Variable Selection for Clustering by Separability Based on Ridgelines

Hyangmin Lee* Jia Li†

Abstract

A new variable selection algorithm is developed for clustering based on mode association. In conventional mixture-model-based clustering, each mixture component is treated as one cluster and the separation between clusters is usually measured by the ratio of between- and within-component dispersion. In this paper, we allow one cluster to contain several components depending on whether they merge into one mode. The extent of separation between clusters is quantified using critical points on the ridgeline between two modes, which reflects the exact geometry of the density function. The computational foundation consists of the recently developed Modal EM (MEM) algorithm which solves the modes of a Gaussian mixture density, and the Ridgeline EM (REM) algorithm which solves the ridgeline passing through the critical points of the mixed density of two uni-mode clusters. Forward selection is used to find a subset of variables that maximizes an aggregated index of pairwise cluster separability. Theoretical analysis of the procedure is provided. We experiment with both simulated and real data sets and compare with several state-of-the-art variable selection algorithms. Supplemental materials including an R-package, data sets, and appendices for proofs are available online.

Keywords: Variable selection, Clustering, Modal EM, Ridgeline EM, Wrapper method, Mixture modeling

1 Introduction

The objective of clustering analysis is to discover well-separated groups within a data set. We refer to such a grouping characteristic of data as the clustering structure. When the number of variables increases, it often occurs that some variables do not contribute to revealing the clustering structure given some other variables, or they may even obscure the hidden clusters because of the curse of dimensionality. For high dimensional data, dimension reduction or variable selection is often performed before or during either classification or clustering. Sometimes variable selection is necessary for multiple reasons, e.g., to obtain better interpretation and to maintain low cost by measuring only a small number of variables.

*Hyangmin Lee acquired her Ph.D in the Department of Statistics at the Pennsylvania State University. Email: hyangminlee@gmail.com

†Jia Li, corresponding author, is an Associate Professor in the Department of Statistics at the Pennsylvania State University, University Park, PA 16802. She is also a Program Director in the Division of Mathematical Sciences, the National Science Foundation, Arlington, VA 22230. Email: jiali@stat.psu.edu

Variable selection has been studied extensively for classification in pattern recognition and machine learning (Narendra and Fukunaga, 1977; Liu and Setiono, 1996; Guyon and Elisseeff, 2003) as well as in statistics (Zou, 2006; Wang and Shen, 2006; Liu and Wu, 2007; Zhang et al., 2008; Yuan et al., 2009). Variable selection for clustering in contrast has been explored relatively recently with considerably less work. Dy and Brodley (2004) provides a comprehensive review of earlier work on this topic.

Explicit variable selection for clustering has been approached from several rather different perspectives: to remove redundancy among variables, to achieve high quality clustering by a certain criterion, and to fit certain statistical models. The first two perspectives are inspired by the *filter* (Narendra and Fukunaga, 1977; Liu and Setiono, 1996) and *wrapper* (Caruana and Freitag, 1994) approaches for selecting variables in classification. However, adapting these approaches to clustering is not straightforward because of the unknown cluster labels. The filter approaches select variables before performing clustering by examining the redundancy among variables (Mitra et al., 2002). The wrapper approaches select variables and conduct clustering simultaneously. Some kind of measurement is adopted to assess the goodness of the clustering result. A subset of variables is then searched to optimize the goodness measure. Because exhaustive search is computationally prohibitive even for moderate dimensions, a greedy search strategy such as forward stepwise addition or backwards deletion is usually used. Dy and Brodley (2004) investigated the wrapper approaches, using the scatter separability and the maximum likelihood criterion to measure the quality of clustering. Friedman and Meulman (2004) also proposed a method to conduct clustering and feature selection simultaneously, but with clear differences from the wrapper approaches. The objective is to minimize the total within cluster dispersion over both the data partition and weights on the variables.

The statistical modeling approaches rely largely on mixture models and fall into two categories. The first category casts variable selection into a model selection problem by exploiting mixture models with specific formulations on how informative and non-informative variables relate to the membership of the mixture components, or equivalently, the cluster labels. Tadesse et al. (2005) proposed a mixture model that treats variables as discriminant and non-discriminant for clusters. The non-discriminant variables are assumed independent from the cluster labels and follow a Gaussian distribution. The discriminant variables follow Gaussian distributions conditioned on the cluster labels. Law et al. (2004) used a similar mixture model. Raftery and Dean (2006) took a more sophisticated view on the relationship between relevant variables, irrelevant variables, and the cluster labels. Instead of being independent from the cluster labels, the irrelevant variables are assumed to be conditionally independent from the cluster labels given the relevant variables. Maugis et al. (2009) extended the work of Raftery and Dean (2006) by allowing the irrelevant variables to depend on a subset of relevant variables.

The second category of mixture-model-based clustering and variable selection methods exploits the mechanism of penalized modeling. A penalty term on the component means is added in the maximum likelihood estimation of the mixture model, possibly shrinking the means across different components to a common value. If the component means of a variable are all equal, under certain setups of the mixture model, this variable becomes non-informative for the clustering structure. Pan and Shen (2007) pioneered the mechanism, using the L_1 norm of the means as a penalty and assuming a common diagonal covariance matrix for the mixture components. The work is extended to cluster-specific diagonal covariance and grouped variables by Xie et al. (2008). Wang and Zhu (2008) noted the limitation of the L_1 norm, which neglects the variable-based grouping of the means, and proposed the L_∞ norm. Extending this work, Guo et al. (2010) explored deeper into the nature of separability between clusters. Instead of dividing variables into informative and non-informative, which either separate at least one

pair of clusters or separate no pair of clusters, they treat each pair of clusters individually when trying to identify informative variables. Witten and Tibshirani (2010) used the Lasso-type penalty to perform sparse clustering (in the sense of variables involved). The framework can be applied to a broad range of optimization based methods. In particular, sparse versions for the extremely popular K-means and hierarchical clustering (dendrogram) have been developed. This recent work shows that the penalty mechanism is valuable beyond mixture modeling.

Although our method also uses mixture models, it differs from all the aforementioned mixture-model-based approaches in its treatment of the relationship between clusters and mixture components. Conventionally, to obtain clustering, a mixture model is estimated first, then every data point is partitioned into the component with the maximum posterior. Here, we employ a different framework for clustering in which a mixture model only serves the purpose of density estimation. To partition data, we solve an ascending path starting from any data point and ending with a local maximum of the density (mode). Data points associated with the same mode are put in one cluster.

Advantages of the new framework are discussed in details with experimental evidence in (Li et al., 2007). To highlight some of these advantages, we note that clusters formed by the new method possess better geometric characteristics. Each cluster is ensured to correspond to a unique mode. In the usual mixture-model-based clustering, several components may be needed to fit a mass of data with high likelihood, but those components may overlap substantially and do not capture distinct groups. This issue is addressed by the mode association clustering (MAC) approach in (Li et al., 2007). In fact, since the mixture components no longer correspond one to one with the clusters, we can use a kernel density estimate which is still in the form of a mixture distribution but eliminates the sensitivity to initialization as encountered in conventional mixture-model-based clustering.

In this paper, we leverage the advantages of the MAC approach for variable selection. When assessing how well clusters separate from each other, we exploit a measure based on the ridgeline between the modes of two clusters, which captures precisely the geometry of the density function. The within or between-cluster scatter matrices, however, lose information about the density function and are motivated largely by Gaussian-type clusters. Our method of variable selection for clustering can be viewed as a wrapper approach based on mixture models. Although the stepwise selection in the wrapper approach lacks the mathematical appeal of an overall optimization, it is a trade off for exploiting the geometry of a mixture density.

The rest of the paper is organized as follows. Section 2 presents preliminaries on the Modal EM, the Ridgeline EM algorithm, and the MAC approach. In Section 3, we introduce a measure for the extent of separation between clusters based on ridgelines and provide theoretical insights. The greedy search procedure and the treatment of outlier clusters are also described. In Section 4, the experimental results based on several simulated and real data sets are presented. Comparisons are made with multiple state-of-the-art variable selection algorithms. We conclude and discuss future work in Section 5.

2 Preliminaries

In this section, we introduce notation and briefly review several related topics: the *Modal EM* (*MEM*) algorithm for finding modes of a mixture distribution, the mode association clustering approach, and the *Ridgeline EM* (*REM*) algorithm for finding ridgelines between two uni-mode clusters. The theoretical derivations of MEM and REM are presented in details in (Li et al., 2007). In that paper, the mixture

model used is a kernel density estimate where each mixture component has an identical and diagonal covariance matrix. Here, we describe these algorithms for general Gaussian mixtures.

Consider the data set $\{x_1, ..., x_n\}$ containing p-dimensional vectors $x_i \in \mathcal{R}^p$. Let K be the number of mixture components, π_k the mixing probability of the kth component with $\sum_{k=1}^{K} \pi_k = 1$, $0 \leq \pi_k \leq 1$, and $f_k(x_i|\theta_k)$ the density of the kth component under parameter θ_k. Then the mixture density $f(x|\theta) = \sum_{k=1}^{K} \pi_k f_k(x|\theta_k), \theta = \{\pi_1, ..., \pi_K, \theta_1, ..., \theta_K\}$.

We are only concerned with Gaussian mixtures in this paper. The parameters of the Gaussian mixture model are usually estimated by the EM algorithm (McLachlan and Peel, 2000). In our studies, we use the R package *Mclust* developed by Fraley and Raftery (1999) as a density estimation tool (Fraley and Raftery, 2002). The *Mclust* package automatically chooses the number of mixture components by BIC. Moreover, the covariance matrices of the mixture components are decomposed into factors of shape, orientation, and volume; and some or all of the factors are allowed to be shared across components. Again, BIC is used to select a particular form of the model.

MEM for Gaussian Mixtures and Modal Clustering: The MEM algorithm solves a local maximum of a mixture density by ascending iterations starting from any initial point. Let the mixture density be $f(x) = \sum_{k=1}^{K} \pi_k f_k(x)$, where $x \in \mathcal{R}^d$, π_k is the prior probability of mixture component k, and $f_k(x)$ is the density of component k. Given any initial value $x^{(0)}$, MEM solves a local maximum of the mixture by alternating the following two steps until a stopping criterion is met (Li et al., 2007): (1) At iteration r, let $p_k = \dfrac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}$, $k = 1, ..., K$; (2) Update $x^{(r+1)} = \underset{x}{\operatorname{argmax}} \sum_{k=1}^{K} p_k \log f_k(x)$.

Consider a Gaussian mixture $f(x) = \sum_{k=1}^{K} \pi_k \phi(x|\mu_k, \Sigma_k)$, where $\phi(\cdot)$ is the normal density function. As proved in Appendix A, MEM becomes

1. E-step: $p_k = \dfrac{\pi_k f_k(x^{(r)})}{f(x^{(r)})}$, where $k = 1, \ ..., \ K$.

2. M-step: $x^{(r+1)} = \left(\sum_{k=1}^{K} p_k \cdot \Sigma_k^{-1} \right)^{-1} \cdot \left(\sum_{k=1}^{K} p_k \cdot \Sigma_k^{-1} \mu_k \right)$.

In (Li et al., 2007), *Mode Association Clustering (MAC)*, referred to in short as *modal clustering*, is proposed based on a Gaussian kernel density estimate. Each data point is treated as an initial point to apply the MEM algorithm. Data points ascending to the same mode are put in the same cluster. In this paper, we use a computationally less intensive approach with a few differences from the clustering method in (Li et al., 2007). First, we replace the Gaussian kernel density estimate by the finite mixture model obtained using the *Mclust* package. It is known that kernel density estimate is highly sensitive to the data dimension and is not recommended even for moderate dimensionality. Second, instead of applying MEM to every data point, we apply it to the means of the mixture components. Components whose mean vectors are brought to the same mode via MEM are merged into one cluster. To obtain a partition of the data set, the data points are initially divided into the mixture components according to the usual maximum a posteriori criterion. These preliminary groups of data points are then subjected to merging depending on whether the corresponding components merge into one mode. The merged groups form the final clusters. To distinguish from the MAC method in (Li et al., 2007), we call the modified method *Component-wise Mode Association Clustering (CMAC)*, which is outlined below.

1. Apply *Mclust* to estimate a density for the data set $\{x_1, ..., x_n\}$: $f(x) = \sum_{k=1}^{K} \pi_k \phi(x \mid \mu_k, \Sigma_k)$.

2. Apply MEM to each component mean $\mu_k, k = 1, ..., K$. Let the distinct modes found by MEM be $\eta_m, m = 1, ..., M$, where $M \leq K$ in general. If the kth component mean μ_k is mapped to the mth mode η_m, we denote the mapping by $\Lambda(k) = m$.

3. Partition x_i's into M clusters by first finding the component k with the maximum posterior probability given x_i and then mapping k to its mode: $x_i \to \Lambda(\text{argmax}_{k=1,...,K}\, \pi_k \phi(x_i \mid \mu_k, \Sigma_k))$.

We do not need any extra mechanism to decide the number of clusters. The number of components in the Gaussian mixture is chosen based on BIC by *Mclust* with the sole purpose of achieving good density estimate. The number of clusters, which depends on the geometry of the acquired density, is then determined by identifying the modes.

REM for Gaussian Mixtures: A ridgeline is a 1-D curve connecting the modes of two uni-mode clusters. The probability density functions (pdfs) of the two clusters are Gaussian mixtures. A ridgeline is geometrically important because it is proved to pass through all the critical points of the mixed density of the two clusters, including modes, antimodes (local minimum) and saddle points (Ray and Lindsay, 2005). Let $g_i(x)$ and $g_j(x), i \neq j$ and $i, j \in \{1, ..., M\}$, be the densities for two clusters. The ridgeline between two clusters is

$$\mathcal{L} = \{x(\alpha) : (1-\alpha)\nabla \log\ g_i(x) + \alpha \nabla \log\ g_j(x) = 0,\ 0 \leq \alpha \leq 1\}. \tag{1}$$

The REM algorithm is derived to solve $\mathcal{L}$ (Li et al., 2007). Specifically for our model, the pdf of a cluster $g_i(x)$ is a Gaussian mixture with T_i components: $g_i(x) = \sum_{k=1}^{T_i} \pi_{i,k} \phi(x|\mu_{i,k}, \Sigma_{i,k})$. When $\alpha = 0$, the starting point $x(0)$ on $\mathcal{L}$ is simply the mode of $g_i(x)$, which can be solved by MEM. Similarly, when $\alpha = 1$, the ending point $x(1)$ on $\mathcal{L}$ is the mode of $g_j(x)$. We solve $\mathcal{L}$ sequentially at grid points $0 = \alpha_0 < \alpha_1 < \alpha_2 \cdots < \alpha_\zeta = 1$. To solve $x(\alpha_l)$, we use the already acquired $x(\alpha_{l-1})$ as the initial value, that is, $x^{(0)} = x(\alpha_{l-1})$, and apply the following iterative procedure:

1. E-step: Compute $p_{i,k} = \dfrac{\pi_{i,k}\ \phi(x^{(r)}|\mu_{i,k}, \Sigma_{i,k})}{\sum_{l=1}^{T_i} \pi_{i,l}\ \phi(x^{(r)}|\mu_{i,l}, \Sigma_{i,l})}$, where $k = 1, ..., T_i$. Similarly, compute $p_{j,k} = \dfrac{\pi_{j,k}\ \phi(x^{(r)}|\mu_{j,k}, \Sigma_{j,k})}{\sum_{l=1}^{T_j} \pi_{j,l}\ \phi(x^{(r)}|\mu_{j,l}, \Sigma_{j,l})}, k = 1, ..., T_j$.

2. M-step: Update $x^{(r+1)} = A^{-1}\bar{\mu}$, where $A = (1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\Sigma_{i,k}^{-1} + \alpha\ \sum_{k=1}^{T_j} p_{j,k}\Sigma_{j,k}^{-1}$, $\bar{\mu} = (1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\Sigma_{i,k}^{-1}\mu_{i,k} + \alpha\ \sum_{k=1}^{T_j} p_{j,k}\Sigma_{j,k}^{-1}\mu_{j,k}$.

The derivation of the M-step is provided in Appendix A. The extracted ridgelines allow us to define a separability measure between clusters according to the precise geometry of the density function.

3 Variable Selection for Clustering

We now introduce the variable selection algorithm based on a cluster separability measure defined using ridgelines. To assess the overall clustering result, the pairwise separability measures are aggregated into a single value. The goal is to find a subset of variables that achieves the maximum aggregated separability. For computational feasibility, we employ the forward selection searching strategy. Two slightly different versions of the algorithm are investigated to meet the needs of different application scenarios.

3.1 Ridgeline-Based Separability

To measure the separability between two clusters, we find the ridgelines between the modes of any two clusters. Let the ridgeline between cluster i and j be $\mathcal{L}_{i,j} = \{x^{(i,j)}(\alpha_l) : l = 0, 1, ..., \zeta; 0 = \alpha_0 < \alpha_1 < ... < \alpha_\zeta = 1\}$, where the point $x^{(i,j)}(\alpha_l)$ is the solution to Eq. (1) at $\alpha = \alpha_l$, solved by REM. The *pairwise separability* measure between two clusters C_i and C_j is defined by

$$S(C_i, C_j) = 1 - \frac{\min_{l=1}^{\zeta}[f(x^{(i,j)}(\alpha_l))]}{\min[f(x^{(i,j)}(0)), f(x^{(i,j)}(1))]} \ . \tag{2}$$

When $i = j$, let $S(C_i, C_i) = 0$. It is obvious that $S(C_i, C_j) = S(C_j, C_i)$ and $0 \leq S(C_i, C_j) \leq 1$.

To measure the overall quality of a clustering result, we propose a so-called *aggregated distinctiveness (AD)* which combines the pairwise separability in a weighted manner. The weights reflect the joint significance of a pair of clusters. Let γ_i be the proportion of data in cluster $C_i, i = 1, ..., M$. The AD, denoted by Δ, is defined by $\Delta = \sum_{i=1}^{M} \sum_{j=1, j\neq i}^{M} \gamma_i \gamma_j S(C_i, C_j)$. When there is only one cluster ($M = 1$), $\Delta = 0$. Obviously, Δ increases when any pairwise separability increases. On the other hand, if $S(C_i, C_j)$ is a fixed constant A, $\Delta = A \cdot \sum_{i=1}^{M} \gamma_i(1 - \gamma_i) = A \cdot (1 - \sum_{i=1}^{M} \gamma_i^2)$. Since γ_i's sum up to 1, $\Delta = A \cdot (1 - \sum_{i=1}^{M} \gamma_i^2)$ is maximized at $\gamma_i = \frac{1}{M}, i = 1, ..., M$, with the maximum value $A \cdot \frac{M-1}{M}$. This coincides with our intuition that when there is no difference in the pairwise separability, the overall clustering structure of the data is considered strongest when the data are partitioned into groups of equal sizes. Moreover, the more clusters there are (larger M), the larger Δ is. When $M \to \infty$, $\Delta \to A$.

It occurs occasionally some clusters contain very few data points. Heuristically, such clusters appear more like outlier data points rather than distinct groups. We propose to adjust the definition of AD to reduce the influence of outlier points. In the definition of AD, we note that the influence of outlier points is suppressed to some extent due to the small weights assigned to outlier clusters. However, if cluster C_i is an outlier point, $S(C_i, C_j)$ tends to be large because the very reason that a single point is identified as a cluster is that it is distant from the rest of the data mass, yielding high separability from other clusters.

Figure 3 shows a clustering result based on two variables of the infant data set, which is introduced in details in the experiment section. In the left scatter plot, eight clusters are shown, three of which contain a single point. The scatter plot on the right zooms in on the five non-singleton clusters. Let $OS(C_i) = min_{j:1\leq j\leq M, j\neq i} S(C_i, C_j)$ be the overall separability of C_i from the other clusters. The three singleton clusters, C_4, C_6, C_7, all have OS above 0.99.

We call a cluster an *effective cluster* (EFC) if it contains more than t data points. In our experiments, we set $t = 1$. Suppose clusters $\{C_1, ..., C_M\}$ are obtained. Let the set of indices for EFCs be $\mathcal{E} = \{i : |C_i| > t, 1 \leq i \leq M\}$. To stress a cluster C_i is an EFC, we denote it explicitly as C_i^{EF}. The modified AD is thus defined by

$$\Delta = \sum_{i\in\mathcal{E}} \sum_{j\in\mathcal{E}, j\neq i} \gamma_i \gamma_j S(C_i^{EF}, C_j^{EF}) \ . \tag{3}$$

For the example shown in Figure 3, the AD without adjustment to outliers is 0.6592, while the AD with adjustment to outliers is 0.5465. We see that the boost of AD due to outliers can be substantial.

3.2 Algorithms for Variable Selection

We present two versions of the variable selection algorithm called Alg. I and Alg. II, in order to target different application scenarios. In both cases, we use forward variable selection based on the corresponding measurement of clustering quality. For Alg. I, mixture modeling and clustering are carried out separately for every examined subset of variables, aiming at seeking a subset with the strongest clustering structure. For Alg. II, the mixture model is estimated in the original (full) space. When a subset of variables is examined, the distribution in the subspace is taken directly as the marginal distribution of the acquired model in the full space. This reduces computation by bypassing mixture modeling in the subspace. The implicit assumption is that the model in the full space is satisfactory. Moreover, the goal of Alg. II is to find a subspace which achieves high separation between clusters already identified in the full space. In other words, the clusters obtained in the full space are treated as the "truth". The definition of AD, as given in Eq. (3), is modified for Alg. II in order to comply to an existing partition of data, which will be described shortly.

To choose a subset of variables, we try to strike a balance between two aspects. When the dimension is high, a clustering structure may be obscured by the non-informative variables. On the other hand, when the dimension is low, some variables useful for revealing the clustering structure may be excluded. Alg. II is suitable only when the clustering structure is clear in the full space, at least not seriously blurred due to the high dimension. It may be favored over Alg. I when all the variables are believed to be useful and the purpose of choosing a subset of them is more for visualization or simplification of data rather than for countering the curse of dimensionality.

Consider a data set $\mathcal{X} = \{x_1, x_2, ..., x_n\}$, $x_i \in \mathcal{R}^p$. Denote the p variables by $(X_1, X_2, ..., X_p)$. Let the set of variables already selected after k iterations of the forward selection be $\mathcal{S}_k = \{X_{j_1}, X_{j_2}, ..., X_{j_k}\}$. By construction, $\mathcal{S}_k$, $k = 1$, ..., p, is a nested sequence. Denote the variables not included in $\mathcal{S}_k$ by $\mathcal{S}_{-k} = \{X_j : X_j \notin \mathcal{S}_k, 1 \leq j \leq p\}$. Denote the subvector of x_i over a subset of variables, e.g., $\mathcal{S}_k$, by $x_i(\mathcal{S}_k)$, and the data set by $\mathcal{X}_{\mathcal{S}_k} = \{x_i(\mathcal{S}_k) : i = 1, ..., n\}$.

At the $k+1$st iteration, Alg. I performs the following steps.

1. For each $X_j \in \mathcal{S}_{-k}$, use *Mclust* to estimate a Gaussian mixture for the data set $\mathcal{X}_{\mathcal{S}_k \cup \{X_j\}}$.
2. Apply CMAC to perform clustering on $\mathcal{X}_{\mathcal{S}_k \cup \{X_j\}}$ using the density obtained in the previous step. Compute the AD by Eq. (3). Denote the AD by $\Delta_{\mathcal{S}_k \cup \{X_j\}}$.
3. Choose the new variable X_{j^*} to add into $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{X_{j^*}\}$: $X_{j^*} = \text{argmax}_{X_j \in \mathcal{S}_{-k}} \Delta_{\mathcal{S}_k \cup \{X_j\}}$.
4. Stop if the number of variables selected reaches a given value or if the increase in AD by adding the new variable is below a threshold ϵ. Otherwise, let $k+1 \to k$ and repeat the above steps.

To introduce Alg. II, let the mixture model obtained by *Mclust* in the original space be $f(x) = \sum_{k=1}^{K} \pi_k \phi(x \mid \mu_k, \Sigma_k)$. Denote the marginal distribution of $f(x)$ on a subset of variables $\mathcal{S}$ by $f_{\mathcal{S}}(x)$. Recall the notation in Section 2. Let the mapping function from a mixture component of $f(x)$ to a cluster (mode) in the original space be Λ_{org}, and that from a component of $f_{\mathcal{S}}(x)$ in the subspace spanned by X_j's in $\mathcal{S}$ to a cluster be $\Lambda_{\mathcal{S}}$. The numbers of clusters (modes) in the original and the subspace may not be the same. Because Alg. II aims at achieving high separation between data in different clusters that are identified in the original space, the separability between mixture components that are mapped to the same cluster by Λ_{org} should not be part of the AD even if they are mapped to

different clusters by $\Lambda_{\mathcal{S}}$ in the subspace. We thus modify the definition of AD in Eq. (3) as follows. Let $\mathcal{E}_{org} = \{k : \text{Cluster } \Lambda_{org}(k) \text{ is not outlier}, 1 \leq k \leq K\}$. The modified AD is

$$\Delta = \sum_{i \in \mathcal{E}_{org}} \sum_{j \in \mathcal{E}_{org}, j \neq i} \pi_i \pi_j I\left(\Lambda_{org}(i) \neq \Lambda_{org}(j)\right) \cdot S\left(\Lambda_{\mathcal{S}}(i), \Lambda_{\mathcal{S}}(j)\right) \ , \tag{4}$$

where $I(\cdot)$ is the indicator function that equals 1 when the argument is true and 0 otherwise, and $S(\cdot, \cdot)$ is the pairwise separability between clusters in the subspace spanned by variables in $\mathcal{S}$. We abuse notation slightly here to use $S(m_1, m_2)$ in replacement of $S(C_{m_1}, C_{m_2})$ in Eq. (3). It is straightforward to see that if $\Lambda_{org}(i) = \Lambda_{\mathcal{S}}(i)$ for all $i = 1, ..., K$, Eq. (4) is equivalent to Eq. (3).

At the $k+1$st iteration, Alg. II performs the following steps.

1. For each $X_j \in \mathcal{S}_{-k}$, find the marginal distribution $f_{\mathcal{S}_k \cup \{X_j\}}(x)$ based on $f(x)$.
2. Apply CMAC to perform clustering on $\mathcal{X}_{\mathcal{S}_k \cup \{X_j\}}$ using density $f_{\mathcal{S}_k \cup \{X_j\}}(x)$. Compute the AD by Eq. (4). Denote the AD by $\Delta_{\mathcal{S}_k \cup \{X_j\}}$.
3. Choose the new variable X_{j^*} to add into $\mathcal{S}_{k+1} = \mathcal{S}_k \cup \{X_{j^*}\}$: $X_{j^*} = \text{argmax}_{X_j \in \mathcal{S}_{-k}} \Delta_{\mathcal{S}_k \cup \{X_j\}}$.
4. Stop if the number of variables selected reaches a given value or if the increase in AD by adding the new variable is below a threshold ϵ. Otherwise, let $k+1 \to k$ and repeat the above steps.

3.3 Theoretical Insights

The geometry of the Gaussian mixture density is known to be difficult to analyze. Ray and Lindsay (2005) made significant advances in the theory about the modes and ridgelines of mixture densities. However, it is still mathematically challenging to obtain analytical solution for the separability (Eq. (2)) between two arbitrary Gaussian distributions, let alone uni-mode mixtures of Gaussian distributions. To gain theoretical insights about our proposed forward selection based on AD, we investigate a basic but simple mixture distribution with two Gaussian components. The theoretical results obtained justify the proposed forward selection procedure.

Lemma 3.1 states that the ridgeline between the two Gaussian components is a curve lying in the subspace spanned only by the informative variables. Lemma 3.2 and 3.3 specify the geometry of the ridgeline and derive all the possible patterns of the density function along the ridgeline. Finally Theorem 3.1 states that under moderate conditions, the forward selection procedure will select only informative variables and in the decreasing order of the extent of separability in each dimension.

Let random variable $X \in \mathcal{R}^p$. Denote $X^{(q)} = (X_1, X_2, ..., X_q)^t$ and $X^{(-q)} = (X_{q+1}, ..., X_p)^t$. For clarity, we also use superscript (q) to indicate in general a q-dimensional vector (or $q \times q$ matrix) and $(-q)$ to mean a $p-q$ dimensional vector (or $(p-q) \times (p-q)$ matrix). Consider two normal distributions in a mixture model for which only the q variables in $X^{(q)}$ are informative and those in $X^{(-q)}$ are non-informative. In particular, the densities of the two components are $g_i(x) = \phi(x^{(q)}|\mu_i^{(q)}, I)\phi(x^{(-q)}|\mu^{(-q)}, I)$, $i = 1, 2$, where $\phi(\cdot|\mu, \Sigma)$ is the normal density with mean μ and covariance Σ. The following lemmas and theorem are proved in Appendix B.

Lemma 3.1 *Let $\mathcal{L}^{(q)}$ be the ridgeline in $\mathcal{R}^q$ between $\phi(x^{(q)}|\mu_i^{(q)}, I)$, $i = 1, 2$. Then the ridgeline between $g_i(x) = \phi(x^{(q)}|\mu_i^{(q)}, I)\phi(x^{(-q)}|\mu^{(-q)}, I)$, $i = 1, 2$, is given by $\mathcal{L} = \{x : x^{(q)} \in \mathcal{L}^{(q)}, x^{(-q)} = \mu^{(-q)}\}$.*

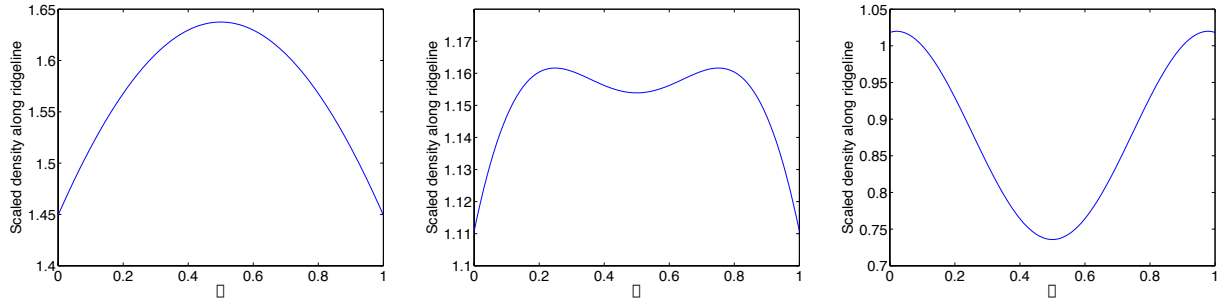
It is trivial to extend the lemma to non-identity covariance matrix. That is, $g_i(x) = \phi(x^{(q)}|\mu_i^{(q)}, \Sigma^{(q)})\phi(x^{(-q)}|\mu^{(-q)}, \Sigma^{(-q)}), i = 1, 2$. Lemma 3.1 implies that the ridgeline in the full space $\mathcal{R}^p$ is obtained from the ridgeline in the informative space $\mathcal{R}^q$ by augmenting the non-informative variables using the constant vector $\mu^{(-q)}$, the common mean of the two Gaussian distributions.

Lemma 3.2 *The ridgeline between two normal distributions $\phi(x|\mu_1, I)$ and $\phi(x|\mu_2, I)$, $x \in \mathcal{R}^q$, is $\mathcal{L} = \{x : x = (1-r)\mu_1 + r\mu_2, r \in (0, 1)\}$.*

By Lemma 3.2, the ridgeline between two normal distributions with the identity covariance matrix is the straight line connecting the two means.

Consider a mixture density with equal priors $f(x) = \frac{1}{2}\phi(x|\mu_1, I) + \frac{1}{2}\phi(x|\mu_2, I)$, $x \in \mathcal{R}^q$. Let $\beta = ||\mu_1 - \mu_2||^2/2$, where $||\cdot||$ is the L_2 norm. Let the density of the mixture distribution along the ridgeline, as given by Lemma 3.2, be $f_r(r) = \frac{1}{2}\phi((1-r)\mu_1 + r\mu_2|\mu_1, I) + \frac{1}{2}\phi((1-r)\mu_1 + r\mu_2|\mu_2, I)$, $r \in [0, 1]$.

Lemma 3.3 *The function $f_r(r), r \in [0, 1]$, is symmetric around $\frac{1}{2}$. When $\beta \leq 2$, $f_r(r)$ increases strictly on $[0, \frac{1}{2}]$, and $r = \frac{1}{2}$ is the global maximum for $f_r(r), r \in [0, 1]$. When $\beta > 2$, $f_r(r)$ increases strictly on $[0, \tilde{r}]$, where $0 < \tilde{r} < \frac{1}{2}$, and decreases strictly on $[\tilde{r}, \frac{1}{2}]$. The points $\tilde{r}$ and $1 - \tilde{r}$ achieve global maximum of $f_r(r)$. The point $r = \frac{1}{2}$ is a local minimum of $f_r(r)$. When $2 < \beta \leq \beta^*$, where $2.4 < \beta^* < 2.5$ is a constant, the global minimum of $f_r(r)$ is achieved at $r = 0, 1$. When $\beta \geq \beta^*$, the global minimum of $f_r(r)$ is achieved at $r = \frac{1}{2}$.*



(a) $\beta = 0.8$ (b) $\beta = 2.2$ (c) $\beta = 4$

Figure 1: The mixture density $f_r(r)$, $r \in [0, 1]$, scaled by $2(2\pi)^{q/2}$, along the ridgeline between $\phi(x|\mu_1, I)$ and $\phi(x|\mu_2, I)$ at several values of $\beta = ||\mu_1 - \mu_2||^2/2$.

According to Lemma 3.3, $f_r(r), r \in [0, 1]$, follows three patterns depending on β. In Figure 1, we plot $f_r(r)$ scaled by $2(2\pi)^{q/2}$ at $\beta = 0.8, 2.2, 4$. When $\beta \leq \beta^*$, $f_r(r)$ achieves the global minimum at the ends of interval $[0, 1]$. When $\beta > \beta^*$, $f_r(r)$ achieves the global minimum at $r = \frac{1}{2}$. Thus, the separability measure between $\phi(x|\mu_1, I)$ and $\phi(x|\mu_2, I)$ is

$$S(\beta) = \begin{cases} 0 & \beta \leq \beta^* \\ 1 - \frac{2e^{-\frac{1}{4}\beta}}{1+e^{-\beta}} & \beta > \beta^* \end{cases}, \qquad \beta = \frac{||\mu_1 - \mu_2||^2}{2}.$$

In the proof for Lemma 3.3, we have shown that $S(\beta)$ is strictly increasing when $\beta > \beta^*$.

Theorem 3.1 *Suppose* $g_i(x) = \phi(x^{(q)}|\mu_i^{(q)}, I)\phi(x^{(-q)}|\mu^{(-q)}, I)$, $i = 1, 2$, $|\mu_{1,1}^{(q)} - \mu_{2,1}^{(q)}| > |\mu_{1,2}^{(q)} - \mu_{2,2}^{(q)}| > \cdots > |\mu_{1,q}^{(q)} - \mu_{2,q}^{(q)}| > 0$, *and* $(\mu_{1,1}^{(q)} - \mu_{2,1}^{(q)})^2/2 > \beta^*$. *Let the mixture density* $f(x) = \frac{1}{2}g_1(x) + \frac{1}{2}g_2(x)$. *Then the forward selection based on the separability between* $g_1(x)$ *and* $g_2(x)$ *will add variables in the order of* X_1, X_2, ..., X_q. *In the first* q *steps, the separability* S *increases strictly, while after* q *steps, it remains constant*.

Theorem 3.1 justifies forward selection by the separability measure for a basic example of mixture density. Note that the AD for a two-component mixture using the true distribution (and without the practical handling of outliers) equals up to a constant scaling factor the separability measure between the two components. For the given $g_i(x)$, it is intuitive to regard a variable X_j as more informative when the difference in component means is larger. By the above theorem, the forward selection procedure adds more informative variables first, and it can detect non-informative variables by examining whether the increase in separability is zero. For variables X_i and X_j, if the difference in component means is the same, their positions in forward selection can be exchanged without affecting the separability. The theorem also justifies the practice of terminating forward selection when the increase in AD is below a threshold ϵ. Ideally, ϵ can be zero. In practice, because the true distribution is unknown, the threshold ϵ is set to be positive.

The exact solution for the separability measure S reveals that S increases strictly at every added informative variable. Because S is upper-bounded by 1, the increase in S converges to zero. Let $\hat{q}$ be the number of variables selected by thresholding the increase in S by $\epsilon \geq 0$. Consider a simplified case when the q informative variables all have the same difference Δ_μ in component means (equally informative). Then the separability achieved by $\hat{q} \leq q$ variables is $1 - \frac{2e^{-\hat{q}\Delta_\mu^2/8}}{1+e^{-\hat{q}\Delta_\mu^2/2}}$. At $\Delta_\mu = 3$, $S > 0.99$ for $\hat{q} \geq 5$. In order to find all the informative variables when $q \to +\infty$, we need to set $\epsilon = 0$. If $\epsilon > 0$, $\hat{q}$ is fixed when q grows. In practice, it is sensible to let ϵ decrease when more variables are selected, although in our experiment, we used a fixed ϵ. We note that in some applications, it is more preferable to find a small set of variables that capture the clustering structure than finding all the informative variables.

4 Experiments

We conduct experiments on four simulated and two real data sets. We compare Alg. I with the wrapper approach based on the *scatter separability criterion (SSC)* introduced by Dy and Brodley (2004), the algorithms of Pan and Shen (2007), Wang and Zhu (2008), and Sparse k-means by Witten and Tibshirani (2010). To measure the overall separability of clusters, each of which corresponds to a mixture component, SSC uses $trace(S_w^{-1}S_b)$, where the within-class scatter matrix $S_w = \sum_{k=1}^K \pi_k \Sigma_k$ and the between class scatter matrix $S_b = \sum_{k=1}^K \pi_k(\mu_k - M_0)(\mu_k - M_0)^T$. $M_0 = \sum_{k=1}^K \pi_k \mu_k$ is the global mean of the data. For selection via SSC, we employ the same density estimation as in Alg. I. Except for our method, all the other algorithms treat every mixture component as a cluster.

In many applications, the desired number of variables may have been pre-determined by practical factors. We thus show in some of the experimental results the forward selection of variables up to the full set. If the number of selected variables needs to be decided by the data based on SSC, Dy and Brodley (2004) suggested the criterion of maximizing $trace(S_w^{-1}S_b)/d$ with d being the dimension of the subspace. In the sequel, this adjusted SSC, simply referred to as SSC, is always reported. For our

algorithms, we choose the dimension by examining the increase in AD. If by adding an extra variable to a selected subset, the increase in AD is lower than a given threshold $\epsilon = 0.01$, we stop adding variables.

4.1 Moderate Dimensional Data

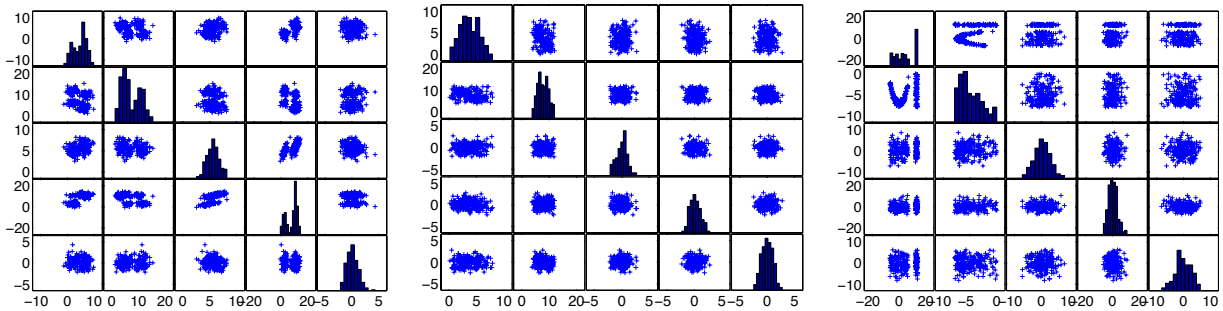
We first consider several simulated data sets when the dimension is moderate. As we will see, some clusters constructed in these data sets deviate significantly from Gaussian-type clusters. It is not obvious what the ground truth for clustering should be. Thus we focus on comparing the selection of variables.

Simulation 1: We generate a data set of size 200 and dimension 8. Four of the eight variables are informative for clustering and the rest are non-informative. Specifically, variables X_1 and X_2 are generated according to a mixture of four Gaussian components with prior probabilities $0.4, 0.2, 0.2$ and 0.2. The Gaussian component means are $(6,4)^t$, $(7,10)^t$, $(2,6)^t$, $(2,12)^t$, and the covariance matrices are $diag(1.5,1.5)$, $diag(2,2)$, $diag(1.5,1.5)$, and $diag(1.5,1.5)$. The variables X_3 and X_4 are generated independently from X_1 and X_2 and according to a two component Gaussian mixture with means $(6,11)^t$ and $(5,3)^t$, prior probabilities $\frac{2}{3}$ and $\frac{1}{3}$, and a non-diagonal common covariance matrix $\Sigma = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. The rest of the variables, $(X_5, X_6, X_7, X_8)^t \sim N(0, I)$ are generated independently from $X_1 \sim X_4$.

Simulation 2: This data set is of size 200 and dimension 8. The variables X_1 and X_2 are generated according to a mixture of two Gaussian components and a uniform distribution intended to weaken the separation between the two Gaussian components. The prior probabilities for the two Gaussians and the uniform distribution are equal. The means of the Gaussians are $(3,9)^t$ and $(5,6)^t$, and the covariance matrices are I. The uniform distribution is on the region $[0,8] \times [4,12]$ and serves as the "background noise". The other six variables $X_3 \sim X_8$ are non-informative and are sampled independently from the standard normal distribution.

Simulation 3: Again, we generate a data set of size 200 and dimension 8. X_1 and X_2 are generated according to the so-called noisy curve data introduced in (Li et al., 2007). Specifically, $(X_1, X_2)^t$ are sampled from a noisy half circle with prior probability $\frac{2}{3}$ and a noisy bar with prior $\frac{1}{3}$. The half circle centers at the origin and is of radius 7; and the bar is a straight line linking coordinates $(13,-8)^t$ and $(13,0)^t$. A point from the noisy circle (or bar) is generated by first sampling uniformly from the half circle (or bar) and then adding a Gaussian noise sampled from $N(0, \frac{1}{4}I)$. The other six variables $X_3 \sim X_8$ are generated independently from X_1 and X_2 according to $N(0, 9I)$.

Figure 2 shows the scatter plot for a data set in each simulation. Due to the space limit, only the first five variables, including at least one non-informative variable, are shown.



(a) Simulated Data 1 (b) Simulated Data 2 (c) Simulated Data 3

Figure 2: Scatter plots for the three simulated data sets.

We first compare the results of forward selection based on AD (i.e., Alg. I) and SSC. Ideally, the informative variables should be selected first, and when forward selection stops, all and only the informative variables are included. Table 1 shows the forward selection results based on one data set in each simulation. For Simulation 1, although both AD and SSC add variables X_4, X_2, X_1, X_3 sequentially, the stopping criterion by AD results in the selection of four variables, while that by SSC results in one variable. For Simulation 2, the first two variables selected by AD are the informative variables X_2 and X_1, while the first two variables by SSC are X_2 and X_3. Moreover, by the stopping criterion, SSC selects one variable, while AD selects three variables. For Simulation 3, both AD and SSC select the two informative variables X_1 and X_2 in the first two steps. They also both indicate the number of informative variables is two.

	Simulated Data 1				Simulated Data 2				Simulated Data 3			
Step	X_i	AD	X_i	SSC	X_i	AD	X_i	SSC	X_i	AD	X_i	SSC
1	4	0.4225	4	7.2078	2	0.0980	2	1.8798	1	0.5699	1	16.5203
2	2	0.5963	2	6.7515	1	0.2799	3	1.1925	2	0.6723	2	21.4296
3	1	0.6461	1	6.2178	8	0.3024	7	0.8637	4	0.6017	3	7.8447
4	3	0.6743	3	6.0443	6	0.2829	4	0.6560	3	0.6066	6	6.0372
5	7	0.6781	7	4.9614	5	0.2962	5	0.5235	6	0.6311	8	4.8973
6	8	0.6788	8	4.1537	3	0.1492	6	0.4273	8	0.6320	7	4.1792
7	6	0.6823	6	3.5731	4	0.1479	8	0.3600	7	0.6306	5	3.0319
8	5	0.6362	5	2.0695	7	0	1	0	5	0.7898	4	0.4393

Table 1: Forward selection based on AD and SSC for three simulated data sets in Simulation 1, 2, and 3 respectively. The variable added at each step is listed with the corresponding values of AD or SSC.

	Simulation 1								Simulation 2				Simulation 3			
	AD				SSC				AD		SSC		AD		SSC	
Step	X_1	X_2	X_3	X_4	X_1	X_2	X_3	X_4	X_1	X_2	X_1	X_2	X_1	X_2	X_1	X_2
1	0	1	0	199	0	5	0	195	26	174	17	182	200	0	192	8
2	0	199	0	1	1	183	11	5	81	25	56	16	0	198	8	188
3	184	0	16	0	85	11	102	0	14	0	24	0	0	1	0	2
4	16	0	126	0	102	1	77	0	5	0	8	0	0	1	0	1
5	0	0	30	0	10	0	7	0	4	0	6	0	0	0	0	1
6	0	0	10	0	1	0	1	0	4	0	7	0	0	0	0	0
7	0	0	9	0	0	0	0	0	11	0	11	0	0	0	0	0
8	0	0	9	0	1	0	2	0	55	1	71	2	0	0	0	0

Table 2: The number of times each variable is chosen at each forward selection step based on 200 sampled data sets using AD and SSC respectively.

To check whether the informative variables are consistently selected in the first few steps by AD and SSC, we repeat the above experiment by sampling a data set with the same size from the same distribution 200 times in each simulation. We then count at every step of forward selection the number of times any $X_i, i = 1, ..., 8$, is chosen across the 200 data sets. Results are shown in Table 2. Due to the lack of space, we only show the number of times an informative variable has been selected at every step.

Simulation		Truth	AD	SSC	Pan and Shen	Wang and Zhu	Sparse k-means
1	$\hat{q}_i$	4	3.56 (0.50)	1.22 (0.46)	4.00 (0.00)	3.99 (0.10)	2.92 (0.86)
	$\hat{q}_n$	0	0.26(0.46)	0.00(0.00)	2.09(1.23)	0.18(0.64)	3.88(1.28)
2	$\hat{q}_i$	2	1.55 (0.50)	1.11 (0.34)	1.61 (0.78)	0.04 (0.28)	2.00 (0.00)
	$\hat{q}_n$	0	1.49(1.04)	0.07(0.26)	0.17(0.49)	0.00(0.00)	5.39(1.33)
3	$\hat{q}_i$	2	1.97 (0.17)	1.83 (0.37)	2.00 (0.00)	1.91 (0.33)	1.99 (0.16)
	$\hat{q}_n$	0	0.29(0.57)	0.35(1.11)	2.96(1.78)	0.52(1.33)	5.82(0.45)

Table 3: Compare the selected variables by several algorithms over the 200 samples in each simulation. The average number of informative variables selected is denoted by $\hat{q}_i$, and the average number of non-informative variables selected is denoted by $\hat{q}_n$. The numbers in parenthesis are standard deviations.

In Simulation 1, for both AD and SSC, in the first two steps, only variables among $X_1 \sim X_4$ are selected. The selection of variables by AD is more consistent across data sets than SSC. By AD, X_4 is selected 199 times in Step 1 and X_2 199 times in Step 2. By SSC, X_4 is selected 195 times in Step 1 and X_2 183 times in Step 2. The better consistency held by AD is also shown by the higher frequencies of choosing X_1 at Step 3 and X_3 at Step 4. Simulation 2 presents data more challenging than Simulation 1 because we have not only non-informative variables but also data points acting as background noise. By AD, in Step 1, either X_1 or X_2 is selected; by SSC, either is selected 199 times. In Step 1 and 2, X_1 and X_2 are simultaneously selected 106 times by AD and 71 times by SSC. In this simulation, possibly due to the background noise, AD performs clearly more robustly than SSC, which relies heavily on the assumption of Gaussian-type clusters. Similarly we see from the table that for Simulation 3 where the clusters deviate substantially from Gaussian, AD performs better than SSC.

We now compare the variable selection results obtained by the algorithms of Pan and Shen (2007), Wang and Zhu (2008), and Sparse k-means by Witten and Tibshirani (2010). Software packages provided by the authors are used. For the algorithms "Pan and Shen" and "Wang and Zhu", BIC is used to select the number of components in the mixture model and the penalty parameter λ. The number of components in both algorithms ranges from 1 to 10. The algorithm "Pan and Shen" is not sensitive to λ. We thus take $[0, 20]$, the default range of λ in the software, and use grid points equally spaced by 2. The algorithm "Wang and Zhu" is sensitive to λ. We use equally spaced values between 0 and 0.1 at step size 0.02 and values between 0.1 and 2 at step size 0.1. We found that at $\lambda = 2$ (usually even at 1), the algorithm "Wang and Zhu" shrinks the component means of all the variables to zero. Hence, it is unnecessary to use larger λ. For Sparse k-means, the permutation method recommended by the authors and provided in the software is used to select the tuning parameter. The number of clusters needs to be specified for Sparse k-means. For Simulation 1 or 2, we set the number to 8 or 2 because there are these many components in the mixture distribution. For Simulation 3, we set the number to 2 because the intention is to partition the noisy half circle and the bar (Figure 2 (c)).

Table 3 shows the average number of selected informative variables, denoted by $\hat{q}_i$, and the average number of non-informative variables selected, denoted by $\hat{q}_n$, along with their standard deviations over the 200 samples. The true number of informative variables is q. For Simulation 1, 2, 3, $q = 4, 2, 2$ respectively. For $\hat{q}_i$, a larger value is better with the maximum possible being q. A smaller value for $\hat{q}_n$ is better with the minimum being zero.

For Simulation 1, AD and "Wang and Zhu" obtain $\hat{q}_i$ close to $q = 4$ and $\hat{q}_n$ close to zero, with "Wang and Zhu" performing slightly better. SSC misses more than half of the informative variables although

the variables it selects are always correct ($\hat{q}_n = 0$). "Pan and Shen" and Sparse K-means select more than 2 non-informative variables on average. "Pan and Shen" always includes the informative variables in the selection ($\hat{q}_i = q$), while Sparse k-means misses more than one informative variable on average. For Simulation 2, "Wang and Zhu" selects no variable for 196 samples out of the 200, which is reflected by the nearly zero $\hat{q}_i$ and $\hat{q}_n$. In the other extreme, Sparse k-means selects all the 8 variables for 147 samples, yielding $\hat{q}_i + \hat{q}_n$ close to 8. "Pan and Shen" yields $\hat{q}_i = 1.61$ and a small $\hat{q}_n = 0.17$, which is the best among the methods. On average, AD selects 1.55 informative variables and 1.49 non-informative ones. SSC again yields a small $\hat{q}_i$ and $\hat{q}_n$. For Simulation 3, all the methods achieve $\hat{q}_i$ close to 2. In terms of $\hat{q}_n$, AD achieves the lowest value of 0.29. "Pan and Shen" selects on average about 3 non-informative variables. Sparse k-means selects all the 8 variables for 162 samples, yielding $\hat{q}_i + \hat{q}_n$ close to 8.

In summary, AD performs well consistently across the three simulations, never seriously missing informative or selecting non-informative variables. SSC misses informative variables severely. "Pan and Shen" selects about two or three non-informative variables for Simulation 1 or 3. "Wang and Zhu" failed nearly on every sample to select any informative variable for Simulation 2. Sparse K-means always selects too many variables. In Simulation 2 and 3, it selects all the variables more than 70% times.

4.2 High Dimensional Data ($\mathbf{n < p}$)

Next, we study a simulation where the dimension p can be larger than the sample size n. In the experiments of Pan and Shen (2007), Wang and Zhu (2008), and Witten and Tibshirani (2010), simulations with $p \gg n$ and $q > n$ are discussed, where q is the number of informative variables. Because our algorithm relies on the geometry of the density, at a certain step of forward selection, the number of variables becomes too large for a reasonable density estimation, and hence it is not meaningful to add more variables. This is an intrinsic limitation of our method. However, we will show through simulation that when q is moderate, even with $p > n$, our method has advantages over the others.

In this simulation, we have $K = 5$ mixture components, each containing n_0 points. Thus, $n = K \cdot n_0$. Given the component label, the variables are independent. The non-informative variables follow $N(0, 1)$. In each component k, the variable $X_k \sim N(\mu, 1)$, while $X_{k'} \sim N(0, 1)$, $k' \in \{1, 2, ..., q\}$, $k' \neq k$. Thus X_1, X_2, ..., X_5 are informative except for small μ (to be elaborated shortly). We compare the variable selection methods under different combinations of μ, p, and n_0. Specifically, we let $n_0 \in \{10, 20\}$, $p \in \{50, 100, 200\}$, and $\mu \in \{1.5, 3, 5\}$. Under every (μ, p, n_0), experiments are repeated on 50 samples.

When $\mu = 1.5$, as shown in Section 3.3, if we consider any pair of components, their equally weighted density functions merge into a single mode mixture. The ridgeline between the component means follows a pattern shown in Figure 1(a). Thus, although there are multiple components in the mixture, there is no clustering structure. The ground truth clustering in this case puts all the data points in one cluster. We also consider that none of the variables is informative ($q = 0$). On the other hand, the squared distance between any pair of component means equals 4.5, close to the threshold $2\beta^* \in (4.8, 5)$ at which two modes occur. It is interesting to investigate how the methods will respond in this near boundary case without a clustering structure. When $\mu = 3, 5$, we can show that the ridgeline between any pair of components follows the pattern shown in Figure 1(c) and the components are well separated. The number of clusters in these cases is $K = 5$, and there are $q = 5$ informative variables.

Recall that for forward selection via AD, we stop adding variables when the increase in AD is below $\epsilon = 0.01$. If in the first step, the best AD is below 0.01, the algorithm selects no variable. Due to the small data size, we only perform a maximum of 10 steps of forward selection. The effect of this restriction is negligible. We find that when $p = 50, 100$, the forward selection has always stopped before reaching the

10th step. When $p = 200$, the 10th step is reached for only about 3% of the sample data sets. Moreover, we constrain the mixture model fitted by *Mclust* to have a common diagonal covariance matrix, as are the models in "Pan and Shen" and "Wang and Zhu".

For algorithms "Pan and Shen" and "Wang and Zhu", we still use BIC to pick the number of mixture components and the penalty parameter λ. When $\mu = 1.5$, we let the number of components range from 1 to 10. When $\mu = 3, 5$, we let the range be 2 to 10. We found that especially for "Wang and Zhu", even when $\mu = 3, 5$, BIC often picks $K = 1$, implying there exists no clustering. We thus give the two algorithms some advantage by restricting the potential $K \geq 2$. For "Pan and Shen", we let λ take values in $[0, 20]$ spaced at step size 1. For "Wang and Zhu", we let λ take values in $[0, 0.1]$ spaced at step size 0.02 and values in $[0.1, 2]$ spaced at step size 0.1. For Sparse k-means, we use the recommended permutation method to choose the tuning parameter, and set $K = 5$. We will not report result for Sparse k-means when $\mu = 1.5$ because if we specify $K = 1$, the algorithm is fully informed about the lack of clustering, but if we specify $K = 5$, the algorithm is misled.

μ	p	n_0	$\hat{q}_i$				$\hat{q}_n$				Adjusted Rand Index			
			AD	Pan	Wang	SK	AD	Pan	Wang	SK	AD	Pan	Wang	SK
1.5	50	10	0.00	0.00	0.00		1.94	0.00	0.00		0.38	1.00	1.00	
		20	0.00	0.00	0.00		0.52	0.00	0.00		0.80	1.00	1.00	
	100	10	0.00	0.00	0.00		3.96	0.00	0.00		0.08	1.00	1.00	
		20	0.00	0.00	0.00		1.32	0.00	0.00		0.58	1.00	1.00	
	200	10	0.00	0.00	0.00		5.38	0.00	0.32		0.02	1.00	0.98	
		20	0.00	0.00	0.00		3.54	0.00	0.00		0.32	1.00	1.00	
3.0	50	10	1.58	0.06	0.56	2.16	2.24	0.02	4.04	17.04	0.25	0.01	0.04	0.08
		20	3.46	0.06	0.62	2.48	1.20	0.00	2.50	17.92	0.68	0.01	0.07	0.14
	100	10	1.08	0.00	0.40	1.34	3.54	0.00	6.58	24.76	0.14	0.00	0.02	0.05
		20	2.82	0.00	0.46	1.48	1.82	0.00	3.58	21.24	0.56	0.00	0.03	0.05
	200	10	0.56	0.00	0.30	1.70	4.36	0.00	9.10	63.32	0.10	0.00	0.01	0.03
		20	1.80	0.00	0.22	1.34	3.12	0.00	7.40	47.08	0.37	0.00	0.02	0.03
5.0	50	10	3.94	4.06	1.00	2.82	0.62	11.84	2.98	15.54	0.92	0.79	0.16	0.28
		20	4.88	5.00	2.32	3.88	0.04	5.74	3.62	17.74	1.00	1.00	0.42	0.53
	100	10	3.12	1.24	0.84	1.62	1.66	5.10	6.08	23.60	0.74	0.21	0.13	0.09
		20	4.88	4.54	1.06	2.28	0.04	37.18	3.78	29.46	1.00	0.87	0.19	0.23
	200	10	1.78	0.14	0.38	1.38	2.50	1.36	9.42	52.46	0.55	0.03	0.04	0.04
		20	4.78	0.64	0.52	1.64	0.08	8.08	5.64	46.92	1.00	0.12	0.05	0.07

Table 4: Comparison of variable selection and clustering by several algorithms. "Pan" refers to the algorithm "Pan and Shen", "Wang" refers to "Wang and Zhu", SK refers to the Sparse k-means. Average results over 50 samples under each setup of (μ, p, n_0) are shown in terms of the number of informative variables selected $\hat{q}_i$, the number of non-informative variables selected $\hat{q}_n$, and the ARI.

Table 4 summarizes the average results over 50 samples under every (μ, p, n_0) in terms of the number of informative variables selected ($\hat{q}_i$), the number of non-informative variables selected ($\hat{q}_n$), and the Adjusted Rand Index (ARI) between the clustering result and the ground truth. For $\hat{q}_i$, the maximum possible (optimal) value is $q = 5$ when $\mu = 3, 5$ and $q = 0$ when $\mu = 1.5$. A larger value for ARI indicates better agreement between clustering results, with 1 implying identical clustering.

When $\mu = 1.5$, "Pan and Shen" selects no variables under all (p, n_0), while "Wang and Zhu" selects

no variables except at $p = 200$ and $n_0 = 10$. Both of these algorithms capture well the fact there exists no clusters. By AD, when $p = 50, n_0 = 20$, on average only 0.52 variables are selected; when $p = 200$, $n_0 = 10$ (more challenging data), 5.38 variables are selected.

When $\mu = 3.0$, AD performs the best in terms of $\hat{q}_i$, $\hat{q}_n$, and ARI. "Pan and Shen" performs the worst. It selects zero or close to zero number of variables, and fails to identify the clusters, as indicated by the nearly zero ARI. Although "Wang and Zhu" selects some variables, they are mostly non-informative. As a result, the ARI values are close to zero, implying that the clusters found are largely random partition of the data. The Sparse k-means on the other hand selects a large number of non-informative variables. Although the number of informative variables selected by Sparse k-means is comparable to that by AD, due to the noise from the many non-informative variables selected, the ARI by Sparse k-means is substantially lower. In fact, the ARI is below 0.1 except for $(p, n_0) = (50, 20)$.

When $\mu = 5$, AD again outperforms all the other algorithms. At $n_0 = 20$ and $p = 50, 100, 200$, $\hat{q}_i \geq 4.78$ (close to $q = 5$) and $\hat{q}_n \leq 0.08$. In fact, at $n_0 = 20$, regardless of p, AD achieves perfect selection on more than 90% of the samples (i.e., $\hat{q}_i = q$, $\hat{q}_n = 0$). The ARI by AD at $n_0 = 20$ and any p is 1.0, indicating identical clustering as the ground truth. Although "Pan and Shen" achieves high values for $\hat{q}_i$ at $(p, n_0) = (50, 10), (50, 20), (100, 20)$, it pays the cost of high values for $\hat{q}_n$. The ARI by "Pan and Shen" is substantially lower than that by AD in most cases. Comparing with "Pan and Shen", "Wang and Zhu" obtains lower values in $\hat{q}_n$ as well as $\hat{q}_i$. The ARI by "Wang and Zhu" is much worse than by "Pan and Shen" under most setups. Sparse k-means over selects variables severely, as shown by the high values of $\hat{q}_n$ across all the setups. The ARI by Sparse k-means is also poor.

Although AD selects some non-informative variables when there is no clustering structure, the average number of non-informative variables selected reduces when the clusters are better separated, for instance, at $\mu = 5$. While for all the other algorithms, the selection of non-informative variables is serious even at the well separated case of $\mu = 5$. In fact, for "Pan and Shen", many more non-informative variables are selected at $\mu = 5$ than at $\mu = 3$.

4.3 Real Data Sets

We experiment with Alg. I, Alg. II, and forward selection based on SSC using two real data sets. For the second real data set, the true class labels are given; and hence we compare the cluster labels with the class labels. We note that however, because there is no reason for different classes to form well separated clusters, a poor agreement between the class labels and the clustering result does not necessarily imply the clustering algorithm is bad. On the other hand, if strong agreement is observed, it indicates that the clustering structure corresponds well with some physical meaning, a justification for the potential usefulness of the clustering analysis.

Infant Attention Time Data: This data set, originally provided by Hoben Thomas in the Department of Psychology at Penn State University, was used by Li et al. (2007) for clustering. The data set records attention time of 51 infants to 11 repeated stimuli spaced equally in time. Each infant corresponds to one data point. The data set is thus of size 51 and dimension 11. It is observed that there is no obvious clustering structure within the 11 dimensional data (Li et al., 2007). Here, we would like to see whether restricting to a subset of variables results in stronger clusters and what are those variables.

Table 5 provides the variable selection results by Alg. I, SSC, and Alg. II. Both Alg. I and SSC yield the same subset of variables in the first four steps of forward selection. The top two variables that exhibit a clustering structure are X_5 and X_{11}. The clustering result based on X_5 and X_{11} obtained by Alg. I is shown in Figure 3. There are five clusters based on these two variables, which can be described

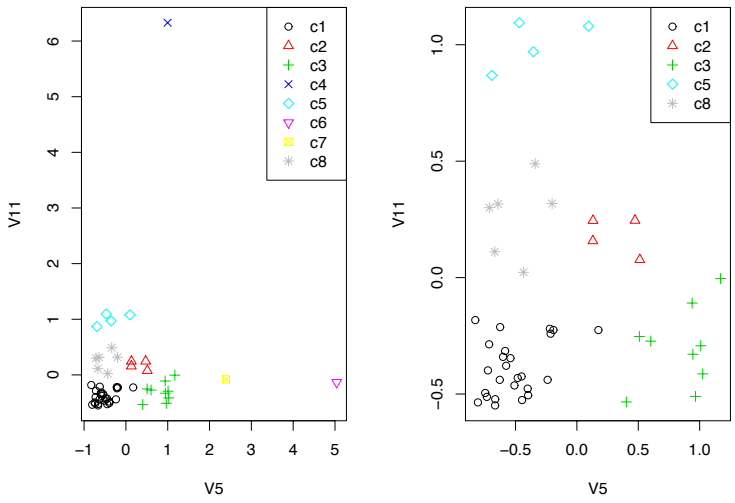


Figure 3: Infant data: The scatter plots show the clustering result before (left) and after (right) removing outlier clusters.

roughly as groups of infants with long attention time towards the end (large X_{11}), with moderate increase of attention towards the end (small X_5 and medium X_{11}), with moderate and stable attention across time (medium X_5 and X_{11}), with little attention across time (small X_5 and X_{11}), and with decreasing attention across time (large X_5 and small X_{11}).

According to results in Table 5, both Alg. I and SSC indicate when the full set of variables is used, a clustering structure hardly exists. It is thus not meaningful to apply Alg. II which attempts to select variables to retain the clustering structure shown by all the variables together. As demonstrated in Table 5, Alg. II fails to select variables for this data set because *Mclust* determines that the data in the original space should be modeled by a two component mixture and the two components happen to share a common mode. Thus in the original space, the two components are merged into one cluster, or equivalently, no clustering structure is found.

Wine Data: This data set is taken from the UCI Machine Learning Repository. The data record chemical element results of wine samples taken from a region of Italy. There are 178 observations, each containing 13 measurements. The wines are labeled into 3 classes. For this data set, we assess the effectiveness of our clustering and variable selection approaches by comparing the cluster labels with the true class labels. Alg. I, II, and forward selection with SSC are all applied in a completely unsupervised manner, that is, the class labels are not used in any way during clustering. For evaluation, a cluster generated by any of the unsupervised methods is assigned with a class label by majority vote on the class labels of data points it contains. We then compute the classification error rate.

We also compare with the usual Gaussian mixture-based clustering approach, referred to as GMM. The number of components in the mixture model is set to 3, the true number of classes. The parameters of each Gaussian component are initialized using data in one of the three classes. The EM algorithm is then used to estimate the mixture model. In this scheme, although the mixture modeling is unsupervised, it does exploit the class labels in a weak manner by initializing each mixture component using within class data. Once the clustering result is obtained, as with Alg. I and II, we also use majority vote to assign

	Alg. I				SSC				Alg. II			
Step	X_i	AD	EFC	Out.	X_i	SSC	EFC	Out.	X_i	AD	EFC	Out.
1	5	0.4784	4	0	5	4.4480	5	0	1	0	1	0
2	11	0.5465	5	3	11	9.5110	5	0	2	0	1	0
3	10	0.5650	6	3	10	6.9341	6	0	3	0	1	0
4	7	0.4285	6	0	7	2.4432	6	0	4	0	1	0
5	4	0.5924	7	0	3	2.4533	5	0	5	0	1	0
6	6	0.4620	6	0	4	0.6861	4	0	6	0	1	0
7	8	0.7197	8	0	6	1.4005	5	0	7	0	1	0
8	3	0.2075	3	0	2	1.5707	5	0	8	0	1	0
9	1	0.1811	3	0	8	0.3653	3	0	9	0	1	0
10	9	0.0002	2	0	9	0.3235	3	0	10	0	1	0
11	2	0	1	0	1	0	2	0	11	0	1	0

Table 5: Infant data: Forward selection by Alg. I, SSC, and Alg. II.

a class to each cluster and compute the classification error rate likewise. Finally, we conduct supervised classification using mixture discriminant analysis (MDA) (package *mda* in R) and compute the error rate by 10-fold cross validation. Since MDA is a supervised classification method, it is expected to perform better than the unsupervised approaches; and is intended as a baseline for comparison.

The variable selection results by Alg. I, SSC, and Alg. II respectively, and the comparison of classification error rates based on selected subsets of variables are provided in Table 6. With Alg. I, AD increases steadily until 7 variables are included. The classification error rate achieved by Alg. I with the 7 variables is 0.0337, which differs little from 0.0331, the 10-fold cross-validation error rate achieved by MDA using the same subset of variables. The error rate by GMM, 0.0618, is higher. If forward selection based on SSC is used, with the subset of 7 variables, the error rate obtained by clustering is 0.0506. SSC suggests the subset of variables $\{12, 7, 13\}$ is most preferable since the maximum value of SSC is achieved at this subset. If we use this subset, the error rate obtained by SSC clustering is 0.0678.

According to the variable selection result of Alg. I, when all the variables are used, the value of AD is close to that based on the subset of 7 variables. We thus expect for this data set, the clustering structure is well retained in the original dimension, which is confirmed by the result based on Alg. II (the third section in Table 6). In fact, in terms of classification error rates, the full set of variables is preferred because the error rate based on Alg. I achieves the minimum value of 0.0169. Table 6 shows that by attempting to retain the clustering structure in the full space, Alg II achieves an error rate of 0.0169 using only 3 variables. The AD based on a subset of 5 variables reaches 0.6239, which is higher than 0.6086 achieved by the full set of variables.

5 Conclusions

In this paper, we developed variable selection methods for clustering by exploiting the geometric features of the density function in the form of a Gaussian mixture model. We measure the separability between two clusters by computing the density along the ridgeline connecting the modes of the clusters. To integrate the pairwise separability, an aggregated distinctiveness (AD) was introduced and used as the criterion in forward selection of variables. Theoretical analysis of the forward selection procedure is

Step	X_i	AD	EFC	Outlier	Error rate Alg. I	Error rate GMM	CV error rate (std) MDA
1	12	0.2277	2	0	0.4269	0.3933	0.2871 (0.0033)
2	13	0.4588	3	0	0.1180	0.1180	0.0860 (0.0017)
3	1	0.5590	4	0	0.0787	0.0674	0.0599 (0.0019)
4	7	0.5994	3	0	0.0562	0.0562	0.0397 (0.0013)
5	8	0.6607	5	0	0.0787	0.0449	0.0407 (0.0012)
6	5	0.6937	6	0	0.0450	0.0449	0.0331 (0.0013)
7	6	0.7293	7	0	0.0337	0.0618	0.0331 (0.0013)
8	3	0.7123	6	0	0.0787	0.0337	0.0243 (0.0009)
9	2	0.7225	6	0	0.0787	0.0337	0.0213 (0.0018)
10	4	0.6743	4	0	0.0618	0.0617	0.0147 (0.0013)
11	10	0.7312	6	0	0.0112	0.0450	0.0018 (0.0007)
12	11	0.6539	3	0	0.0169	0.0225	0.0004 (0.0004)
13	9	0.7102	6	0	0.0169	0.0225	0.0007 (0.0005)
Step	X_i	SSC	EFC	Outlier	SSC	GMM	MDA
1	12	3.0342	2	0	0.4270	0.4382	0.2879 (0.0037)
2	13	2.6844	3	0	0.1180	0.1180	0.0860 (0.0017)
3	7	5.1539	5	1	0.0678	0.0847	0.0688 (0.0017)
4	6	4.7493	6	1	0.0621	0.0904	0.0592 (0.0010)
5	5	4.5807	8	1	0.1921	0.1186	0.0484 (0.0015)
6	3	3.7119	8	1	0.0452	0.0621	0.0507 (0.0015)
7	1	2.9367	7	0	0.0506	0.0337	0.0276 (0.0011)
8	10	2.5157	6	1	0.0452	0.0169	0.0251 (0.0021)
9	9	3.4021	7	1	0.0339	0.0169	0.0207 (0.0010)
10	4	2.0914	6	0	0.0225	0.0169	0.0088 (0.0012)
11	8	2.4421	7	0	0.1685	0.1517	0.0007 (0.0005)
12	2	1.3785	4	0	0.0843	0.0225	0.0007 (0.0005)
13	11	1.4392	7	0	0.0169	0.0225	0.0011 (0.0006)
Step	X_i	AD	EFC	Outlier	Alg. II	GMM	MDA
1	7	0.0559	2	0	0.5169	0.5842	0.2901 (0.0036)
2	10	0.3971	4	0	0.1292	0.0955	0.1165 (0.0026)
3	13	0.5215	5	0	0.0169	0.0674	0.0665 (0.0022)
4	1	0.5851	5	0	0.0169	0.1011	0.0335 (0.0016)
5	11	0.6239	5	0	0.0169	0.0787	0.0243 (0.0012)
6	12	0.6467	6	0	0.0169	0.0337	0.0228 (0.0015)
7	2	0.6631	5	0	0.0169	0.0281	0.0103 (0.0017)
8	4	0.6723	5	0	0.0169	0.0393	0.0114 (0.0012)
9	8	0.6787	5	0	0.0169	0.0112	0.0018 (0.0007)
10	3	0.6732	5	0	0.0169	0.0112	0.0007 (0.0005)
11	6	0.6587	6	0	0.0169	0.0112	0.0015 (0.0006)
12	5	0.6284	6	0	0.0169	0.0112	0.0022 (0.0007)
13	9	0.7102	6	0	0.0169	0.0225	0.0015 (0.0006)

Table 6: Wine data: Variable selection by Alg. I, forward selection with SSC, and Alg. II respectively, and the classification error rates based on selected subsets of variables

provided. We conducted experiments on both simulated and real data sets, and compared with several state-of-the-art variable selection methods for clustering. A software package in R for the variable selection algorithms can be accessed at *http://www.stat.psu.edu/∼jiali/varsel_ridgeline*.

Two versions of variable selection have been investigated with slightly different practical purposes. Alg. I attempts to find a subset of variables under which the clustering structure is strongest, as measured by AD. The essence of Alg. II is to search a subset of variables under which a given partition of data is well preserved. In our current study, this given partition is assumed to be the clustering result in the full space. Hence, in its present form, Alg. II is meaningful only when the clustering structure is clear in the full space and is intended to be retained. As a potential future work, Alg. II can be used to select variables for classification, where the class labels are given, thus providing a partition.

Supplemental Materials

Appendix to Variable Selection for Clustering by Separability Based on Ridgelines: In Appendix A, we derive the M-step in the MEM and REM algorithms for a general Gaussian mixture model. Appendix B provides the proofs for Lemman 3.1, 3.2, 3.3, and Theorem 3.1. (appendix.pdf)

R-package for the two variable selection algorithms by ridgeline-based separability: The R-package contains codes for Alg I, Alg II, and for generating the simulated data sets. Files for the real data sets are also included. A "Readme" file is provided in the package. (varsel_ridgeline.tar, tar file)

Acknowledgment

We would like to thank Bruce Lindsay and Don Richards for valuable discussions. The research is supported by NSF DMS-0705210 and CCF-0936948. We run most of the experiments using the CyberSTAR cluster computers at Penn State, supported by NSF OCI-0821527. The final revision of manuscript was done when Jia Li was a Program Director at the National Science Foundation in 2011. Any opinions, findings, and conclusions or recommendations expressed in this article are those of the authors and do not necessarily reflect the views of the Foundation.

Appendix A

We now prove the equations in the M-steps of the MEM and REM algorithms for a general Gaussian mixture. According to the MEM algorithm, the update of $x^{(r+1)}$ is solved by:

$$x^{(r+1)} = \arg\max_x \sum_{k=1}^{K} p_k \cdot \log f_k(x) = \arg\min_x \sum_{k=1}^{K} p_k \cdot (x-\mu_k)^t \Sigma_k^{-1} (x-\mu_k)\,.$$

Since $\sum_{k=1}^{K} p_k \cdot (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)$ is a convex function of x, there exists a unique minimum solved by setting the first order derivative to zero:

$$\frac{\partial \sum_{k=1}^{K} p_k \cdot (x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)}{\partial x} = 2\sum_{k=1}^{K} p_k \cdot \Sigma_k^{-1}(x-\mu_k) = 0\,.$$

Hence $x^{(r+1)} = \left(\sum_{k=1}^{K} p_k \cdot \Sigma_k^{-1}\right)^{-1} \cdot \left(\sum_{k=1}^{K} p_k \cdot \Sigma_k^{-1}\mu_k\right)$.

For clusters with density $g_i(x) = \sum_{k=1}^{T_i} \pi_{i,k}\phi(x|\mu_{i,k}, \Sigma_{i,k})$, $i = 1, ..., M$, REM updates $x^{(r+1)}$ on the ridgeline between cluster i and j by the equation (Li et al., 2007):

$$x^{(r+1)} = \arg\max_x (1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\log\phi(x|\mu_{i,k},\Sigma_{i,k}) + \alpha\sum_{k=1}^{T_j} p_{j,k}\log\phi(x|\mu_{j,k},\Sigma_{j,k})\,.$$

Again, $x^{(r+1)}$ is solved by setting the first order derivative of the objective function to zero:

$$(1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\Sigma_{i,k}^{-1}(x^{(r)}-\mu_{i,k}) + \alpha\sum_{k=1}^{T_j} p_{j,k}\Sigma_{j,k}^{-1}(x^{(r)}-\mu_{j,k}) = 0\,.$$

We thus have $x^{(r+1)} = A^{-1}\bar{\mu}$, where

$$A = (1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\Sigma_{i,k}^{-1} + \alpha\sum_{k=1}^{T_j} p_{j,k}\Sigma_{j,k}^{-1}\,,$$

$$\bar{\mu} = (1-\alpha)\sum_{k=1}^{T_i} p_{i,k}\Sigma_{i,k}^{-1}\mu_{i,k} + \alpha\sum_{k=1}^{T_j} p_{j,k}\Sigma_{j,k}^{-1}\mu_{j,k}\,.$$

Appendix B

We now prove Lemma 3.1, 3.2, 3.3, and Theorem 3.1.

Proof of Lemma 3.1

By (1), $\mathcal{L} = \{x : (1-\alpha)\nabla\log g_1(x) + \alpha\nabla\log g_2(x) = 0\}$. Substituting in $g_i(x)$, $i = 1, 2$, we can show that $(1-\alpha)\nabla\log g_1(x) + \alpha\nabla\log g_2(x) = 0$ is equivalent to

$$(1-\alpha)\nabla\log\phi(x^{(q)}|\mu_1^{(q)}, I) + \alpha\nabla\log\phi(x^{(q)}|\mu_2^{(q)}, I) = 0, \text{ and } \quad \nabla\log\phi(x^{(-q)}|\mu^{(-q)}, I) = 0\,.$$

The former implies $x^{(q)} \in \mathcal{L}^{(q)}$, and the latter implies $x^{(-q)} = \mu^{(-q)}$. Thus

$$\mathcal{L} = \left\{x : x^{(q)} \in \mathcal{L}^{(q)}, x^{(-q)} = \mu^{(-q)}\right\}.$$

Proof of Lemma 3.2

We only need to show that (a) for $\forall \alpha \in [0,1]$, the solution of $(1-\alpha)\nabla \log \phi(x|\mu_1, I) + \alpha \nabla \log \phi(x|\mu_2, I) = 0$ bears the form of $(1-r)\mu_1 + r\mu_2$ for some $r \in [0,1]$, and (b) for $\forall r \in [0,1]$, $\exists \alpha \in [0,1]$, such that the ridgeline equation holds: $(1-\alpha)\nabla \log \phi(x|\mu_1, I) + \alpha \nabla \log \phi(x|\mu_2, I) \mid_{x=(1-r)\mu_1 + r\mu_2} = 0$.

To prove condition (a), we have

$$(1-\alpha)\nabla \log \phi(x|\mu_1, I) + \alpha \nabla \log \phi(x|\mu_2, I) = \zeta_1(x-\mu_1) + \zeta_2(x-\mu_2),$$

where ζ_i's are non-negative scalars with

$$\zeta_1 = (1-\alpha)e^{\frac{-(x-\mu_1)^t(x-\mu_1)}{2}}, \quad \zeta_2 = \alpha e^{\frac{-(x-\mu_2)^t(x-\mu_2)}{2}}.$$

Note that $\zeta_1 + \zeta_2 > 0$. Then $\zeta_1(x-\mu_1) + \zeta_2(x-\mu_2) = 0$ is solved by

$$x = \frac{\zeta_1}{\zeta_1 + \zeta_2}\mu_1 + \frac{\zeta_2}{\zeta_1 + \zeta_2}\mu_2.$$

Thus the solution to the ridgeline equation bears the form of $(1-r)\mu_1 + r\mu_2$ with $r \in [0,1]$.

We now prove condition (b). Obviously, when $r = 0$ (or 1), the ridgeline equation holds at $\alpha = 0$ (or 1). Hence we only need to consider $r \in (0,1)$. The ridgeline equation can be simplified to

$$-(1-\alpha)re^{-r^2\beta} + \alpha(1-r)e^{-(1-r)^2\beta} = 0,$$

which is equivalent to

$$\frac{1-\alpha}{\alpha} = \frac{1-r}{r}e^{(2r-1)\beta},$$

where $\beta = \frac{||\mu_1 - \mu_2||^2}{2}$. Because for $r \in (0,1)$, $\frac{1-r}{r}e^{(2r-1)\beta} > 0$, and $\frac{1-\alpha}{\alpha}$ is a continuous function with $\lim_{\alpha \to 0^+} \frac{1-\alpha}{\alpha} = +\infty$ and $\frac{1-\alpha}{\alpha}|_{\alpha=1} = 0$, for $\forall r \in (0,1)$, $\exists \alpha \in (0,1)$ such that the ridgeline equation holds.

Proof of Lemma 3.3

We can show that $f_r(r) = ae^{-r^2\beta} + ae^{-(1-r)^2\beta}$, where $a = 1/[2(2\pi)^{q/2}]$ and $\beta = \frac{||\mu_1 - \mu_2||^2}{2}$. It is obvious that $f_r(r)$ is symmetric around $\frac{1}{2}$ on $[0,1]$. Since $f_r(r)$ is a smooth function on $[0,1]$, any interior point that is a global minimum or maximum should have zero first order derivative.

$$f_r'(r) = -2\beta a[re^{-r^2\beta} - (1-r)e^{-(1-r)^2\beta}].$$

Setting $f_r'(r) = 0$, we get

$$\beta = \log \frac{1-r}{r}/(1-2r)\,.$$

By symmetry of $f_r(r)$ around $\frac{1}{2}$, we only need to analyze $r \in [0, \frac{1}{2}]$. Consider function $h(r) = \log \frac{1-r}{r}/(1-2r)$.

$$h'(r) = \left[\frac{2r-1}{r(1-r)} - 2\log\frac{r}{1-r}\right]/(1-2r)^2\,.$$

Let

$$\kappa(r) = \frac{2r-1}{r(1-r)} - 2\log\frac{r}{1-r}\,.$$

Because

$$\kappa'(r) = (2r-1)^2/[r(1-r)]^2 \geq 0$$

for $r \in (0, \frac{1}{2}]$, $\kappa(r)$ is strictly increasing on $(0, \frac{1}{2}]$. Because $\kappa(\frac{1}{2}) = 0$,

$$h'(r) = \kappa(r)/(1-2r)^2 < 0, \text{for } r \in (0, \frac{1}{2})\,.$$

Hence $h(r)$ is strictly decreasing on $(0, \frac{1}{2})$. Because $\lim_{r\to\frac{1}{2}^-} h(r) = 2$, $h(r) > 2$ for $r \in (0, \frac{1}{2})$. Because $\lim_{r\to 0^+} h(r) = +\infty$, for $\forall \beta > 2$, there exists a unique $r \in (0, \frac{1}{2})$, such that $\beta = \log \frac{1-r}{r}/(1-2r)$, that is, $f_r'(r) = 0$; and for $\forall \beta \leq 2$, there exists no $r \in (0, \frac{1}{2})$ such that $f_r'(r) = 0$.

It is obvious for $\forall \beta$,

$$f_r'(r)\mid_{r=\frac{1}{2}} = 0\,, \quad f_r'(r)\mid_{r=0} > 0\,.$$

When $\beta \leq 2$, because $f_r'(r) = 0$ has no solution on $(0, \frac{1}{2})$ and $f_r'(r)$ is a smooth function, $f_r'(r) > 0$ for $r \in [0, \frac{1}{2})$. Hence, $f_r(r)$ increases strictly on $[0, \frac{1}{2}]$. By symmetry of $f_r(r)$, $r = \frac{1}{2}$ is the global maximum for $f_r(r)$, $r \in [0, 1]$. When $\beta > 2$, let the unique solution of $r \in (0, \frac{1}{2})$ for $f_r'(r) = 0$ be $\tilde{r}$. Then $f_r'(r) > 0$ for $r \in [0, \tilde{r})$. When $\beta > 2$, because

$$f_r''(r)\mid_{r=\frac{1}{2}} = 2\beta a(\beta-2)e^{-\frac{1}{4}\beta} > 0\,, \quad f_r'(r)\mid_{r=\tilde{r}} = 0\,, \quad f_r'(r)\mid_{r=\frac{1}{2}} = 0\,,$$

and $f_r'(r) = 0$ has no solution on $(\tilde{r}, \frac{1}{2})$, we can show that $f_r'(r) < 0$ on $(\tilde{r}, \frac{1}{2})$. Hence, $f_r(r)$ increases strictly on $[0, \tilde{r}]$, decreases strictly on $[\tilde{r}, \frac{1}{2}]$. A global maximum of $f_r(r)$ is achieved at $\tilde{r}$ and its symmetric point $1-\tilde{r}$, and a local minimum is achieved at $\frac{1}{2}$.

$r = \frac{1}{2}$ is a global minimum of $f_r(r)$ if $f_r(\frac{1}{2}) < f(0)$. Note that

$$f_r(\frac{1}{2}) = 2ae^{-\frac{1}{4}\beta}\,, \quad f_r(0) = a(1+e^{-\beta})\,.$$

Consider function

$$\tau(\beta) = f_r(\frac{1}{2})/f_r(0) = \frac{2e^{-\frac{1}{4}\beta}}{1+e^{-\beta}}\,.$$

Note

$$\tau'(\beta) = \frac{-\frac{1}{2}e^{-\frac{1}{4}\beta} + \frac{3}{2}e^{-\frac{5}{4}\beta}}{(1+e^{-\beta})^2}\,.$$

Because $-\frac{1}{2}e^{-\frac{1}{4}\beta} + \frac{3}{2}e^{-\frac{5}{4}\beta} = 0$ has a unique solution at $\log 3 < 2$ and $-\frac{1}{2}e^{-\frac{1}{4}\beta} + \frac{3}{2}e^{-\frac{5}{4}\beta}\,|_{\beta=2} < 0$, we have $-\frac{1}{2}e^{-\frac{1}{4}\beta} + \frac{3}{2}e^{-\frac{5}{4}\beta} < 0$ for all $\beta \geq 2$. Thus $\tau(\beta)$ decreases strictly for $\beta > 2$. Because $\tau(2) > 1$, $\lim_{\beta\to+\infty}\tau(\beta) = 0$, $\tau(\beta) = 1$ has a unique solution at $\beta^* > 2$. Numerical calculation shows $2.4 < \beta^* < 2.5$. We thus conclude when $2 < \beta \leq \beta^*$, $r = 0, 1$ are the global minimum of $f_r(r)$ and when $\beta \geq \beta^*$, $r = \frac{1}{2}$ is the global minimum.

Proof of Theorem 3.1

Consider the selection of any p' variables from the p variables. Assume $q' \leq p'$ selected variables are informative. Denote the subset of selected variables by V, e.g., $V = \{X_1, ..., X_{q'}, X_{q+1}, ..., X_{q+p'-q'}\}$. Let the selected subvector of x be x_V. Denote the marginal densities of $g_i(x)$ in the selected subspace by $g_{V,i}(x_V)$. Then

$$g_{V,i}(x_V) = \phi(x_V^{(q')}|\mu_{V,i}^{(q')}, I)\phi(x_V^{(-q')}|\mu_V^{(-q')}), \quad i = 1, 2\,.$$

By Lemma 3.1 and 3.2, the ridgeline between $g_{V,1}(x_V)$ and $g_{V,2}(x_V)$ is given by

$$\mathcal{L}_V = \{x_V : x_V^{(q')} = (1-r)\mu_{V,1}^{(q')} + r\mu_{V,2}^{(q')}, r \in [0,1], x_V^{(-q')} = \mu_V^{(-q')}\}\,.$$

By Lemma 3.3, when $\beta = ||\mu_{V,1}^{(q')} - \mu_{V,2}^{(q')}||^2/2 < \beta^*$, the separability between $g_{V,1}(x_V)$ and $g_{V,2}(x_V)$ is zero. When $\beta \geq \beta^*$, the separability measure is $1 - \tau(\beta)$, where $\tau(\beta)$ is defined in the proof of Lemma 3.3 and shown to be strictly decreasing for $\beta > 2$. Thus the separability increases strictly with β, when $\beta > \beta^*$.

Consider the first step in forward selection by AD, which is equivalent to the separability measure when there are two components. The separability obtained from any non-informative variables is zero, and the separability obtained from an informative variable is maximized by variable X_1 because $|\mu_{1,1}^{(q)} - \mu_{2,1}^{(q)}| > |\mu_{1,2}^{(q)} - \mu_{2,2}^{(q)}| > \cdots > |\mu_{1,q}^{(q)} - \mu_{2,q}^{(q)}| > 0$, and $(\mu_{1,1}^{(q)} - \mu_{2,1}^{(q)})^2/2 > \beta^*$. When $V = \{X_1\}$, the separability measure is positive. We thus have shown that in the first step of forward selection, X_1 will be chosen. Assume at step $k \leq q$, X_1, X_2, ..., X_{k-1}, have been selected. The variable selected at step k should yield the largest separability among the variables left. By the monotonicity of the separability in β, variable X_k will be selected. Thus we have proved by induction, in the first q steps, variables X_1, X_2, ..., X_q will be selected sequentially. Because β increases strictly in the first q steps, the separability measure increases strictly. After the first q steps, β becomes constant, so does the separability measure.

References

Aha, D. W., and Bankert, R. L. (1994), "Feature selection for case-based classification of cloud types: An empirical comparison," *Proc. 1994 AAAI Workshop on Case-Based Reasoning*, 106-112, Seattle, WA, AAAI Press.

Caruana, R., and Freitag, D. (1994), "Greedy attribute selection," *Proc. 11th Int. Conf. Machine Learning*, 28-36, New Brunswick, NJ.

Dy, J. G., and Brodley, C. E. (2004), "Feature selection for unsupervised learning," *Journal of Machine Learning Research*, 5, 845-889.

Fraley, C. and Raftery, A. E. (1999), "MCLUST: Software for Model-Based Cluster Analysis", *Journal of Classification*, 16, 297-306.

Fraley, C. and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation", *Journal of the American Statistical Association*, 97, 661-631.

Friedman, J. and Meulman, J. J. (2004), "Clustering objects on subsets of attributes," *J. R. Statist. Soc. B.*, 66, 815-849.

Guo, J., Levina, E., Michailidis, G., Zhu, J. (2010), "Pairwise variable selection for high-dimensional model-based clustering," *Biometrics*, 66(3), 793-804.

Guyon, I, and Elisseeff, A. (2003), "An introduction to variable and feature selection," *Journal of Machine Learning Research*, 3, 1157-1182.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, Springer.

Johnson, R. A., and Wichern, D. W. (1998), *Applied Multivariate Statistical Analysis*, Prentice-Hall, 4ed.

Law, M. H.C., Figueiredo, M. A.T., and Jain, A. K. (2004), "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. PAMI*, 26, 1154-1166.

Li, J. (2005) "Clustering based on a multi-layer mixture model," *Journal of Computational and Graphical Statistics*, 14, 547-568.

Li, J., Ray, S., and Lindsay, B. G. (2007) "A nonparametric statistical approach to clustering via mode identification", *Journal of Machine Learning Research*, 8, 1687-1723.

Liu, H., and Setiono, R. (1996), "Dimension reduction via discretization," *Knowledge-Based Systems*, 9, 67-72.

Liu, Y., and Wu, Y. (2007), "Variable selection via a combination of the L0 and L1 penalties," *Journal of Computational and Graphical Statistics*, 16, 782-798.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009), "Variable selection for clustering with Gaussian mixture models," *Biometrics*, 65, 701-709.

McLachlan, G., and Peel, D. (2000) *Finite Mixture Models*, Wiley US.

Mitra, P., Murthy, C. A., and Pal, S. K. (2002), "Unsupervised feature selection using feature similarity," *IEEE Trans. PAMI*, 24, 301-312.

Narendra, P. M., and Fukunaga, K. (1977), "A branch and bound algorithm for feature subset selection," *IEEE Trans. Computers*, C-26, 917-922.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* 8, 1145–1164.

Raftery, A., and Dean, N. (2006), "Variable selection for model-based clustering," *JASA*, 101, 168-178.

Ray, S., and Lindsay, B. G. (2005) "The topography of multivariate normal mixtures", *The Annals of Statistics*, 33, 2042-2065.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005), "Bayesian variable selection in clustering high-dimensional data," *JASA*, 100, 602-617.

Wang, L. and Shen, X. (2006), "Multicategory support vector machines, feature selection and solution path," *Statistics Sinica*, 16, 617-634.

Wang, S. and Zhu, J. (2008), "Variable selection for model-based high-dimensional clustering and its application to microarray data," *Biometrics*, 64, 440-448.

Witten, D. M. and Tibshirani, R. (2010), "A framework for feature selection in clustering," *JASA*, 105, 713-726.

Xie, B., Pan, W., and Shen, X. (2008), "Penalized model-based clustering with cluster-specific diagonal covariance matrices and grouped variables," *Electronic Journal of Statistics*, 2, 168-212.

Yuan, M., Joseph, V. R., and Zou, H. (2009), "Structured variable selection and estimation," *Ann. Appl. Statist.*, 3, 1738-1757.

Zhang, H. H., Liu, Y., Wu, Y., and Zhu, J. (2008), "Variable selection for multicategory SVM via sup-norm regularization," *Electron. J. Statist.*, 2, 149-167.

Zou, H. (2006), "The adaptive lasso and its oracle properties," *J. Amer. Statist. Assoc.*, 101, 1418-1429.