning parameters that lead to the most stable clustering algorithm. The advantages of the proposed methods are demonstrated in a variety of the numerical examples, as well as their asymptotic estimation and selection consistencies.

The rest of the article is organized as follows. Section 2 introduces the reduced-rank Gaussian mixture model, as well as the computing algorithms including the expectation-maximization (EM) algorithm and the linearized ADMM algorithm. Section 3 presents the stability-based selection criterion for optimally determining the tuning parameters. Section 4 establishes the asymptotic estimation and selection consistencies of the proposed methods. Numerical simulation and real applications are presented in Section 5. Section 6 gives conclusions. The appendix is devoted to technical proofs and a counter example for the likelihood-based criteria.

2. Reduced-rank Model-based Clustering

In a typical clustering setup, the training set consists of n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ independently and identically distributed. The primary goal is to cluster $\mathbf{x}_i$'s into a set of subgroups so that the observations within the same subgroup are more similar than those in different subgroups.

2.1. Reduced-rank Gaussian mixture model

Classical model-based clustering models each subgroup by a distinct probability distribution, say Gaussian distribution, and assumes that the data follows

$$\mathbf{X} \sim \sum_{k=1}^{K} \pi_k N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\pi_k \geq 0$ is the mixture weight, $\sum_{k=1}^{K} \pi_k = 1$, and $N_p(\cdot, \cdot)$ denotes a p-variate Gaussian distribution. Note that this model consists of of $K-1+Kp+Kp(p+1)/2$ unknown parameters, which may require high computational cost when p is relatively large.

To shrink the number of parameters, the reduced-rank Gaussian mixture model assumes that the data follows

$$\mathbf{B}^T\mathbf{X} \sim \sum_{k=1}^{K} \pi_k N_r(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{2.1}$$

where $\mathbf{B}$ is a $p \times r$ orthogonal reduced-rank matrix with $r \ll p$. The number of its unknown parameters becomes $pr+K-1+Kr+Kr(r+1)/2$, which can substantially reduce the number of parameters when $r \ll p$ and thus facilitate the high-dimensional model-based clustering.

Note that the reduced-rank model may be unidentifiable since the model in (2.1) is equivalent to $(\mathbf{B}\,\boldsymbol{\Gamma})^T\mathbf{X} \sim \sum_{k=1}^{K} \pi_k N_r(\boldsymbol{\Gamma}^T\,\boldsymbol{\mu}_k, \boldsymbol{\Gamma}^T\,\boldsymbol{\Sigma}_k\,\boldsymbol{\Gamma})$ for any $r \times r$ orthogonal matrix $\boldsymbol{\Gamma}$. Furthermore, $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ can be permuted without changing the Gaussian mixture model. However, the primary goal of the reduced-rank model-based clustering is to detect the clustering structure, and the orthogonal matrix $\boldsymbol{\Gamma}$ or the permutation of $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ does not change the Mahalanobis distance and thus the clustering structure.

2.2. EM algorithm

Let $\boldsymbol{\Theta} = \mathrm{vec}(\mathbf{B}, \pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ denote the vector of all unknown parameters in (2.1). Given the training data, the log-likelihood of the reduced-rank Gaussian mixture model is

$$l(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \phi(\mathbf{B}^T\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right), \tag{2.2}$$

where $\phi(\mathbf{u}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-r/2} |\,\boldsymbol{\Sigma}\,|^{-1/2} \exp\big\{ -(\mathbf{u}-\boldsymbol{\mu})^T\,\boldsymbol{\Sigma}^{-1}(\mathbf{u}-\boldsymbol{\mu})/2 \big\}$ denotes the probability density function of a r-dimensional Gaussian distribution. In literature, the EM algorithm [10] is commonly used to maximize the log-likelihood function in (2.2). It introduces a set of z_{ik}'s indicating whether the observation $\mathbf{x}_i$ is from the k-th cluster. Specifically, $z_{ik} = 1$ if $\mathbf{x}_i$ is from the k-th cluster, and 0 otherwise. Apparently, z_{ik}'s are missing from the training set. If z_{ik}'s were observed, the complete log-likelihood of $(\mathbf{x}_i, \mathbf{z}_i)$ is

$$l_{comp}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} z_{ik} \big(\log \pi_k + \log \phi(\mathbf{B}^T\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) \right). \tag{2.3}$$

When p is large, it is generally believed that only a small portion of the covariates are informative for the clustering structure while others are redundant. To achieve the sparsity in the estimation, sparsity-encouraging penalty terms on $\mathbf{B}$ can be employed and the penalized log-likelihood functions become

$$l_P(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \phi(\mathbf{B}^T\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right) - \lambda J(\mathbf{B}), \tag{2.4}$$

$$l_{comp,P}(\boldsymbol{\Theta}) = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} z_{ik} \big(\log \pi_k + \log \phi(\mathbf{B}^T\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) \right) - \lambda J(\mathbf{B}). \tag{2.5}$$

The penalty terms can take various forms, including the Lasso penalty [25], the SCAD penalty [12], the adaptive Lasso penalty [30], the group Lasso penalty [29] on each row of $\mathbf{B}$, and the

truncated l_1 penalty [22]. In this paper, we adopt the adaptive group Lasso penalty

$$J(\mathbf{B}) = \sum_{j=1}^{p} \frac{\|\mathbf{B}_j\|_2}{\|\widetilde{\mathbf{B}}_j\|_2},$$

where $\mathbf{B}_j$ denotes the j-th row of $\mathbf{B}$, and $\widetilde{\mathbf{B}}_j$ is a consistent estimate of $\mathbf{B}_j$ and is obtained by maximizing (2.2). The adaptive group Lasso penalty selects the informative variables in an "all-in-or-all-out" fashion, in the sense that the j-th variable is regarded as non-informative only when all components of $\mathbf{B}_j$ are zero, or equivalently, $\|\mathbf{B}_j\|_2 = 0$.

The EM algorithm consists of the E-step and the M-step, and proceeds as follows. The E-step computes

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}) = E_{\boldsymbol{\Theta}^{(t)}}(l_{comp,P}(\boldsymbol{\Theta})|\mathbf{x}_i),$$

where the expectation is taken with respect to the missing data z_{ik}'s. In specific,

$$Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}) = \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \tau_{ik}^{(t)} \big(\log \pi_k + \log \phi(\mathbf{B}^T \mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \big) \right) - \lambda J(\mathbf{B}), \tag{2.6}$$

where

$$\tau_{ik}^{(t)} = E_{\boldsymbol{\Theta}^{(t)}}(Z_{ik}|\mathbf{x}_i) = \frac{\pi_k^{(t)} \phi((\mathbf{B}^{(t)})^T \mathbf{x}_i; \boldsymbol{\mu}_k^{(t)}, \boldsymbol{\Sigma}_k^{(t)})}{\sum_{j=1}^{K} \pi_j^{(t)} \phi((\mathbf{B}^{(t)})^T \mathbf{x}_i; \boldsymbol{\mu}_j^{(t)}, \boldsymbol{\Sigma}_j^{(t)})}$$

is the estimated posterior probability of $\mathbf{x}_i$ coming from the k-th Gaussian component.

The M-step maximizes $Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)})$ with respect to $\boldsymbol{\Theta}$ to get the updated parameter estimate $\boldsymbol{\Theta}^{(t+1)}$. In specific,

$$\boldsymbol{\Theta}^{(t+1)} = \underset{\boldsymbol{\Theta}}{\operatorname{argmax}} \ \{Q(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(t)}) \mid \mathbf{B}^T \mathbf{B} = \mathbf{I}_r\}. \tag{2.7}$$

We propose to solve (2.7) iteratively by alternating between $\mathbf{B}$ and $(\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$. The detailed procedure is as follows.

We use ℓ to denote the iteration counter. When $\mathbf{B}^{(t+1,\ell)}$ is fixed, by denoting $N_k = \sum_{i=1}^{n} \tau_{ik}^{(t)}$, we update $\pi_k^{(t+1,\ell+1)}$, $\boldsymbol{\mu}_k^{(t+1,\ell+1)}$ and $\boldsymbol{\Sigma}_k^{(t+1,\ell+1)}$ as follows:

$$\begin{aligned}
\pi_k^{(t+1,\ell+1)} &:= \frac{N_k}{n}, \\
\boldsymbol{\mu}_k^{(t+1,\ell+1)} &:= \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik}^{(t)} (\mathbf{B}^{(t+1,\ell)})^T \mathbf{x}_i, \\
\boldsymbol{\Sigma}_k^{(t+1,\ell+1)} &:= \frac{1}{N_k} \sum_{i=1}^{n} \tau_{ik}^{(t)} \Big((\mathbf{B}^{(t+1,\ell)})^T \mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1,\ell)} \Big) \Big((\mathbf{B}^{(t+1,\ell)})^T \mathbf{x}_i - \boldsymbol{\mu}_k^{(t+1,\ell+1)} \Big)^T.
\end{aligned} \tag{2.8}$$

When $(\pi_k^{(t+1,\ell+1)}, \boldsymbol{\mu}_k^{(t+1,\ell+1)}, \boldsymbol{\Sigma}_k^{(t+1,\ell+1)})$ is fixed, we update $\mathbf{B}^{(t+1,\ell+1)}$ by solving

$$\mathbf{B}^{(t+1,\ell+1)} := \underset{\mathbf{B}^T \mathbf{B} = \mathbf{I}_r}{\operatorname{argmax}} \ \sum_{i=1}^{n} \sum_{k=1}^{K} \tau_{ik}^{(t)} \log \phi(\mathbf{B}^T \mathbf{x}_i; \boldsymbol{\mu}_k^{(t+1,\ell+1)}, \boldsymbol{\Sigma}_k^{(t+1,\ell+1)}) - \lambda J(\mathbf{B}). \tag{2.9}$$

As a result, the overall EM algorithm can be described as in Algorithm 2.1.

Algorithm 2.1. The EM Algorithm

Require: Given data $\mathbf{X}$, and initial point $\boldsymbol{\Theta}^{(0)}$
  for $t = 0, 1, \ldots$ **do**
    Set $(\mathbf{B}^{(t+1,0)}, \pi_k^{(t+1,0)}, \boldsymbol{\mu}_k^{(t+1,0)}, \boldsymbol{\Sigma}_k^{(t+1,0)}) := \boldsymbol{\Theta}^{(t)}$
    while Stopping criterion is not met **do**
      Update $(\pi^{(t+1,\ell+1)}, \boldsymbol{\mu}^{(t+1,\ell+1)}, \boldsymbol{\Sigma}^{(t+1,\ell+1)})$ by (2.8)
      Update $\mathbf{B}^{(t+1,\ell+1)}$ by (2.9)
      $\ell \leftarrow \ell + 1$
    end while
    Set $\boldsymbol{\Theta}^{(t+1)} := (\mathbf{B}^{(t+1,\ell)}, \pi_k^{(t+1,\ell)}, \boldsymbol{\mu}_k^{(t+1,\ell)}, \boldsymbol{\Sigma}_k^{(t+1,\ell)})$
  end for

Note that the only difficult task in Algorithm 2.1 is to solve (2.9). In the next section, we propose a linearized ADMM algorithm for solving it.

2.3. A Linearized ADMM for Solving (2.9)

In this section, we propose a linearized ADMM algorithm to solve (2.9). Note that (2.9) is a nonconvex problem, so usually one only expects to find a stationary point. This problem is difficult to solve because it involves both nonsmooth function $J(\mathbf{B})$ and nonconvex constraints $\mathbf{B}^T\mathbf{B} = I$. One way to deal with this kind of difficulty is to introduce auxiliary variable to split the difficulties. Specifically, for ease of notation, we rewrite (2.9) as the following form:

$$\min_{\mathbf{B}} f(\mathbf{B}) + \lambda J(\mathbf{B}), \quad \text{s.t.}, \mathbf{B}^T\mathbf{B} = I, \tag{2.10}$$

where $f(\mathbf{B}) := -\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{(t)}\log\phi(\mathbf{B}^T\mathbf{x}_i; \boldsymbol{\mu}_k^{(t+1,\ell+1)}, \boldsymbol{\Sigma}_k^{(t+1,\ell+1)})$, and also note that f is smooth. By introducing an auxiliary variable $\mathbf{C}$, (2.10) can be equivalently written as:

$$\begin{array}{ll} \min_{\mathbf{B},\mathbf{C}} & f(\mathbf{C}) + \lambda J(\mathbf{C}) \\ \text{s.t.}, & \mathbf{B} = \mathbf{C}, \\ & \mathbf{B}^T\mathbf{B} = I. \end{array} \tag{2.11}$$

By associating an Lagrange multiplier Λ to the linear equality constraint $\mathbf{B} = \mathbf{C}$, the augmented Lagrangian function of (2.11) is defined as

$$\mathcal{L}(\mathbf{B}, \mathbf{C}; \Lambda) := f(\mathbf{C}) + \lambda J(\mathbf{C}) - \langle \Lambda, \mathbf{B} - \mathbf{C}\rangle + \frac{\beta}{2}\| \mathbf{B} - \mathbf{C} \|_F^2$$

where $\beta > 0$ is a penalty parameter. A typical iteration of the linearized ADMM for solving (2.11) can be described as:

$$\begin{aligned} \mathbf{B}^{\ell+1} &:= \operatorname*{argmin}_{\mathbf{B}^T\mathbf{B}=I} \mathcal{L}(\mathbf{B}, \mathbf{C}^\ell; \Lambda^\ell), \\ \mathbf{C}^{\ell+1} &:= \operatorname*{argmin}_{\mathbf{C}} \ \frac{1}{2\xi}\| \mathbf{C} - (\mathbf{C}^\ell - \xi\nabla_{\mathbf{C}}\tilde{\mathcal{L}}(\mathbf{B}^{\ell+1}, \mathbf{C}^\ell; \Lambda^\ell))\|_F^2 + \lambda J(\mathbf{C}), \\ \Lambda^{\ell+1} &:= \Lambda^\ell - \beta(\mathbf{B}^{\ell+1} - \mathbf{C}^{\ell+1}), \end{aligned} \tag{2.12}$$

where $\xi > 0$ denotes a step size, and $\tilde{\mathcal{L}}(\mathbf{B}, \mathbf{C}; \Lambda)$ denotes the smooth part of $\mathcal{L}(\mathbf{B}, \mathbf{C}; \Lambda)$, i.e.,

$$\tilde{\mathcal{L}}(\mathbf{B}, \mathbf{C}; \Lambda) := f(\mathbf{C}) - \langle \Lambda, \mathbf{B} - \mathbf{C} \rangle + \frac{\beta}{2} \| \mathbf{B} - \mathbf{C} \|_F^2.$$

Note that the two subproblems in (2.12) can both be easily solved. In particular, the $\mathbf{B}$-subproblem in (2.12) can be reduced to projecting the matrix $\mathbf{C}^\ell + \Lambda^\ell/\beta$ onto the constraint set $\mathbf{B}^T \mathbf{B} = I$, which can be done by $\mathbf{B}^{\ell+1} := UV^\top$, where $U\mathrm{diag}(\sigma)V^\top$ is the SVD of $\mathbf{C}^\ell + \Lambda^\ell/\beta$. The $\mathbf{C}$-subproblem in (2.12) can be reduced to computing the proximal mapping of $J(\cdot)$, which can be computed by using the SLEP package [15].

3. Stability-Based Tuning

The performance of the proposed reduced-rank model-based clustering method in (2.5) largely relies on the tuning parameters K, r and λ. In order to optimize the performance, it is of great importance to develop an appropriate tuning criterion.

In literature, likelihood-based information criteria have been widely used for selecting tuning parameters, such as AIC [3], BIC [21], and so on. However, the likelihood-based criteria may not be suitable for selecting tuning parameters, particularly r, in the reduced-rank models, since the likelihood functions in different spaces with different ranks are no longer comparable. A simple example illustrating this point is discussed in Appendix B.

This section proposes to use a clustering stability criterion for determining the tuning parameters (K, r) and variable selection stability for λ. The reason is that (K, r) control the balance between the cluster quality and the model complexity, whereas λ controls the balance between the cluster quality and the model sparsity. As the tuning process is split into two stages, we first fix $\lambda = 0$ and find the optimal $\hat{K}$ and $\hat{r}$ yielding the smallest clustering instability, and then fix $\hat{K}$ and $\hat{r}$ and find the optimal $\hat{\lambda}$ yielding the smallest variable selection instability. Furthermore, when clustering methods with different K's and/or r's lead to the same minimum instability, the larger K and smaller r are preferred as suggested in [19] and [26]. This two-stage tuning scheme delivers superior performance based on our limited experience in the numerical examples.

The concept of clustering stability has been extensively discussed in literature [5, 6, 26]. Its main idea is that if multiple samples are independently drawn from the same population, a good reduced-rank clustering method with appropriate tuning parameters shall produce clustering assignments that are similar from one sample to another. In this paper, the reduced-rank clustering methods are indexed by the tuning parameters K, r and λ, denoted as $\Psi_{K,r,\lambda}$. For simplicity, denote $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_n\}$ as the training sample. We then adopt the cross-validation scheme as introduced in [26] to choose (K, r). It splits $\mathbf{X}$ into two training sets $\mathbf{X}_1^c$, $\mathbf{X}_2^c$ and one validation set $\mathbf{X}_3^c$, where two training sets are used to generate two clustering methods via the reduced-rank clustering model with given tuning parameters, denoted as $\widehat{\psi}_1^c = \Psi_{K,r,0}(\mathbf{X}_1^c)$, and $\widehat{\psi}_2^c = \Psi_{K,r,0}(\mathbf{X}_2^c)$. The instability is then estimated as the distance between $\widehat{\psi}_1^c$ and $\widehat{\psi}_2^c$ on the validation set. Specifically,

$$\widehat{\mathrm{ins}}^c(\Psi_{K,r,0}) = \binom{n-2m}{2}^{-1} \sum_{\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}_3^c} I\Big(I(\widehat{\psi}_1^c(\mathbf{x}_i^c) = \widehat{\psi}_1^c(\mathbf{x}_j^c)) \neq I(\widehat{\psi}_2^c(\mathbf{x}_i^c) = \widehat{\psi}_2^c(\mathbf{x}_j^c)) \Big),$$

where the sizes of $\mathbf{X}_1^c$, $\mathbf{X}_2^c$ and $\mathbf{X}_3^c$ are m, m and $n - 2m$, respectively, and the function $I(\cdot)$ is the indicator function, i.e., $I(x) = 1$ if x is true, and $I(x) = 0$ if x is false. The process can be

replicated for $c = 1, \ldots, C$, and the final estimation of $\text{ins}(\Psi_{K,r,0})$ is

$$\widehat{\text{ins}}(\Psi_{K,r,0}) = C^{-1} \sum_{c=1}^{C} \widehat{\text{ins}}^{c}(\Psi_{K,r,0}). \tag{3.1}$$

The optimal $(\hat{K}, \hat{r})$ are then selected by the ones minimizing $\widehat{\text{ins}}(\Psi_{K,r,0})$, and the search is conducted by an exhaustive grid search scheme.

Similar to the clustering stability criterion, variable selection stability [24] is employed to select λ given $(\hat{K}, \hat{r})$. The variable selection instability measures the disagreement between two estimated active sets, which are obtained by applying the candidate variable selection algorithm to two independent training subsets. The tuning process randomly splits the training set into two subsets, estimate active sets $\widehat{\mathcal{A}}_{1b}^{(\hat{K},\hat{r},\lambda)}$ and $\widehat{\mathcal{A}}_{2b}^{(\hat{K},\hat{r},\lambda)}$ by fitting the proposed model with given $(\hat{K}, \hat{r})$, where $\widehat{\mathcal{A}} = \{j : \|\widehat{\mathbf{B}}_j\|_2 > 0\}$. The variable selection instability is estimated as

$$\widehat{\text{ins}}_{\lambda} = 1 - \frac{1}{B} \sum_{b=1}^{B} \kappa\Big(\widehat{\mathcal{A}}_{1b}^{(\hat{K},\hat{r},\lambda)}, \widehat{\mathcal{A}}_{2b}^{(\hat{K},\hat{r},\lambda)}\Big),$$

where B is the number of splittings in cross validation, and $\kappa(\cdot, \cdot)$ is the standard Cohen's kappa statistic measuring the agreement between two sets. The tuning parameter λ is then selected as the one minimizing $\widehat{\text{ins}}_{\lambda}$.

4. Asymptotic Theory

This section studies the asymptotic behavior of the proposed reduced-rank model-based clustering method regarding its consistency in estimating the true parameters and selecting the truly informative variables. Denote the regularized negative log-likelihood function as

$$l_p(\boldsymbol{\Theta}) = -\sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k \phi(\mathbf{B}^T \mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\right) + \lambda_n \sum_{j=1}^{p} \frac{\|\mathbf{B}_j\|_2}{\|\widetilde{\mathbf{B}}_j\|_2},$$

where the subscript of λ_n denotes its dependency on n, $\widetilde{\mathbf{B}}_j$ is obtained by maximizing (2.2), and $\| \mathbf{B}_j \|_2$ is the L_2-norm of $\mathbf{B}_j$. Denote $\mathcal{T}$ as the set of the equivalent true parameters leading to the same Gaussian mixture distribution as in (2.1). The estimation accuracy of $\widehat{\boldsymbol{\Theta}}$ is measured by its distance to $\mathcal{T}$, denoted as $d(\widehat{\boldsymbol{\Theta}}, \mathcal{T}) = \inf_{\boldsymbol{\Theta}^* \in \mathcal{T}} \|\widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}^*\|_2$.

The following technical assumptions are made.

Assumption 1. There exists $c_1 > 0$ such that $\|\boldsymbol{\mu}_k\|_2 \leq c_1$ for all k's.

Assumption 2. There exist $c_2, c_3 > 0$ such that $c_2 \leq \psi_{min}(\boldsymbol{\Sigma}_k) \leq \psi_{max}(\boldsymbol{\Sigma}_k) \leq c_3$ for all k's, where $\psi_{min}(\boldsymbol{\Sigma}_k)$ and $\psi_{max}(\boldsymbol{\Sigma}_k)$ denote the smallest and largest eigenvalues of $\boldsymbol{\Sigma}_k$, respectively.

Assumptions 1 and 2 are necessary to assure each covariance matrix to be nonsingular, and the Fisher information matrix of the reduced-rank Gaussian Mixture model is positive definite and bounded.

Theorem 4.1. *(Estimation consistency) Under Assumptions* 1 *and* 2*, there exists a local minimizer* $\widehat{\boldsymbol{\Theta}}$ *of* $l_p(\boldsymbol{\Theta})$ *such that* $d(\widehat{\boldsymbol{\Theta}}, \mathcal{T}) \xrightarrow{p} 0$ *when* $\frac{\lambda_n}{\sqrt{n}} \to 0$*, where* $\xrightarrow{p}$ *indicates convergence in probability.*

Theorem 4.1 establishes the estimation consistency of the proposed reduced-rank model-based clustering method in estimating the true parameters. In fact, the $\sqrt{n}$-consistency of $\widehat{\boldsymbol{\Theta}}$ can be established following the proof of Theorem 4.1.

Next, we study the asymptotic variable selection consistency in the high-dimensional setting.

Theorem 4.2. *(Variable selection consistency) Without loss of generality, we assume* $\|\mathbf{B}_j^*\|_2 > 0$ *when* $1 \le j \le p_0$, *and* $\|\mathbf{B}_j^*\|_2 = 0$ *when* $j > p_0$. *Then under the same assumptions of Theorem* 4.1, *when* $\frac{\lambda_n}{\sqrt{n}} \to 0$ *and* $\lambda_n \to \infty$, $P(\widehat{\mathcal{A}} = \{1, \ldots, p_0\}) \to 1$, *where* $\widehat{\mathcal{A}} = \{j : \|\widehat{\mathbf{B}}_j\|_2 > 0\}$.

Theorem 4.2 shows that the proposed method is able to identify the truly informative variables for high-dimensional model-based clustering with probably tending to 1.

5. Numerical Experiments

This section examines the effectiveness of the proposed reduced-rank model-based clustering, and compares against the diagonal EM [18], the regularized K-means [23], and the sparse FisherEM [7]. The performance is measured by the clustering error evaluated on the given dataset,

$$\mathrm{ce}(\hat{\psi}) = \binom{n_{data}}{2}^{-1} \sum_{\text{dataset}} I\big(I(\hat{\psi}(\mathbf{x}_i) = \hat{\psi}(\mathbf{x}_j)) \neq I(\psi^*(\mathbf{x}_i) = \psi^*(\mathbf{x}_j))\big),$$

where n_{data} is the size of the given dataset, and ψ^* is the true clustering assignment.

5.1. Simulated examples

The simulated dataset is generated from the model in (2.1). Specifically, the number of mixture is set as $K^* = 4$, the mixture weights $\pi_k = 1/4$ for $k = 1, \ldots, 4$, the reduced-rank matrix $\mathbf{B}$ is randomly generated under the constraint that $\mathbf{B}^T\mathbf{B} = \mathbf{I}_r$, the mean vectors are

$$\boldsymbol{\mu}_k = (-\mu\mathbf{1}_{\frac{r}{2}}^T, \mu\mathbf{1}_{\frac{r}{2}}^T)^T I(k=1) + \mu\mathbf{1}_r^T I(k=2) + (\mu\mathbf{1}_{\frac{r}{2}}^T, -\mu\mathbf{1}_{\frac{r}{2}}^T)^T I(k=3) - \mu\mathbf{1}_r^T I(k=4)$$

with $\mathbf{1}_r$ being the vector of all ones, and positive definite $\boldsymbol{\Sigma}_k$ are generated with diagonal $(1, 2, \ldots, r)^T$ and equal correlation 0.5. A total of n observations are generated, which are used for the clustering analysis.

The tuning parameters of all clustering methods are determined by the stability-based selection criterion in Section 3. Note that there is no need to include the sparsity penalty for this set of examples as all variables are informative. The search is then conducted through a grid search scheme, where the grid points are set as $K \in \{2, ..., 10\}$ and $r \in \{2, ..., 10\}$. Various scenarios with different values of n, p, r and μ are examined. Each scenario is replicated 50 times, and the averaged clustering errors and selected numbers of cluster are summarized in Tables 5.1 and 5.2.

In Table 5.1, it is evident that the proposed reduced-rank model-based clustering method outperforms its competitors in these three scenarios with low-rank structure. It is also interesting to note that diagonal EM and the FisherEM yield competitive performance, whereas the regularized K-means seems less satisfactory as it ignores the low rank structure. Furthermore, Table 5.2 demonstrates the effectiveness of the stability-based criterion for tuning the proposed method. It selects almost all correct numbers of cluster in the first three scenarios, while its

Table 5.1: The averaged clustering error and corresponding estimated standard deviation for various clustering algorithm. Here RR, DiagEM, RKmeans and FisherEM denote the proposed reduced-rank method, the diagonal EM, the regularized K-means and the Fisher EM methods, respectively.

(n, p, r, μ)	RR	DiagEM	RKmeans	FisherEM
(500,100,4,3)	**.009(.0009)**	.030(.0024)	.155(.0162)	.027(.0015)
(500,200,2,4)	**.010(.0013)**	.015(.0023)	.175(.0007)	.032(.0012)
(500,500,2,4.5)	.002(.0011)	**.001(.0005)**	.175(.0001)	.035(.0015)

Table 5.2: Frequencies of the selected number of clusters for the reduced-rank clustering algorithm.

(n, p, r, μ)	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
(500,100,4,3)	1/50	0/50	**49/50**	0/50	0/50	0/50	0/50	0/50	0/50
(500,200,2,3)	0/50	0/50	**50/50**	0/50	0/50	0/50	0/50	0/50	0/50
(500,500,2,4.5)	1/50	1/50	**48/50**	0/50	0/50	0/50	0/50	0/50	0/50
(500,100,100,3)	0/50	0/50	0/50	0/50	0/50	10/50	13/50	**22/50**	5/50

performance in the last scenario become less superior. This may be due to the fact that the true rank is 100 but we only search the grid up to $r = 10$. The performance might be improved if we enlarge the search grid at the cost of increasing computational cost.

In the second set of examples, the simulated dataset is generated from $\mathbf{X} = [\mathbf{X}_1^T \ \mathbf{X}_2^T]^T$, with

$$\mathbf{B}^T\mathbf{X}_1 \sim \sum_{k=1}^{K} \pi_k N_r(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where $\mathbf{X}$ is p-dimensional, $\mathbf{X}_1$ is p_0-dimensional and consists of the informative variables and $\mathbf{X}_2$ contains the other non-informative variables generated from $N_{p-p_0}(0, \mathbf{I})$. Other parameters are defined similarly as in the first set of simulated examples. Note that the grid search now needs to be conducted with respect to $K \in \{2, ..., 10\}$, $r \in \{2, ..., 10\}$ and $s \in \{10^{-2+4(t-1)/19}; t = 1, ..., 20\}$. Each scenario is replicated 50 times, and the averaged clustering errors and selected numbers of variables are summarized in Table 5.3.

Table 5.3: The averaged numbers of the selected variables, the averaged clustering error and their corresponding estimated standard deviation for various of clustering algorithm.

	RR	DiagEM	RKmeans	FisherEM
n=500, p=100, p_0=50, r=4, $\mu = 3$				
Clust. Error	**.001(.0000)**	.028(.0020)	.147(.0091)	.025(.0021)
Dimension	64.2(2.3546)	56.8(.7572)	40.1(1.9114)	64.1(1.2333)
n=500, p=200, p_0=50, r=2, $\mu = 4$				
Clust. Error	**.001(.0001)**	.006(.0019)	.161(.0134)	.019(.0006)
Dimension	75.6(2.6336)	64.6(1.7922)	45.5(.6872)	172.4(1.8690)
n=500, p=500, p_0=50, r=2, $\mu = 4.5$				
Clust. Error	**.000(.0000)**	.004(.0002)	.175(.0001)	.013(.0005)
Dimension	340.6(3.5581)	295.8(2.9110)	45.4(.6494)	200.8(2.0450)

Table 5.3 shows that the proposed method delivers superior clustering accuracy compared with the other existing alternatives. The regularized K-means selects the smallest number of informative variables but yields the largest clustering error, suggesting that it may miss some

truly informative variables. The performance of the diagonal EM and the sparse FisherEM methods are also less competitive, as they select slightly less number of informative variables with a deteriorated clustering accuracy.

5.2. Real applications

In this section, we apply the proposed reduced-rank model-based clustering to two real applications, wine recognition and human activity recognition. Both datasets are publicly available in UCI machine learning repository $http://archive.ics.uci.edu/ml/$.

Wine recognition dataset is used to distinguish different types of wine according to 13 attributes, such as malic acid, ash, alcalinity of ash, magnesium, total phenols, flavanoids, nonflavanoid phenols, proanthocyanins, color intensity, hue, OD280/OD315 of diluted wines, proline. The dataset contains 178 observations, among which 59 are the first type of wine, 71 are the sencond type and 48 are the third type.

Human activity recognition is an artificial intelligence task, and the main purpose is to enable smart phones to automatically recognize and react to human activities [4]. The experiment is carried out with a group of 30 volunteers whose ages are between 19 and 48. Each person performs six activities, including walking, walking upstairs, walking downstairs, sitting, standing and lying, with the smart phone on their waist. The dataset consists of 10,299 observations and 561 variables are recorded for each observation. For illustration, we only used 1000 randomly selected observations to conduct the cluster analysis.

For each real example, the stability-based selection criteria is used to determine the tuning parameters (K, r, λ) as in Section 5.1. The clustering errors evaluated on the dataset and the selected numbers of variables are summarized in Table 5.4. As there are only 13 attributes in the wine recognition example, variable selection is not conducted for this dataset.

Table 5.4: Performance of various clustering algorithms for two real applications.

	RR	DiagEM	RKmeans	FisherEM
Wine recognition				
K	**3**	**3**	**3**	**3**
Dimension	-	-	-	-
Clust. Error	**.015**	.030	.027	.016
Human activity recognition				
K	**2**	**2**	**2**	**2**
Dimension	10	457	416	440
Clust. Error	**.000**	.005	**.000**	.002

It is evident that the proposed reduced-rank clustering method yields the smallest clustering error as well as competitive variable selection performance in both two real examples. Particularly, in the human activity recognition dataset, the proposed method selects only 10 informative variables while achieving the smallest clustering error.

Furthermore, for the wine recognition, the proposed method identifies three clusters, which agree with the three types of wines. For the human activity recognition dataset, the proposed method identifies two clusters that appear to be different from the six different activities in the dataset. We then plot the dataset on its first two principal components in Figure 5.1. It is clear that each of the two identified clusters consists of three similar activities related to walking or

standing, respectively, whereas each set of the three related activities severely overlap with each other.

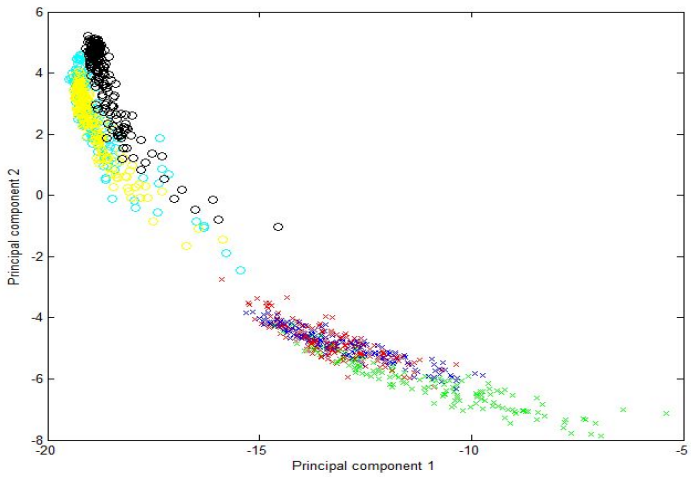

Fig. 5.1. The human activity recognition dataset displayed on the first two principal components. Different colors correspond to the original class labels and different symbols correspond to the estimated clustering structure.

6. Summary

This article proposes a reduced-rank model-based clustering method that is able to simultaneously cluster high-dimensional observations and select informative variables. An EM algorithm is developed for the model fitting, and one of whose difficult subproblems is solved by a linearized ADMM. A stability-based model selection criterion is used for the model tuning. The proposed clustering method delivers superior performance in both cluster analysis and variable selection, and outperforms its popular competitors in simulated and real experiments. A possible future direction is to extend the Gaussian mixture model to a much more flexible family that allows other distributions for the mixture components.

Acknowledgments. The authors would like to thank the referees and the editor for insightful suggestions that have improved this paper greatly. J. Wang is supported by Hong Kong Research Grants Council General Research Funds 11302615 and 11331016. S. Ma is supported by a startup package in Department of Mathematics at UC Davis.

Appendix A: Technical Proofs

Proof for Theorem 4.1:
For simplicity, denote the negative likelihood function as:

$$l_{-}(\boldsymbol{\Theta}) = -\sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_k \phi(\mathbf{B}^T \mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right),$$

and then $l_p(\boldsymbol{\Theta}) = l_-(\boldsymbol{\Theta}) + \lambda_n \sum_{j=1}^{p} \| \mathbf{B}_j \|_2$. The Taylor expansion of $l_-(\boldsymbol{\Theta})$ at $\boldsymbol{\Theta}^*$ yields that

$$l_-(\boldsymbol{\Theta}) = l_-(\boldsymbol{\Theta}^*) + l'_-(\boldsymbol{\Theta}^*)(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) + \frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T H(\bar{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*).$$

where $H(\bar{\boldsymbol{\Theta}})$ is an Hessian matrix evaluated at $\bar{\boldsymbol{\Theta}}$, and $\bar{\boldsymbol{\Theta}}$ is a matrix between $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}^*$.

For any $\boldsymbol{\Theta}^* \in \mathcal{T}$, denote $I_1(\boldsymbol{\Theta}^*)$ as the Fisher information matrix induced by reduced-rank Mixture Gaussian model for parameter $\boldsymbol{\Theta}^*$, $B_n(\alpha; \boldsymbol{\Theta}^*) = \{\boldsymbol{\Theta} : \|I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\|_2 \leq \frac{\alpha}{\sqrt{n}}\}$, and then $B_n(\alpha) \to \mathcal{T}$ as $n \to 0$. Then for any $\boldsymbol{\Theta}$ on the boundary of $B_n(\alpha; \boldsymbol{\Theta}^*)$, we have $\|I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\|_2 = \frac{\alpha}{\sqrt{n}}$ and

$$\begin{aligned}
& l_p(\boldsymbol{\Theta}) - l_p(\boldsymbol{\Theta}^*) \\
&= l_-(\boldsymbol{\Theta}) - l_-(\boldsymbol{\Theta}^*) + \lambda_n \sum_{j=1}^{p} \frac{\left(\|\mathbf{B}_j\|_2 - \|\mathbf{B}_j^*\|_2\right)}{\|\widetilde{\mathbf{B}}_j\|} \\
&\geq l'_-(\boldsymbol{\Theta}^*)^T(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) + \frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T H(\bar{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) - \lambda_n \sum_{j=1}^{p_0} \frac{\|\mathbf{B}_j - \mathbf{B}_j^*\|_2}{\|\widetilde{\mathbf{B}}_j\|_2} \\
&\geq l'_-(\boldsymbol{\Theta}^*)^T(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) + \frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T H(\bar{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) \\
&\quad - \frac{\lambda_n p_0}{\sqrt{n} \min_{1 \leq j \leq p_0} \|\widetilde{\mathbf{B}}_j\|_2} \|I_1(\boldsymbol{\Theta}^*)^{-1/2}\|_2 \|\sqrt{n} I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\|_2.
\end{aligned}$$

We now bound the terms in the last inequality separately. First, it follows from the definition of $B_n(\alpha; \boldsymbol{\Theta}^*)$ that $l'_-(\boldsymbol{\Theta}^*)^T(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) = \left(I_n(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\right)^T I_n(\boldsymbol{\Theta}^*)^{-1/2} l'_-(\boldsymbol{\Theta}^*) \geq -\alpha \|I_n(\boldsymbol{\Theta}^*)^{-1/2} l'_-(\boldsymbol{\Theta}^*)\|_2$, where $I_n(\boldsymbol{\Theta}^*) = n I_1(\boldsymbol{\Theta}^*)$. By Markov's inequality

$$\begin{aligned}
& P\left(\alpha \|I_n(\boldsymbol{\Theta}^*)^{-1/2} l'_-(\boldsymbol{\Theta}^*)\|_2 \leq \alpha^2/2\right) \\
& \geq 1 - 4E\|I_n(\boldsymbol{\Theta}^*)^{-1/2} l'_-(\boldsymbol{\Theta}^*)\|_2^2/\alpha^2 = 1 - 4p/\alpha^2.
\end{aligned}$$

Therefore, we have $P\left(l'_-(\boldsymbol{\Theta}^*)^T(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) \geq -\alpha^2/2\right) \geq 1 - 4p/\alpha^2$.

The second term can be bounded as

$$\begin{aligned}
& \frac{1}{2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)^T H(\bar{\boldsymbol{\Theta}})(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*) \\
&= \frac{1}{2n}\left(\sqrt{n} I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\right)^T I_1(\boldsymbol{\Theta}^*)^{-1/2} H(\bar{\boldsymbol{\Theta}}) I_1(\boldsymbol{\Theta}^*)^{-1/2}\left(\sqrt{n} I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\right) \xrightarrow{p} \alpha^2/2
\end{aligned}$$

The last inequality due to the fact that $\frac{1}{n} H(\bar{\boldsymbol{\Theta}}) \xrightarrow{p} I_1(\boldsymbol{\Theta}^*)$ as $n \to \infty$.

To bound the last term, it follows from Assumption 2 that the eigenvalues of each covariance matrix is bounded, therefore there exists a constant $c_4 \geq 0$ such that $\|I_1(\boldsymbol{\Theta}^*)^{-1/2}\|_2 \leq c_4$ and $\min_{1 \leq j \leq p_0} \|\widetilde{\mathbf{B}}_j\|_2$ is also bounded by another constant c_5, which implies that

$$\frac{\lambda_n p_0}{\sqrt{n} \min_{1 \leq j \leq p_0} \|\widetilde{\mathbf{B}}_j\|_2} \|I_1(\boldsymbol{\Theta}^*)^{-1/2}\|_2 \|\sqrt{n} I_1(\boldsymbol{\Theta}^*)^{1/2}(\boldsymbol{\Theta} - \boldsymbol{\Theta}^*)\|_2 \leq c_4 \alpha p_0 \lambda_n / \sqrt{n} c_5 \to 0,$$

provided that $\lambda_n / \sqrt{n} \to 0$ as $n \to \infty$.

Combining all the above bounds, for any $\epsilon > 0$, we can always choose a sufficiently large α such that $P(l_p(\boldsymbol{\Theta}) - l_p(\boldsymbol{\Theta}^*) > 0) > 1 - \epsilon$. Therefore, with probability tending to one, there exists a local minimizer $\widehat{\boldsymbol{\Theta}}$ such that $d(\widehat{\boldsymbol{\Theta}}, \boldsymbol{\Theta}^*) \leq \alpha/\sqrt{n}$ for any $\boldsymbol{\Theta}^* \in \mathcal{T}$. The desired result then follows immediately. □

Proof for Theorem 4.2:

We prove it by contradiction. Suppose that there exists some $j_0 > p_0$ such that $\|\widehat{\mathbf{B}}_{j_0}\|_2 > 0$. Denote $\mathbf{G} = \frac{\partial l_p(\mathbf{B})}{\partial \mathbf{B}}$, then the first order Karush-Kuhn-Tucker condition on stiefel manifold $\mathcal{M}$ yields $\widehat{\mathbf{G}}\widehat{\mathbf{B}}^T = \widehat{\mathbf{B}}\widehat{\mathbf{G}}^T$ [27], leading to $\widehat{\mathbf{G}} = \widehat{\mathbf{B}}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}$ because $\widehat{\mathbf{B}}^T\widehat{\mathbf{B}} = \mathbf{I}_r$. This implies $\widehat{\mathbf{G}}_k = \widehat{\mathbf{B}}_k\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}$ for any k, where $\widehat{\mathbf{G}}_k$ is the k-th row of $\widehat{\mathbf{G}}$. On one hand, the derivative of $l_p(\mathbf{B})$ at $\widehat{\mathbf{B}}_{j_0}$ yields that

$$\begin{aligned}\widehat{\mathbf{G}}_{j_0} &= l_p'(\widehat{\mathbf{B}}_{j_0}) = l_-'(\widehat{\mathbf{B}}_{j_0}) + \lambda_n \frac{\widehat{\mathbf{B}}_{j_0}}{\|\widehat{\mathbf{B}}_{j_0}\|_2\|\widetilde{\mathbf{B}}_{j_0}\|_2} \\ &= l_-'(\mathbf{B}^*_{j_0}) + H(\bar{\mathbf{B}}_{j_0})(\widehat{\mathbf{B}}_{j_0} - \mathbf{B}^*_{j_0}) + \lambda_n \frac{\widehat{\mathbf{B}}_{j_0}}{\|\widehat{\mathbf{B}}_{j_0}\|_2\|\widetilde{\mathbf{B}}_{j_0}\|_2},\end{aligned}$$

where $\mathbf{B}^*_{j_0} = \mathbf{0}_r$ and $\bar{\mathbf{B}}_{j_0}$ is between $\widehat{\mathbf{B}}_{j_0}$ and $\mathbf{B}^*_{j_0}$. Here $\mathbf{0}_r$ is the r dimensional vector with all elements 0.

Next, note that $n^{-1}H(\bar{\mathbf{B}}_{j_0}) - I_1(\mathbf{B}^*_{j_0}) = O_p(1/\sqrt{n})$ and $n^{-1}l_-'(\mathbf{B}^*_{j_0}) - S_1(\mathbf{B}^*_{j_0}) = O_p(1/\sqrt{n})$, where the score function $S_1(\mathbf{B}^*_{j_0}) = 0$ at $\mathbf{B}^*_{j_0}$. Therefore,

$$O_p(\sqrt{n}) + \left(nI_1(\mathbf{B}^*_{j_0}) + O_p(\sqrt{n}) + \frac{\lambda_n}{\sqrt{n}\|\widehat{\mathbf{B}}_{j_0}\|_2\|\widetilde{\mathbf{B}}_{j_0}\|_2}\right)\widehat{\mathbf{B}}_{j_0} = \widehat{\mathbf{G}}_{j_0}. \tag{A.1}$$

Since $\sqrt{n}\|\widetilde{\mathbf{B}}_{j_0}\|_2 = O_p(1)$ by the proof of Theorem 4.1, $I_1(\mathbf{B}^*_{j_0})$ is positive definite and $\frac{\lambda}{\sqrt{n}} \to 0$, we have $\|\widehat{\mathbf{G}}_{j_0}\|_2$ has the same order as $O_p(n)\|\widehat{\mathbf{B}}_{j_0}\|_2$.

On the other hand,

$$\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}} = \widehat{\mathbf{B}}_{j_0}\Big(O_p(\sqrt{n}) + \big(nI_1(\mathbf{B}^*) + O_r(\sqrt{n})\big)(\widehat{\mathbf{B}} - \mathbf{B}^*)^T + \lambda_n\frac{\partial J(\widehat{\mathbf{B}})}{\partial \mathbf{B}^T}\Big)\widehat{\mathbf{B}},$$

which implies

$$\|\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}\|_2 \le \|\widehat{\mathbf{B}}_{j_0}\|_2\left\|\Big(O_p(\sqrt{n}) + \big(nI_1(\mathbf{B}^*) + O_p(\sqrt{n})\big)(\widehat{\mathbf{B}} - \mathbf{B}^*)^T + \lambda_n\frac{\partial J(\widehat{\mathbf{B}})}{\partial \mathbf{B}^T}\Big)\widehat{\mathbf{B}}\right\|_F.$$

By Theorem 4.1, we have $\|\widehat{\mathbf{B}} - \mathbf{B}^*\|_F = O_p(1/\sqrt{n})$, thus

$$\|\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}\|_2 \le O_p(\lambda_n\sqrt{n})\|\widehat{\mathbf{B}}_{j_0}\|_2,$$

which implies

$$\|\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}\|_2 \le O_p(\lambda_n/\sqrt{n})\|\widehat{\mathbf{G}}_{j_0}\|_2.$$

Because $\lambda_n/\sqrt{n} \to 0$, $\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}}$ and $\widehat{\mathbf{G}}_{j_0}$ have the different order, contradicts with the fact $\widehat{\mathbf{B}}_{j_0}\widehat{\mathbf{G}}^T\widehat{\mathbf{B}} = \widehat{\mathbf{G}}_{j_0}$. Therefore, $\|\widehat{\mathbf{B}}_{j_0}\|_2 = 0$ for any $j_0 > p_0$. □

Appendix B: A Counter Example for Likelihood-Based Criteria

Now we use a simple example to illustrate likelihood-based criteria may not be appropriate for comparing likelihoods obtained in spaces with different rank. The counter example is constructed to show that the likelihood may increase when the rank r decreases, regardless of the true rank. As a consequence, likelihood-based criteria may mistakenly prefer the smallest r in comparison.

Suppose $\mathbf{x}_1, ..., \mathbf{x}_n$ are i.i.d sampled from $\mathbf{B}^T\mathbf{X} \sim \sum_{k=1}^{K} \pi_k N_r(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where the true rank $r = 2$ and $\boldsymbol{\Sigma}_k = \mathrm{diag}(\sigma_1^2, \sigma_2^2)$ with $\sigma_2^2 > 1/2\pi$. For simplicity, write $\mathbf{B} = [\mathbf{B}_1\ \mathbf{B}_2]$ and $\boldsymbol{\mu}_k = (\mu_{k1}, \mu_{k2})^T$. The likelihood of the true parameters is then

$$
\begin{aligned}
& L(\mathbf{x}; \mathbf{B}, \pi, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\
&= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k (2\pi)^{-1} |\boldsymbol{\Sigma}_k|^{-1/2} \exp\left\{ -\frac{(\mathbf{B}^T\mathbf{x}_i - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}_k^{-1} (\mathbf{B}^T\mathbf{x}_i - \boldsymbol{\mu}_k)}{2} \right\} \right) \\
&= \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k (2\pi\sigma_1^2)^{-1/2} \exp\left\{ -\frac{(\mathbf{B}_1^T\mathbf{x}_i - \boldsymbol{\mu}_{k1})^T (\mathbf{B}_1^T\mathbf{x}_i - \boldsymbol{\mu}_{k1})}{2\sigma_1^2} \right\} \right) \times \\
&\qquad \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k (2\pi\sigma_2^2)^{-1/2} \exp\left\{ -\frac{(\mathbf{B}_2^T\mathbf{x}_i - \boldsymbol{\mu}_{k2})^T (\mathbf{B}_2^T\mathbf{x}_i - \boldsymbol{\mu}_{k2})}{2\sigma_2^2} \right\} \right) \\
&\le \prod_{i=1}^{n} \left(\sum_{k=1}^{K} \pi_k (2\pi\sigma_1^2)^{-1/2} \exp\left\{ -\frac{(\mathbf{B}_1^T\mathbf{x}_i - \boldsymbol{\mu}_{k1})^T (\mathbf{B}_1^T\mathbf{x}_i - \boldsymbol{\mu}_{k1})}{2\sigma_1^2} \right\} \right) \\
&= L(\mathbf{x}; \mathbf{B}_1, \pi, \boldsymbol{\mu}_1, \sigma_1^2).
\end{aligned}
$$

The second last inequality is due to the fact

$$
(2\pi\sigma_2^2)^{-1/2} \exp\left\{ -\frac{(\mathbf{B}_2^T\,\mathbf{x}_i - \boldsymbol{\mu}_{k2})^T \sigma_2^{-2} (\mathbf{B}_2^T\,\mathbf{x}_i - \boldsymbol{\mu}_{k2})}{2} \right\} < 1
$$

for each $\mathbf{x}_i$. Clearly, the likelihood for the true parameters is less than the likelihood for other parameters with lower rank.

References

[1] P. Absil, C. Baker and K. Gallivan, Trust-region methods on Riemannian manifolds, *Foundations of Computational Mathematics*, **7** (2007), 303-330.

[2] P. Absil, R. Mahony and R. Sepulchre, Optimization Algorithms on Matrix Manifolds, Princeton University Press. 2008.

[3] H. Akaike, Information theory and an extension of the maximum likelihood pricinple, *2nd. Inter. Symposium on Information Theory*, Budapest Akademiai Kiado, (1973), 267-281.

[4] D. Anguita, A. Ghio, L. Oneto, X. Parra and L. Jorge, Human activity recognition on smartphones using a multiclass handware-friendly support vector machine, *IWAAL 2012*, 2012

[5] S. Ben-David, U. Von Luxburg and D. Pal, A sober look at stability of clustering, *19th Annal Conference on Comutational Learning Theory*, (2006), pp. 5-19, Berlin: Springer.

[6] A. Ben-Hur, A. Elisseeff and I. Guyon, A stability based model for discovering structure in clustered data, *Pacific Symposium on Biocomputing*, **7** (2002), 6-17.

[7] C. Bouveyron and C. Brunet, Simultaneous model-based clustering and visualization in the Fisher discriminative subspace, *Statistics and Computing*, **22** (2012), 301-324.

[8] C. Bouveyron and C. Brunet, Model-based clustering of high-dimensional data: A review, *Computational Statistics and Data Analysis*, **71** (2014), 52-78.

[9] L. Chen and J. Huang, Reduced-rank regression for simultaneous dimension reduction and variable selection in multivariate regression, *Journal of the American Statistical Association*, **107** (2012), 1533-1545.

[10] A. Dempster, N. Laird and D. Rubin, Maximize likelihood from incomplete data via EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39** (1977), 1-38.

[11] D. Donoho and J. Jin, Higher criticism thresholding: Optimal feature selection when useful features are rare and weak, *Proceedings of the National Academy of Science*, **105** (2008), 14790-14795.
[12] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96** (2001), 1348-1360.
[13] C. Fraley and A. Raftery, Model based clustering, discriminant analysis and density estimation, *Journal of the American Statistical Association*, **97** (2002), 611-631.
[14] D. Ghosh and A. Chinnaiyan, Mixture modeling of gene expression data from microarray experiments, *Bioinformatics*, **18** (2002), 275-286.
[15] J. Liu, S. Ji and J. Ye, SLEP: sparse learning with efficient projections, *Technical report in Arizona State University*, 2009.
[16] J. Liu, J. Zhang, M. J. Palumbo and C. Lawrence, Bayesian clustering with variable and transformation selection (with discussion), *Bayesian Statistics*, **7** (2003), 249-275.
[17] A. Ng, M. Jordan and Y. Weiss, On spectral clustering: Analysis and an algorithm, *Advances in Neural Information Processing Systems*, **14** (2001), 849-856.
[18] W. Pan and X. Shen, Penalized model-based clustering with application to variable selection, *Journal of Machine Learning Research*, **8** (2007), 1145-1164.
[19] W. Pedrycz, Interpretation of clusters in the framework of shadowed sets, *Pattern Recognition Letters*, **26** (2005), 2439-2449.
[20] A. Raftery and N. Dean, Variable selection for model-based clustering, *Journal of the American Statistical Association*, **101** (2006), 168-178.
[21] G. Schwars, Estimating the dimension of a model, *Annal of Statistics*, **6** (1973), 461-464.
[22] X. Shen, W. Pan and Y. Zhu, Likelihood-based selection and sharp parameters estimation, *Journal of the American Statistical Association*, **107** (2012), 223-232.
[23] W. Sun and J. Wang, Regularized k-means clustering of high-dimensional data and its asymptotic consistency, *Electronic Journal of Statistics*, **6** (2012), 148-167.
[24] W. Sun, J. Wang and Y. Fang, Consistent selection of tuning parameters via variable selection stability, *Journal of Machine Learning Research*, **14** (2013), 3419-3440.
[25] R. Tibshirani, Regression shrinkage and selection via the LASSO penalty, *Journal of the Royal Statistical Society, Series B*, **58** (1996), 267-288.
[26] J. Wang, Consistent selection of the number of clusters via cross validation, *Biometrika*, **58** (2010), 1-12.
[27] Z. Wen and W. Yin, A feasible method for optimization with orthogonality constraints, *Mathematical Programming*, **142** (2013), 397-434.
[28] J. Wolfe, Object cluster analysis of social areas, Master's thesis, University of California, Berkeley, 1963.
[29] M. Yuan and Y. Lin, Model selection and estimation in regression with group variables, *Journal of the American Statistical Association*, **93** (2006), 120-131.
[30] H. Zou, The adaptive Lasso and its oracal properties, *Journal of the American Statistical Association*, **101** (2006), 1418-1429.