The Pennsylvania State University
The Graduate School
Eberly College of Science

VISUAL ANALYTICS THROUGH GAUSSIAN MIXTURE MODELS WITH SUBSPACE CONSTRAINED COMPONENT MEANS

A Thesis in
Statistics
by
Mu Qiao

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

December 2017

The thesis of Mu Qiao was reviewed and approved* by the following:

Jia Li
Professor of Statistics and (by courtesy) Computer Science and Engineering
Thesis Advisor

Lynn Lin
Assistant Professor of Statistics

David Hunter
Professor of Statistics
Head of the Department of Statistics

*Signatures are on file in the Graduate School.

Abstract

We develop a new method for high dimensional data visualization via the Gaussian mixture model (GMM) with the component means constrained in a pre-selected subspace. 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 class means and the modes of the density. This motivates us to find a subspace by applying weighted principal component analysis to the class means and the modes. A dimension reduction property is proved in the sense of being informative for classification or clustering. Experiments on real data sets indicate that our method with the simple technique of spanning the subspace only by class means often outperforms the reduced rank mixture discriminant analysis (MDA) when the subspace dimension is very low. Visualization results on independent test data show that our proposed method exhibits more distinct class-wise separation of high dimensional data in $2d$ or $3d$ subspaces in comparison with reduced rank MDA.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my advisor Dr. Jia Li for her great guidance, encouragement, and support. This thesis would not have been completed without her help. Dr. Li led me into the statistical learning research field and taught me many essential research skills, for which I am very grateful. I would also like to thank Dr. Lynn Lin for spending time reading this thesis and providing valuable suggestions.

Part of this thesis work was supported by the NSF grant CCF-0936948. The content is solely the responsibility of the author and does not necessarily represent the official views of the NSF.

Chapter 1 Introduction

1.1 High Dimensional Data Visualization

High dimensional data is prevalent in the big data era. We are surrounded by all types of data with high dimensional representations, such as images, videos, audios, and textual documents. Visualization plays an important role in making sense of these data and exploring their underlying patterns. A lot of visualization methods have been proposed in the past decades. Dimension reduction is one of the most widely used techniques for visualizing high dimensional data in lower dimensional spaces. Dimension reduction can be roughly divided into two categories: feature transformation and feature selection. Many classical feature transformation methods, such as principal component analysis (PCA) (Pearson, 1901), multidimensional scaling (MDS) (Young, 1987), and linear discriminant analysis (LDA) (Fisher, 1936), project high dimensional data onto low dimensional spaces. These methods create linear combinations of the original dimensions and attempt to uncover the latent structure. Feature selection methods, on the other hand, do not intend to create any new dimensions, but rather select a subset of informative original dimensions, discarding irrelevant or redundant dimensions. Subspace clustering, as an extension of feature selection, seeks to identify the clustering structure of high dimensional data in different subspaces. For example, Baumgartner et al. (2004) proposed the SURFING algorithm to find and rank subspaces interesting for clustering. Tatu et al. (2012) developed a pipeline to efficiently explore large sets of subspace candidates and allow users to compare and relate subspaces by comparing their topological and dimensional similarities. Most recently, feature transformation has also been used in subspace clustering. Anand et al. (2012) applied the random projection method to find interesting low dimensional substructures in high dimensional data. A data-driven approach was proposed by Lehmann and Theisel (2016) to discover important data patterns through a minimal number of projections. Elhamifar and Vidal (2013) proposed a subspace clustering algorithm using sparse representation techniques to cluster a collection of data points lying in or close to a union of low-dimensional linear subspaces. Liu et al. (2015) developed a graph embedding approach based on spectral clustering to discriminate the different subspaces and improve the resilience to outliers.

1.2 Subspace Constrained Models for Visualization

Linear discriminant analysis (LDA), as one of the most well-known dimension reduction methods, has gained wide popularity in visualization. LDA projects the high dimensional data onto a low dimensional linear subspace, where the projected centroids are spread out as much as possible in terms of variance. This amounts 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 (Hastie and Tibshirani, 1996). 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. LDA has been applied in many visual analytical works. For example, Choo et al. (2010) developed iVisClassifier, a visual analytics system for classification, which allows users to interact with all the reduced dimensions obtained by LDA through parallel coordinates and a scatter plot. Zhou et al. (2016) introduced LDA into their visual analytical pipeline to help users construct subspaces which display informative cluster structures.

LDA is essentially a restricted Gaussian classifier. Given a high dimensional data, suppose each class density is modeled as a multivariate Gaussian distribution, LDA is the special case when we assume that the classes share common covariance matrix. The Gaussian mixture model (GMM) is a popular and effective tool for clustering and classification. When applied to clustering high dimensional data, usually each cluster is modeled by a multivariate Gaussian distribution. 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 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 μ_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. Reduced rank MDA allows the visualization of high dimensional data in two-dimensional or three-dimensional subspaces when the reduced rank is extremely low. While reduced rank LDA has been widely applied in visual analytics, we have not seen much work in visualization using reduced rank MDA, which shows superior classification performance in low dimensional spaces.

1.3 New Method

In this thesis, we propose a new visualization method by regularizing the component means of GMM in low dimensional subspaces, 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 which best preserves the classification and clustering characteristics.

Another related line of research is regularizing the component means of a GMM in a latent factor space, i.e., the use of factor analysis in GMM. It was originally proposed by Ghahramani and Hinton (1997) to perform concurrent clustering and dimension reduction using mixture of factor analyzers (see also McLachlan and Peel (2000b) and McLachlan et al. (2003)). Factor analysis was later used to regularize the component means of a GMM in each state of the Hidden Markov Model (HMM) for speaker verification (P. Kenny et al., 2008; D. Povey et al., 2011). In those models, the total number of parameters is significantly reduced due to the regularization, which effectively prevents over fitting. Usually the EM type of algorithm is applied to estimate the parameters and find the latent subspace.

The role of the subspace constraining the means differs intrinsically between our approach, the reduced rank MDA and factor analysis based mixture models, resulting in mathematical solutions of quite different nature. 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. The subspace depends on the component-based clustering of data in each iteration. Similarly, the subspaces in factor analysis based mixture models are found through the iterations of the EM algorithm, as part of the model estimation. However, 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. 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 find the unknown constrained subspace. It is easy to see that the class means reside in the same constrained subspace as the component means. Therefore, in the first approach, we generate the constrained subspace using class means. Traditional methods, such as reduced rank MDA, use a rather sophisticated approach to find the subspace that best preserves the classification structure of high dimensional data in lower dimensional subspace. In contrast, our proposed method with the simple approach of finding the subspace based on class means shows superior performance. We also prove that the modes (i.e., local maxima) of the mixture probability density lie in the same constrained subspace as the component means. As a second approach, the subspace is therefore generated using modes. Specifically, the *modal EM*

(MEM) algorithm (Li et al., 2007) is applied to find the modes. Note that, in our method, each GMM is a full model for the original data, although the component means can be constrained in different subspaces. We therefore can compare the constrained subspaces by the estimated likelihoods under each model.

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

Chapter 2
Preliminaries and Notation

2.1 Gaussian Mixture Model

Let $\boldsymbol{X} = (X_1, X_2, ..., X_p)^t$, where p is the dimension of the data. A sample of $\boldsymbol{X}$ is denoted by $\boldsymbol{x} = (x_1, x_2, ..., x_p)^t$. 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 $\boldsymbol{X}$ be $Y \in \mathcal{K} = \{1, 2, ..., K\}$. For classification purpose, the joint distribution of $\boldsymbol{X}$ and Y under a Gaussian mixture is

$$f(\boldsymbol{X} = \boldsymbol{x}, Y = k) = a_k f_k(\boldsymbol{x}) = a_k \sum_{r=1}^{R_k} \pi_{kr} \phi(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \,, \tag{2.1}$$

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 $f_k(\boldsymbol{x})$ is the within-class density for $\boldsymbol{X}$. 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 π_{kr} be the mixing proportions for the rth 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 $\boldsymbol{X} = \boldsymbol{x}$, the Bayes classification rule is used: $\widehat{y} = \text{argmax}_k f(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \text{argmax}_k f(\boldsymbol{X} = \boldsymbol{x}, Y = k)$.

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

$$f(\boldsymbol{X} = \boldsymbol{x}) = \sum_{r=1}^{R} \pi_r \phi(\boldsymbol{x}|\boldsymbol{\mu}_r, \boldsymbol{\Sigma}) \,, \tag{2.2}$$

where R is the total number of mixture components and π_r is the mixing proportions for the rth component. $\boldsymbol{\mu}_r$ and $\boldsymbol{\Sigma}$ denote the rth 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. We propose a new algorithm to find the constrained subspace. The motivations of using modes to find subspace are outlined in Chapter 3.2. Before we present the new algorithm, we will first introduce the modal EM algorithm (Li et al., 2007) which solves the local maxima, that is, modes, of a mixture density.

2.2 Modal EM

Given a mixture density $f(\boldsymbol{X} = \boldsymbol{x}) = \sum_{r=1}^{R} \pi_r f_r(\boldsymbol{x})$, as in model (2.2), starting from any initial data point $\boldsymbol{x}^{(0)}$, 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 $p_r = \frac{\pi_r f_r(\boldsymbol{x}^{(t)})}{f(\boldsymbol{x}^{(t)})}$, $r = 1, ..., R$.
2. Update $\boldsymbol{x}^{(t+1)} = \text{argmax}_{\boldsymbol{x}} \sum_{r=1}^{R} p_r \log f_r(\boldsymbol{x})$.

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 $\boldsymbol{x}^{(t)}$ is computed. The second step is the "maximization" step. $\sum_{r=1}^{R} p_r \log f_r(\boldsymbol{x})$ has a unique maximum, if the $f_r(\boldsymbol{x})$'s are normal densities. In the special case of a mixture of Gaussians with common covariance matrix, that is, $f_r(\boldsymbol{x}) = \phi(\boldsymbol{x} \mid \boldsymbol{\mu}_r, \boldsymbol{\Sigma})$, we simply have $\boldsymbol{x}^{(t+1)} = \sum_{r=1}^{R} p_r \boldsymbol{\mu}_r$. The data points are grouped into one cluster if they climb to the same mode. We call the mode as the cluster representative.

We use the *Mclust* package in R (Fraley et al., 2012) as the model density estimator. Mclust produces a density estimate for each data point using a Gaussian finite mixture model. The modal EM algorithm is then applied to find the modes of the density by alternating the above two steps.

2.3 Reduced Rank Mixture Discriminant Analysis

Since our method is along the line of reduced rank mixture discriminant analysis (MDA), we also briefly introduce it in this section. 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 A, 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 A.

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 containing a set of functions to fit MDA, FDA, and PDA by multiple regressions is provided by Hastie and Tibshirani (1996).

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 this work, 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 and the implementation via optimal scoring from the R package provided by Hastie and Tibshirani (1996) are used.

Chapter 3
GMM with Means Constrained in Subspace

The Gaussian mixture model with subspace constrained means is presented in this chapter. 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:

$$f_k(\boldsymbol{x}) = \sum_{r=1}^{R_k} \pi_{kr} \phi(\boldsymbol{x}|\boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \tag{3.1}$$

$$\boldsymbol{v}_j^t \cdot \boldsymbol{\mu}_{k1} = \boldsymbol{v}_j^t \cdot \boldsymbol{\mu}_{k2} = \cdots = \boldsymbol{v}_j^t \cdot \boldsymbol{\mu}_{kR_k} = c_j \ , \tag{3.2}$$

where $\boldsymbol{v}_j$'s are linearly independent vectors, $j = 1, ..., q$, $q < p$, and c_j is a constant, invariant to different classes. Without loss of generality, we can assume $\{\boldsymbol{v}_1, ..., \boldsymbol{v}_q\}$ span an orthonormal basis. Augment it to full rank by $\{\boldsymbol{v}_{q+1}, ..., \boldsymbol{v}_p\}$. Suppose $\boldsymbol{\nu} = \{\boldsymbol{v}_{q+1}, ..., \boldsymbol{v}_p\}$, $\boldsymbol{\nu}^\perp = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_q\}$, and $\boldsymbol{c} = (c_1, c_2, ..., c_q)^t$. 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 $\mathbf{Proj}_{\boldsymbol{\nu}^\perp}^{\boldsymbol{\mu}_{kr}} = \boldsymbol{c}$ 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 $\boldsymbol{c}$. 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 (3.2) 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 Class Means Spanned Subspace

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 rth component in class k. It is easy to see that the class means lie in the same subspace as the Gaussian mixture component means. A weighted principal component analysis is proposed to find the constrained subspace using class means.

Suppose the set of class means is $\mathcal{J} = \{\mathcal{M}_1, \mathcal{M}_2, \cdots, \mathcal{M}_K\}$. We first assign a weight a_k to the kth 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. 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 $\mu_{\mathcal{J}} = \sum_{r=1}^{K} a_k \mathcal{M}_k$.

The principal components are then obtained by performing an eigenvalue decomposition on $\Sigma_{\mathcal{J}}$. Recall that 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 $\boldsymbol{\nu}^{\perp}$. We refer to this method as MEAN-PCA.

3.2 Modes Spanned Subspace

We propose another approach to generate the constrained subspace using modes. We prove in Appendix B the following theorem.

Theorem 3.2.1. *For a Gaussian mixture model with component means constrained in a subspace $\boldsymbol{\nu} = \{\boldsymbol{v}_{q+1}, ..., \boldsymbol{v}_p\}$, $q < p$, 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 3.2.1, the modes and component means of Gaussian mixtures reside in the same constrained subspace. We apply the aforementioned MEM algorithm introduced in Chapter 2.2 to find the modes of 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 by $\mathcal{G} = \{\mathcal{M}'_1, \mathcal{M}'_2, ..., \mathcal{M}'_{|\mathcal{G}|}\}$. A weight w_r is assigned to the rth 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 $\mathcal{G}$:

$$\Sigma_{\mathcal{G}} = \sum_{r=1}^{|\mathcal{G}|} w_r (\mathcal{M}'_r - \mu_{\mathcal{G}})^T (\mathcal{M}'_r - \mu_{\mathcal{G}}) \ ,$$

where $\mu_{\mathcal{G}} = \sum_{r=1}^{|\mathcal{G}|} w_r \mathcal{M}'_r$.

An eigenvalue decomposition on $\Sigma_{\mathcal{G}}$ is then performed to obtain all the principal components. The constrained subspace is spanned by the first d most significant principal components. We refer to this method as MODE-PCA.

3.3 Dimension Reduction

The mixture model with component means under constraint (3.2) implies a dimension reduction property for the classification purpose, formally stated below.

Theorem 3.3.1. *For a Gaussian mixture model with a common covariance matrix $\mathbf{\Sigma}$, suppose all the component mean $\boldsymbol{\mu}_{kr}$'s are constrained in a subspace spanned by $\boldsymbol{\nu} = \{\boldsymbol{v}_{q+1}, ..., \boldsymbol{v}_p\}$, $q < p$, up to a translation, only a linear projection of the data $\boldsymbol{x}$ onto a subspace spanned by $\{\mathbf{\Sigma}^{-1}\boldsymbol{v}_j | j = q+1, ..., p\}$ (the same dimension as $\boldsymbol{\nu}$) is informative for classification.*

In Appendix C, we provide the detailed proof for Theorem 3.3.1. If the common covariance matrix $\mathbf{\Sigma}$ is an identity matrix (or a scalar matrix), the class label Y only depends on the projection of $\boldsymbol{x}$ onto the constrained subspace $\boldsymbol{\nu}$. However, in general, $\mathbf{\Sigma}$ is non-identity. Hence the spanning vectors, $\mathbf{\Sigma}^{-1}\boldsymbol{v}_j$, $j = q+1, ..., p$, for the subspace informative for classification are not orthogonal in general as well. In Appendix C, we use the column vectors of $orth(\{\mathbf{\Sigma}^{-1}\boldsymbol{v}_j | j = q+1, ..., p\})$ to span this subspace. To differentiate it from the constrained subspace in which the component means lie, we call it as *discriminant subspace*. The dimension of the discriminant subspace is referred to 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 C we prove Theorem 3.3.1 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 $\boldsymbol{\nu}$ to compute the posterior probability of the data belonging to a component (aka cluster).

3.4 Model Estimation

We describe the estimation algorithm in this section, where the constraints on the component means are characterized by (3.2). We 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 $\mathbf{\Sigma}$ 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 π_{kr}, and the Gaussian parameters $\boldsymbol{\mu}_{kr}$, $\mathbf{\Sigma}$, $r = 1, 2, ..., R_k$. Denote the training data by $\{(\boldsymbol{x}_i, y_i) : i = 1, ..., n\}$. 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. *Expectation-step*: Given the current parameters, for each class k, compute the component posteriori probability for each data point $\boldsymbol{x}_i$ within class k:

$$q_{i,kr} \propto \pi_{kr} \phi(\boldsymbol{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \ , \quad \text{subject to} \sum_{r=1}^{R_k} q_{i,kr} = 1 \ . \tag{3.3}$$

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

$$\sum_{k=1}^{K} \sum_{r=1}^{R_k} \left(\sum_{i=1}^{n_k} q_{i,kr} \right) \log \pi_{kr} + \sum_{k=1}^{K} \sum_{r=1}^{R_k} \sum_{i=1}^{n_k} q_{i,kr} \log \phi(\boldsymbol{x}_i | \boldsymbol{\mu}_{kr}, \boldsymbol{\Sigma}) \tag{3.4}$$

   under the constraint (3.2).

In the maximization step, the optimal π_{kr}'s are not affected by the constraint (3.2) and are solved separately from $\boldsymbol{\mu}_{kr}$'s and $\boldsymbol{\Sigma}$:

$$\pi_{kr} \propto \sum_{i=1}^{n_k} q_{i,kr} \ , \quad \sum_{r=1}^{R_k} \pi_{kr} = 1 \ . \tag{3.5}$$

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. The key steps are summarized here. For detailed derivation, we refer interested readers to Appendix D. In constraint (3.2), we have $\boldsymbol{v}_j^t \cdot \boldsymbol{\mu}_{kr} = c_j$, i.e., identical across all the k and r for $j = 1, ..., q$. It is easy to see that $\boldsymbol{c} = (c_1, ..., c_q)^t$ 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 $\boldsymbol{c}$ 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 $\boldsymbol{\Sigma} = V_{\boldsymbol{\Sigma}} D_{\boldsymbol{\Sigma}} V_{\boldsymbol{\Sigma}}^t$, then $\boldsymbol{\Sigma}^{\frac{1}{2}} = D_{\boldsymbol{\Sigma}}^{\frac{1}{2}} V_{\boldsymbol{\Sigma}}^t$. Let $\boldsymbol{V}_{null}$ be a $p \times q$ orthonormal matrix $(\boldsymbol{v}_1, ..., \boldsymbol{v}_q)$, the column vectors of which span the null space $\boldsymbol{\nu}^{\perp}$. Suppose $\boldsymbol{B} = \boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{V}_{null}$. Perform a singular value decomposition (SVD) on $\boldsymbol{B}$, i.e., $\boldsymbol{B} = \boldsymbol{U}_{\boldsymbol{B}} \boldsymbol{D}_{\boldsymbol{B}} \boldsymbol{V}_{\boldsymbol{B}}^t$, where $\boldsymbol{U}_{\boldsymbol{B}}$ is a $p \times q$ matrix, the column vectors of which form an orthonormal basis for the space spanned by the column vectors of $\boldsymbol{B}$. Let $\hat{\boldsymbol{U}}$ be a column augmented orthonormal matrix of $\boldsymbol{U}_{\boldsymbol{B}}$. Denote $\sum_{i=1}^{n_k} q_{i,kr}$ by l_{kr}. Let $\bar{\boldsymbol{x}}_{kr} = \sum_{i=1}^{n_k} q_{i,kr} \boldsymbol{x}_i / l_{kr}$, i.e., the weighted sample mean of the component r in class k, and $\breve{\boldsymbol{x}}_{kr} = \hat{\boldsymbol{U}}^t \left(\boldsymbol{\Sigma}^{-\frac{1}{2}} \right)^t \cdot \bar{\boldsymbol{x}}_{kr}$. Define $\breve{\boldsymbol{\mu}}_{kr}^*$ by the following Eqs. (3.6) and (3.7):

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

$$\breve{\mu}_{kr,j}^* = \frac{\sum_{k'=1}^{K} \sum_{r'=1}^{R_{k'}} l_{k'r'} \breve{x}_{k'r',j}}{n} \ , \quad \text{identical over } r \text{ and } k \ ; \tag{3.6}$$

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

$$\breve{\mu}^*_{kr,j} = \breve{x}_{kr,j} \ . \tag{3.7}$$

That is, the first q constrained coordinates are optimized using component-pooled sample mean (components from all the classes) while those $p-q$ unconstrained coordinates are optimized separately within each component using the component-wise sample mean. Note that we abuse the term "sample mean" here to mean $\breve{\boldsymbol{x}}_{kr}$, instead of $\bar{\boldsymbol{x}}_{kr'}$. In the maximization step, the parameter $\boldsymbol{\mu}_{kr}$ is finally solved by:

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

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

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

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.

Chapter 4
Experiments

We present the experimental results on several real data sets in this chapter. The proposed Gaussian mixture model with subspace constrained means and reduced rank MDA are compared on the classification of data sets with moderate to high dimensions. We also visualize the classification results in lower dimensional subspaces. The methods compared in the experiments are summarized as follows:

- **GMM-MEAN-PCA** The Gaussian mixture model with subspace constrained means, where the subspace is obtained by MEAN-PCA.
- **GMM-MODE-PCA** The Gaussian mixture model with subspace constrained means, where the subspace is obtained by MODE-PCA.
- **MDA-RR** The reduced rank mixture discriminant analysis (MDA), which is our own implementation based on weighted rank reduction of the full MDA.
- **MDA-RR-OS** The reduced rank mixture discriminant analysis (MDA) via optimal scoring, which is the implementation from the R package provided by (Hastie and Tibshirani, 1996).

4.1 Data and Experimental Setup

Six real data sets are used in the experiment. We summarize their detailed information in the following:

- 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 **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 16×16 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 32×32 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 robot, satellite and semeion data are from the UCI machine learning repository. Among the above data sets, the semeion and yaleB data have high dimensions, while 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 data, whose accuracy is the average over ten runs of simulations, the same setting used in (Hastie et al., 2001). For the semeion and yaleB data, the randomly split training and test samples are used to compute their classification accuracy instead of cross validation due to the relatively high computational cost. For all the tested methods, we assume a common full covariance matrix across all the components in all the classes. Since the semeion and yaleB data are of high dimensions, a common diagonal covariance matrix across all the components in all the classes is assumed. For simplicity, we use the same number of mixture components to model each class in all the methods.

In GMM-MODE-PCA, we first apply the Mclust package to estimate the probability density of the mixture model. The MEM algorithm is then used to find the modes based on the estimated density of the model. In Mclust, we have the option of estimating the mixture density by using common diagonal covariance matrices or common full covariance matrices across all the mixture components. In practice, we find that the modes found by using the probability density estimation via common diagonal covariance matrices are more robust, which is evidenced by the better classification performance of GMM-MODE-PCA. We therefore assume a common diagonal covariance matrix across all the components for all the data in Mclust. Note that for the yaleB data, Mclust cannot fit a mixture model due to the numerical issue in high dimensional covariance matrix estimation. We thus skip the running of GMM-MODE-PCA on yaleB. The BIC criterion is used to select the optimal number of mixture components in Mclust. We empirically set the range of the numbers of mixture components for which the BIC is calculated to be 1:20, a relatively large search range.

4.2 Classification Results

We show the classification results of all the methods in this section. Table 4.1 shows the classification error rates of these methods on six data sets, where the final discriminant dimension d is

set to 2 and 3, significantly lower than their original number of dimensions. As we can see, when d is 2, as the number of mixture components varies, the proposed new method GMM-MEAN-PCA significantly outperforms MDA-RR and MDA-RR-OS on all the data sets, except for the waveform and semeion data. The classification performance of all the tested methods are very close on the waveform data with no clear winner on the semeion data. When d is 3, GMM-MEAN-PCA significantly outperforms MDA-RR and MDA-RR-OS on the robot, imagery, and yaleB data, under different number of mixture components. All the tested methods achieve very similar classification accuracy on the satelllite data. For the semeion data, when d is 3, MDA-RR-OS achieves the lowest error rate while GMM-MEAN-PCA still outperforms MDA-RR.

Since the number of classes in waveform is 3, MEAN-PCA can not be applied to obtain a $3d$ subspace using class means. Mclust has numerical issues for the yaleB data in estimating the high dimensional covariance matrix of the mixture model. Therefore, GMM-MODE-PCA can not be applied. In addition, since the dimension of the yaleB data is significantly larger than its sample size, MDA-RR-OS can not be used due to the lack of degrees of freedom in the underlying regression approach. Therefore, for the aforementioned methods, their corresponding runs are marked as "NA" in the reported experiment results.

Among all the methods, when $d = 2$, under different number of mixture components, GMM-MODE-PCA achieves the lowest classification error rate on the satellite data. Comparing with GMM-MEAN-PCA, the relatively worse classification performance of GMM-MODE-PCA may be due to the inaccurate estimation of modes.

4.3 Subspace Comparison

In our method, the fitted GMM is a full model on the original data, although the component means can be constrained in different subspaces. One advantage of our subspace constrained model fitting is that different subspaces can be compared directly by likelihood since the ultimate model is on the full dimensions. MDA-RR is also a model fitted on the original data. We therefore can compare the constrained subspaces based on the estimated likelihood under each model. Table 4.2 shows the training log-likelihoods of all the methods on six data sets. Specifically, the likelihood is computed as the sum of the training log-likelihood of each fold in a five-fold cross validation for all the data sets, except for semeion and yaleB, where the likelihood is on a single training data. Since the R package for MDA-RR-OS cannot provide the full training log-likelihood of the model, we only report the likelihood of MDA-RR. As we can see, MDA-RR has the highest training log-likelihood among all the methods on most data sets. Comparing with the results in Table 4.1, MDA-RR fits the training data better, but has higher classification error rates on the test data.

In our proposed method, the search for a constrained linear subspace and the model fitting is separate. Therefore, we can fit the model by constraining its component means to any subspace. We also experiment with fitting GMM models with the component means constrained by the subspace provided by reduced rank MDA. Specifically, we obtain the discriminant subspace found by reduced rank MDA and use it as the constrained subspace in the new method. Table 4.3 and

Table 4.1: Classification error rates (%) of all the methods on six data sets

(a)

Num of components		Robot		Imagery	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	**30.99**	**27.88**	**44.21**	**40.57**
	GMM-MODE-MPCA	41.11	42.10	54.43	50.21
	MDA-RR	39.59	32.73	53.64	45.79
	MDA-RR-OS	41.04	36.22	52.21	43.71
4	GMM-MEAN-PCA	**28.63**	**23.85**	**45.79**	**39.43**
	GMM-MODE-PCA	41.97	42.56	54.71	51.36
	MDA-RR	40.76	38.84	53.57	49.57
	MDA-RR-OS	40.14	38.82	52.36	44.79
5	GMM-MEAN-PCA	**24.69**	**23.13**	**45.79**	**40.50**
	GMM-MODE-PCA	39.59	36.47	55.79	52.00
	MDA-RR	34.86	35.10	53.14	46.21
	MDA-RR-OS	40.14	38.82	51.86	46.07

(b)

Num of components		Satellite		Waveform	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	16.94	14.17	16.12	NA
	GMM-MODE-PCA	**16.43**	14.30	16.12	**16.32**
	MDA-RR	35.18	**14.08**	**16.00**	17.66
	MDA-RR-OS	25.94	13.66	15.62	16.62
4	GMM-MEAN-PCA	17.31	13.97	15.84	NA
	GMM-MODE-PCA	**16.81**	13.80	15.84	**16.24**
	MDA-RR	35.06	**13.63**	**15.80**	17.06
	MDA-RR-OS	19.50	13.43	15.90	16.98
5	GMM-MEAN-PCA	16.77	13.46	16.26	NA
	GMM-MODE-PCA	**16.02**	13.44	16.26	17.00
	MDA-RR	27.43	**13.18**	16.76	17.62
	MDA-RR-OS	17.39	13.40	**16.02**	**16.58**

(c)

Num of components		Semeion		YaleB	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	49.54	34.37	**31.43**	**22.86**
	GMM-MODE-PCA	56.66	39.94	NA	NA
	MDA-RR	**45.51**	35.91	84.29	60.00
	MDA-RR-OS	48.41	**32.64**	NA	NA
4	GMM-MEAN-PCA	51.39	36.53	**34.29**	**30.00**
	GMM-MODE-PCA	55.73	40.56	NA	NA
	MDA-RR	**49.23**	37.77	85.71	62.86
	MDA-RR-OS	49.52	**31.19**	NA	NA
5	GMM-MEAN-PCA	**43.03**	35.29	**37.14**	**27.14**
	GMM-MODE-PCA	52.63	42.41	NA	NA
	MDA-RR	48.92	38.70	85.71	62.86
	MDA-RR-OS	47.58	**33.07**	NA	NA

Table 4.2: Training log-likelihoods of all the methods on six data sets

(a)

Num of components		Robot		Imagery	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	-774041.92	**-765524.47**	-900921.72	-900074.89
	GMM-MODE-PCA	-781711.78	-781843.18	-902842.10	-901945.83
	MDA-RR	**-767462.15**	-767739.00	**-900119.80**	**-898863.62**
4	GMM-MEAN-PCA	-772996.82	**-755607.10**	-900765.50	-899741.12
	GMM-MODE-PCA	-778320.34	-770265.75	-902534.54	-901551.62
	MDA-RR	**-770755.65**	-759997.26	**-899614.02**	**-898454.26**
5	GMM-MEAN-PCA	**-762956.99**	**-754740.25**	-900548.33	-899452.00
	GMM-MODE-PCA	-777841.50	-769453.55	-902437.79	-901194.25
	MDA-RR	-763511.32	-757370.89	**-899367.42**	**-897785.91**

(b)

Num of components		Satellite		Waveform	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	-2688352.13	-2661632.66	**-95206.10**	NA
	GMM-MODE-PCA	-2689838.43	-2661735.05	-95377.41	-95094.66
	MDA-RR	**-2677474.30**	**-2661488.23**	-95222.29	**-94988.97**
4	GMM-MEAN-PCA	**-2684871.31**	-2656874.25	-95077.30	NA
	GMM-MODE-PCA	-2686277.70	**-2656868.04**	-95142.53	-94830.62
	MDA-RR	-2674159.82	-2657300.01	**-94978.41**	**-94733.52**
5	GMM-MEAN-PCA	-2682968.91	-2653378.08	-95061.15	NA
	GMM-MODE-PCA	2684174.97	-2653541.69	-95061.15	-94699.01
	MDA-RR	**-2672099.16**	**-2653030.44**	**-94886.82**	**-94536.59**

(c)

Num of components		Semeion		YaleB	
		$d = 2$	$d = 3$	$d = 2$	$d = 3$
3	GMM-MEAN-PCA	112835.228798	119987.12	907183.15	910705.56
	GMM-MODE-PCA	106386.673	114788.55	NA	NA
	MDA-RR	**113850.0619**	**122102.92**	**978296.38**	**987087.38**
4	GMM-MEAN-PCA	113221.84	120617.05	907410.38	911095.84
	GMM-MODE-PCA	106579.67	115468.05	NA	NA
	MDA-RR	**114460.33**	**123208.48**	**988147.68**	**997575.79**
5	GMM-MEAN-PCA	113400.50	120992.63	907671.24	911214.23
	GMM-MODE-PCA	106691.33	115781.50	NA	NA
	MDA-RR	**114763.28**	**123906.97**	**993932.81**	**1003748.58**

Table 4.4 show the classification error rates and the training log-likelihoods of reduced rank MDA and the new method constrained by the subspace found by reduced rank MDA, which is denoted by GMM-MPCA. We also use the final estimated component means and covariance matrices from reduced rank MDA to initialize the parameters of GMM-MPCA. Similar as before, all the experiments are conducted using five-fold cross validation. As we can see, GMM-MPCA achieves better classification accuracy on the robot and satellite data under certain discriminant dimension and component number combinations. However, for most cases, the classification

Table 4.3: Classification error rates (%) of the new method constrained by the subspace provided by reduced rank MDA

(a) Robot data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	**31.56**	**30.75**
	MDA-RR	37.13	32.74
4	GMM-MPCA	44.54	38.61
	MDA-RR	**43.00**	**36.14**
5	GMM-MPCA	44.59	41.70
	MDA-RR	**40.76**	**34.53**

(b) Imagery data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	72.07	72.93
	MDA-RR	**53.36**	**46.14**
4	GMM-MPCA	73.36	74.14
	MDA-RR	**52.17**	**48.86**
5	GMM-MPCA	73.29	74.79
	MDA-RR	**53.43**	**47.43**

(c) Satellite data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	**32.21**	30.74
	MDA-RR	35.18	**14.06**
4	GMM-MPCA	**31.67**	29.10
	MDA-RR	35.07	**13.66**
5	GMM-MPCA	31.51	30.46
	MDA-RR	**27.66**	**13.22**

accuracy of GMM-MPCA is worse than that of MDA-RR. MDA-RR also has higher training log-likelihood than GMM-MPCA. These results indicate that the mixture component means may not reside in the discriminant subspace obtained by MDA-RR. Therefore, fitting a GMM with such a constraint yields relatively worse performance.

4.4 Visualization

We visualize the classification of high dimensional data in both two and three dimensional subspaces using GMM-MEAN-PCA and MDA-RR. The visualization results are on independent test data from one fold in the five-fold cross validation. The models are trained on 80% of the data and then applied to visualize the remaining 20% test data with color coding the predicted class labels. We empirically set the number of components in each class to be 3. Figure 4.1 and Figure 4.2 show the visualization results of the robot, imagery, and satellite data in two and three dimensional discriminant subspaces, respectively. As shown in Figure 4.1, the test error rate of GMM-MEAN-PCA is significantly lower than that of MDA-RR in their corresponding

Table 4.4: Training log-likelihoods of the new method constrained by the subspace provided by reduced rank MDA

(a) Robot data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	-768488.82	-771144.04
	MDA-RR	**-766727.33**	**-767739.00**
4	GMM-MPCA	-776969.74	-771181.73
	MDA-RR	**-769073.50**	**-758053.47**
5	GMM-MPCA	-770047.24	-769195.29
	MDA-RR	**-760697.31**	**-755490.63**

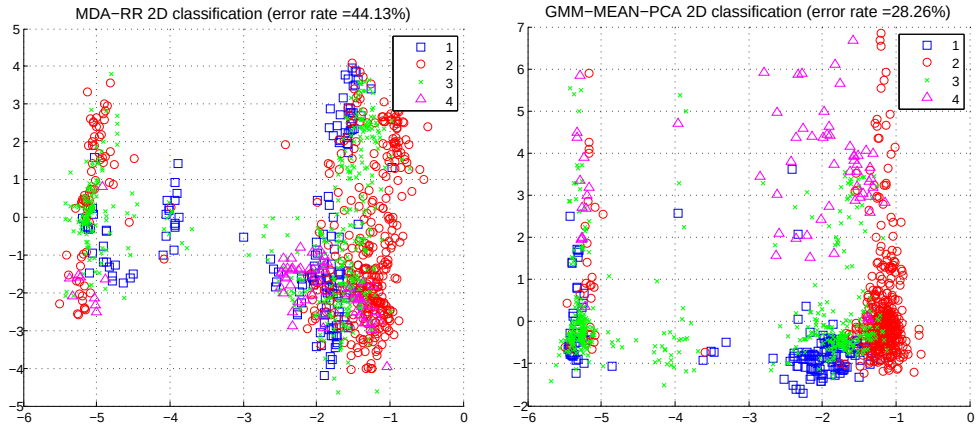
(b) Imagery data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	-905680.82	-905526.87
	MDA-RR	**-900138.34**	**-898870.71**
4	GMM-MPCA	-905482.56	-905311.42
	MDA-RR	**-899544.90**	**-898364.33**
5	GMM-MPCA	-905440.19	-905191.04
	MDA-RR	**-899358.85**	**-897797.37**

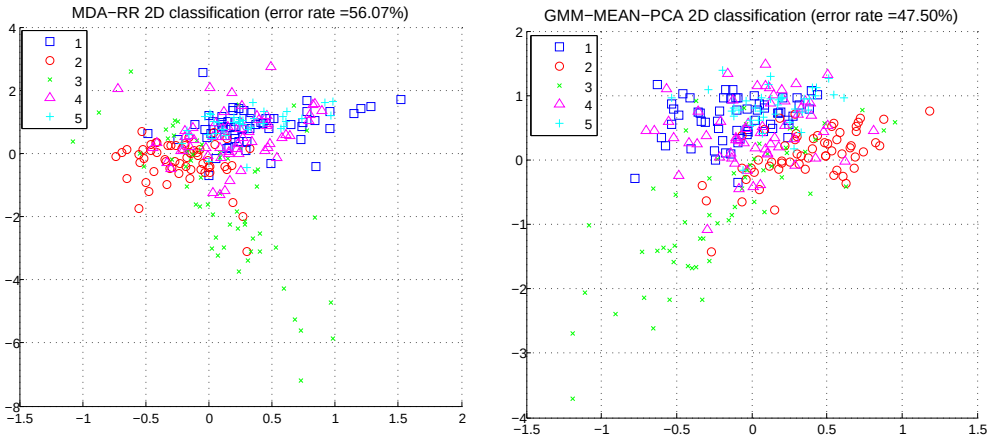
(c) Satellite data

Num of components		$d = 2$	$d = 3$
3	GMM-MPCA	-2729987.28	-2730086.25
	MDA-RR	**-2677474.31**	**-2661500.16**
4	GMM-MPCA	-2728208.16	-2727811.85
	MDA-RR	**-2674159.88**	**-2657253.84**
5	GMM-MPCA	-2725817.72	-2725718.37
	MDA-RR	**-2672124.56**	**-2653030.31**

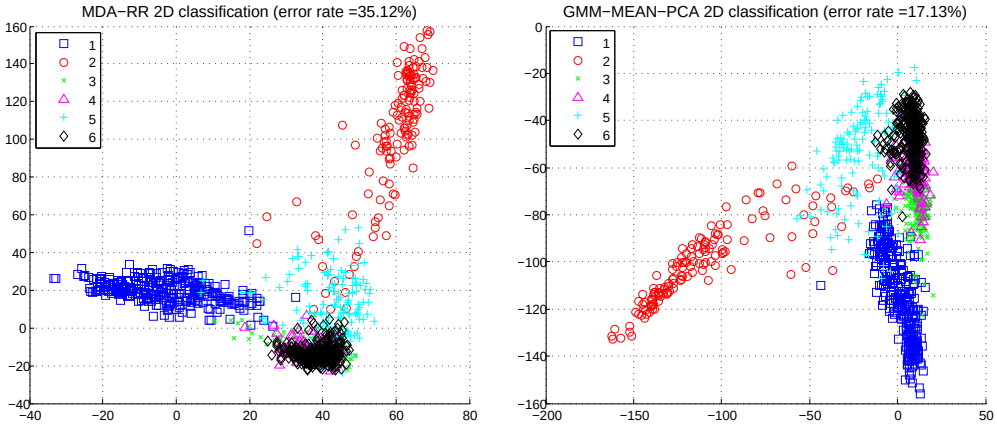
$2d$ subspaces. The data points of different classes from GMM-MEAN-PCA are more spread out or distinct, comparing with that from MDA-RR. In Figure 4.2, the test error rate of GMM-MEAN-PCA in $3d$ subspace is also lower than that of MDA-RR. Comparing with Figure 4.1, the performance difference between these two methods becomes relatively small. Users can therefore visually explore the patterns of these high dimensional data in $2d$ or $3d$ subspaces, which best preserve their classification characteristics from high dimensional space.



(a) Robot data

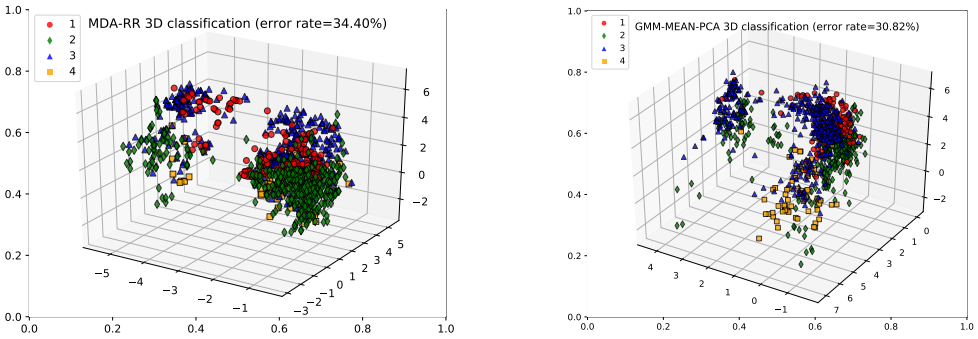


(b) Imagery data

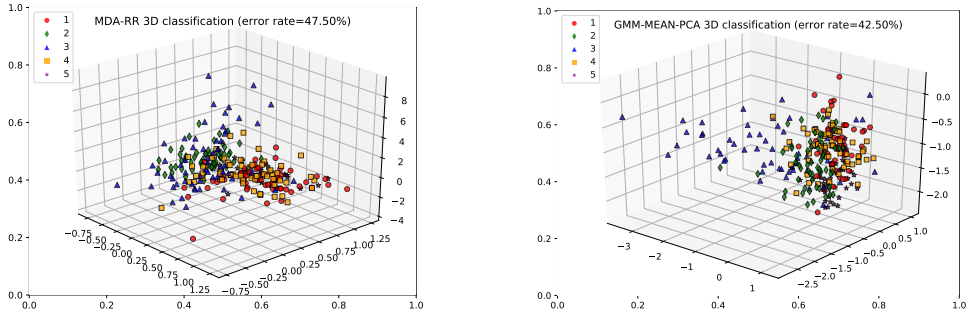


(c) Satellite data

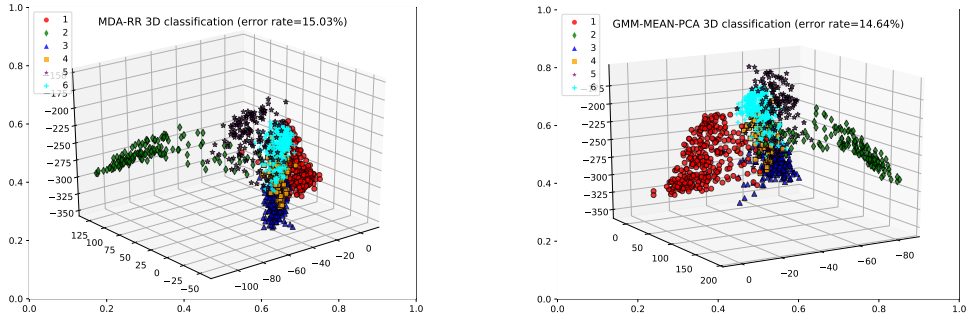
Figure 4.1: Two-dimensional plot for the classification of test data, color-coding the predicted classes.



(a) Robot data



(b) Imagery data



(c) Satellite data

Figure 4.2: Three-dimensional plot for the classification of test data, color-coding the predicted classes.

Chapter 5
Conclusion

In this thesis, we propose a new method for visualizing high dimensional data via the Gaussian mixture model, with component means constrained in a pre-selected subspace. We prove that the class means, modes, and the component means of a Gaussian mixture all lie in the same constrained subspace. We develop two approaches to obtain the subspace by applying weighted PCA to the class means or modes. The constrained method results in a dimension reduction property, which allows us to view the classification or clustering structure of the data in a much lower dimensional space. An EM-type algorithm is derived to estimate the model, given any constrained subspace. 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 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. Visualization results on independent test data show that our proposed method exhibits more clear and distinct classification structure in lower dimensional subspaces in comparison with reduced rank MDA.

In the future, it may be desired to incorporate some prior knowledge in finding the constrained subspace. For instance, the proposed 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
Reduced Rank MDA

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., rank $\{\mu_{kr}\} = L$.

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. (3.3), which are conditional on the current (reduced rank) version of component means and common covariance matrix. In the M-step, π_{kr}'s are still maximized using Eq. (3.5). 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 lth 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

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

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 $V^* = (v_1^*, v_2^*, ..., v_p^*)$. 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 rank $\{\boldsymbol{\mu}_{kr}\} = L$, the solutions are

$$\hat{\boldsymbol{\mu}}_{kr} = WVV^T(\boldsymbol{\mu}_{kr}-\boldsymbol{\mu}) + \boldsymbol{\mu} \ , \tag{A.1}$$

$$\hat{\boldsymbol{\Sigma}} = W + \frac{\sum_{k=1}^{K}\sum_{r=1}^{R_k} q'_{kr}(\boldsymbol{\mu}_{kr}-\hat{\boldsymbol{\mu}}_{kr})^t(\boldsymbol{\mu}_{kr}-\hat{\boldsymbol{\mu}}_{kr})}{\sum_{k=1}^{K}\sum_{r=1}^{R_k} q'_{kr}} \ . \tag{A.2}$$

As a summary, in the M-step of reduced rank MDA, the parameters, π_{kr}, $\boldsymbol{\mu}_{kr}$ and $\boldsymbol{\Sigma}$, are maximized by Eqs. (3.5), (A.1), and (A.2), respectively.

Note that the discriminant subspace is spanned by the column vectors of $V = W^{-\frac{1}{2}}V^*$, with the lth 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.

Appendix B
Proof of Theorem 3.2.1

We prove Theorem 3.2.1 here. Consider a mixture of Gaussians with a common covariance matrix $\mathbf{\Sigma}$ shared across all the components as in (2.2):

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

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

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

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

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 $\widetilde{\boldsymbol{z}}$ by $\widetilde{\boldsymbol{z}}^{(1)}$,

and the last $p-q$ dimensions by $\widetilde{\boldsymbol{z}}^{(2)}$. We can write $g(\widetilde{\boldsymbol{z}})$ as

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

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

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

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

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

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

and equivalently

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

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

Appendix C
Proof of Theorem 3.3.1

We prove Theorem 3.3.1 here. Assume $\boldsymbol{\nu} = \{\boldsymbol{v}_{q+1}, ..., \boldsymbol{v}_p\}$ is the constrained subspace where $\boldsymbol{\mu}_{kr}$'s reside, and $\boldsymbol{\nu}^{\perp} = \{\boldsymbol{v}_1, ..., \boldsymbol{v}_q\}$ is the corresponding null subspace, as defined in Section 3. We use the Bayes classification rule to classify a sample x: $\widehat{y} = \text{argmax}_k\, f(Y = k|\mathbf{X} = \mathbf{x}) = \text{argmax}_k\, f(\mathbf{X} = \mathbf{x}, Y = k)$.

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

Let $\boldsymbol{V} = \begin{pmatrix} \boldsymbol{v}_1^t \\ \vdots \\ \boldsymbol{v}_p^t \end{pmatrix}$. Matrix $\boldsymbol{V}$ is orthonormal because $\boldsymbol{v}_j$'s are orthonormal by construction. Consider the following cases of $\boldsymbol{\Sigma}$.

C.1 Σ is an identity matrix

From Eq. (C.1), we have

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

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

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

Therefore,

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

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

C.2 Σ is a non-identity matrix

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

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

where $\tilde{\boldsymbol{\mu}}_{kr} = (\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{\mu}_{kr}$. According to our assumption, $\boldsymbol{v}_j^t \cdot \boldsymbol{\mu}_{kr} = c_j$, i.e., identical across all k and r for $j = 1, ..., q$. Plugging into $\boldsymbol{\mu}_{kr} = (\boldsymbol{\Sigma}^{\frac{1}{2}})^t \tilde{\boldsymbol{\mu}}_{kr}$, we get $(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{v}_j)^t \cdot \tilde{\boldsymbol{\mu}}_{kr} = c_j$, $j = 1, ..., q$. This means for the transformed data, the component means $\tilde{\boldsymbol{\mu}}_{kr}$'s have a null space spanned by $\{\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{v}_j | j = 1, ..., q\}$. Correspondingly, the constrained subspace is spanned by $\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\}$. It is easy to verify that the new null space and constrained subspace are orthogonal, since $(\boldsymbol{\Sigma}^{\frac{1}{2}} \boldsymbol{v}_j)^t \cdot (\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_{j'} = \boldsymbol{v}_j{}^t \cdot \boldsymbol{v}_{j'} = 0$, $j = 1, ...q$ and $j' = q+1, ..., p$. The spanning vectors for the constrained subspace, $(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j$, $j = q+1, ..., p$, are not orthonormal in general, but there exists an orthonormal basis that spans the same subspace. With a slight abuse of notation, we use $\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\}$ to denote a $p \times (p-q)$ matrix containing the column vector $(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j$. 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 C.1, for the transformed data $\boldsymbol{Z}$, we only need the projection of $\boldsymbol{Z}$ onto a subspace spanned by the column vectors of $orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\})$ to compute the class posterior. Note that $\boldsymbol{Z} = (\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{X}$. So the subspace that matters for classification for the original data $\boldsymbol{X}$ is spanned by the column vectors of $(\boldsymbol{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\boldsymbol{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\})$. 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 $orth((\mathbf{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\mathbf{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\}))$. Since $orth((\mathbf{\Sigma}^{-\frac{1}{2}}) \times orth(\{(\mathbf{\Sigma}^{-\frac{1}{2}})^t \boldsymbol{v}_j | j = q+1, ..., p\})) = orth(\{\mathbf{\Sigma}^{-1} \boldsymbol{v}_j | j = q+1, ..., p\})$,[1] the subspace that matters for classification is thus spanned by the column vectors of $orth(\{\mathbf{\Sigma}^{-1} \boldsymbol{v}_j | j = q+1, ..., p\})$.

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

[1] 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 $orth(A \times orth(B)) = orth(A \times B)$.

Appendix D
Derivation of μ_{kr} in GEM

We derive the optimal $\boldsymbol{\mu}_{kr}$'s under constraint (3.2) for a given $\boldsymbol{\Sigma}$. Note that the term in Eq. (3.4) that involves $\boldsymbol{\mu}_{kr}$'s is:

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

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

$$\sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})\ . \tag{D.2}$$

To solve the above optimization problem under constraint (3.2), 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 $\bar{\boldsymbol{x}}_{kr}$ and $\boldsymbol{\mu}_{kr}$. Once this is achieved, the optimal solution will simply be given by setting those unconstrained coordinates within 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 $\boldsymbol{\Sigma}$, that is, $\boldsymbol{\Sigma}=(\boldsymbol{\Sigma}^{\frac{1}{2}})^t\boldsymbol{\Sigma}^{\frac{1}{2}}$. If the eigen decomposition of $\boldsymbol{\Sigma}$ is $\boldsymbol{\Sigma}=V_{\boldsymbol{\Sigma}}D_{\boldsymbol{\Sigma}}V_{\boldsymbol{\Sigma}}^t$, then, $\boldsymbol{\Sigma}^{\frac{1}{2}}=D_{\boldsymbol{\Sigma}}^{\frac{1}{2}}V_{\boldsymbol{\Sigma}}^t$. Now perform the following change of variables:

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})^t\boldsymbol{\Sigma}^{-1}(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})\\
=\ &\sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}\left[\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^t(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})\right]^t\left[\left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^t(\bar{\boldsymbol{x}}_{kr}-\boldsymbol{\mu}_{kr})\right]\\
=\ &\sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}(\tilde{\boldsymbol{x}}_{kr}-\tilde{\boldsymbol{\mu}}_{kr})^t(\tilde{\boldsymbol{x}}_{kr}-\tilde{\boldsymbol{\mu}}_{kr})\ ,
\end{aligned} \tag{D.3}$$

where $\tilde{\boldsymbol{\mu}}_{kr} = \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^t \cdot \boldsymbol{\mu}_{kr}$, and $\tilde{\boldsymbol{x}}_{kr} = \left(\boldsymbol{\Sigma}^{-\frac{1}{2}}\right)^t \cdot \bar{\boldsymbol{x}}_{kr}$. Correspondingly, the constraint in (3.2) becomes

$$\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{v}_j\right)^t \tilde{\boldsymbol{\mu}}_{kr} = \text{constant over } r \text{ and } k\ , \quad j = 1, ..., q\ . \tag{D.4}$$

Let $\boldsymbol{b}_j = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{v}_j$ and $\boldsymbol{B} = (\boldsymbol{b}_1, \boldsymbol{b}_2, ..., \boldsymbol{b}_q)$. Note that the rank of $\boldsymbol{V} = (\boldsymbol{v}_1, ..., \boldsymbol{v}_q)$ is q. Since $\boldsymbol{\Sigma}^{\frac{1}{2}}$ is of full rank, $\boldsymbol{B} = \boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{V}$ also has rank q. The constraint in (D.4) becomes

$$\boldsymbol{B}^t\tilde{\boldsymbol{\mu}}_{kr} = \boldsymbol{B}^t\tilde{\boldsymbol{\mu}}_{k'r'}\ , \text{for any } r,\ r' = 1, ..., R_k, \text{ and any } k,\ k' = 1, ..., K\ . \tag{D.5}$$

Now perform a singular value decomposition (SVD) on $\boldsymbol{B}$, i.e., $\boldsymbol{B} = \boldsymbol{U_B}\boldsymbol{D_B}\boldsymbol{V_B}^t$, where $\boldsymbol{V_B}$ is a $q \times q$ orthonormal matrix, $\boldsymbol{D_B}$ is a $q \times q$ diagonal matrix, which is non-singular since the rank of $\boldsymbol{B}$ is q, and U_B is a $p \times q$ orthonormal matrix. Substituting the SVD of $\boldsymbol{B}$ in (D.5), we get

$$\boldsymbol{V_B}\boldsymbol{D_B}\boldsymbol{U_B}^t\tilde{\boldsymbol{\mu}}_{kr} = \boldsymbol{V_B}\boldsymbol{D_B}\boldsymbol{U_B}^t\tilde{\boldsymbol{\mu}}_{k'r'}, \quad \text{for any } r,\ r' = 1, ..., R_k, \text{ and any } k,\ k' = 1, ..., K\ ,$$

which is equivalent to

$$\boldsymbol{U_B}^t\tilde{\boldsymbol{\mu}}_{kr} = \boldsymbol{U_B}^t\tilde{\boldsymbol{\mu}}_{k'r'}, \quad \text{for any } r,\ r' = 1, ..., R_k, \text{ and any } k,\ k' = 1, ..., K, \tag{D.6}$$

because $\boldsymbol{V_B}$ and $\boldsymbol{D_B}$ have full rank. We can augment $\boldsymbol{U_B}$ to a $p \times p$ orthonormal matrix, $\hat{\boldsymbol{U}} = (\boldsymbol{u}_1, ..., \boldsymbol{u}_q, \boldsymbol{u}_{q+1}, ..., \boldsymbol{u}_p)$, where $\boldsymbol{u}_{q+1}$, ..., $\boldsymbol{u}_p$ are augmented orthonormal vectors. Since $\hat{\boldsymbol{U}}$ is orthonormal, the objective function in Eq. (D.3) can be written as

$$\sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}[\hat{\boldsymbol{U}}^t(\tilde{\boldsymbol{x}}_{kr} - \tilde{\boldsymbol{\mu}}_j)]^t \cdot [\hat{\boldsymbol{U}}^t(\tilde{\boldsymbol{x}}_{kr} - \tilde{\boldsymbol{\mu}}_{kr})] = \sum_{k=1}^{K}\sum_{r=1}^{R_k} l_{kr}(\breve{\boldsymbol{x}}_{kr} - \breve{\boldsymbol{\mu}}_{kr})^t(\breve{\boldsymbol{x}}_{kr} - \breve{\boldsymbol{\mu}}_{kr})\ , \tag{D.7}$$

where $\breve{\boldsymbol{x}}_{kr} = \hat{\boldsymbol{U}}^t\tilde{\boldsymbol{x}}_{kr}$ and $\breve{\boldsymbol{\mu}}_{kr} = \hat{\boldsymbol{U}}^t\tilde{\boldsymbol{\mu}}_{kr}$. If we denote $\breve{\boldsymbol{\mu}}_{kr} = (\breve{\mu}_{kr,1}, \breve{\mu}_{kr,2}, ..., \breve{\mu}_{kr,p})^t$, then the constraint in (D.6) simply becomes

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

That is, the first q coordinates of $\breve{\boldsymbol{\mu}}$ have to be common over all the k and r. The objective function (D.7) can be separated coordinate wise:

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

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

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

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

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

After $\breve{\boldsymbol{\mu}}^*_{kr}$ is calculated, we finally get $\boldsymbol{\mu}_{kr}$'s under the constraint(3.2):

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

Bibliography

T. W. Anderson. *An introduction to multivariate statistical analysis*. Wiley, 2000.

A. Anand, L. Wilkinson, and T. N. Dang. Visual pattern discovery using random projections. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 43-52, 2012.

L. Breiman and R. Ihaka. Nonlinear discriminant analysis via scaling and ACE. Technical Report 40, Department of Statistics, University of California, Berkeley, California, 1984.

C. Baumgartner, C. Plant, K. Railing, H.-P. Kriegel, and P. Kroger, Subspace selection for clustering high-dimensional data. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 11-18, 2004.

J. Choo, H. Lee, J. Kihm, and H. Park. iVisClassifier: An interactive visual analytics system for classification based on supervised dimension reduction. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 27-34, 2010.

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.

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

R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179-188, 1936.

C. Fraley, A. Raftery, T. B. Murphy, and L. Scrucca. MCLUST version 4 for R: normal mixture modeling for model-based clustering, classification, and density Estimation. *Technical report no. 597, Department of Statistics, University of Washington*, 2012.

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. Georghiades, 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.

Z. Ghahramani, G. E. Hinton. The EM algorithm for factor analyzers. Technical Report CRG-TR-96-1, The University of Toronto, Toronto, 1997.

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.

P. Kenny, P. Ouellet, N. Dehak, V. Gupta. A study of interspeaker variability in speaker verification. *IEEE Transactions on Audio, Speech and Language Processing*, 16(5):980-987, 2008.

K. C. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):684-698, 2005.

D. J. Lehmann and H. Theisel. Optimal sets of projections of high-dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 22(1): 609-618, 2016.

S. Liu, B. Wang, J. J. Thiagarajan, P. T. Bremer, and V. Pascucci. Visual exploration of high-dimensional data through subspace analysis and dynamic Projections. *Computer Graph Forum*, 34: 271-280, 2015.

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.

G. J. McLachlan and D. Peel. *Finite mixture models*. Wiley, 2000.

G. J. McLachlan, D. Peel, and R. W. Bean. Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* 41:379-388, 2003

G. J. McLachlan and D. Peel. Mixtures of factor analyzers. In *Proceeding of the International Conference on Machine Learning*, pages 599-606, 2000.

D. Povey, et al. The subspace gaussian mixture model - a structured model for speech recognition. *Computer Speech and Language*, 25(2):404-439, 2011.

K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(11):559-572, 1901.

M. Qiao, J. Li. Two-way gaussian mixture models for high dimensional classification. *Statistical Analysis and Data Mining*, 3(4):259-271, 2010.

A. Tatu, F. Maaß, I. Färber, et al. Subspace search and visualization to make sense of alternative clusterings in high-dimensional data. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 63-72, 2012.

F. W. Young. *Multidimensional scaling: history, theory, and applications*. Lawrence Erlbaum Associates, 1987.

F. Zhou, J. Li, Wei Huang, Y. Zhao, X. Yuan, X. Liang, and Y. Shi. Dimension reconstruction for visual exploration of subspace clusters in high-dimensional data. In *Proceedings of the IEEE Pacific Visualization Symposium (PacificVis)*, pages 128-135, 2016.