

# Gaussian Mixture Models with Component Means Constrained in Pre-selected Subspaces

**Mu Qiao**

*Department of Computer Science and Engineering  
The Pennsylvania State University  
University Park, PA 16802, USA*

MUQ103@CSE.PSU.EDU

**Jia Li**

*Department of Statistics  
The Pennsylvania State University  
University Park, PA 16802, USA*

JIALI@STAT.PSU.EDU

**Editor:**

## Abstract

We investigate a Gaussian mixture model (GMM) with component means constrained in a pre-selected subspace. Applications to classification and clustering are explored. An EM-type estimation algorithm is derived. We prove that the subspace containing the component means of a GMM with a common covariance matrix also contains the modes of the density and the class means. This motivates us to find a subspace by applying weighted principal component analysis to the modes of a kernel density and the class means. To circumvent the difficulty of deciding the kernel bandwidth, we acquire multiple subspaces from the kernel densities based on a sequence of bandwidths. The GMM constrained by each subspace is estimated; and the model yielding the maximum likelihood is chosen. A dimension reduction property is proved in the sense of being informative for classification or clustering. Experiments on real and simulated data sets are conducted to examine several ways of determining the subspace and to compare with the reduced rank mixture discriminant analysis (MDA). Our new method with the simple technique of spanning the subspace only by class means often outperforms the reduced rank MDA when the subspace dimension is very low, making it particularly appealing for visualization.

**Keywords:** Gaussian mixture model, subspace constrained, modal PCA, dimension reduction, visualization

## 1. Introduction

The Gaussian mixture model (GMM) is a popular and effective tool for clustering and classification. When applied to clustering, usually each cluster is modeled by a Gaussian distribution. Because the cluster labels are unknown, we face the issue of estimating a GMM. A thorough treatment of clustering by GMM is referred to (McLachlan and Peel, 2000). Hastie and Tibshirani (1996) proposed the mixture discriminant analysis (MDA) for classification, which assumes a GMM for each class. Fraley and Raftery (2002) examined the roles of GMM for clustering, classification, and density estimation.

As a probability density, GMM enjoys great flexibility comparing with parametric distributions. Although GMM can approximate any smooth density by increasing the number

of components  $R$ , the number of parameters in the model grows quickly with  $R$ , especially for high dimensional data. The regularization of GMM has been a major research topic on mixture models. Early efforts focused on controlling the complexity of the covariance matrices, partly driven by the frequent occurrences of singular matrices in estimation. For example, Fraley and Raftery (2002) proposed a series of mixture models with the covariance matrices constrained in various ways: diagonal, scalar matrices, or common among components in terms of shape, orientation, or volume. More recently, it is noted that for data with very high dimensions, e.g.,  $n > p$ , a mixture model with parsimonious covariance structures, for example, common diagonal matrices, may still have high complexity due to the component means alone. Methods to regularize the component means have been proposed from quite different perspectives. Li and Zha (2006) developed the so-called two-way mixture of Poisson distributions in which the variables are grouped and the means of the variables in the same group within any component are assumed identical. The grouping of the variables reduces the number of parameters in the component means dramatically. As a result, the model is successfully applied to data with dimensions in the thousands and greater than the sample size. In the same spirit, Qiao and Li (2010) developed the two-way mixture of Gaussians. Pan and Shen (2007) explored the penalized likelihood method with  $L_1$  norm penalty on the component means. The method aims at shrinking the component means of some variables to a common value. Variable selection for clustering is achieved because the variables with common means across all the components are non-informative for cluster labels. Wang and Zhu (2008) proposed the  $L_\infty$  norm as a penalty instead.

In this paper, we propose another approach for regularizing the component means in GMM, which is more along the line of reduced rank MDA (Hastie and Tibshirani, 1996) but with profound differences. We search for a linear subspace in which the component means reside and estimate a GMM under such a constraint. The constrained GMM has a dimension reduction property. It is proved that with the subspace restriction on the component means and under common covariance matrices, only a linear projection of the data with the same dimension as the subspace matters for classification and clustering. The method is especially useful for visualization when we want to view data in a low dimensional space (usually two dimensional) which best preserves the classification and clustering characteristics.

The idea of restricting component means to a linear subspace was first explored in the linear discriminant analysis (LDA). Fisher (1936) proposed to find a subspace of rank  $r < K$ , where  $K$  is the number of classes, so that the projected class means are spread apart maximally, as measured by the ratio of between- and within-class variances. The coordinates of the optimal subspace are derived by successively maximizing the between-class variance relative to the within-class variance, known as *canonical* or *discriminant* variables. Although LDA does not involve the estimation of a mixture model, the marginal distribution of the observation without the class label is a mixture distribution. The idea of reduced rank LDA was used by Hastie and Tibshirani (1996) for GMM. It was proved in Hastie and Tibshirani (1996) that reduced rank LDA can be viewed as a Gaussian maximum likelihood solution with the restriction that the means of Gaussians lie in a  $L$ -dimension subspace, i.e.,  $\text{rank}\{\mu_k\}_1^K = L < \max(K - 1, p)$ , where  $\mu_k$ 's are the means of Gaussians and  $p$  is the dimension of the data. Hastie and Tibshirani (1996) extended this concept and proposed a reduced rank version of the mixture discriminant analysis (MDA), which performed a reduced rank weighted LDA in each iteration of the EM algorithm.

The role of the subspace constraining the means differs intrinsically between our approach and the reduced rank MDA, resulting in mathematical solutions of quite different nature. As explained previously, within each iteration of the EM algorithm for estimating a GMM, the reduced rank MDA finds the subspace with a given dimension that yields the maximum likelihood under the current partition of the data into the mixture components (soft partition by the posteriori probabilities). The subspace depends on the component-based clustering of data in each iteration. In our method, we treat the seek of the subspace and the estimation of the model separately. The subspace is fixed throughout the estimation of the GMM. Mathematically speaking, we try to solve the maximum likelihood estimation of GMM under the condition that the component means lie in a given subspace.

Our formulation of the model estimation problem allows us to exploit multiple and better choices of density estimate when we seek the constraining subspace. For instance, if we want to visualize the data in a plane while the component means are not truly constrained to a plane, fitting a GMM with means constrained to a plane may lead to poor density estimation. As a result, the plane sought during the estimation will be problematic. It is thus sensible to find the plane based on a density estimate without the constraint. Afterward, we can fit a GMM under the constraint purely for the purpose of visualization. When it comes to decide the number of components in the GMM, we can focus on the need of clustering rather than best approximating the density. Moreover, the subspace may be specified based on prior knowledge. For instance, in multi-dimensional data visualization, we may already know that the component (or cluster) means of data lie in a subspace spanned by several dimensions of the data. Therefore, the subspace is required to be fixed.

We propose two approaches to finding the unknown subspace. The first approach is the so-called *modal PCA (MPCA)*. We prove that the modes (local maxima) lie in the same constrained subspace as the component means. We use the *modal EM (MEM)* algorithm (Li et al., 2007) to find the modes. By exploiting the modes, we are no longer restricted to the GMM as a tool for density estimation. Instead, we use the kernel density estimate which avoids sensitivity to initialization. There is an issue of choosing the bandwidth, which is easier than usual in our framework by the following strategy. We take a sequence of subspaces based on density estimates resulting from different kernel bandwidths. We then estimate GMMs under the constraint of each subspace and finally choose a model yielding the maximum likelihood. Note that, each GMM is a full model for the original data, although the component means are constrained in a different subspace. We therefore can compare the estimated likelihood under each model. This framework in fact extends beyond kernel density estimation. As discussed in (Li et al., 2007), modes can be found using modal EM for any density in the form of a mixture distribution. For instance, we could use the *Mclust* package in R (Fraley and Raftery, 2006). The second approach is an extension of MPCA which exploits class means or a union set of modes and class means. It is easy to see that the class means also reside in the same constrained subspace as the component means. Comparing with modes, class means do not depend on the kernel bandwidth and are more robust to estimate, but the subspace they span has a maximum dimension of  $K - 1$ , where  $K$  is the number of classes.

Experiments on the classification of several real and simulated data sets with moderate to high dimensions show that reduced rank MDA does not always have good performance. When the constraining subspace of the component means is of a very low dimension, our

proposed method with the simple technique of finding the subspace based on class means often outperforms the reduced rank MDA, which solves a discriminant subspace via a much more sophisticated approach. In addition, we compare our methods with standard MDA on the data projected to the subspace containing the component means. For data with moderately high dimensions, our proposed methods are better. Besides classification, our method easily applies to clustering.

The rest of the paper is organized as follows. In Section 2, we review some background and notation. We present a Gaussian mixture model with subspace constrained component means, the MPCA algorithm and its extension for finding the subspace in Section 3. We also present several properties of the constrained subspace, with detailed proofs in the appendix. In Section 4, we describe the estimation algorithm for the proposed model. Experimental results are provided in Section 5. Finally, we conclude and discuss future work in Section 6.

## 2. Preliminaries and Notation

Let  $\mathbf{X} = (X_1, X_2, \dots, X_p)^t$ , where  $p$  is the dimension of the data. A sample of  $\mathbf{X}$  is denoted by  $\mathbf{x} = (x_1, x_2, \dots, x_p)^t$ . We present the notations for a general Gaussian mixture model before introducing the mixture model with component means constrained to a given subspace. Gaussian mixture model can be applied to both classification and clustering. Let the class label of  $\mathbf{X}$  be  $Y \in \mathcal{K} = \{1, 2, \dots, K\}$ . For classification purpose, the joint distribution of  $\mathbf{X}$  and  $Y$  under a Gaussian mixture is

$$f(\mathbf{X} = \mathbf{x}, Y = k) = a_k f_k(\mathbf{x}) = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(\mathbf{x} | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}), \quad (1)$$

where  $a_k$  is the prior probability of class  $k$ , satisfying  $0 \leq a_k \leq 1$  and  $\sum_{k=1}^K a_k = 1$ , and  $f_k(\mathbf{x})$  is the within-class density for  $\mathbf{X}$ .  $R_k$  is the number of mixture components used to model class  $k$ , and the total number of mixture components for all the classes is  $R = \sum_{k=1}^K R_k$ . Let  $\pi_{kr}$  be the mixing proportions for the  $r$ th component in class  $k$ ,  $0 \leq \pi_{kr} \leq 1$ ,  $\sum_{r=1}^{R_k} \pi_{kr} = 1$ .  $\phi(\cdot)$  denotes the pdf of a Gaussian distribution:  $\boldsymbol{\mu}_{kr}$  is the mean vector for component  $r$  in class  $k$  and  $\boldsymbol{\Sigma}$  is the common covariance matrix shared across all the components in all the classes. To classify a sample  $\mathbf{X} = \mathbf{x}$ , the Bayes classification rule is used:  $\hat{y} = \operatorname{argmax}_k f(Y = k | \mathbf{X} = \mathbf{x}) = \operatorname{argmax}_k f(\mathbf{X} = \mathbf{x}, Y = k)$ .

In the context of clustering, the Gaussian mixture model is now simplified as

$$f(\mathbf{X} = \mathbf{x}) = \sum_{r=1}^R \pi_r \phi(\mathbf{x} | \boldsymbol{\mu}_r, \boldsymbol{\Sigma}), \quad (2)$$

where  $R$  is the total number of mixture components and  $\pi_r$  is the mixing proportions for the  $r$ th component.  $\boldsymbol{\mu}_r$  and  $\boldsymbol{\Sigma}$  denote the  $r$ th component mean and the common covariance matrix for all the components. The clustering procedure involves first fitting the above mixture model and then computing the posterior probability of each mixture component given a sample point. The component with the highest posterior probability is chosen for that sample point, and all the points belonging to the same component form one cluster.

In this work, we assume that the Gaussian component means reside in a given linear subspace and estimate a GMM with subspace constrained means. A new algorithm, namely

the *modal PCA (MPCA)*, is proposed to find the constrained subspace. The motivations of using modes to find subspace are outlined in Section 3.1. Before we present MPCA, we will first introduce the modal EM algorithm (Li et al., 2007) which solves the local maxima, that is, modes, of a mixture density.

### 2.1 Modal EM

Given a mixture density  $f(\mathbf{X} = \mathbf{x}) = \sum_{r=1}^R \pi_r f_r(\mathbf{x})$ , as in model (2), starting from any initial data point  $\mathbf{x}^{(0)}$ , the modal EM algorithm finds a mode of the density by alternating the following two steps until a stopping criterion is met. Start with  $t = 0$ .

1. Let

$$p_r = \frac{\pi_r f_r(\mathbf{x}^{(t)})}{f(\mathbf{x}^{(t)})}, \quad r = 1, \dots, R.$$

2. Update

$$\mathbf{x}^{(t+1)} = \operatorname{argmax}_{\mathbf{x}} \sum_{r=1}^R p_r \log f_r(\mathbf{x}).$$

The above two steps are similar to the expectation and the maximization steps in EM (Dempster et al., 1977). The first step is the “expectation” step where the posterior probability of each mixture component  $r$ ,  $1 \leq r \leq R$ , at the current data point  $\mathbf{x}^{(t)}$  is computed. The second step is the “maximization” step.  $\sum_{r=1}^R p_r \log f_r(\mathbf{x})$  has a unique maximum, if the  $f_r(\mathbf{x})$ ’s are normal densities. In the special case of a mixture of Gaussians with common covariance matrix, that is,  $f_r(\mathbf{x}) = \phi(\mathbf{x} \mid \boldsymbol{\mu}_r, \boldsymbol{\Sigma})$ , we simply have  $\mathbf{x}^{(t+1)} = \sum_{r=1}^R p_r \boldsymbol{\mu}_r$ . In modal EM, the probability density function of the data is estimated nonparametrically using Gaussian kernels, which are in the form of a Gaussian mixture distribution:

$$f(\mathbf{X} = \mathbf{x}) = \sum_{i=1}^n \frac{1}{n} \phi(\mathbf{x} \mid \mathbf{x}_i, \boldsymbol{\Sigma}),$$

where the Gaussian density function is

$$\phi(\mathbf{x} \mid \mathbf{x}_i, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{x}_i)^t \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{x}_i)\right).$$

We use a spherical covariance matrix  $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma^2, \sigma^2, \dots, \sigma^2)$ . The standard deviation  $\sigma$  is also referred to as the *bandwidth* of the Gaussian kernel. When the bandwidth of Gaussian kernels increases, the density estimate becomes smoother, and more data points tend to ascend to the same mode. Different numbers of modes can thus be found by gradually increasing the bandwidth of Gaussian kernels. The data points are grouped into one cluster if they climb to the same mode. We call the mode as the cluster representative.

In (Li et al., 2007), a hierarchical clustering approach, namely, *Hierarchical Mode Association Clustering (HMAC)*, is proposed based on mode association and kernel bandwidth growth. Given a sequence of bandwidths  $\sigma_1 < \sigma_2 < \dots < \sigma_\eta$ , HMAC starts with every point  $\mathbf{x}_i$  being a cluster by itself, which corresponds to the extreme case that  $\sigma_1$  approaches 0.

At any bandwidth  $\sigma_l (l > 1)$ , the modes, that is, cluster representatives, obtained from the preceding bandwidth are input into the modal EM algorithm. The modes identified then form a new set of cluster representatives. This procedure is repeated across all  $\sigma_l$ 's. For details of HMAc, we refer interested readers to (Li et al., 2007). We therefore obtain modes at different levels of bandwidth by HMAc. The clustering performed by HMAc is only for the purpose of finding modes across different bandwidths and should not be confused with the clustering or classification based on the Gaussian mixture model we propose here.

### 3. Gaussian Mixture Model with Subspace Constrained Means

We present the Gaussian mixture model with subspace constrained means in this section. For brevity, we focus on the constrained mixture model in a classification set-up, since clustering can be treated as a ‘‘one-class’’ modeling and is likewise solved.

We propose to model the within-class density by a Gaussian mixture with component means constrained to a pre-selected subspace:

$$f_k(\mathbf{x}) = \sum_{r=1}^{R_k} \pi_{kr} \phi(\mathbf{x} | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \quad (3)$$

subject to

$$\mathbf{v}_j^t \cdot \boldsymbol{\mu}_{k1} = \mathbf{v}_j^t \cdot \boldsymbol{\mu}_{k2} = \cdots = \mathbf{v}_j^t \cdot \boldsymbol{\mu}_{kR_k} = c_j, \quad (4)$$

where  $\mathbf{v}_j$ 's are linearly independent vectors,  $j = 1, \dots, q$ ,  $q < p$ , and  $c_j$  is a constant, invariant to different classes. Without loss of generality, we can assume  $\{\mathbf{v}_1, \dots, \mathbf{v}_q\}$  span an orthonormal basis. Augment it to full rank by  $\{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$ . Suppose  $\boldsymbol{\nu} = \{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$ ,  $\boldsymbol{\nu}^\perp = \{\mathbf{v}_1, \dots, \mathbf{v}_q\}$ , and  $\mathbf{c} = (c_1, c_2, \dots, c_q)^t$ . Denote the projection of a vector  $\boldsymbol{\mu}$  or a matrix  $U$  onto an orthonormal basis  $S$  by  $\mathbf{Proj}_S^\boldsymbol{\mu}$  or  $\mathbf{Proj}_S^U$ . We have  $\mathbf{Proj}_{\boldsymbol{\nu}^\perp}^{\boldsymbol{\mu}_{kr}} = \mathbf{c}$  over all the  $k$  and  $r$ . That is, the projections of all the component means  $\boldsymbol{\mu}_{kr}$ 's onto the subspace  $\boldsymbol{\nu}^\perp$  coincide at  $\mathbf{c}$ . We refer to  $\boldsymbol{\nu}$  as the *constrained subspace* where  $\boldsymbol{\mu}_{kr}$ 's reside (or more strictly,  $\boldsymbol{\mu}_{kr}$ 's reside in the subspace up to a translation), and  $\boldsymbol{\nu}^\perp$  as the corresponding *null subspace*. Suppose the dimension of the constrained subspace  $\boldsymbol{\nu}$  is  $d$ , then  $d = p - q$ . With the constraint (4) and the assumption of a common covariance matrix across all the components in all the classes, essentially, we assume that the data within each component have identical distributions in the null space  $\boldsymbol{\nu}^\perp$ . In the following section, we will explain how to find an appropriate constrained subspace  $\boldsymbol{\nu}$ .

#### 3.1 Modal PCA

We introduce in this section the modal PCA (MPCA) algorithm that finds a constrained subspace for the component means of a Gaussian mixture and the properties of the found subspace. We prove in Appendix A the following theorem.

**Theorem 1** *For a Gaussian mixture model with component means constrained in a subspace  $\boldsymbol{\nu} = \{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$ ,  $q < p$ , and a common covariance matrix across all the components, the modes of the mixture density are also constrained in the same subspace  $\boldsymbol{\nu}$ .*

According to Theorem 1, the modes and component means of Gaussian mixtures reside in the same constrained subspace. We use the aforementioned MEM algorithm introduced in Section 2 to find the modes of the density. To avoid sensitivity to initialization and the number of components, we use the Gaussian kernel density estimate instead of a finite mixture model for the density. It is well known that mixture distributions with drastically different parameters may yield similar densities. We are thus motivated to exploit modes which are geometric characteristics of the densities.

Let us denote the set of modes found by MEM under the kernel bandwidth  $\sigma$  by  $\mathcal{G} = \{\mathcal{M}_{\sigma,1}, \mathcal{M}_{\sigma,2}, \dots, \mathcal{M}_{\sigma,|\mathcal{G}|}\}$ . A weighted principal component analysis is proposed to find the constrained subspace. We assign a weight  $w_{\sigma,r}$  to the  $r$ th mode, which is the proportion of sample points in the entire data ascending to that mode. We therefore have a weighted covariance matrix of all the modes in  $\mathcal{G}$ :

$$\Sigma_{\mathcal{G}} = \sum_{r=1}^{|\mathcal{G}|} w_{\sigma,r} (\mathcal{M}_{\sigma,r} - \mu_{\mathcal{G}})^T (\mathcal{M}_{\sigma,r} - \mu_{\mathcal{G}}),$$

where  $\mu_{\mathcal{G}} = \sum_{r=1}^{|\mathcal{G}|} w_{\sigma,r} \mathcal{M}_{\sigma,r}$ . The principal components are then obtained by performing an eigenvalue decomposition on  $\Sigma_{\mathcal{G}}$ . Recall the dimension of the constrained subspace  $\boldsymbol{\nu}$  is  $d$ . Since the leading principal components capture the most variation in the data, we use the first  $d$  most significant principal components to span the constrained subspace  $\boldsymbol{\nu}$ , and the remaining principal components to span the corresponding null space  $\boldsymbol{\nu}^{\perp}$ .

Given a sequence of bandwidths  $\sigma_1 < \sigma_2 < \dots < \sigma_{\eta}$ , we can obtain the modes at different levels of bandwidth using the HMAC algorithm introduced in Section 2. At each level, we apply the weighted PCA to the modes, and obtain a new constrained subspace by their first  $d$  most significant principal components. In practice, if the number of modes found at a particular level of bandwidth is smaller than 3, we will skip the modes at that level. For the extreme case, when  $\sigma = 0$ , the subspace is actually spanned by the principal components of the original data points. We therefore obtain a collection of subspaces,  $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_{\eta}$ , resulting from a sequence of bandwidths through HMAC.

### 3.2 Extension of Modal PCA

In this section, we propose another approach to generate the constrained subspace, which is an extension of MPCA. Suppose the mean of class  $k$  is  $\mathcal{M}'_k$ , we have  $\mathcal{M}'_k = \sum_{r=1}^{R_k} \pi_{kr} \boldsymbol{\mu}_{kr}$ , where  $\boldsymbol{\mu}_{kr}$  is the  $r$ th component in class  $k$ . It is easy to see that the class means lie in the same subspace as the Gaussian mixture component means. From Theorem 1, we know that in Gaussian mixtures, the modes and component means also reside in the same constrained subspace. So the class means, modes and component means all lie in the same constrained subspace. Comparing with the modes, class means are more robust to estimate. It is thus natural to incorporate class means to find the subspace. In the new approach, if the dimension  $d$  of the constrained subspace is smaller than  $K$ , the subspace is spanned by applying weighted PCA only to class means. Otherwise, it is spanned by applying weighted PCA to a union set of modes and class means.

Similar to modal PCA, we first assign a weight  $a_k$  to the  $k$ th class mean  $\mathcal{M}'_k$ , which is the proportion of the number of sample points in class  $k$  over the entire data, i.e., the

prior probability of class  $k$ . Suppose the set of class means is  $\mathcal{J} = \{\mathcal{M}'_1, \mathcal{M}'_2, \dots, \mathcal{M}'_K\}$ . If  $d < K$ , we have a weighted covariance matrix of all the class means:

$$\Sigma_{\mathcal{J}} = \sum_{r=1}^K a_k (\mathcal{M}'_r - \mu_{\mathcal{J}})^T (\mathcal{M}'_r - \mu_{\mathcal{J}}),$$

where  $\mu_{\mathcal{J}} = \sum_{r=1}^K a_k \mathcal{M}'_k$ . We then perform an eigenvalue decomposition on  $\Sigma_{\mathcal{J}}$  to obtain all the principal components. Similar to MPCA, the constrained subspace is spanned by the first  $d$  most significant principal components. If  $d \geq K$ , we will put together all the class means and modes and assign different weights to them. Suppose  $\gamma$  is a value between 0 and 100, we allocate a total of  $\gamma\%$  of weight to the class means, and the remaining  $(100 - \gamma)\%$  weights allocated proportionally to the modes. That is, the weights assigned to the class mean  $\mathcal{M}'_k$  and the mode  $\mathcal{M}_{\sigma,r}$  are  $\gamma a_k\%$  and  $(100 - \gamma)w_{\sigma,r}\%$ , respectively. Then the weighted covariance matrix of the union set of class means and modes becomes

$$\Sigma_{\mathcal{G} \cup \mathcal{J}} = \sum_{r=1}^K \gamma a_k\% (\mathcal{M}'_r - \mu_{\mathcal{J}})^T (\mathcal{M}'_r - \mu_{\mathcal{J}}) + \sum_{r=1}^{|\mathcal{G}|} (100 - \gamma)w_{\sigma,r}\% (\mathcal{M}_{\sigma,r} - \mu_{\mathcal{G}})^T (\mathcal{M}_{\sigma,r} - \mu_{\mathcal{G}}).$$

Different weights can be allocated to the class means and the modes. For instance, if we want the class means to play a more important role in spanning subspaces, we can set  $\gamma > 50$ . Again, an eigenvalue decomposition is performed on  $\Sigma_{\mathcal{G} \cup \mathcal{J}}$  to obtain all the principal components and the first  $d$  most significant principal components span the constrained subspace. To differentiate this method from MPCA, we denote it by MPCA-MEAN.

### 3.3 Dimension Reduction

The mixture model with component means under constraint (4) implies a dimension reduction property for the classification purpose, formally stated below.

**Theorem 2** *For a Gaussian mixture model with a common covariance matrix  $\Sigma$ , suppose all the component mean  $\mu_{kr}$ 's are constrained in a subspace spanned by  $\nu = \{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$ ,  $q < p$ , up to a translation, only a linear projection of the data  $\mathbf{x}$  onto a subspace spanned by  $\{\Sigma^{-1}\mathbf{v}_j | j = q + 1, \dots, p\}$  (the same dimension as  $\nu$ ) is informative for classification.*

In Appendix B, we provide the detailed proof for Theorem 1. If the common covariance matrix  $\Sigma$  is an identity matrix (or a scalar matrix), the class label  $Y$  only depends on the projection of  $\mathbf{x}$  onto the constrained subspace  $\nu$ . However, in general,  $\Sigma$  is non-identity. Hence the spanning vectors,  $\Sigma^{-1}\mathbf{v}_j$ ,  $j = q + 1, \dots, p$ , for the subspace informative for classification are not orthogonal in general as well. In Appendix B, we use the column vectors of *orth*( $\{\Sigma^{-1}\mathbf{v}_j | j = q + 1, \dots, p\}$ ) to span this subspace. To differentiate it from the constrained subspace in which the component means lie, we call it as *discriminant subspace*. We refer to the dimension of the discriminant subspace as *discriminant dimension*, which is the dimension actually needed for classification. The discriminant subspace is of the same dimension as the constrained subspace. When the discriminant dimension is small, significant dimension reduction is achieved. Our method can thus be used as a data reduction tool for visualization when we want to view the classification of data in a two or three dimensional space.

Although in Appendix B we prove Theorem 1 in the context of classification, the proof can be easily modified to show that the dimension reduction property applies to clustering as well. That is, we only need the data projected onto a subspace with the same dimension as the constrained subspace  $\nu$  to compute the posterior probability of the data belonging to a component (aka cluster). Similarly, we name the subspace that matters for clustering as *discriminant subspace* and its dimension as *discriminant dimension*.

#### 4. The Algorithm

Let us first summarize the work flow of our proposed method:

1. Given a sequence of kernel bandwidths  $\sigma_1 < \sigma_2 < \dots < \sigma_\eta$ , apply HMAC to find the modes of the density estimation at each bandwidth  $\sigma_l$ .
2. Apply MPCA or MPCA-MEAN to the modes or a union set of modes and class means at each kernel bandwidth and obtain a sequence of constrained subspaces.<sup>1</sup>
3. Estimate the Gaussian mixture model with component means constrained in each subspace and select the model yielding the maximum likelihood.
4. Perform classification on the test data or clustering on the overall data, with the selected model from Step 3.

Remarks:

1. In our method, the seek of subspace and the estimation of the mixture model are separate. We first search for a sequence of subspaces and then estimate the model constrained in each subspace separately.
2. In Step 1, the identified modes are from the density estimation of the overall data (in clustering) or the overall training data (in classification).
3. Some prior knowledge may be exploited to yield an appropriate subspace. Then, we can estimate GMM under the constraint of the given subspace directly.

Now we will derive an EM algorithm to estimate a GMM under the constraint of a given subspace. The estimation method for classification is introduced first. A common covariance matrix  $\Sigma$  is assumed across all the components in all the classes. In class  $k$ , the parameters to be estimated include the class prior probability  $a_k$ , the mixture component prior probabilities  $\pi_{kr}$ , and the Gaussian parameters  $\mu_{kr}$ ,  $\Sigma$ ,  $r = 1, 2, \dots, R_k$ . Denote the training data by  $\{(\mathbf{x}_i, y_i) : i = 1, \dots, n\}$ . Let  $n_k$  be the number of data points in class  $k$ . The total number of data points  $n$  is  $\sum_{k=1}^K n_k$ . The class prior probability  $a_k$  is estimated by the empirical frequency  $n_k / \sum_{k'=1}^K n_{k'}$ . The EM algorithm comprises the following two steps:

---

1. For MPCA-MEAN, if the dimension  $d$  of the constrained subspace is smaller than  $K$ , the subspace is spanned only by class means and is therefore fixed. We do not need to choose the subspace.

1. *Expectation-step*: Given the current parameters, for each class  $k$ , compute the component posteriori probability for each data point  $\mathbf{x}_i$  within class  $k$ :

$$q_{i,kr} \propto \pi_{kr} \phi(\mathbf{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}), \quad \text{subject to } \sum_{r=1}^{R_k} q_{i,kr} = 1. \quad (5)$$

2. *Maximization-step*: Update  $\pi_{kr}$ ,  $\boldsymbol{\mu}_{kr}$ , and  $\boldsymbol{\Sigma}$ , which maximize the following objective function (the  $i$  subscript indicates  $\mathbf{x}_i$  with  $y_i = k$ ):

$$\sum_{k=1}^K \sum_{r=1}^{R_k} \left( \sum_{i=1}^{n_k} q_{i,kr} \right) \log \pi_{kr} + \sum_{k=1}^K \sum_{r=1}^{R_k} \sum_{i=1}^{n_k} q_{i,kr} \log \phi(\mathbf{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \quad (6)$$

under the constraint (4).

In the maximization step, the optimal  $\pi_{kr}$ 's are not affected by the constraint (4) and are solved separately from  $\boldsymbol{\mu}_{kr}$ 's and  $\boldsymbol{\Sigma}$ :

$$\pi_{kr} \propto \sum_{i=1}^{n_k} q_{i,kr}, \quad \sum_{r=1}^{R_k} \pi_{kr} = 1. \quad (7)$$

Since there are no analytic solutions to  $\boldsymbol{\mu}_{kr}$ 's and  $\boldsymbol{\Sigma}$  in the above constrained optimization, we adopt the generalized EM (GEM) algorithm. Specifically, we use a conditional maximization approach. In every maximization step of GEM, we first fix  $\boldsymbol{\Sigma}$ , and then update the  $\boldsymbol{\mu}_{kr}$ 's. Then we update  $\boldsymbol{\Sigma}$  conditioned on the  $\boldsymbol{\mu}_{kr}$ 's held fixed. This iteration will be repeated multiple times.

Given  $\boldsymbol{\Sigma}$ , solving  $\boldsymbol{\mu}_{kr}$  is non-trivial. We summarize the key steps here. For detailed derivation, we refer interested readers to Appendix C. In constraint (4), we have  $\mathbf{v}_j^t \cdot \boldsymbol{\mu}_{kr} = c_j$ , i.e., identical across all the  $k$  and  $r$  for  $j = 1, \dots, q$ . It is easy to see that  $\mathbf{c} = (c_1, \dots, c_q)^t$  is equal to the projection of the mean of the overall data onto the null space  $\boldsymbol{\nu}^\perp$ . However, in practice, we do not need the value of  $\mathbf{c}$  in the parameter estimation. Before we give the equation to solve  $\boldsymbol{\mu}_{kr}$ , let us define a few notations first. Assume  $\boldsymbol{\Sigma}$  is non-singular and hence positive definite, we can write  $\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}^{\frac{1}{2}})^t (\boldsymbol{\Sigma}^{\frac{1}{2}})$ , where  $\boldsymbol{\Sigma}^{\frac{1}{2}}$  is of full rank. If the eigen decomposition of  $\boldsymbol{\Sigma}$  is  $\boldsymbol{\Sigma} = V_\Sigma D_\Sigma V_\Sigma^t$ , then  $\boldsymbol{\Sigma}^{\frac{1}{2}} = D_\Sigma^{\frac{1}{2}} V_\Sigma^t$ . Let  $V_{null}$  be a  $p \times q$  orthonormal matrix ( $\mathbf{v}_1, \dots, \mathbf{v}_q$ ), the column vectors of which span the null space  $\boldsymbol{\nu}^\perp$ . Suppose  $\mathbf{B} = \boldsymbol{\Sigma}^{\frac{1}{2}} V_{null}$ . Perform a singular value decomposition (SVD) on  $\mathbf{B}$ , i.e.,  $\mathbf{B} = \mathbf{U}_B \mathbf{D}_B \mathbf{V}_B^t$ , where  $\mathbf{U}_B$  is a  $p \times q$  matrix, the column vectors of which form an orthonormal basis for the space spanned by the column vectors of  $\mathbf{B}$ . Let  $\hat{\mathbf{U}}$  be a column augmented orthonormal matrix of  $\mathbf{U}_B$ . Denote  $\sum_{i=1}^{n_k} q_{i,kr}$  by  $l_{kr}$ . Let  $\bar{\mathbf{x}}_{kr} = \sum_{i=1}^{n_k} q_{i,kr} \mathbf{x}_i / l_{kr}$ , i.e., the weighted sample mean of the component  $r$  in class  $k$ , and  $\check{\mathbf{x}}_{kr} = \hat{\mathbf{U}}^t \left( \boldsymbol{\Sigma}^{-\frac{1}{2}} \right)^t \cdot \bar{\mathbf{x}}_{kr}$ . Define  $\check{\boldsymbol{\mu}}_{kr}^*$  by the following Eqs. (8) and (9):

1. for the first  $q$  coordinates,  $j = 1, \dots, q$ :

$$\check{\boldsymbol{\mu}}_{kr,j}^* = \frac{\sum_{k'=1}^K \sum_{r'=1}^{R_{k'}} l_{k'r'} \check{\mathbf{x}}_{k'r',j}}{n}, \quad \text{identical over } r \text{ and } k; \quad (8)$$

2. for the remaining  $p - q$  coordinates,  $j = q + 1, \dots, p$ :

$$\check{\mu}_{kr,j}^* = \check{x}_{kr,j} . \quad (9)$$

That is, the first  $q$  constrained coordinates are optimized using component-pooled sample mean<sup>2</sup> (components from all the classes) while those  $p - q$  unconstrained coordinates are optimized separately within each component using the component-wise sample mean. In the maximization step, the parameter  $\mu_{kr}$  is finally solved by:

$$\mu_{kr} = (\Sigma^{\frac{1}{2}})^t \hat{U} \check{\mu}_{kr}^* .$$

Given the  $\mu_{kr}$ 's, it is easy to solve  $\Sigma$ :

$$\Sigma = \frac{\sum_{k=1}^K \sum_{r=1}^{R_k} \sum_{i=1}^{n_k} q_{i,kr} (\mathbf{x}_i - \mu_{kr})^t (\mathbf{x}_i - \mu_{kr})}{n} .$$

To initialize the estimation algorithm, we first choose  $R_k$ , the number of mixture components for each class  $k$ . For simplicity, an equal number of components are assigned to each class. The constrained model is initialized by the estimated parameters from a standard Gaussian mixture model with the same number of components.

We have so far discussed the model estimation in a classification set-up. We assume a common covariance matrix and a common constrained subspace for all the components in all the classes. Similar parameter estimations can also be applied to the clustering model. Specifically, all the data are put in one ‘‘class’’. In this ‘‘one-class’’ estimation problem, all the parameters can be estimated likewise, by omitting the ‘‘ $k$ ’’ subscript for classes. For brevity, we skip the details here.

#### 4.1 Variation of the Algorithm

We have introduced the Gaussian mixture model with component means from different classes constrained in the same subspace. It is natural to modify the previous constraint in (4) to

$$\mathbf{v}_j^t \cdot \mu_{k1} = \mathbf{v}_j^t \cdot \mu_{k2} = \dots = \mathbf{v}_j^t \cdot \mu_{kR_k} = c_{k,j} , \quad (10)$$

where  $\mathbf{v}_j$ 's are linearly independent vectors spanning an orthonormal basis,  $j = 1, \dots, q$ ,  $q < p$ , and  $c_{k,j}$  depends on class  $k$ . That is, the projections of all the component means within class  $k$  onto the null space  $\nu^\perp$  coincide at the constant  $\mathbf{c}_k$ , where  $\mathbf{c}_k = (c_{k,1}, c_{k,2}, \dots, c_{k,q})^t$ . In the new constraint (10),  $\{\mathbf{v}_1, \dots, \mathbf{v}_q\}$  is the same set of vectors as used in constraint (4), which spans the null space  $\nu^\perp$ . Because  $c_k$  varies with class  $k$ , the subspace in which the component means from each class reside differs from each other by a translation, that is, these subspaces are parallel.

We train a constrained model for each class separately, and assume a common covariance matrix across all the components in all the classes. In the new constraint (10),  $\mathbf{c}_k$  is actually equal to the projection of the class mean  $\mathcal{M}'_k$  onto the null space  $\nu^\perp$ . Similar to the previous estimation, in practice, we do not need the value of  $\mathbf{c}_k$  in the parameter estimation. With

---

2. We abuse the term ‘‘sample mean’’ here to mean  $\check{x}_{kr}$ , instead of  $\bar{x}_{kr}$ .

the constraint (10), essentially, the component means in each class are now constrained in a shifted subspace parallel to  $\boldsymbol{\nu}$ . The shifting of subspace for each class is determined by  $\mathbf{c}_k$ , or the class mean  $\mathcal{M}'_k$ . Suppose the dimension of the constrained subspace is  $d$ . In general, the dimension that matters for classification in this variation of the algorithm is  $d + K - 1$ , assuming that the class means already span a subspace of dimension  $K - 1$ .

We first subtract the class specific means from the data in the training set, that is, do a class specific centering of the data. Similarly as the algorithm outlined in Section 4, we put all the centered data from all the classes into one training set, find all the modes under different kernel bandwidths, and then apply MPCA to generate a sequence of constrained subspaces. The reason that we remove the class specific means first is that they have already played a role in spanning the subspace containing all the component means. When applying MPCA, we only want to capture the dominant directions for the variation within the classes.

Comparing with the parameter estimation in Section 4, the only change that we need to make is that the constrained  $q$  coordinates in  $\check{\boldsymbol{\mu}}_{kr}^*$  are now identical over  $r$ , but not over class  $k$ . For the first  $q$  coordinates,  $j = 1, \dots, q$ , we have:

$$\check{\mu}_{kr,j}^* = \frac{\sum_{r'=1}^{R_k} l_{kr'} \check{x}_{kr',j}}{n_k}, \quad \text{identical over } r \text{ in class } k.$$

That is, the first  $q$  constrained coordinates are optimized using component-pooled sample mean in class  $k$ . All the other equations in the estimation remain the same.

## 5. Experiments

In this section, we present experimental results on several real and simulated data sets. The mixture model with subspace constrained means, reduced rank MDA, and standard MDA on the projection of data onto the constrained subspace, are compared for the classification of data with moderate to high dimensions. We also visualize and compare the classification and clustering results of our proposed method and the reduced rank MDA on several simulated data sets.

The detailed methods tested in the experiments and their name abbreviations are summarized as follows:

- **GMM-MPCA** The mixture model with subspace constrained means, in which the subspace is obtained by MPCA.
- **GMM-MPCA-MEAN** The mixture model with subspace constrained means, in which the subspace is obtained by MPCA-MEAN, as introduced in Section 3.2.
- **GMM-MPCA-SEP** The mixture model with component means constrained by separately shifted subspace for each class, as introduced in Section 4.1.
- **MDA-RR** The reduced rank mixture discriminant analysis (MDA), which is a weighted rank reduction of the full MDA.
- **MDA-RR-OS** The reduced rank mixture discriminant analysis (MDA), which is based on optimal scoring (Hastie and Tibshirani, 1996), a multiple linear regression approach.

- **MDA-DR-MPCA** The standard MDA on the projection of data onto the same constrained subspace selected by GMM-MPCA.
- **MDA-DR-MPCA-MEAN** The standard MDA on the projection of data onto the same constrained subspace selected by GMM-MPCA-MEAN.

Remarks:

1. Since the most relevant work to our proposed method is reduced rank mixture discriminant analysis (MDA), we briefly introduce MDA-RR and MDA-RR-OS in Section 5.1.
2. In MDA-DR-MPCA or MDA-DR-MPCA-MEAN, the data are projected onto the constrained subspace which has yielded the largest training likelihood in GMM-MPCA or GMM-MPCA-MEAN. Note that this constrained subspace is spanned by  $\boldsymbol{\nu} = \{\boldsymbol{v}_{q+1}, \dots, \boldsymbol{v}_p\}$ , which is found by MPCA or MPCA-MEAN, rather than the discriminant subspace informative for classification. We then apply standard MDA (assume a common covariance matrix across all the components in all the classes) to the projected training data, and classify the test data projected onto the same subspace. Note that, if we project the data onto the discriminant subspace spanned by  $\{\boldsymbol{\Sigma}^{-1}\boldsymbol{v}_j | j = q + 1, \dots, p\}$ , and then apply standard MDA to classification, it is theoretically equivalent to GMM-MPCA or GMM-MPCA-MEAN (ignoring the variation caused by model estimation). The reason that we conduct these comparisons is multi-fold: first, we want to see if there is advantage of the proposed method as compared to a relative naive dimension reduction scheme; second, when the dimension of the data is high, we want to investigate if the proposed method has robust estimation of  $\boldsymbol{\Sigma}$ ; third, we want to investigate the difference between the constrained subspace and the discriminant subspace.

### 5.1 Reduced Rank Mixture Discriminant Analysis

Reduced rank MDA is a data reduction method which allows us to have a low dimensional view on the classification of data in a discriminant subspace, by controlling the within-class spread of component means relative to the between class spread. We outline its estimation method in Appendix D, which is a weighted rank reduction of the full mixture solution proposed by Hastie and Tibshirani (1996). We also show how to obtain the discriminant subspace of the reduced rank method in Appendix D.

Hastie and Tibshirani (1996) applied the optimal scoring approach (Breiman and Ihaka, 1984) to fit reduced rank MDA, which converted the discriminant analysis to a nonparametric multiple linear regression problem. By expressing the problem as a multiple regression, the fitting procedures can be generalized using more sophisticated regression methods than linear regression (Hastie and Tibshirani, 1996), for instance, flexible discriminant analysis (FDA) and penalized discriminant analysis (PDA). The use of optimal scoring also has some computational advantages, for instance, using fewer observations than the weighted rank reduction. A software package<sup>3</sup> containing a set of functions to fit MDA, FDA, and PDA by multiple regressions is provided by Hastie and Tibshirani (1996).

---

3. Available at <http://cran.r-project.org/web/packages/mda/index.html>

Although the above benefits for estimating reduced rank MDA are gained from the optimal scoring approach, there are also some restrictions. For instance, it can not be easily extended to fit a mixture model for clustering since the component means and covariance are not estimated explicitly. In addition, when the dimension of the data is larger than the sample size, optimal scaling can not be used due to the lack of degrees of freedom in regression. In the following experiment section, we will compare our proposed methods with reduced rank MDA. Both our own implementation of reduced rank MDA based on weighted rank reduction of the full mixture, i.e., MDA-RR, and the implementation via optimal scoring from the software package provided by Hastie and Tibshirani (1996), i.e., MDA-RR-OS, are tested.

## 5.2 Classification

Eight data sets from various sources are used for classification. We summarize the detailed information of these data below.

- The **sonar** data set consists of 208 patterns of sonar signals. Each pattern has 60 dimensions and the number of classes is two. The sample sizes of the two classes are (111, 97).
- The **robot** data set has 5456 navigation instances, with 24 dimensions and four classes (826, 2097, 2205, 328).
- The **waveform** data (Hastie et al., 2001) is a simulated three-classes data of 21 features, with a waveform function generating both training and test sets (300, 500).
- The **imagery** semantics data set (Qiao and Li, 2010) contains 1400 images each represented by a 64 dimensional feature vector. These 1400 images come from five classes with different semantics (300, 300, 300, 300, 200).
- The **parkinsons** data set is composed of 195 individual voice recordings, which are of 21 dimensions and divided into two classes (147, 48).
- The **satellite** data set consists of 6435 instances which are square neighborhoods of pixels, with 36 dimensions and six classes (1533, 703, 1358, 626, 707, 1508).
- The **semeion** handwritten digit data have 1593 binary images from ten classes (0-9 digits) with roughly equal sample size in each class. Each image is of  $16 \times 16$  pixels and thus has 256 dimensions. Four fifths of the images are randomly selected to form a training set and the remaining as testing.
- The **yaleB** face image data (Georghiades et al., 2001; Lee et al., 2005; He et al., 2005) contains gray scale human face images for 38 individuals. Each individual has 64 images, which are of  $32 \times 32$  pixels, normalized to unit vectors. We randomly select the images of five individuals, and form a data set of 250 training images and 70 test images, with equal sample size for each individual.

The sonar, robot, parkinsons, satellite and semeion data are from the UCI machine learning repository. Among the above data sets, the semeion and yaleB data have high dimensions. The other data sets are of moderately high dimensions.

For the data sets with moderately high dimensions, five-fold cross validation is used to compute their classification accuracy except for the waveform, whose accuracy is the average over ten simulations, the same setting used in (Hastie et al., 2001). We assume a full common covariance matrix across all the components in all the classes. For the semeion and yaleB data sets, the randomly split training and test samples are used to compute their classification accuracy instead of cross validation due to the high computational cost. Since these two data sets are of high dimensions, for all the tested methods, we assume common diagonal covariance matrices across all the components in all the classes. For simplicity, the same number of mixture components is used to model each class for all the methods.

In our proposed methods, the constrained subspaces are found by MPCA or MPCA-MEAN, introduced in Section 3.1 and 3.2. Specifically, in MPCA, a sequence of subspaces are identified from the training data by gradually increasing the kernel bandwidth  $\sigma_l$ , i.e.,  $\sigma_1 < \sigma_2 < \dots < \sigma_\eta$ ,  $l = 1, 2, \dots, \eta$ . In practice, we set  $\eta = 20$  and choose  $\sigma_l$ 's equally spaced from  $[0.1\hat{\sigma}, 2\hat{\sigma}]$ , where  $\hat{\sigma}$  is the largest sample standard deviation of all the dimensions in the data. HMAC is used to obtain the modes at different bandwidths. Note that in HMAC, some  $\sigma_l$  may result in the same clustering as  $\sigma_{l-1}$ , indicating that the bandwidth needs to be increased substantially so that some existing cluster representatives can be merged. In our experiments, only the modes at the bandwidth resulting in different clustering from the preceding bandwidth are employed to span the subspace. For the high dimensional data, since the previous kernel bandwidth range  $[0.1\hat{\sigma}, 2\hat{\sigma}]$  does not yield a sequence of distinguishable subspaces, we therefore increase their bandwidths. Specifically, for the semeion and yaleB data, the kernel bandwidth  $\sigma_l$  is now chosen equally spaced from  $[4\hat{\sigma}, 5\hat{\sigma}]$  and  $[2\hat{\sigma}, 3\hat{\sigma}]$ ,<sup>4</sup> respectively, with the interval being  $0.1\hat{\sigma}$ . In MPCA-MEAN, if the dimension of the constrained subspace is smaller than the class number  $K$ , the subspace is obtained by applying weighted PCA only to class means. Otherwise, at each bandwidth, we obtain the subspace by applying weighted PCA to a union set of class means and modes, with 60% weight allocated proportionally to the means and 40% to the modes, that is,  $\gamma = 60$ . The subspace yielding the largest likelihood on the training data is finally chosen as the constrained subspace.

### 5.2.1 CLASSIFICATION RESULTS

We show the classification results of the tested methods in this section. The classification error rates on data sets of moderately high dimensions are shown in Tables 1, 2, and 3. We vary the discriminant dimension  $d$  and also the number of mixture components used for modeling each class. Similarly, Table 4 shows the classification error rates on the semeion and yaleB data, which are of high dimensions. For all the methods except GMM-MPCA-SEP, the dimension of the discriminant subspace equals the dimension of the constrained subspace, denoted by  $d$ . For GMM-MPCA-SEP, the dimension of the discriminant space is actually  $K - 1 + d$ . In order to compare on a common ground, for GMM-MPCA-SEP, we change the notation for the dimension of the constrained subspace to  $d'$ , and still denote the dimension of the discriminant subspace by  $d = K - 1 + d'$ . The minimum number

---

4. In GMM-MPCA-SEP, the modes are identified from a new set of class mean removed data, where the kernel bandwidth  $\sigma_l$  is chosen equally spaced from  $[3.1\hat{\sigma}, 5\hat{\sigma}]$ , with the interval being  $0.1\hat{\sigma}$ , for both the semeion and yaleB data. For the other data sets,  $\sigma_l$  is still chosen equally spaced from  $[0.1\hat{\sigma}, 2\hat{\sigma}]$ .

of dimensions used for classification in GMM-MPCA-SEP is therefore  $K - 1$ . In all these tables, if  $d$  is set to be smaller than  $K - 1$ , we do not have the classification results of GMM-MPCA-SEP, which are marked by “NA”. In addition, in Table 4(b), the classification error rates of MDA-RR-OS on yaleB data are not reported since the dimension  $p$  of the data is significantly larger than the number of samples  $n$ . The reduced rank MDA based on optimal scoring approach cannot be employed due to the lack of degree freedom in the regression step for the small  $n$  large  $p$  problem. The minimum error rate in each column is in bold font. From these tables, we can see that for the three Gaussian mixture models with subspace constrained means, GMM-MPCA-MEAN and GMM-MPCA-SEP usually outperform GMM-MPCA, except on the waveform data. Since the class means are involved in spanning the constrained subspace in GMM-MPCA-MEAN and determine the shifting of the subspace for each class in GMM-MPCA-SEP, the observed advantage of GMM-MPCA-MEAN and GMM-MPCA-SEP indicates that class means are valuable for finding a good subspace.

Comparing the proposed methods and the reduced rank MDA methods, we see that, when the discriminant dimension is low, GMM-MPCA-MEAN and GMM-MPCA-SEP usually perform better than MDA-RR and MDA-RR-OS. When the dimension becomes higher, we do not observe a clear winner among different methods. The results are very data-dependent. Note that in GMM-MPCA-MEAN, when the discriminant dimension is smaller than  $K - 1$ , the subspace is obtained by applying weighted PCA only to the class means. For most data sets, when the discriminant dimension is very low, GMM-MPCA-MEAN performs best or close to best.

We also report the classification results of MDA-DR-MPCA and MDA-DR-MPCA-MEAN. For the data sets of moderately high dimensions, when the discriminant dimension is very low, they are usually inferior to GMM-MPCA and GMM-MPCA-MEAN. As the dimension increases, with certain component numbers, MDA-DR-MPCA and MDA-DR-MPCA-MEAN may have a better classification accuracy than GMM-MPCA and GMM-MPCA-MEAN. In addition, if the data set is of high dimension, for instance, the yaleB data, MDA-DR-MPCA/MDA-DR-MPCA-MEAN may perform better than GMM-MPCA/GMM-MPCA-MEAN even at lower discriminant dimension. As discussed in Remark 2 of this section, for MDA-DR-MPCA/MDA-DR-MPCA-MEAN and GMM-MPCA/GMM-MPCA-MEAN, we essentially do classification on the data in two different subspaces, i.e., the constrained subspace and the discriminant subspace. For GMM-MPCA/GMM-MPCA-MEAN, under the subspace constraint, we need to estimate a common covariance matrix, which affects the discriminant subspace, as shown in Section 3.3. Generally speaking, when the discriminant dimension becomes higher or the data set itself is of high dimensions, it becomes more difficult to accurately estimate the covariance matrix. For instance, for the high dimensional data, we assume a common diagonal covariance matrix, so that the covariance estimation becomes feasible and avoids singularity issue. However, this may result in a poor discriminant subspace, which leads to worse classification accuracy. On the other hand, for the data sets of moderately high dimensions, when the discriminant dimension is very low, the estimated covariance matrix is more accurate and the discriminant subspace informative for classification is empirically better than the constrained subspace. As a final note, when the discriminant dimension is low, MDA-DR-MPCA-MEAN generally outperforms MDA-DR-MPCA.

Table 1: Classification error rates (%) for the data with moderately high dimensions (I)

(a) Robots data

Num of components	$d = 2$	$d = 5$	$d = 7$	$d = 9$	$d = 11$	$d = 13$	$d = 15$	$d = 17$
3	GMM-MPCA	41.39	35.78	31.93	31.73	31.25	31.65	31.60
	GMM-MPCA-MEAN	<b>30.32</b>	32.06	<b>30.11</b>	30.52	31.19	31.40	31.29
	GMM-MPCA-SEP	NA	30.86	30.68	<b>30.42</b>	<b>29.58</b>	30.28	30.97
	MDA-RR	41.22	<b>30.32</b>	30.85	30.57	29.95	<b>29.95</b>	<b>29.95</b>
	MDA-RR-OS	40.16	32.73	32.44	30.35	30.66	30.43	30.26
	MDA-DR-MPCA	44.10	40.30	35.04	32.72	33.03	33.56	33.39
	MDA-DR-MPCA-MEAN	41.22	36.42	34.71	33.83	32.75	32.44	33.47
4	GMM-MPCA	40.74	31.91	30.77	30.15	29.40	29.05	28.24
	GMM-MPCA-MEAN	<b>26.56</b>	31.49	<b>29.71</b>	29.98	28.43	28.02	28.39
	GMM-MPCA-SEP	NA	<b>31.41</b>	30.19	28.45	28.92	30.13	29.54
	MDA-RR	40.45	33.63	30.41	<b>28.28</b>	<b>27.77</b>	<b>27.09</b>	<b>27.18</b>
	MDA-RR-OS	40.91	31.87	31.36	30.24	27.88	29.01	28.59
	MDA-DR-MPCA	42.26	36.16	34.64	31.95	30.06	28.90	29.77
	MDA-DR-MPCA-MEAN	39.41	34.38	34.53	31.96	29.73	29.45	28.28
5	GMM-MPCA	37.72	29.67	29.25	29.31	27.86	27.91	<b>26.28</b>
	GMM-MPCA-MEAN	<b>28.72</b>	27.86	<b>26.98</b>	26.69	26.83	<b>25.90</b>	26.37
	GMM-MPCA-SEP	NA	<b>26.48</b>	27.05	27.46	27.22	26.76	26.74
	MDA-RR	40.39	29.01	26.52	<b>26.08</b>	<b>26.03</b>	26.61	26.52
	MDA-RR-OS	39.96	30.99	29.38	28.24	28.48	27.59	28.24
	MDA-DR-MPCA	41.07	35.69	32.44	30.86	29.03	27.99	28.52
	MDA-DR-MPCA-MEAN	38.34	33.10	32.18	30.13	28.56	26.70	27.05

(b) Waveform data

Num of components	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	$d = 16$
3	GMM-MPCA	15.70	15.64	16.12	17.10	17.76	17.80	18.64
	GMM-MPCA-MEAN	16.12	16.14	16.82	17.38	17.76	17.92	18.84
	GMM-MPCA-SEP	NA	17.08	17.04	17.22	17.44	17.50	18.34
	MDA-RR	16.00	18.48	18.64	18.58	18.58	18.58	18.58
	MDA-RR-OS	15.50	17.20	18.14	17.98	18.00	17.84	18.08
	MDA-DR-MPCA	<b>14.74</b>	<b>15.28</b>	15.78	<b>16.14</b>	<b>16.58</b>	17.12	17.62
	MDA-DR-MPCA-MEAN	<b>14.74</b>	15.50	<b>15.76</b>	16.50	17.00	<b>16.94</b>	<b>17.26</b>
4	GMM-MPCA	15.56	16.28	16.06	16.94	17.84	<b>17.54</b>	18.58
	GMM-MPCA-MEAN	15.84	16.70	16.90	17.28	17.96	18.34	18.36
	GMM-MPCA-SEP	NA	16.34	17.14	17.56	17.56	18.02	18.16
	MDA-RR	15.80	18.12	18.28	19.06	19.26	19.66	19.66
	MDA-RR-OS	15.50	17.54	18.36	18.36	19.34	18.92	18.72
	MDA-DR-MPCA	15.18	15.78	<b>16.00</b>	<b>16.36</b>	17.12	17.64	<b>17.64</b>
	MDA-DR-MPCA-MEAN	<b>15.12</b>	<b>15.86</b>	16.16	16.70	<b>17.00</b>	17.56	17.66
5	GMM-MPCA	16.44	16.72	16.42	16.96	17.56	17.86	18.52
	GMM-MPCA-MEAN	16.26	16.30	17.32	17.72	18.04	17.68	19.04
	GMM-MPCA-SEP	NA	17.24	16.96	17.32	17.40	17.66	17.68
	MDA-RR	16.76	18.18	18.26	19.14	19.16	19.70	19.78
	MDA-RR-OS	15.80	17.78	18.62	19.02	19.30	18.92	18.92
	MDA-DR-MPCA	15.34	15.86	<b>15.98</b>	16.66	<b>17.16</b>	<b>16.90</b>	17.90
	MDA-DR-MPCA-MEAN	<b>15.08</b>	<b>15.70</b>	16.76	<b>16.16</b>	17.30	17.90	<b>17.56</b>

Table 2: Classification error rates (%) for the data with moderately high dimensions (II)

(a) Sonar data

Num of components	$d = 2$	$d = 3$	$d = 5$	$d = 7$	$d = 9$	$d = 11$	$d = 13$	$d = 15$
3	GMM-MPCA	39.29	39.78	24.56	25.48	24.51	21.61	21.13
	GMM-MPCA-MEAN	<b>35.92</b>	23.57	23.54	24.04	23.09	22.12	21.63
	GMM-MPCA-SEP	NA	27.85	25.45	24.54	24.06	24.55	23.56
	MDA-RR	36.48	28.82	22.08	22.08	22.08	22.08	22.08
	MDA-RR-OS	45.16	25.87	22.61	24.05	20.68	23.60	22.59
	MDA-DR-MPCA	42.31	38.45	19.71	<b>18.33</b>	19.77	22.18	<b>20.71</b>
	MDA-DR-MPCA-MEAN	39.43	<b>23.56</b>	<b>18.77</b>	<b>18.33</b>	<b>18.83</b>	<b>19.78</b>	21.20
	<b>16.81</b>							
4	GMM-MPCA	40.53	38.88	<b>20.19</b>	20.72	18.32	18.75	17.33
	GMM-MPCA-MEAN	<b>35.08</b>	25.45	20.20	<b>17.83</b>	<b>17.37</b>	<b>18.26</b>	<b>17.31</b>
	GMM-MPCA-SEP	NA	26.51	22.62	22.16	20.25	19.75	20.71
	MDA-RR	46.21	27.91	23.07	19.27	19.27	19.27	19.27
	MDA-RR-OS	42.80	26.35	26.44	19.23	21.62	22.10	19.25
	MDA-DR-MPCA	37.50	37.42	22.11	18.33	18.82	21.21	21.23
	MDA-DR-MPCA-MEAN	40.85	<b>22.11</b>	20.24	19.28	20.24	19.76	20.73
	19.31							
5	GMM-MPCA	44.77	39.78	24.56	25.48	24.51	21.61	21.13
	GMM-MPCA-MEAN	<b>35.42</b>	27.89	<b>21.15</b>	19.73	<b>18.78</b>	19.71	<b>18.26</b>
	GMM-MPCA-SEP	NA	32.31	29.35	20.21	20.22	20.21	19.23
	MDA-RR	43.70	27.38	25.91	22.06	19.67	<b>19.67</b>	19.67
	MDA-RR-OS	35.55	29.34	24.86	22.12	20.68	22.19	21.18
	MDA-DR-MPCA	36.05	35.07	21.20	19.29	20.73	23.09	20.71
	MDA-DR-MPCA-MEAN	38.37	<b>26.44</b>	21.21	<b>18.34</b>	23.10	24.56	21.64
	<b>18.74</b>							

(b) Imagery data

Num of components	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	$d = 16$
3	GMM-MPCA	55.36	48.00	40.36	38.64	38.36	37.43	36.07
	GMM-MPCA-MEAN	<b>44.50</b>	<b>36.21</b>	36.86	37.07	36.36	36.79	36.71
	GMM-MPCA-SEP	NA	NA	<b>35.21</b>	<b>34.07</b>	35.57	35.79	35.14
	MDA-RR	52.57	43.14	40.21	35.86	35.86	35.71	35.29
	MDA-RR-OS	52.36	42.50	38.50	<b>34.07</b>	<b>35.29</b>	<b>35.50</b>	<b>34.93</b>
	MDA-DR-MPCA	59.93	49.36	42.21	41.71	41.00	39.50	37.00
	MDA-DR-MPCA-MEAN	49.36	44.14	40.86	40.93	38.64	38.50	37.79
	38.07							
4	GMM-MPCA	57.00	48.29	39.79	38.14	36.57	36.93	35.64
	GMM-MPCA-MEAN	<b>45.00</b>	<b>37.00</b>	39.21	36.57	35.36	35.43	35.86
	GMM-MPCA-SEP	NA	NA	<b>35.00</b>	<b>35.43</b>	35.07	35.50	35.43
	MDA-RR	52.21	40.64	38.93	35.79	37.50	36.50	35.29
	MDA-RR-OS	51.64	43.57	37.64	35.50	<b>34.50</b>	<b>32.36</b>	<b>34.50</b>
	MDA-DR-MPCA	59.71	50.00	40.14	40.36	38.29	37.86	36.29
	MDA-DR-MPCA-MEAN	49.71	42.36	39.71	39.71	38.64	37.43	37.21
	37.93							
5	GMM-MPCA	57.79	48.50	40.36	37.57	37.36	39.07	36.07
	GMM-MPCA-MEAN	<b>45.64</b>	<b>36.57</b>	38.64	36.14	37.00	36.64	35.64
	GMM-MPCA-SEP	NA	NA	<b>35.79</b>	35.14	34.43	34.36	35.57
	MDA-RR	53.21	43.36	39.00	36.07	35.86	34.43	33.93
	MDA-RR-OS	52.07	42.57	39.71	<b>34.21</b>	<b>32.64</b>	<b>34.21</b>	<b>33.50</b>
	MDA-DR-MPCA	58.50	48.93	39.79	38.21	39.57	39.07	36.00
	MDA-DR-MPCA-MEAN	50.00	42.86	39.21	38.36	39.00	37.57	37.64
	36.21							

Table 3: Classification error rates (%) for the data with moderately high dimensions (III)

(a) Parkinsons data

Num of components	$d = 2$	$d = 3$	$d = 5$	$d = 7$	$d = 9$	$d = 11$	$d = 13$	$d = 15$	
3	GMM-MPCA	17.96	17.42	14.33	14.84	16.98	14.92	15.47	12.84
	GMM-MPCA-MEAN	18.90	14.84	<b>11.75</b>	13.33	13.88	<b>12.88</b>	13.89	12.85
	GMM-MPCA-SEP	NA	<b>11.75</b>	13.29	<b>12.26</b>	13.34	13.85	<b>11.25</b>	13.34
	MDA-RR	19.42	15.96	13.88	13.88	13.88	13.88	13.88	13.88
	MDA-RR-OS	<b>16.88</b>	16.42	12.31	13.89	<b>12.31</b>	13.89	13.37	<b>12.31</b>
	MDA-DR-MPCA	19.47	17.90	13.81	14.35	15.37	14.83	15.38	16.41
	MDA-DR-MPCA-MEAN	19.47	17.90	13.81	14.33	15.37	15.35	15.35	15.35
4	GMM-MPCA	17.88	14.77	14.31	14.81	12.84	11.30	10.28	12.32
	GMM-MPCA-MEAN	<b>14.81</b>	14.31	13.83	12.81	<b>9.25</b>	<b>9.28</b>	<b>8.74</b>	10.76
	GMM-MPCA-SEP	NA	<b>11.28</b>	11.33	12.29	11.80	10.29	9.76	<b>9.79</b>
	MDA-RR	16.85	12.81	11.79	<b>10.29</b>	9.79	9.79	9.79	<b>9.79</b>
	MDA-RR-OS	18.41	15.38	<b>10.74</b>	10.79	11.84	11.83	12.85	10.32
	MDA-DR-MPCA	19.47	18.47	12.29	12.35	10.77	11.23	10.72	12.30
	MDA-DR-MPCA-MEAN	19.47	17.43	12.29	11.81	9.72	10.24	10.72	11.76
5	GMM-MPCA	19.39	18.39	14.84	17.37	15.31	12.81	11.25	12.79
	GMM-MPCA-MEAN	<b>18.39</b>	16.34	13.26	14.83	11.78	11.25	10.25	11.29
	GMM-MPCA-SEP	NA	<b>14.81</b>	<b>10.75</b>	12.25	11.75	<b>10.75</b>	10.25	10.78
	MDA-RR	19.94	16.30	12.27	13.83	11.28	10.78	11.28	10.79
	MDA-RR-OS	18.96	16.43	14.30	<b>11.22</b>	10.25	12.33	<b>9.71</b>	<b>9.74</b>
	MDA-DR-MPCA	18.96	18.47	13.80	11.81	11.28	12.77	10.70	10.76
	MDA-DR-MPCA-MEAN	18.96	19.52	12.26	11.31	<b>9.75</b>	12.27	10.20	<b>9.74</b>

(b) Satellite data

Num of components	$d = 2$	$d = 4$	$d = 7$	$d = 9$	$d = 11$	$d = 13$	$d = 15$	$d = 17$	
3	GMM-MPCA	<b>16.74</b>	15.01	14.16	14.67	14.06	13.95	13.97	13.63
	GMM-MPCA-MEAN	16.94	14.10	13.53	13.77	13.95	13.78	13.66	13.68
	GMM-MPCA-SEP	NA	NA	15.48	13.58	13.80	13.58	13.71	13.67
	MDA-RR	35.18	14.41	12.84	<b>12.96</b>	13.46	13.60	13.66	13.53
	MDA-RR-OS	34.90	<b>13.95</b>	<b>13.01</b>	13.09	<b>12.82</b>	<b>13.04</b>	<b>13.35</b>	13.29
	MDA-DR-MPCA	17.20	14.83	13.61	13.91	13.80	13.35	13.60	13.69
	MDA-DR-MPCA-MEAN	17.09	14.42	13.58	14.12	14.06	13.41	13.38	<b>13.13</b>
4	GMM-MPCA	<b>17.02</b>	14.13	13.61	13.80	13.58	12.90	12.93	12.88
	GMM-MPCA-MEAN	17.31	13.41	13.50	13.53	13.24	12.94	12.91	12.87
	GMM-MPCA-SEP	NA	NA	15.40	12.93	13.08	13.35	13.38	13.54
	MDA-RR	35.06	<b>13.35</b>	12.60	12.77	12.74	12.73	12.63	13.05
	MDA-RR-OS	34.28	13.49	<b>11.95</b>	<b>12.17</b>	<b>11.90</b>	12.49	<b>11.97</b>	<b>12.14</b>
	MDA-DR-MPCA	17.54	14.14	13.21	13.52	13.05	12.74	12.46	12.45
	MDA-DR-MPCA-MEAN	17.37	13.36	13.53	13.57	13.07	<b>12.43</b>	12.45	12.82
5	GMM-MPCA	<b>16.25</b>	13.66	12.90	13.29	12.79	12.26	11.92	12.24
	GMM-MPCA-MEAN	16.77	12.93	12.85	12.96	12.40	12.18	11.89	12.24
	GMM-MPCA-SEP	NA	NA	15.48	13.21	12.56	12.70	12.59	12.49
	MDA-RR	27.43	13.27	12.85	12.34	12.15	12.18	12.29	12.28
	MDA-RR-OS	30.16	13.30	<b>12.31</b>	<b>12.23</b>	<b>11.73</b>	<b>11.89</b>	11.98	<b>11.97</b>
	MDA-DR-MPCA	16.61	13.80	12.70	12.82	12.74	12.09	<b>11.79</b>	12.42
	MDA-DR-MPCA-MEAN	16.58	<b>12.82</b>	12.99	12.93	12.66	11.92	11.92	12.31

Table 4: Classification error rates (%) for the data with high dimensions

		(a) Semeion data							
Num of components		$d = 2$	$d = 4$	$d = 8$	$d = 11$	$d = 13$	$d = 15$	$d = 17$	$d = 19$
3	GMM-MPCA	53.56	29.72	18.27	19.81	18.89	19.20	18.89	18.27
	GMM-MPCA-MEAN	49.54	29.10	13.31	14.86	16.72	14.86	16.72	16.10
	GMM-MPCA-SEP	NA	NA	NA	13.31	12.07	13.00	16.10	14.86
	MDA-RR	<b>45.51</b>	26.93	15.79	14.86	13.93	15.79	14.24	13.62
	MDA-RR-OS	48.36	<b>24.92</b>	13.93	<b>12.41</b>	<b>10.60</b>	<b>11.10</b>	<b>10.59</b>	<b>11.41</b>
	MDA-DR-MPCA	49.54	27.86	19.20	17.03	17.03	15.79	16.72	16.41
	MDA-DR-MPCA-MEAN	48.30	26.01	<b>12.07</b>	13.62	13.31	12.07	14.24	14.55
4	GMM-MPCA	53.56	26.32	17.03	16.10	16.10	16.10	16.10	15.17
	GMM-MPCA-MEAN	51.39	25.70	<b>11.46</b>	11.76	12.69	13.00	14.24	15.17
	GMM-MPCA-SEP	NA	NA	NA	13.31	12.07	12.07	13.62	11.76
	MDA-RR	49.23	26.01	13.93	13.62	13.00	12.07	13.62	12.07
	MDA-RR-OS	48.70	<b>24.83</b>	14.21	<b>11.60</b>	<b>10.59</b>	<b>11.16</b>	<b>9.97</b>	<b>11.09</b>
	MDA-DR-MPCA	46.75	26.32	17.34	16.41	16.41	15.17	16.10	15.17
	MDA-DR-MPCA-MEAN	<b>44.58</b>	26.32	13.00	11.76	15.17	13.31	13.00	13.00
5	GMM-MPCA	51.70	<b>24.46</b>	15.79	13.62	15.17	15.17	13.93	13.00
	GMM-MPCA-MEAN	<b>43.03</b>	26.63	11.15	11.46	12.38	12.07	13.31	13.00
	GMM-MPCA-SEP	NA	NA	NA	13.00	11.46	13.00	12.38	13.00
	MDA-RR	48.92	25.39	13.00	12.69	11.46	<b>10.53</b>	12.07	12.69
	MDA-RR-OS	49.16	26.53	14.21	11.10	<b>10.60</b>	<b>10.53</b>	9.84	9.96
	MDA-DR-MPCA	48.61	27.24	18.58	13.93	14.24	13.62	13.93	10.84
	MDA-DR-MPCA-MEAN	46.13	25.08	<b>10.22</b>	<b>10.84</b>	10.84	10.53	<b>8.98</b>	<b>9.29</b>
		(b) YaleB data							
Num of components		$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	$d = 16$
3	GMM-MPCA	84.29	64.29	64.29	55.71	45.71	38.57	40.00	34.29
	GMM-MPCA-MEAN	31.43	<b>17.14</b>	52.86	51.43	38.57	30.00	28.57	27.14
	GMM-MPCA-SEP	NA	NA	<b>27.14</b>	20.00	<b>21.43</b>	20.00	20.00	20.00
	MDA-RR	87.14	42.86	<b>27.14</b>	<b>17.14</b>	28.57	<b>8.57</b>	<b>11.43</b>	<b>11.43</b>
	MDA-DR-MPCA	82.86	58.57	50.00	44.29	37.14	42.86	25.71	32.86
	MDA-DR-MPCA-MEAN	<b>30.00</b>	<b>17.14</b>	60.00	37.14	40.00	21.43	17.14	14.29
4	GMM-MPCA	84.29	67.14	68.57	55.71	44.29	44.29	40.00	37.14
	GMM-MPCA-MEAN	34.29	22.86	64.29	50.00	35.71	30.00	35.71	30.00
	GMM-MPCA-SEP	NA	NA	<b>31.43</b>	25.71	28.57	27.14	24.29	25.71
	MDA-RR	85.71	60.00	41.43	<b>24.29</b>	<b>14.29</b>	<b>10.00</b>	12.86	<b>11.43</b>
	MDA-DR-MPCA	90.00	55.71	50.00	42.86	37.14	41.43	28.57	27.14
	MDA-DR-MPCA-MEAN	<b>25.71</b>	<b>21.43</b>	60.00	35.71	32.86	11.43	<b>11.43</b>	12.86
5	GMM-MPCA	85.71	65.71	65.71	55.71	50.00	45.71	42.86	40.00
	GMM-MPCA-MEAN	37.14	<b>14.29</b>	60.00	51.43	47.14	42.86	41.43	38.57
	GMM-MPCA-SEP	NA	NA	<b>31.43</b>	35.71	<b>30.00</b>	32.86	28.57	35.71
	MDA-RR	85.71	61.43	42.86	42.86	32.86	30.00	22.86	24.29
	MDA-DR-MPCA	87.14	67.14	52.86	38.57	34.29	34.29	<b>17.14</b>	22.86
	MDA-DR-MPCA-MEAN	<b>27.14</b>	18.57	50.00	<b>34.29</b>	<b>27.14</b>	<b>21.43</b>	20.00	<b>7.14</b>

### 5.3 Sensitivity of Subspace to Bandwidths

Different kernel bandwidths may result in different sets of modes by HMAC, which again may yield different constrained subspaces. We investigate in this section the sensitivity of constrained subspaces to kernel bandwidths.

Assume two subspaces  $\boldsymbol{\nu}_1$  and  $\boldsymbol{\nu}_2$  are spanned by two sets of orthonormal basis vectors  $\{\mathbf{v}_1^{(1)}, \dots, \mathbf{v}_d^{(1)}\}$  and  $\{\mathbf{v}_1^{(2)}, \dots, \mathbf{v}_d^{(2)}\}$ , where  $d$  is the dimension. To measure the closeness between two subspaces, we project the basis of one subspace onto the other. Specifically, the closeness between  $\boldsymbol{\nu}_1$  and  $\boldsymbol{\nu}_2$  is defined as  $\text{closeness}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2) = \sum_{i=1}^d \sum_{j=1}^d (\mathbf{v}_i^{(1)t} \cdot \mathbf{v}_j^{(2)})^2$ . If  $\boldsymbol{\nu}_1$  and  $\boldsymbol{\nu}_2$  span the same subspace,  $\sum_{j=1}^d (\mathbf{v}_i^{(1)t} \cdot \mathbf{v}_j^{(2)})^2 = 1$ , for  $i = 1, 2, \dots, d$ . If they are orthogonal to each other,  $\sum_{j=1}^d (\mathbf{v}_i^{(1)t} \cdot \mathbf{v}_j^{(2)})^2 = 0$ , for  $i = 1, 2, \dots, d$ . Therefore, the range of  $\text{closeness}(\boldsymbol{\nu}_1, \boldsymbol{\nu}_2)$  is  $(0, d)$ . The higher the value, the closer the two subspaces are.

In our proposed methods, a collection of constrained subspaces are obtained through MPCA or MPCA-MEAN at different kernel bandwidth  $\sigma_l$ 's,  $l = 1, 2, \dots, \eta$ , and  $\sigma_1 < \sigma_2 < \dots < \sigma_\eta$ . To measure the sensitivity of subspaces to different bandwidths, we compute the mean closeness between the subspace found at  $\sigma_l$  and all the other subspaces at preceding bandwidths  $\sigma_{l'}, l' = 1, 2, \dots, l-1$ . A large mean closeness indicates that the current subspace is close to preceding subspaces. Table 5 and 6 list the mean closeness of subspaces by MPCA and MPCA-MEAN at different bandwidth levels for the sonar and imagery data (the training set from one fold in the previous five-fold cross validation setup). We vary the dimension of the constrained subspace. The number of modes identified at each level is also shown in the tables. As Table 5 and 6 show, for both methods, the subspaces found at the first few levels are close to each other, indicated by their large mean closeness values, which are close to  $d$ , the dimension of the subspace. As the bandwidth  $\sigma_l$  increases, the mean closeness starts to decline, which indicates that the corresponding subspace changes. When  $\sigma_l$  is small, the number of modes identified by HMAC is large. The modes and their associated weights do not change much. As a result, the generated subspaces at these bandwidths are relatively stable. As  $\sigma_l$  increases, the kernel density estimate becomes smoother, and more data points tend to ascend to the same mode. We thus have a smaller number of modes with changing weights, which may yield a substantially different subspace. Additionally, the subspace by MPCA-MEAN is spanned by applying weighted PCA to a union set of modes and class means. In our experiment, we have allocated a larger weight proportionally to class means (in total, 60%) and the class means remain unchanged in the union set at each kernel bandwidth. Therefore, the differences between subspaces by MPCA-MEAN are smaller than that by MPCA.

### 5.4 Model Selection

In our proposed method, the following model selection strategy is adopted. We take a sequence of subspaces resulting from different kernel bandwidths, and then estimate a mixture model constrained by each subspace and finally choose a model yielding the maximum likelihood. In this section, we examine our model selection criteria, and the relationships among test classification error rates, training likelihoods and kernel bandwidths.

Figure 1 shows the test classification error rates at different levels of kernel bandwidth for several data sets (from one fold in the previous five-fold cross validation setup), when

the number of mixture components for each class is set to three. The error rates are close to each other at the first few levels. As the kernel bandwidth increases, the error rates start to change. Except for the waveform, on which the error rates of GMM-MPCA and GMM-MPCA-MEAN are very close, for the other data sets in Figure 1, the error rate of GMM-MPCA-MEAN at each bandwidth level is lower than that of GMM-MPCA. Similarly, at each kernel bandwidth level, the error rate of GMM-MPCA-SEP is also lower than that of GMM-MPCA, except for the robot data. We also show the training log-likelihoods of these methods with respect to different kernel bandwidth levels in Figure 2. The training log-likelihoods are also stable at the first few levels and start to fluctuate as the bandwidth increases. This is due to the possible big change in subspaces under large kernel bandwidths.

In our model selection strategy, the subspace which results in the maximum log likelihood of the training model is selected and then we apply the model under the constraint of that specific subspace to classify the test data. In Figure 1, the test error rate of the model which has the largest training likelihood is indicated by an arrow. As we can see, for each method, this error rate is mostly ranked in the middle among all the error rates at different levels of bandwidth, which indicate that our model selection strategy helps find a reasonable training model.

Table 5: Mean closeness of subspaces by MPCA at different levels of kernel bandwidth

(a) Sonar data

Bandwidth level	2	4	6	8	10	12	14
Num of modes	158	144	114	86	60	15	8
$d = 2$	2.000	2.000	1.997	1.993	1.980	1.773	1.427
$d = 4$	4.000	3.999	3.994	3.976	3.838	2.625	2.093
$d = 6$	5.999	5.990	5.951	5.906	5.408	4.483	3.484
$d = 8$	8.000	7.996	7.969	7.862	6.983	6.201	4.291

(b) Imagery data

Bandwidth level	2	4	6	8	10	12	14
Num of modes	1109	746	343	144	60	35	14
$d = 6$	6.000	5.965	5.321	5.340	5.207	5.017	3.634
$d = 8$	8.000	7.958	7.541	7.305	6.815	6.274	4.825
$d = 10$	9.997	9.797	9.561	9.249	8.450	7.711	5.862
$d = 12$	12.000	11.957	11.479	10.286	10.056	9.487	7.419

#### 5.4.1 CLASSIFICATION ON SIMULATED DATA

A simulation study is conducted in this section to further investigate the classification performance of our proposed methods and the reduced rank MDA. We generate the synthetic data using a Gaussian mixture model with pre-determined subspace constrained means.

Specifically, we take the training set of the imagery data (1120 samples, 64 dimensions, and five classes) from one fold in the previous five-fold cross validation setup and estimate its distribution by fitting a mixture model via GMM-MPCA. The number of components

Table 6: Mean closeness of subspaces by MPCA-MEAN at different levels of kernel bandwidth

(a) Sonar data

Bandwidth level	2	4	6	8	10	12	14
Num of modes	158	144	114	86	60	15	8
$d = 2$	2.000	2.000	1.997	1.992	1.978	1.884	1.802
$d = 4$	4.000	4.000	3.996	3.985	3.943	3.078	2.964
$d = 6$	6.000	5.993	5.987	5.879	5.468	4.642	4.134
$d = 8$	8.000	7.996	7.968	7.885	7.004	6.402	5.175

(b) Imagery data

Bandwidth level	2	4	6	8	10	12	14
Num of modes	1109	746	343	144	60	35	14
$d = 6$	6.000	5.987	5.855	5.605	5.317	5.318	5.153
$d = 8$	8.000	7.962	7.890	7.755	7.433	6.850	6.610
$d = 10$	9.999	9.517	9.476	9.227	9.005	8.423	7.554
$d = 12$	12.000	11.958	11.495	10.893	10.332	9.868	8.976

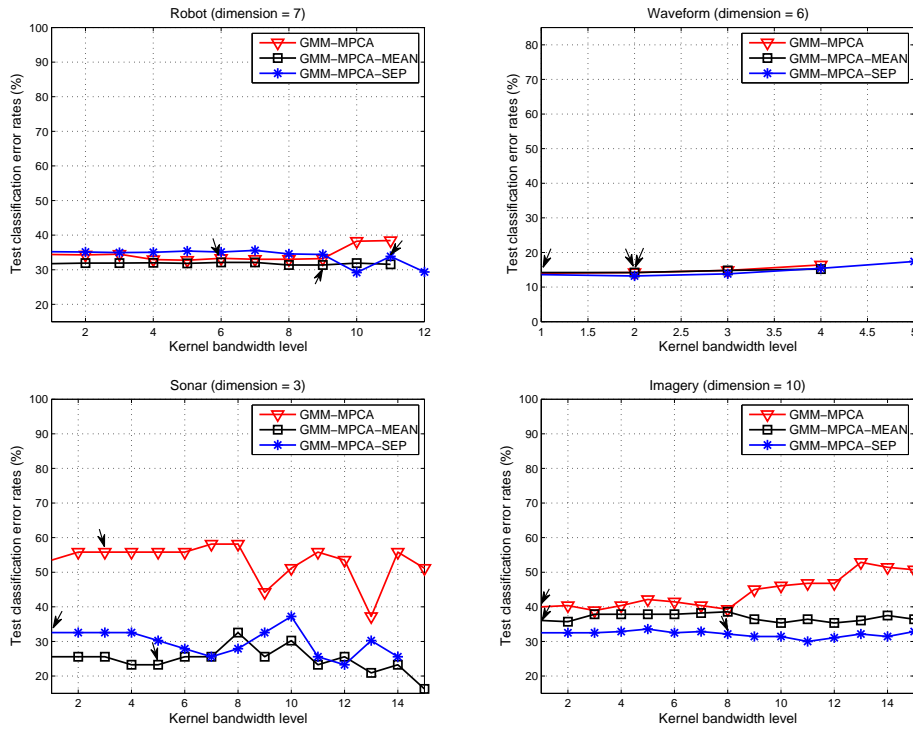


Figure 1: The test classification error rates at different levels of kernel bandwidth

Table 7: Classification error rates (%) for simulation data with different dispersions

(a) Simulation data I (low dispersion)

Num of components	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	
3	GMM-MPCA	<b>17.00</b>	<b>17.00</b>	<b>17.73</b>	<b>18.00</b>	<b>17.73</b>	<b>18.13</b>	<b>18.00</b>
	GMM-MPCA-MEAN	17.07	17.60	18.13	19.07	18.53	18.87	18.87
	MDA-RR	17.27	18.27	18.67	20.13	20.47	21.27	21.47
	MDA-RR-OS	17.20	17.27	18.73	21.40	20.80	21.60	20.80
4	GMM-MPCA	<b>16.93</b>	<b>17.47</b>	<b>17.33</b>	<b>17.67</b>	<b>17.40</b>	<b>17.33</b>	<b>17.53</b>
	GMM-MPCA-MEAN	17.20	18.47	18.33	18.60	19.40	19.07	19.40
	MDA-RR	17.07	18.07	18.47	19.93	19.87	19.53	21.80
	MDA-RR-OS	17.13	17.40	19.20	20.27	19.87	20.73	21.33
5	GMM-MPCA	<b>16.93</b>	<b>17.33</b>	<b>17.73</b>	<b>18.00</b>	<b>18.47</b>	<b>17.93</b>	<b>18.53</b>
	GMM-MPCA-MEAN	17.00	18.87	18.53	18.93	19.20	19.27	20.27
	MDA-RR	17.13	19.27	19.47	21.00	20.73	21.93	21.33
	MDA-RR-OS	16.93	18.67	19.13	21.47	21.07	21.13	21.33

(b) Simulation data II (middle dispersion)

Num of components	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	
3	GMM-MPCA	<b>12.47</b>	<b>12.80</b>	<b>12.60</b>	<b>12.80</b>	<b>12.47</b>	<b>12.73</b>	<b>12.93</b>
	GMM-MPCA-MEAN	12.60	13.67	14.13	13.80	14.27	14.20	14.07
	MDA-RR	12.67	13.60	13.47	15.00	17.53	15.67	15.73
	MDA-RR-OS	12.67	14.07	13.53	17.87	14.80	17.13	16.00
4	GMM-MPCA	<b>12.47</b>	<b>12.53</b>	<b>12.07</b>	<b>12.53</b>	<b>12.33</b>	<b>12.80</b>	<b>12.73</b>
	GMM-MPCA-MEAN	12.53	14.00	14.60	13.87	14.40	14.13	14.00
	MDA-RR	<b>12.47</b>	14.00	15.13	14.40	14.67	15.20	14.67
	MDA-RR-OS	<b>12.47</b>	13.53	13.87	15.53	14.67	14.27	16.67
5	GMM-MPCA	12.27	<b>12.67</b>	<b>11.87</b>	<b>12.33</b>	<b>12.67</b>	<b>13.60</b>	<b>13.53</b>
	GMM-MPCA-MEAN	<b>12.07</b>	13.73	14.60	14.47	14.53	14.67	15.20
	MDA-RR	12.40	13.53	14.40	13.87	14.27	16.20	15.40
	MDA-RR-OS	12.53	14.00	15.00	15.67	15.60	14.80	15.73

(c) Simulation data III (high dispersion)

Num of components	$d = 2$	$d = 4$	$d = 6$	$d = 8$	$d = 10$	$d = 12$	$d = 14$	
3	GMM-MPCA	<b>1.80</b>	<b>1.80</b>	<b>1.87</b>	<b>2.00</b>	2.27	<b>2.13</b>	<b>2.07</b>
	GMM-MPCA-MEAN	<b>1.80</b>	1.87	2.00	2.13	<b>2.07</b>	<b>2.13</b>	2.20
	MDA-RR	<b>1.80</b>	2.00	1.93	2.53	2.47	2.47	2.73
	MDA-RR-OS	2.40	2.93	2.27	3.27	2.53	3.07	3.07
4	GMM-MPCA	<b>1.73</b>	1.80	<b>1.80</b>	<b>1.80</b>	2.20	<b>1.80</b>	<b>2.00</b>
	GMM-MPCA-MEAN	<b>1.73</b>	<b>1.73</b>	2.20	2.13	<b>2.07</b>	2.20	2.07
	MDA-RR	<b>1.73</b>	2.27	2.33	2.27	2.47	3.07	3.00
	MDA-RR-OS	1.93	2.27	2.33	2.47	2.73	3.00	2.73
5	GMM-MPCA	<b>1.67</b>	<b>1.87</b>	<b>2.00</b>	<b>1.93</b>	<b>1.87</b>	<b>1.87</b>	<b>2.13</b>
	GMM-MPCA-MEAN	<b>1.67</b>	<b>1.87</b>	2.13	2.20	2.53	2.40	2.27
	MDA-RR	<b>1.67</b>	2.07	2.13	2.27	2.67	2.27	3.13
	MDA-RR-OS	2.07	1.93	1.93	2.20	2.07	2.60	3.00

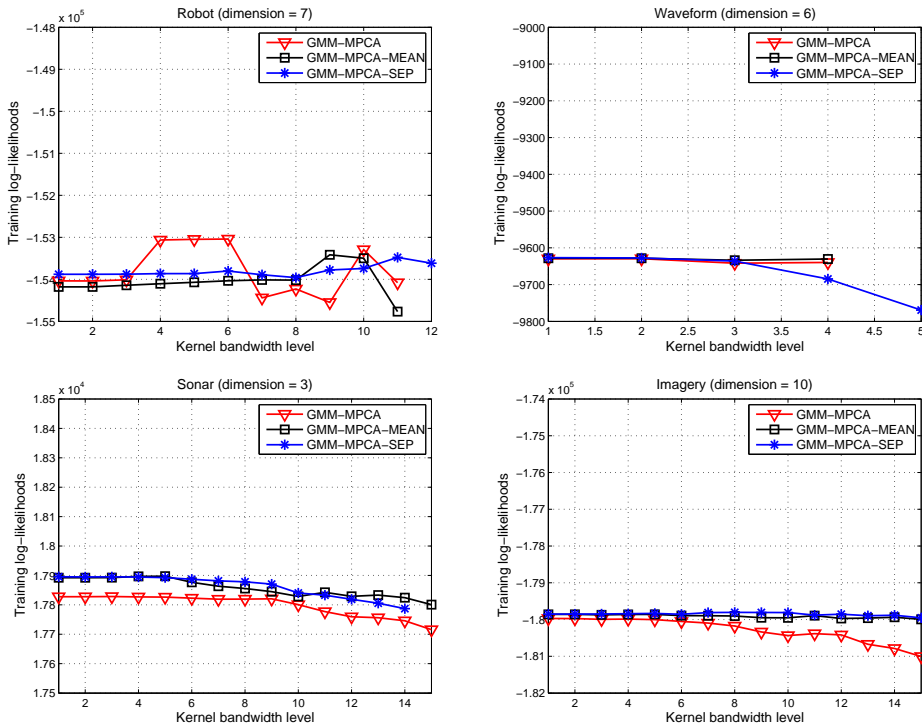


Figure 2: Training log-likelihoods at different kernel bandwidth levels

used to model each class is set to be three and the dimension of constrained subspace is two. We then obtain the estimated component means from all the classes which are ensured to be constrained in a two dimensional subspace. For each class, a set of 300 samples are randomly drawn from a multivariate Gaussian mixture distribution with the previously estimated component means as the sample means. For simplicity, we set an equal prior for each component in the mixture, that is,  $1/3$ . A common diagonal covariance is assumed for the multivariate Gaussian mixture distribution over all the classes, in which each dimension has an equal covariance. Five sets of samples can be generated in this way, forming a data set of 1500 samples. Since both the constrained subspace and the common covariance are pre-determined, we know the true discriminant subspace of the data.

In order to investigate the performance of the proposed methods and reduced rank MDA on the classification with different complexity, we generate three sets of simulated data with various levels of dispersion between classes. The lower the dispersion, the more difficult the classification task. Suppose the three mixture component means within class  $k$  from the pre-determined subspace constrained model are  $\mu_{k1}, \mu_{k2}, \mu_{k3}$ , we can randomly draw samples from the Gaussian mixture distribution with these component means as sample means and generate a new set of simulated data. Suppose the class mean of the simulated data in class  $k$  is  $\mu_k$ , we have  $\mu_k = \sum_{r=1}^3 \pi_r \mu_{kr}$ , where  $\pi_r$  is the prior for component  $r$ , which is set to be  $1/3$  in our simulation. We can change the dispersion between classes by multiplying  $\mu_k$  by a constant  $a$ . Let the new class mean be  $\mu'_k = a\mu_k$ . We have the new mixture component as  $\mu'_{kr} = \mu'_k + (\mu_{kr} - \mu_k)$ . That is, we shift the mixture component means so that they

still center around the new class mean without changing their original distance to the class mean. This will keep the dispersion within a class unchanged, but change the dispersion between classes. If  $a > 1.0$ , we will increase the dispersion between classes. Otherwise, the dispersion will remain unchanged or reduced. Note that these linear transformations on the component means will not change their constrained subspaces. In our experiments, we set  $a$  to be 0.5, 1.5, and 3.5, which result in simulated data with low, middle, and high level dispersions, respectively.

In Table 7, we show the five-fold cross validation classification accuracy of our proposed methods and the reduced rank MDA methods on three simulated data sets, with different number of discriminant dimension and mixture components. GMM-MPCA achieves the lowest error rates under almost all the discriminant dimensions for the three data sets. When  $d = 2$ , the same dimension as the true discriminant subspace, the classification error rates of all the tested methods are very close, although the error rates of GMM-MPCA are equal to or slightly lower than those of MDA-RR and MDA-RR-OS. As the dimension  $d$  increases, the error rates of all the methods start to change. Comparing with reduced rank MDA, the performance of our proposed methods are more stable. We select the data in one fold and visualize the classification results of the test data. Figure 3(a) shows the visualization of three simulated data sets in the true two dimensional discriminant subspace, color-coding the true classes. The top, middle, and bottom plots correspond to data with low, middle and high level dispersions, respectively. As the simulation data I are with low level dispersion between classes, all the tested methods have higher classification error rates, shown in Table 7(a), comparing with those on simulation data III which has high level dispersion between classes. In addition, Figure 3(b) and Figure 3(c) show the projections of test data onto the two-dimensional discriminant subspace by GMM-MPCA and MDA-RR, color-coding the predicted classes. For all the methods, we assume three mixture components in each class. The misclassified data points are colored by gray. The classification error rates by GMM-MPCA (listed in the title above each plot) in Figure 3(b) are very close to those by MDA-RR in Figure 3(c). For all the simulated data sets, we calculate the closeness between the true discriminant subspace and the discriminant subspaces found by GMM-MPCA and MDA-RR, shown in Table 8(a). As we can see, both the subspaces by GMM-MPCA and MDA-RR are similar to the true discriminant subspace, indicated by their large closeness values, which are close to the dimension of the true discriminant subspace.

Table 8: Closeness between subspaces in simulation studies

(a) Classification with different dispersions

Closeness	low	middle	high
GMM-MPCA	1.925	1.912	1.887
MDA-RR	1.919	1.908	1.885

(b) Clustering with different dispersions

Closeness	low	middle	high
GMM-MPCA	1.769	1.820	1.881
MDA-RR	1.552	1.760	1.866

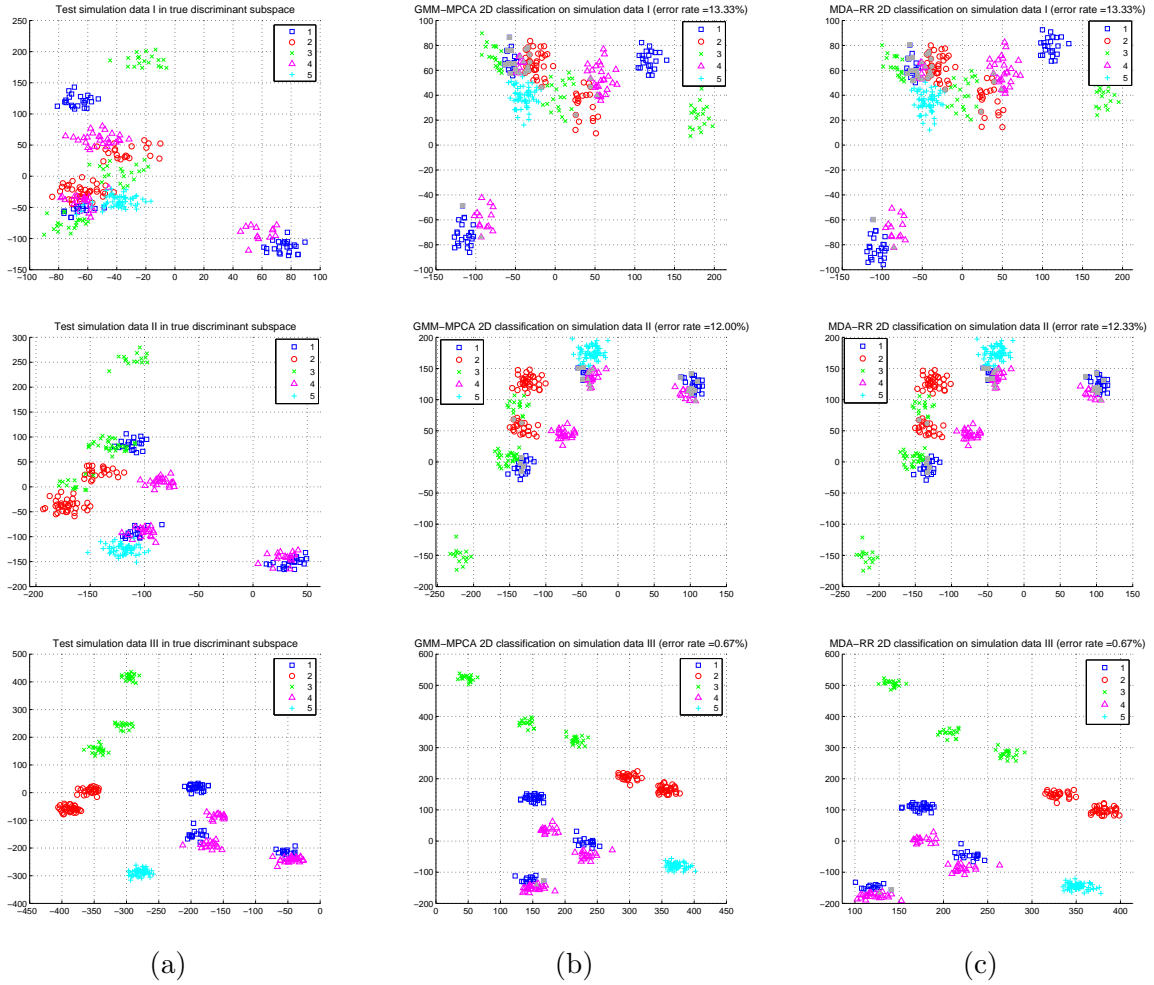


Figure 3: Two-dimensional plots for the classifications of test data from synthetic data sets, color-coding the classes. (a) Projections onto the true discriminant subspace, with true class labels. (b) Projections onto the two-dimensional discriminant subspace by GMM-MPCA, with predicted class labels. (c) Projections onto the two-dimensional discriminant subspace by MDA-RR, with predicted class labels.

## 5.5 Clustering

We present the clustering results of GMM-MPCA and MDA-RR on several synthetic data sets and visualize the results in a low-dimensional subspace. The previous model selection criteria is also used in clustering. After fitting a subspace constrained Gaussian mixture model, all the data points having the highest posterior probability belonging to a particular component form one cluster. We outline the data simulation process as follows.

Similar to the simulation study in Section 5.4.1, the data is generated from some existing subspace constrained model. Again, we take the training set of the imagery data from one fold in the previous five-fold cross validation setup and estimate its distribution by fitting a mixture model via GMM-MPCA. We will obtain five estimated component means which are ensured to be constrained in a two dimensional subspace. A set of 200 samples are randomly drawn from a multivariate Gaussian distribution with the previously estimated component means as the sample means. A common identity covariance is assumed for the Gaussian multivariate distributions. We generate five sets of samples in this way, which form a data set of 1000 samples. We scale the component means by different factors  $a$  so that the clusters of data have different levels of dispersion. Specifically,  $a$  is set to be 0.125, 0.150, and 0.250, respectively, generating three simulated data with low, middle and high level dispersion between clusters.

Figure 4 shows the clustering results of three simulated data sets by GMM-MPCA and MDM-RR, in two-dimensional plots, color-coding the clusters. The data projected onto the true discriminant subspace with true cluster labels are shown in Figure 4(a). In addition, Figure 4(b) and Figure 4(c) show the data projected onto the two-dimensional discriminant subspaces by GMM-MPCA and MDA-RR. For all the synthetic data sets, both GMM-MPCA and MDA-RR can effectively reveal the clustering structure in a low-dimensional subspace. To evaluate their clustering performance, we compute the clustering accuracy by comparing their predicted and true clustering labels. Suppose the true cluster label of data point  $\mathbf{x}_i$  is  $t_i$  and the predicted cluster label is  $p_i$ , the clustering error rate is calculated as  $1 - \sum_{i=1}^n I(t_i, \text{map}(p_i))/n$ , where  $n$  is the total number of data points,  $I(x, y)$  is an indicator function that is equal to one if  $x = y$  otherwise zero, and  $\text{map}(p_i)$  is a permutation function which maps the predicted label to an equivalent label in the data set. Specifically, we use the Kuhn-Munkres algorithm to find the best matching (Lovász and Plummer, 1986). The clustering error rates are listed in the titles above the plots in Figure 4. The mis-clustered data points are in gray. When the dispersion between clusters is low or middle, the clustering error rates of GMM-MPCA are smaller than those of MDA-RR. When the dispersion is high, the task becomes relatively easy and the clustering accuracy of these two methods are the same. In Table 8(b), we also show the closeness between the true discriminant subspace and the discriminant subspaces found by GMM-MPCA and MDA-RR. Comparing with MDA-RR, for all the three data sets, the closeness between the subspace by GMM-MPCA and the true subspace are smaller.

## 6. Conclusion

In this paper, we propose a Gaussian mixture model with the component means constrained in a pre-selected subspace. We prove that the modes, the component means of a Gaussian mixture, and the class means all lie in the same constrained subspace. Several approaches

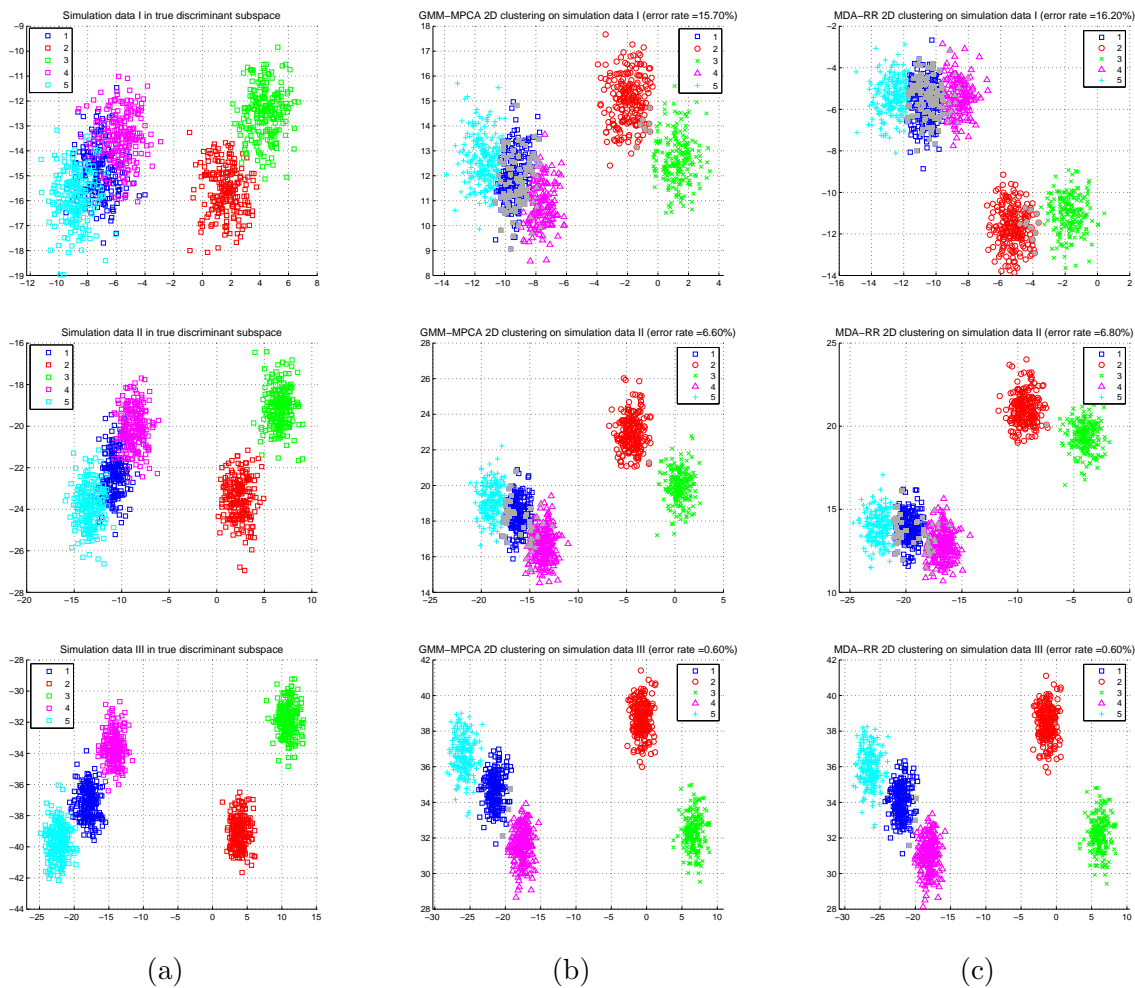


Figure 4: Two-dimensional plot for the clustering of synthetic data, color-coding the clusters. (a) Projections onto the two-dimensional true discriminant subspace, with true cluster labels. (b) Projections onto the two-dimensional discriminant subspace by GMM-MPCA, with predicted cluster labels. (c) Projections onto the two-dimensional discriminant subspace by MDA-RR, with predicted cluster labels.

to finding the subspace are proposed by applying weighted PCA to the modes, class means, or a union set of modes and class means. The constrained method results in a dimension reduction property, which allows us to view the classification or clustering structure of the data in a lower dimensional space. An EM-type algorithm is derived to estimate the model, given any constrained subspace. In addition, the Gaussian mixture model with the component means constrained by separate parallel subspace for each class is investigated. Although reduced rank MDA is a competitive classification method by constraining the class means to an optimal discriminant subspace within each EM iteration, experiments on several real and simulated data sets of moderate to high dimensions show that when the dimension of the discriminant subspace is very low, it is often outperformed by our proposed method with a simple technique of spanning the constrained subspace using only class means.

We select the constrained subspace which has the largest training likelihood among a sequence of subspaces resulting from different kernel bandwidths. If the number of candidate subspaces is large, it may be desired to narrow down the search by incorporating some prior knowledge, for instance, the new method may have a potential in visualization when users already know that only a certain dimensions of the data matter for classification or clustering, i.e., a constrained subspace can be obtained beforehand. Finally, we expect this subspace constrained method can be extended to other parametric mixtures, for instance, mixture of Poisson for discrete data.

## Appendix A.

We prove Theorem 1 in Section 3.1. Consider a mixture of Gaussians with a common covariance matrix  $\Sigma$  shared across all the components as in (2):

$$f(\mathbf{X} = \mathbf{x}) = \sum_{r=1}^R \pi_r \phi(\mathbf{x} | \boldsymbol{\mu}_r, \Sigma) .$$

Once  $\Sigma$  is identified, a linear transform (a “whitening” operation) can be applied to  $\mathbf{X}$  so that the transformed data follow a mixture with component-wise diagonal covariance, more specifically, the identity matrix  $\mathbf{I}$ . Assume  $\Sigma$  is non-singular and hence positive definite, we can find the matrix square root of  $\Sigma$ , that is,  $\Sigma = (\Sigma^{\frac{1}{2}})^t \Sigma^{\frac{1}{2}}$ . If the eigen decomposition of  $\Sigma$  is  $\Sigma = V_{\Sigma} D_{\Sigma} V_{\Sigma}^t$ , then,  $\Sigma^{\frac{1}{2}} = D_{\Sigma}^{\frac{1}{2}} V_{\Sigma}^t$ . Let  $W = ((\Sigma^{\frac{1}{2}})^t)^{-1}$  and  $\mathbf{Z} = W\mathbf{X}$ . The density of  $\mathbf{Z}$  is  $g(\mathbf{Z} = \mathbf{z}) = \sum_{r=1}^R \pi_r \phi(\mathbf{z} | W\boldsymbol{\mu}_r, \mathbf{I})$ . Any mode of  $g(\mathbf{z})$  corresponds to a mode of  $f(\mathbf{x})$  and vice versa. Hence, without loss of generality, we can assume  $\Sigma = \mathbf{I}$ .

Another linear transform on  $\mathbf{Z}$  can be performed using the orthonormal basis  $V = \boldsymbol{\nu} \cup \boldsymbol{\nu}^{\perp} = \{\mathbf{v}_1, \dots, \mathbf{v}_p\}$ , where  $\boldsymbol{\nu} = \{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$  is the constrained subspace where  $\boldsymbol{\mu}_{kr}$ 's reside, and  $\boldsymbol{\nu}^{\perp} = \{\mathbf{v}_1, \dots, \mathbf{v}_q\}$  is the corresponding null subspace, as defined in Section 3. Suppose  $\tilde{\mathbf{Z}} = \mathbf{Proj}_{\tilde{V}}^{\mathbf{Z}}$ . For the transformed data  $\tilde{\mathbf{z}}$ , the covariance matrix is still  $\mathbf{I}$ . Again, there is a one-to-one correspondence (via the orthonormal linear transform) between the modes of  $g_k(\tilde{\mathbf{z}})$  and the modes of  $g_k(\mathbf{z})$ . The density of  $\tilde{\mathbf{z}}$  is

$$g(\tilde{\mathbf{Z}} = \tilde{\mathbf{z}}) = \sum_{r=1}^R \pi_r \phi(\tilde{\mathbf{z}} | \boldsymbol{\theta}_r, \mathbf{I}) ,$$

where  $\boldsymbol{\theta}_r$  is the projection of  $W\boldsymbol{\mu}_r$  onto the orthonormal basis  $V$ , i.e.,  $\boldsymbol{\theta}_{kr} = \mathbf{Proj}_V^W \boldsymbol{\mu}_r$ . Split  $\boldsymbol{\theta}_r$  into two parts,  $\boldsymbol{\theta}_{r,1}$  being the first  $q$  dimensions of  $\boldsymbol{\theta}_r$  and  $\boldsymbol{\theta}_{r,2}$  being the last  $p - q$  dimensions. Since the projections of  $\boldsymbol{\mu}_r$ 's onto the null subspace  $\boldsymbol{\nu}^\perp$  are the same,  $\boldsymbol{\theta}_{r,1}$  are identical for all the components, which is hence denoted by  $\boldsymbol{\theta}_{\cdot,1}$ . Also denote the first  $q$  dimensions of  $\tilde{\mathbf{z}}$  by  $\tilde{\mathbf{z}}^{(1)}$ , and the last  $p - q$  dimensions by  $\tilde{\mathbf{z}}^{(2)}$ . We can write  $g(\tilde{\mathbf{z}})$  as

$$g(\tilde{\mathbf{Z}} = \tilde{\mathbf{z}}) = \sum_{r=1}^R \pi_r \phi(\tilde{\mathbf{z}}^{(1)} | \boldsymbol{\theta}_{\cdot,1}, \mathbf{I}_q) \phi(\tilde{\mathbf{z}}^{(2)} | \boldsymbol{\theta}_{r,2}, \mathbf{I}_{p-q}) .$$

where  $\mathbf{I}_q$  indicates a  $q \times q$  identity matrix. Since  $g(\tilde{\mathbf{z}})$  is a smooth function, its modes have zero first order derivatives. Note

$$\frac{\partial g(\tilde{\mathbf{z}})}{\partial \tilde{\mathbf{z}}^{(1)}} = \frac{\partial \phi(\tilde{\mathbf{z}}^{(1)} | \boldsymbol{\theta}_{\cdot,1}, \mathbf{I}_q)}{\partial \tilde{\mathbf{z}}^{(1)}} \sum_{r=1}^R \pi_r \phi(\tilde{\mathbf{z}}^{(2)} | \boldsymbol{\theta}_{r,2}, \mathbf{I}_{p-q}) ,$$

$$\frac{\partial g(\tilde{\mathbf{z}})}{\partial \tilde{\mathbf{z}}^{(2)}} = \phi(\tilde{\mathbf{z}}^{(1)} | \boldsymbol{\theta}_{\cdot,1}, \mathbf{I}_q) \sum_{r=1}^R \pi_r \frac{\partial \phi(\tilde{\mathbf{z}}^{(2)} | \boldsymbol{\theta}_{r,2}, \mathbf{I}_{p-q})}{\partial \tilde{\mathbf{z}}^{(2)}} .$$

By setting the first partial derivative to zero and using the fact  $\sum_{r=1}^R \pi_r \phi(\tilde{\mathbf{z}}^{(2)} | \boldsymbol{\theta}_{r,2}, \mathbf{I}_{p-q}) > 0$ , we get

$$\frac{\partial \phi(\tilde{\mathbf{z}}^{(1)} | \boldsymbol{\theta}_{\cdot,1}, \mathbf{I}_q)}{\partial \tilde{\mathbf{z}}^{(1)}} = 0 ,$$

and equivalently

$$\tilde{\mathbf{z}}^{(1)} = \boldsymbol{\theta}_{\cdot,1} , \quad \text{the only mode of a Gaussian density.}$$

For any modes of  $g(\tilde{\mathbf{z}})$ , the first part  $\tilde{\mathbf{z}}^{(1)}$  all equal to  $\boldsymbol{\theta}_{\cdot,1}$ , that is, the projections of the modes onto the null subspace  $\boldsymbol{\nu}^\perp$  coincide at  $\boldsymbol{\theta}_{\cdot,1}$ . Hence the modes and component means lie in the same constrained subspace  $\boldsymbol{\nu}$ .

## Appendix B.

We prove Theorem 2 in Section 3.3 here. Assume  $\boldsymbol{\nu} = \{\mathbf{v}_{q+1}, \dots, \mathbf{v}_p\}$  is the constrained subspace where  $\boldsymbol{\mu}_{kr}$ 's reside, and  $\boldsymbol{\nu}^\perp = \{\mathbf{v}_1, \dots, \mathbf{v}_q\}$  is the corresponding null subspace, as defined in Section 3. We use the Bayes classification rule to classify a sample  $x$ :  $\hat{y} = \operatorname{argmax}_k f(Y = k | \mathbf{X} = \mathbf{x}) = \operatorname{argmax}_k f(\mathbf{X} = \mathbf{x}, Y = k)$ .

$$f(\mathbf{X} = \mathbf{x}, Y = k) = a_k f_k(\mathbf{x}) \propto a_k \sum_{r=1}^{R_k} \pi_{kr} \exp(-(\mathbf{x} - \boldsymbol{\mu}_{kr})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_{kr})) . \quad (11)$$

Let  $\mathbf{V} = \begin{pmatrix} \mathbf{v}_1^t \\ \vdots \\ \mathbf{v}_p^t \end{pmatrix}$ . Matrix  $\mathbf{V}$  is orthonormal because  $\mathbf{v}_j$ 's are orthonormal by construction.

Consider the following cases of  $\boldsymbol{\Sigma}$ .

### B.1 $\Sigma$ is an identity matrix

From Eq. (11), we have

$$\begin{aligned}
 & \sum_{r=1}^{R_k} \pi_{kr} \exp(-(\mathbf{x} - \boldsymbol{\mu}_{kr})^t \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}_{kr})) \\
 = & \sum_{r=1}^{R_k} \pi_{kr} \exp(-(\mathbf{x} - \boldsymbol{\mu}_{kr})^t (\mathbf{V}^t \mathbf{V}) (\mathbf{x} - \boldsymbol{\mu}_{kr})) \\
 = & \sum_{r=1}^{R_k} \pi_{kr} \exp(-(\mathbf{V} \mathbf{x} - \mathbf{V} \boldsymbol{\mu}_{kr})^t (\mathbf{V} \mathbf{x} - \mathbf{V} \boldsymbol{\mu}_{kr})) \\
 = & \sum_{r=1}^{R_k} \pi_{kr} \exp(-\sum_{j=1}^p (\check{x}_j - \check{\mu}_{kr,j})^2), \tag{12}
 \end{aligned}$$

where  $\check{x}_j = \mathbf{v}_j^t \cdot \mathbf{x}$ ,  $\check{\mu}_{kr,j} = \mathbf{v}_j^t \cdot \boldsymbol{\mu}_{kr}$ ,  $j = 1, 2, \dots, p$ . Because  $\check{\mu}_{kr,j} = c_j$ , identical across all  $k$  and  $r$  for  $j = 1, \dots, q$ , the first  $q$  terms in the sum of exponent in Eq. (12) are all constants. We have

$$\begin{aligned}
 & \sum_{r=1}^{R_k} \pi_{kr} \exp(-\sum_{j=1}^p (\check{x}_j - \check{\mu}_{kr,j})^2) \\
 \propto & \sum_{r=1}^{R_k} \pi_{kr} \exp(-\sum_{j=q+1}^p (\check{x}_j - \check{\mu}_{kr,j})^2).
 \end{aligned}$$

Therefore,

$$f(\mathbf{X} = \mathbf{x}, Y = k) \propto a_k \sum_{r=1}^{R_k} \pi_{kr} \exp(-\sum_{j=q+1}^p (\check{x}_j - \check{\mu}_{kr,j})^2).$$

That is, to classify a sample  $\mathbf{x}$ , we only need the projection of  $\mathbf{x}$  onto the constrained subspace  $\boldsymbol{\nu}^\perp = \{\mathbf{v}_1, \dots, \mathbf{v}_q\}$ .

### B.2 $\Sigma$ is a non-identity matrix

We can perform a linear transform (a “whitening” operation) on  $\mathbf{X}$  so that the transformed data have an identity covariance matrix  $\mathbf{I}$ . Find the matrix square root of  $\Sigma$ , that is,  $\Sigma = (\Sigma^{\frac{1}{2}})^t \Sigma^{\frac{1}{2}}$ . If the eigen decomposition of  $\Sigma$  is  $\Sigma = V_\Sigma D_\Sigma V_\Sigma^t$ , then  $\Sigma^{\frac{1}{2}} = D_\Sigma^{\frac{1}{2}} V_\Sigma^t$ . Let  $\mathbf{Z} = (\Sigma^{-\frac{1}{2}})^t \mathbf{X}$ . The distribution of  $\mathbf{Z}$  is

$$g(\mathbf{Z} = \mathbf{z}, Y = k) = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(\mathbf{z} | \tilde{\boldsymbol{\mu}}_{kr}, \mathbf{I}),$$

where  $\tilde{\boldsymbol{\mu}}_{kr} = (\Sigma^{-\frac{1}{2}})^t \boldsymbol{\mu}_{kr}$ . According to our assumption,  $\mathbf{v}_j^t \cdot \boldsymbol{\mu}_{kr} = c_j$ , i.e., identical across all  $k$  and  $r$  for  $j = 1, \dots, q$ . Plugging into  $\boldsymbol{\mu}_{kr} = (\Sigma^{\frac{1}{2}})^t \tilde{\boldsymbol{\mu}}_{kr}$ , we get  $(\Sigma^{\frac{1}{2}} \mathbf{v}_j)^t \cdot \tilde{\boldsymbol{\mu}}_{kr} = c_j$ ,

$j = 1, \dots, q$ . This means for the transformed data, the component means  $\tilde{\boldsymbol{\mu}}_{kr}$ 's have a null space spanned by  $\{\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{v}_j | j = 1, \dots, q\}$ . Correspondingly, the constrained subspace is spanned by  $\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\}$ . It is easy to verify that the new null space and constrained subspace are orthogonal, since  $(\boldsymbol{\Sigma}^{\frac{1}{2}}\mathbf{v}_j)^t \cdot (\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_{j'} = \mathbf{v}_j^t \cdot \mathbf{v}_{j'} = 0$ ,  $j = 1, \dots, q$  and  $j' = q+1, \dots, p$ . The spanning vectors for the constrained subspace,  $(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j$ ,  $j = q+1, \dots, p$ , are not orthonormal in general, but there exists an orthonormal basis that spans the same subspace. With a slight abuse of notation, we use  $\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\}$  to denote a  $p \times (p-q)$  matrix containing the column vector  $(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j$ . For any matrix  $A$  of dimension  $p \times d$ ,  $d < p$ , let the notation  $orth(A)$  denote a  $p \times d$  matrix whose column vectors are orthonormal and span the same subspace as the column vectors of  $A$ . According to B.1, for the transformed data  $\mathbf{Z}$ , we only need the projection of  $\mathbf{Z}$  onto a subspace spanned by the column vectors of  $orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\})$  to compute the class posterior. Note that  $\mathbf{Z} = (\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{X}$ . So the subspace that matters for classification for the original data  $\mathbf{X}$  is spanned by the column vectors of  $(\boldsymbol{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\})$ . Again, these column vectors are not orthonormal in general, but there exists an orthonormal basis that spans the same subspace. This orthonormal basis is hence spanned by the column vectors of  $orth((\boldsymbol{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\}))$ . Since  $orth((\boldsymbol{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \mathbf{v}_j | j = q+1, \dots, p\})) = orth(\{\boldsymbol{\Sigma}^{-1} \mathbf{v}_j | j = q+1, \dots, p\})$ ,<sup>5</sup> the subspace that matters for classification is thus spanned by the column vectors of  $orth(\{\boldsymbol{\Sigma}^{-1} \mathbf{v}_j | j = q+1, \dots, p\})$ .

In summary, only the linear projection of the data onto a subspace with the same dimension as  $\boldsymbol{\nu}$  matters for classification.

### Appendix C. Derivation of $\mu_{kr}$ in GEM

We derive the optimal  $\boldsymbol{\mu}_{kr}$ 's under constraint (4) for a given  $\boldsymbol{\Sigma}$ . Note that the term in Eq. (6) that involves  $\boldsymbol{\mu}_{kr}$ 's is:

$$-\frac{1}{2} \sum_{k=1}^K \sum_{r=1}^{R_k} \sum_{i=1}^{n_k} q_{i,kr} (\mathbf{x}_i - \boldsymbol{\mu}_{kr})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}_{kr}). \quad (13)$$

Denote  $\sum_{i=1}^{n_k} q_{i,kr}$  by  $l_{kr}$ . Let  $\bar{\mathbf{x}}_{kr} = \sum_{i=1}^{n_k} q_{i,kr} \mathbf{x}_i / l_{kr}$ , i.e., the weighted sample mean of the component  $r$  in class  $k$ . To maximize Eq. (13) is equivalent to minimizing the following term (Anderson, 2000):

$$\sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr})^t \boldsymbol{\Sigma}^{-1} (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr}). \quad (14)$$

To solve the above optimization problem under constraint (4), we need to find a linear transform such that in the transformed space, the constraint is imposed on individual coordinates (rather than linear combinations of them), and the objective function is a weighted sum of squared Euclidean distances between the transformed  $\bar{\mathbf{x}}_{kr}$  and  $\boldsymbol{\mu}_{kr}$ . Once this is achieved, the optimal solution will simply be given by setting those unconstrained coordinates within

5. Let matrix  $A$  be a  $p \times p$  square matrix and  $B$  be a  $p \times d$  matrix,  $d < p$ , it can be proved that  $orth(A \times orth(B)) = orth(A \times B)$ .

each component by the component-wise sample mean, and the constrained coordinates by the component-pooled sample mean. We will discuss the detailed solution in the following.

Find the matrix square root of  $\Sigma$ , that is,  $\Sigma = (\Sigma^{\frac{1}{2}})^t \Sigma^{\frac{1}{2}}$ . If the eigen decomposition of  $\Sigma$  is  $\Sigma = V_\Sigma D_\Sigma V_\Sigma^t$ , then,  $\Sigma^{\frac{1}{2}} = D_\Sigma^{\frac{1}{2}} V_\Sigma^t$ . Now perform the following change of variables:

$$\begin{aligned}
 & \sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr})^t \Sigma^{-1} (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr}) \\
 &= \sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} \left[ \left( \Sigma^{-\frac{1}{2}} \right)^t (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr}) \right]^t \left[ \left( \Sigma^{-\frac{1}{2}} \right)^t (\bar{\mathbf{x}}_{kr} - \boldsymbol{\mu}_{kr}) \right] \\
 &= \sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\tilde{\mathbf{x}}_{kr} - \tilde{\boldsymbol{\mu}}_{kr})^t (\tilde{\mathbf{x}}_{kr} - \tilde{\boldsymbol{\mu}}_{kr}), \tag{15}
 \end{aligned}$$

where  $\tilde{\boldsymbol{\mu}}_{kr} = \left( \Sigma^{-\frac{1}{2}} \right)^t \cdot \boldsymbol{\mu}_{kr}$ , and  $\tilde{\mathbf{x}}_{kr} = \left( \Sigma^{-\frac{1}{2}} \right)^t \cdot \bar{\mathbf{x}}_{kr}$ . Correspondingly, the constraint in (4) becomes

$$\left( \Sigma^{\frac{1}{2}} \mathbf{v}_j \right)^t \tilde{\boldsymbol{\mu}}_{kr} = \text{constant over } r \text{ and } k, \quad j = 1, \dots, q. \tag{16}$$

Let  $\mathbf{b}_j = \Sigma^{\frac{1}{2}} \mathbf{v}_j$  and  $\mathbf{B} = (\mathbf{b}_1, \mathbf{b}_2, \dots, \mathbf{b}_q)$ . Note that the rank of  $\mathbf{V} = (\mathbf{v}_1, \dots, \mathbf{v}_q)$  is  $q$ . Since  $\Sigma^{\frac{1}{2}}$  is of full rank,  $\mathbf{B} = \Sigma^{\frac{1}{2}} \mathbf{V}$  also has rank  $q$ . The constraint in (16) becomes

$$\mathbf{B}^t \tilde{\boldsymbol{\mu}}_{kr} = \mathbf{B}^t \tilde{\boldsymbol{\mu}}_{k'r'}, \text{ for any } r, r' = 1, \dots, R_k, \text{ and any } k, k' = 1, \dots, K. \tag{17}$$

Now perform a singular value decomposition (SVD) on  $\mathbf{B}$ , i.e.,  $\mathbf{B} = \mathbf{U}_B \mathbf{D}_B \mathbf{V}_B^t$ , where  $\mathbf{V}_B$  is a  $q \times q$  orthonormal matrix,  $\mathbf{D}_B$  is a  $q \times q$  diagonal matrix, which is non-singular since the rank of  $\mathbf{B}$  is  $q$ , and  $\mathbf{U}_B$  is a  $p \times q$  orthonormal matrix. Substituting the SVD of  $\mathbf{B}$  in (17), we get

$$\mathbf{V}_B \mathbf{D}_B \mathbf{U}_B^t \tilde{\boldsymbol{\mu}}_{kr} = \mathbf{V}_B \mathbf{D}_B \mathbf{U}_B^t \tilde{\boldsymbol{\mu}}_{k'r'}, \quad \text{for any } r, r' = 1, \dots, R_k, \text{ and any } k, k' = 1, \dots, K,$$

which is equivalent to

$$\mathbf{U}_B^t \tilde{\boldsymbol{\mu}}_{kr} = \mathbf{U}_B^t \tilde{\boldsymbol{\mu}}_{k'r'}, \quad \text{for any } r, r' = 1, \dots, R_k, \text{ and any } k, k' = 1, \dots, K, \tag{18}$$

because  $\mathbf{V}_B$  and  $\mathbf{D}_B$  have full rank. We can augment  $\mathbf{U}_B$  to a  $p \times p$  orthonormal matrix,  $\hat{\mathbf{U}} = (\mathbf{u}_1, \dots, \mathbf{u}_q, \mathbf{u}_{q+1}, \dots, \mathbf{u}_p)$ , where  $\mathbf{u}_{q+1}, \dots, \mathbf{u}_p$  are augmented orthonormal vectors. Since  $\hat{\mathbf{U}}$  is orthonormal, the objective function in Eq. (15) can be written as

$$\sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} [\hat{\mathbf{U}}^t (\tilde{\mathbf{x}}_{kr} - \tilde{\boldsymbol{\mu}}_j)]^t \cdot [\hat{\mathbf{U}}^t (\tilde{\mathbf{x}}_{kr} - \tilde{\boldsymbol{\mu}}_{kr})] = \sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\check{\mathbf{x}}_{kr} - \check{\boldsymbol{\mu}}_{kr})^t (\check{\mathbf{x}}_{kr} - \check{\boldsymbol{\mu}}_{kr}), \tag{19}$$

where  $\check{\mathbf{x}}_{kr} = \hat{\mathbf{U}}^t \tilde{\mathbf{x}}_{kr}$  and  $\check{\boldsymbol{\mu}}_{kr} = \hat{\mathbf{U}}^t \tilde{\boldsymbol{\mu}}_{kr}$ . If we denote  $\check{\boldsymbol{\mu}}_{kr} = (\check{\mu}_{kr,1}, \check{\mu}_{kr,2}, \dots, \check{\mu}_{kr,p})^t$ , then the constraint in (18) simply becomes

$$\check{\mu}_{kr,j} = \check{\mu}_{k'r',j}, \text{ for any } r, r' = 1, \dots, R_k, \text{ and any } k, k' = 1, \dots, K, j = 1, \dots, q.$$

That is, the first  $q$  coordinates of  $\check{\boldsymbol{\mu}}$  have to be common over all the  $k$  and  $r$ . The objective function (19) can be separated coordinate wise:

$$\sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\check{\boldsymbol{x}}_j - \check{\boldsymbol{\mu}}_{kr})^t (\check{\boldsymbol{x}}_{kr} - \check{\boldsymbol{\mu}}_{kr}) = \sum_{j=1}^p \sum_{k=1}^K \sum_{r=1}^{R_k} l_{kr} (\check{x}_{kr,j} - \check{\mu}_{kr,j})^2 .$$

For the first  $q$  coordinates, the optimal  $\check{\mu}_{kr,j}$ ,  $j = 1, \dots, q$ , is solved by

$$\check{\mu}_{kr,j}^* = \frac{\sum_{k'=1}^K \sum_{r'=1}^{R_{k'}} l_{k'r'} \check{x}_{k'r',j}}{\sum_{k'=1}^K \sum_{r'=1}^{R_{k'}} l_{k'r'}} = \frac{\sum_{k'=1}^K \sum_{r'=1}^{R_{k'}} l_{k'r'} \check{x}_{k'r',j}}{n} , \quad \text{identical over } r \text{ and } k .$$

For the remaining coordinates,  $\check{\mu}_{kr,j}$ ,  $j = q + 1, \dots, p$ :

$$\check{\mu}_{kr,j}^* = \check{x}_{kr,j} .$$

After  $\check{\boldsymbol{\mu}}_{kr}^*$  is calculated, we finally get  $\boldsymbol{\mu}_{kr}$ 's under the constraint(4):

$$\boldsymbol{\mu}_{kr} = (\boldsymbol{\Sigma}^{\frac{1}{2}})^t \hat{\boldsymbol{U}} \check{\boldsymbol{\mu}}_{kr}^* .$$

## Appendix D. Reduced Rank Mixture Discriminant Analysis

The rank restriction can be incorporated into the mixture discriminant analysis (MDA). It is known that the rank- $L$  LDA fit is equivalent to a Gaussian maximum likelihood solution, where the means of Gaussians lie in a  $L$ -dimension subspace (Hastie and Tibshirani, 1996). Similarly, in MDA, the log-likelihood can be maximized with the restriction that all the  $R = \sum_{k=1}^K R_k$  centroids are confined to a rank- $L$  subspace, i.e.,  $\text{rank} \{ \boldsymbol{\mu}_{kr} \} = L$ .

The EM algorithm is used to estimate the parameters of the reduced rank MDA, and the M-step is a weighted version of LDA, with  $R$  ‘‘classes’’. The component posterior probabilities  $q_{i,kr}$ 's in the E-step are calculated in the same way as in Eq. (5), which are conditional on the current (reduced rank) version of component means and common covariance matrix. In the M-step,  $\pi_{kr}$ 's are still maximized using Eq. (7). The maximizations of  $\boldsymbol{\mu}_{kr}$  and  $\boldsymbol{\Sigma}$  can be viewed as weighted mean and pooled covariance maximum likelihood estimates in a weighted and augmented  $R$ -class problem. Specifically, we augment the data by replicating the  $n_k$  observations in class  $k$   $R_k$  times, with the  $l$ th such replication having the observation weight  $q_{i,kl}$ . This is done for each of the  $K$  classes, resulting in an augmented and weighted training set of  $\sum_{k=1}^K n_k R_k$  observations. Note that the sum of all the weights is  $n$ . We now impose the rank restriction. For all the sample points  $\boldsymbol{x}_i$ 's within class  $k$ , the weighted component mean is

$$\boldsymbol{\mu}_{kr} = \frac{\sum_{i=1}^{n_k} q_{i,kr} \boldsymbol{x}_i}{\sum_{i=1}^{n_k} q_{i,kr}} .$$

Let  $q'_{kr} = \sum_{i=1}^{n_k} q_{i,kr}$ . The overall mean is

$$\boldsymbol{\mu} = \frac{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr} \boldsymbol{\mu}_{kr}}{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr}} .$$

The pooled within-class covariance matrix is

$$W = \frac{\sum_{k=1}^K \sum_{r=1}^{R_k} \sum_{i=1}^{n_k} q_{i,kr} (\mathbf{x}_i - \boldsymbol{\mu}_{kr})^t (\mathbf{x}_i - \boldsymbol{\mu}_{kr})}{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr}}.$$

The between-class covariance matrix is

$$B = \frac{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr} (\boldsymbol{\mu}_{kr} - \boldsymbol{\mu})^t (\boldsymbol{\mu}_{kr} - \boldsymbol{\mu})}{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr}}.$$

Define  $B^* = (W^{-\frac{1}{2}})^T B W^{-\frac{1}{2}}$ . Now perform an eigen-decomposition on  $B^*$ , i.e.,  $B^* = V^* D_B V^{*T}$ , where  $V^* = (v_1^*, v_2^*, \dots, v_p^*)$ . Let  $V$  be a matrix consisting of the leading  $L$  columns of  $W^{-\frac{1}{2}} V^*$ . Considering maximizing the Gaussian log-likelihood subject to the constraints  $\text{rank} \{\boldsymbol{\mu}_{kr}\} = L$ , the solutions are

$$\hat{\boldsymbol{\mu}}_{kr} = W V V^T (\boldsymbol{\mu}_{kr} - \boldsymbol{\mu}) + \boldsymbol{\mu}, \quad (20)$$

$$\hat{\boldsymbol{\Sigma}} = W + \frac{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr} (\boldsymbol{\mu}_{kr} - \hat{\boldsymbol{\mu}}_{kr})^t (\boldsymbol{\mu}_{kr} - \hat{\boldsymbol{\mu}}_{kr})}{\sum_{k=1}^K \sum_{r=1}^{R_k} q'_{kr}}. \quad (21)$$

As a summary, in the M-step of reduced rank MDA, the parameters,  $\pi_{kr}$ ,  $\boldsymbol{\mu}_{kr}$  and  $\boldsymbol{\Sigma}$ , are maximized by Eqs. (7), (20), and (21), respectively.

Note that the discriminant subspace is spanned by the column vectors of  $V = W^{-\frac{1}{2}} V^*$ , with the  $l$ th discriminant variable as  $W^{-\frac{1}{2}} v_l^*$ . In general,  $W^{-\frac{1}{2}} v_l^*$ 's are not orthogonal, but we can find an orthonormal basis that spans the same subspace.

## References

- T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley, 2000.
- L. Breiman and R. Ihaka. Nonlinear discriminant analysis via scaling and ACE. Technical Report 40, Department of Statistics, University of California, Berkeley, California, 1984.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1-21, 1977.
- R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7:179-188, 1936.
- C. Fraley and A. E. Raftery. MCLUST Version 3 for R: Normal Mixture Modeling and Model-Based Clustering. Technical Report 504, Department of Statistics, University of Washington, Washington, 2006.
- C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97:611-631, 2002.

- A. S. Georgiades, P. N. Belhumeur and D. J. Kriegman. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):643-660, 2001.
- T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, 2001.
- T. Hastie and R. Tibshirani. Discriminant Analysis by Gaussian Mixtures. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):155-176, 1996.
- X. He, S. Yan, Y. Hu, P. Niyogi and H. Zhang. Face Recognition Using Laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3):328-340, 2005.
- K. C. Lee, J. Ho, D. Kriegman. Acquiring Linear Subspaces for Face Recognition under Variable Lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684-698, 2005.
- J. Li, S. Ray, B. G. Lindsay. A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8(8):1687-1723, 2007.
- J. Li and H. Zha. Two-way Poisson mixture models for simultaneous document classification and word clustering. *Computational Statistics and Data Analysis*, 50:163-180, 2006.
- L. Lovász and M. D. Plummer. *Matching Theory*. Akadémiai Kiadó - North Holland, Budapest, 1986.
- G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2000.
- W. Pan and X. Shen. Penalized Model-Based Clustering with Application to Variable Selection. *Journal of Machine Learning Research*, 8:1145-1164, 2007.
- M. Qiao, J. Li. Two-way Gaussian Mixture Models for High Dimensional Classification. *Statistical Analysis and Data Mining*, 3(4):259-271, 2010.
- S. Wang and J. Zhu. Variable Selection for Model-Based High-Dimensional Clustering and Its Application to Microarray Data. *Biometrics*, 64:440-448, 2008.