Graphical Tools for Model-based Mixture Discriminant Analysis

Luca Scrucca

Università degli Studi di Perugia

September 4, 2018

Abstract

The paper introduces a methodology for visualizing on a dimension reduced subspace the classification structure and the geometric characteristics induced by an estimated Gaussian mixture model for discriminant analysis. In particular, we consider the case of mixture of mixture models with varying parametrization which allow for parsimonious models. The approach is an extension of an existing work on reducing dimensionality for model-based clustering based on Gaussian mixtures. Information on the dimension reduction subspace is provided by the variation on class locations and, depending on the estimated mixture model, on the variation on class dispersions. Projections along the estimated directions provide summary plots which help to visualize the structure of the classes and their characteristics. A suitable modification of the method allows us to recover the most discriminant directions, i.e., those that show maximal separation among classes. The approach is illustrated using simulated and real data.

Keywords: Dimension reduction, Model-based discriminant analysis, Gaussian mixtures, Canonical variates for mixture modeling

1 Introduction

Discriminant analysis or supervised learning indicates a broad set of statistical methods aimed at classifying a categorical outcome variable Y, an indicator with K classes, on the basis of a $(p \times 1)$ vector of features $\boldsymbol{x}$. Among the several methods available for continuous features, one of the most popular approach is classical linear discriminant analysis (LDA). This method has been extended to quadratic discriminant analysis (QDA), and, more generally, to models based on finite mixture modeling of Gaussian densities.

Independently from the statistical method adopted, visualization and graphics can play an important role in the understanding of the classification results. Typically, canonical variates are computed when the dimension of the features space is large. This allows us to visualize the classes on a reduced subspace, often bi-dimensional. However, canonical variates are tailored to LDA. Some methods have been proposed for QDA, while graphical methods for finite mixture modeling is still a research area to be explored. From a different point of view, Hennig (2004) has proposed an asymmetric discriminant projection method by looking at the projections where a class appears as homogeneous as possible and separated from the remaining groups.

In this paper a dimension reduction method for visualizing and summarizing the fit of a model-based mixture discriminant analysis is discussed. The approach is an extension of the method proposed by Scrucca (2010) for model-based clustering. The estimated subspace is found by looking at the variation in class means and class covariances depending upon the assumed parameterization of the fitted Gaussian mixture model. The resulting projection subspace is able to capture most of the classification structure available in the data. The proposal reduces to LDA canonical variates for a specific parameterization of the mixture model, while it is equivalent to a recently proposed method for QDA. In all the other cases, the proposed visualization method

is able to show the main geometric characteristics of the fitted mixture model. Furthermore, the proposal can be adapted to allow for the visualization of the separation among the classes.

The paper is organized as follows. In Section 2 a brief review of classification and graphical methods based on the Gaussian distribution is provided. Section 3 describes the Gaussian mixture models for discriminant analysis we consider in this paper. In Section 4 the methodology is introduced and the main properties are described. Section 5 presents examples based on both simulated and real data. In Section 6 the proposed method is extended to allow to recover the most discriminant directions, i.e., those that show maximal separation among classes. Concluding remarks appear in the final Section.

2 Classification based on the Gaussian distribution and existing graphical methods

All the models discussed in this paper are probabilistic, i.e., based on the assumption that the observations in the kth class ($k = 1, \ldots, K$) are generated by a probability distribution $f_k(\boldsymbol{x})$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_p)^\top$ is a column vector of p observed features. Most discriminant analysis methods for continuous variables are based on the assumption that observations in each class are multivariate normal, so that $f_k(\boldsymbol{x}) = \phi(\boldsymbol{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, where ϕ is the p-variate Gaussian density with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

Linear discriminant analysis (LDA) assumes normal populations with equal class covariance matrices, i.e., $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_W$ for all $k = 1, \ldots, K$, where $\boldsymbol{\Sigma}_W = \sum_{i=k}^{K} \pi_k \boldsymbol{\Sigma}_k$ is the within-class covariance matrix with class prior probabilities π_k. The resulting discriminant function is linear in the feature vector $\boldsymbol{x}$ and the acceptance regions for the classes are separated in $\mathbb{R}^p$ by means of hyperplanes. However, Fisher's (1936) original proposal did not rely on the Gaussian distribution. Based on geometric arguments, he looked for a vector of d linear combinations $\boldsymbol{\beta}^\top \boldsymbol{x}$, with $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$, such that the between-class covariance, $\boldsymbol{\Sigma}_B = \sum_{k=1}^{K} \pi_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top$ with $\boldsymbol{\mu} = \sum_{k=1}^{K} \pi_k \boldsymbol{\mu}_k$, is maximized relative to the within-class covariance, $\boldsymbol{\Sigma}_W$. This amounts to maximize the so called *Rayleigh quotient*, i.e.,

$$\arg\max_{\boldsymbol{\beta}} \frac{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_B \boldsymbol{\beta}}{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_W \boldsymbol{\beta}},$$

or, equivalently, find $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ which maximizes $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_B \boldsymbol{\beta}$ subject to $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}_W \boldsymbol{\beta} = \boldsymbol{I}_d$, where $\boldsymbol{I}_d$ is the identity matrix of dimension $(d \times d)$. The problem is solved by the generalized eigenvalue decomposition of $\boldsymbol{\Sigma}_B$ with respect to $\boldsymbol{\Sigma}_W$.

The directions given by the d columns of $\boldsymbol{\beta}$ form the basis of the d-dimensional reduction subspace, $\mathcal{S}(\boldsymbol{\beta})$, which shows the maximal separation among classes, and decision boundaries are linear in the projected features subspace. The dimension of this subspace is $d = \min(p, K - 1)$, so just one direction can be estimated in two-class problems. Fisher's or LDA canonical variates, $\boldsymbol{\beta}^\top \boldsymbol{x}$, express the projection onto this subspace, and provide a graphical counterpart to LDA (Mardia et al, 1979, Chap. 11).

There exists some connection between LDA canonical variates and other dimension reduction methods. In particular, it has been shown that for a categorical response variable sliced inverse regression (SIR; Li, 1991) is equivalent to LDA, except for a different scaling. In fact, SIR covariates are scaled to have unit covariance while the LDA canonical variates are scaled to have unit within-class covariance (Chen and Li, 2001). See also Kent (1991) for more discussion on the connection between SIR and LDA.

Quadratic discriminant analysis (QDA) is obtained by removing the assumption of a common class covariance matrix. The resulting discriminant function is quadratic, and the decision boundaries are quadratic surfaces over the features subspace. However, in this case there appears to be no standard canonical variates analysis for QDA as there is for LDA. Some authors have

considered dimension reduction methods for quadratic discrimination in normal populations with different covariance matrices. Pardoe et al (2007) proposed the use of sliced average variance estimation (SAVE; Cook and Weisberg, 1991) as a graphical representation in quadratic discriminant analysis. Velilla (2008, 2010) discussed the concept of quadratic subspace as a tool for dimension reduction in QDA.

Other extensions to LDA are regularized discriminant analysis (RDA, Friedman, 1989), flexible discriminant analysis (FDA, Hastie et al, 1994), and penalized discriminant analysis (PDA, Hastie et al, 1995). RDA represents a compromise between LDA and QDA; it uses a tuning parameter α for class covariance matrix estimation, i.e., the covariance matrix for class k is estimated by the convex combination

$$\boldsymbol{\Sigma}_k(\alpha) = \alpha \boldsymbol{\Sigma}_k + (1 - \alpha) \boldsymbol{\Sigma}_W.$$

FDA fits by optimal scoring a linear regression model using a basis expansion $h(\boldsymbol{x})$ of the feature vector $\boldsymbol{x}$. PDA also uses optimal scoring on a basis expansion $h(\boldsymbol{x})$ as in FDA, but with a quadratic penalty on the coefficients, i.e., solving the following optimization problem

$$\arg\max_{\boldsymbol{\beta}} \ \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_B \boldsymbol{\beta} \quad \text{subject to } \boldsymbol{\beta}^\top (\boldsymbol{\Sigma}_W + \lambda \boldsymbol{\Omega}) \boldsymbol{\beta} = \boldsymbol{I}_d,$$

where $\boldsymbol{\Omega}$ is a $(p \times p)$ symmetric, nonnegative definite, penalty matrix. All these methods have no direct graphical representation associated, so usually canonical variates are computed as in LDA using the estimated class means.

Finally, we mention the LAD proposal by Cook and Forzani (2009), a likelihood-based dimension reduction method under the assumption of conditional normality of predictors given the response. This model closely resembles the family of models we adopted, but the estimation procedure is quite different. In fact, no closed-form solution to the maximum likelihood estimation of the central subspace (the parameter of interest) is available. Thus, numerical optimization is used for maximization of the log-likelihood on Grassman manifolds.

3 Finite mixture modelling in discriminant analysis

Mixture discriminant analysis generalizes the previous approaches by allowing the density for each class conditional density to be expressed by a finite mixture of normals. Thus, a Gaussian mixture model for the k-th class $(k = 1, \ldots, K)$ has density

$$f_k(\boldsymbol{x}) = \sum_{g=1}^{G_k} \pi_{gk} \phi(\boldsymbol{x}; \boldsymbol{\mu}_{gk}, \boldsymbol{\Sigma}_{gk}), \tag{1}$$

where π_{gk} are the mixing probabilities $(\pi_{gk} > 0, \sum_{g=1}^{G_k} \pi_{gk} = 1)$, $\boldsymbol{\mu}_{gk}$ is the mean of component g in class k, and $\boldsymbol{\Sigma}_{gk}$ is the covariance matrix of component g in class k. Thus, Gaussian components are ellipsoidal, centered at $\boldsymbol{\mu}_{gk}$, and with other geometric features, such as volume, shape and orientation, determined by $\boldsymbol{\Sigma}_{gk}$.

Hastie and Tibshirani (1996) introduced mixture discriminant analysis (MDA) assuming a common full covariance matrix, i.e. $\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Sigma}$ for all g, k, with known fixed number of mixture components for each class.

In a procedure named eigenvalue decomposition discriminant analysis (EDDA), Bensmail and Celeux (1996) proposed the use of Gaussian finite mixture modeling for discriminant analysis in which each class is modeled by a single Gaussian term, i.e., $G_k = 1$ for all k, with the same (possibly parsimonious) class covariance structure factorized as

$$\boldsymbol{\Sigma}_k = \lambda_k \boldsymbol{D}_k \boldsymbol{A}_k \boldsymbol{D}_k^\top,$$

where λ_k is a scalar value controlling the volume of the ellipsoid, $\boldsymbol{A}_k$ is a diagonal matrix specifying the shape of the density contours, and $\boldsymbol{D}_k$ is an orthogonal matrix which determines the orientation of the ellipsoid (Banfield and Raftery, 1993; Celeux and Govaert, 1995). Table 1 shows the MCLUST family of mixture models supported by the `mclust` package (Fraley et al, 2012) for the R software (R Core Team, 2013).

Table 1: Parametrizations of covariance matrices available in the `mclust` software (Fraley et al, 2012) and related geometric characteristics.

Model	$\boldsymbol{\Sigma}_k$	Distribution	Volume	Shape	Orientation
E	σ	Univariate	equal		
V	σ_k	Univariate	variable		
EII	$\lambda\boldsymbol{I}$	Spherical	equal	equal	
VII	$\lambda_k\boldsymbol{I}$	Spherical	variable	equal	
EEI	$\lambda\boldsymbol{A}$	Diagonal	equal	equal	coordinate axes
VEI	$\lambda_k\boldsymbol{A}$	Diagonal	variable	equal	coordinate axes
EVI	$\lambda\boldsymbol{A}_k$	Diagonal	equal	variable	coordinate axes
VVI	$\lambda_k\boldsymbol{A}_k$	Diagonal	variable	variable	coordinate axes
EEE	$\lambda\boldsymbol{D}\boldsymbol{A}\boldsymbol{D}^\top$	Ellipsoidal	equal	equal	equal
EEV	$\lambda\boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	equal	equal	variable
VEV	$\lambda_k\boldsymbol{D}_k\boldsymbol{A}\boldsymbol{D}_k^\top$	Ellipsoidal	variable	equal	variable
VVV	$\lambda_k\boldsymbol{D}_k\boldsymbol{A}_k\boldsymbol{D}_k^\top$	Ellipsoidal	variable	variable	variable

A generalization of the previous two approaches is the method called MclustDA (Fraley and Raftery, 2002), where a density estimate for the data is obtained by a Gaussian finite mixture model with a different number of components and a different (possibly parsimonious) covariance matrix for each class. The corresponding family is thus very flexible allowing the distribution of each class to be approximated by a mixture of Gaussian components.

Maximum likelihood estimates for finite mixture models can be computed via the EM algorithm (Dempster et al, 1977). Model selection, which requires choosing the number of mixture components and the covariance parameterization for each class, is usually based on penalized criteria, such as the Bayesian information criterion (BIC, Schwartz, 1978) or the integrated complete likelihood (ICL, Biernacki et al, 2000).

4 Dimension reduction for model-based discriminant analysis

4.1 Methodology

Suppose we would like to find a suitable reduced number of projections which, depending on the estimated Gaussian mixture model, are able to visualize variation both in groups location and dispersion. Following the proposal of Scrucca (2010) for model-based clustering, consider the following matrices:

$$\boldsymbol{M}_{\mathsf{I}} = \sum_{k=1}^{K}\sum_{g=1}^{G_k} \omega_{gk}(\boldsymbol{\mu}_{gk} - \boldsymbol{\mu})(\boldsymbol{\mu}_{gk} - \boldsymbol{\mu})^\top,$$

where $\omega_{gk} = \pi_k\pi_{gk}$ ($\omega_{gk} > 0$, $\sum_{k,g}\omega_{gk} = 1$), $\boldsymbol{\mu} = \sum_{k=1}^{K}\pi_k\boldsymbol{\mu}_k = \sum_{k,g}\omega_{gk}\boldsymbol{\mu}_{gk}$ is the marginal mean vector, $\boldsymbol{\mu}_k = \sum_{g=1}^{G_k}\pi_{gk}\boldsymbol{\mu}_{gk}$ is the mean vector for class k, and

$$\boldsymbol{M}_{\mathsf{II}} = \sum_{k=1}^{K}\sum_{g=1}^{G_k} \omega_{gk}(\boldsymbol{\Sigma}_{gk} - \bar{\boldsymbol{\Sigma}})\boldsymbol{\Sigma}_X^{-1}(\boldsymbol{\Sigma}_{gk} - \bar{\boldsymbol{\Sigma}})^\top,$$

where $\bar{\boldsymbol{\Sigma}} = \sum_{k,g} \omega_{gk} \boldsymbol{\Sigma}_{gk}$ is the pooled within-class covariance matrix, and $\boldsymbol{\Sigma}_X = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^\top$ is the marginal covariance matrix.

The matrix $\boldsymbol{M}_{\mathsf{I}}$ contains information on class-component means variation, while $\boldsymbol{M}_{\mathsf{II}}$ contains information on class-component covariances variation. The two types of information can be summarized using the following kernel matrix

$$\boldsymbol{M} = \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_{\mathsf{I}} + \boldsymbol{M}_{\mathsf{II}}. \tag{2}$$

The matrix $\boldsymbol{\beta} \in \mathbb{R}^{p \times d}$ ($d = \min(p, \sum_{k=1}^{K} G_k - 1)$) spanning the desired subspace is the solution of the following optimization

$$\arg\max_{\boldsymbol{\beta}} \ \boldsymbol{\beta}^\top \boldsymbol{M} \boldsymbol{\beta} \qquad \text{subject to} \quad \boldsymbol{\beta}^\top \boldsymbol{\Sigma}_X \boldsymbol{\beta} = \boldsymbol{I}_d, \tag{3}$$

where $\boldsymbol{I}_d$ is the $(d \times d)$ identity matrix. The solution of (3) is obtained through the generalized eigen-decomposition of $\boldsymbol{M}$ with respect to $\boldsymbol{\Sigma}_X$. Hence, the basis $\boldsymbol{\beta}$ of the dimension reduction subspace $\mathcal{S}(\boldsymbol{\beta})$ is obtained as $\boldsymbol{\Sigma}_X^{-1/2}$ times the eigenvectors of $\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{M} \boldsymbol{\Sigma}_X^{-1/2}$, with directions ordered according to the corresponding eigenvalues. The projections onto such subspace is then given by $\boldsymbol{z} = \boldsymbol{\beta}^\top \boldsymbol{x}$. In analogy with the name of the method proposed in Scrucca (2010), these are called GMMDRC (Gaussian Mixture Model Dimension Reduction for Classification) variables and provide a graphical method to display the classification structure resulting from a Gaussian mixture model.

Note that in the presentation made so far we have assumed that the parameters of the population are known. Usually, however, they are unknown and must be estimated from training data as discussed in Section 4.3.

4.2 Properties

For MDA models the subspace spanned by $\boldsymbol{M}$ is equivalent to that spanned by $\boldsymbol{M}_{\mathsf{I}}$. This because under the MDA assumption of common class covariance, i.e., $\boldsymbol{\Sigma}_{gk} = \boldsymbol{\Sigma}$ for all g, k, we get $\boldsymbol{M}_{\mathsf{II}} = 0$, so no contribution comes from the variation on class covariances. The same also happens for those EDDA models which assume constant class covariance matrices (i.e., models EII, EEI, and EEE – see Table 1), because here $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}$ for all k. In all the other cases, i.e., for those mixture models which allow different class covariance matrices, $\boldsymbol{M}_{\mathsf{II}}$ adds further information for the identification of the reduction subspace.

In two specific cases the subspace identified by GMMDRC reduces to simple known situations as described in the following propositions, whose proofs are contained in the Appendix.

Proposition 1 *Consider the EDDA mixture model with common full class covariance matrix (EEE). The subspace $\mathcal{S}(\boldsymbol{\beta})$, obtained by solving the GMMDRC constrained optimization in (3) with $\boldsymbol{M} = \boldsymbol{M}_I \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_I$, is identical to the subspace $\mathcal{S}(\boldsymbol{\beta}^{SIR})$ spanned by SIR, and also to the subspace $\mathcal{S}(\boldsymbol{\beta}^{LDA})$ spanned by LDA canonical directions.*

Based on Proposition 1 we claim that using canonical LDA variables is only relevant when the adopted mixture model for classification assumes a single component with common covariance matrix for each class (see also Chen and Li, 2001).

Proposition 2 *Consider the EDDA mixture model with different full class covariance matrices (VVV). The subspace $\mathcal{S}(\boldsymbol{\beta})$, obtained by solving the GMMDRC constrained optimization in (3) with $\boldsymbol{M}$ as in (2), is identical to the subspace $\mathcal{S}(\boldsymbol{\beta}^{SAVE})$ spanned by SAVE.*

Noting that an EDDA mixture model with a different full covariance matrix for each class is essentially equivalent to QDA, Proposition 2 supports the use of SAVE as a graphical counterpart to QDA.

5

Proposition 3 *Let $l_1 \geq l_2 \geq \cdots \geq l_d > 0$ be the eigenvalues from the generalized eigen-decomposition of the kernel $\boldsymbol{M}$, i.e., $\boldsymbol{M}\boldsymbol{\beta}_j = l_j\boldsymbol{\Sigma}_X\boldsymbol{\beta}_j$, for $j = 1, \ldots, d$. Each eigenvalue l_j, corresponding to the direction $\boldsymbol{\beta}_j$ of the projection subspace $\mathcal{S}(\boldsymbol{\beta})$, can be written as*

$$l_j = \mathrm{Var}(\mathrm{E}(\boldsymbol{\beta}_j^\top \boldsymbol{x}|Y))^2 + \mathrm{E}(\mathrm{Var}(\boldsymbol{\beta}_j^\top \boldsymbol{x}|Y)^2) \qquad for\ j = 1, \ldots, d. \tag{4}$$

Thus, the eigenvalues can be decomposed in the sum of the contributions given by

- *the squared variance of the between class-component means,*

- *the average of the squared within class-component variances,*

along the corresponding directions.

This result provides an interpretation for the contribution of each direction to the visualization of the classification structure. Along each GMMDRC direction classes can be separated by location, by dispersion, or both. In addition, those directions associated with zero or approximately zero eigenvalues can be neglected since their contribution to class location or dispersion is negligible. A formal assessment of dimensionality could be pursued, for instance, by implementing a permutation test as described in Li (1991), however, is beyond the scope of this paper and deserves further study and investigation.

4.3 Estimation

For a $(n \times p)$ sample data matrix $\boldsymbol{X}$ and the corresponding $(n \times 1)$ vector Y containing the observed classes, the sample version $\widehat{\boldsymbol{M}}$ of (2) is obtained using the corresponding estimates from the fit of a Gaussian finite mixture model among those discussed in Section 3. Then, sample GMMDRC directions are calculated from the generalized eigen-decomposition of $\widehat{\boldsymbol{M}}$ with respect to $\widehat{\boldsymbol{\Sigma}}_X$, the sample marginal covariance matrix.

5 Examples

5.1 Waveform data

This is an artificial three-class problem with $p = 21$ variables, often used in machine learning and considered to be a difficult pattern recognition problem (Breiman et al, 1984; Hastie and Tibshirani, 1996). Consider the following three shifted triangular waveforms defined as

$$w_1(j) = \max(6 - |j - 11|, 0), \quad w_2(j) = w_1(j - 4), \quad w_3(j) = w_1(j + 4),$$

for $j = 1, \ldots, 21$. Then, the variables X_j are generated within each class Y as a random convex combination of two basic waveforms with noise added:

$$X_j = \begin{cases} u_1 w_1(j) + (1 - u_1)w_2(j) + \epsilon_j & \text{for } Y = 1 \\ u_2 w_2(j) + (1 - u_2)w_3(j) + \epsilon_j & \text{for } Y = 2 \\ u_3 w_3(j) + (1 - u_3)w_1(j) + \epsilon_j & \text{for } Y = 3 \end{cases},$$

where $j = 1, 2, \ldots, 21$, $w_h = (w_h(1), \ldots, w_h(21))^\top$ for $h = 1, 2, 3$, (u_1, u_2, u_3) be independent random variables uniformly distributed on $[0, 1]$, and ϵ_j following a standard normal distribution.

Figure 1 shows a scatterplot of data points projected onto the directions estimated for the EDDA mixture model with EEE covariance structure, i.e., assuming a common class covariance. As already mentioned, this is equivalent to a plot of LDA canonical variates. Panel (a) contains the density contours for the three classes, which have the same shape, orientation, and volume. The graph on panel (b) displays the corresponding decision boundaries with associated
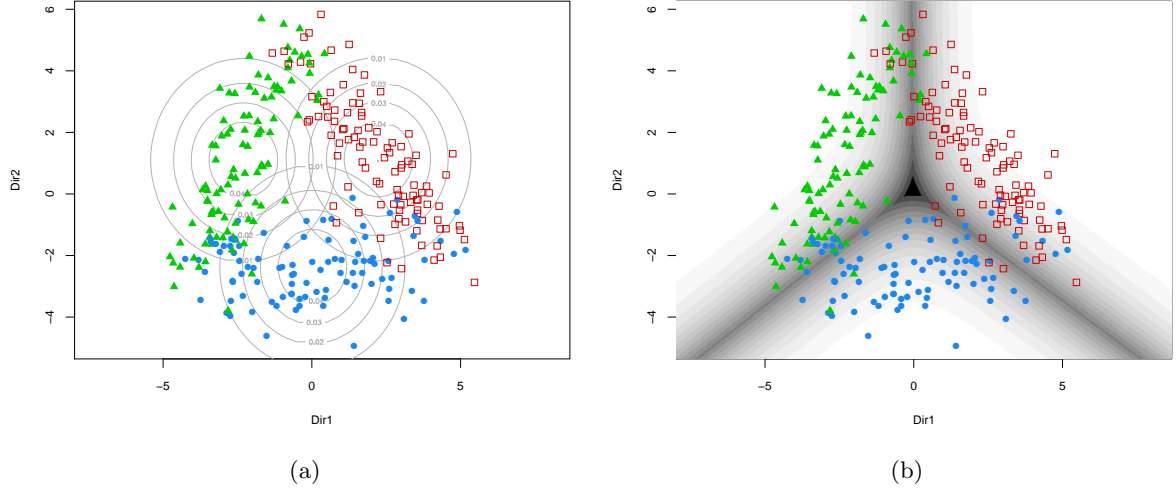
Figure 1: Plot of waveform data projected onto the first two directions for the EDDA model with common class covariance. Panel (a) shows the class density contours, while panel (b) the decision boundary with corresponding uncertainty.

classification uncertainty, with the uncertainty shown using a greyscale where darker regions indicating higher uncertainty. As expected the decision boundaries are linear.

Moving to a more complex model, we fitted an EDDA mixture model with VVV covariance structure, i.e., different class covariances. The corresponding projection is shown in Figure 2. In this case, the contours have different orientation (see panel (a)) because no restrictions were placed on the class covariance matrices, hence the estimated model provides a better approximation to the data distribution. The triangular form of the data appears more clearly than in the previous case. The plot on panel (b) contains the classification boundaries given by quadratic polygons, which shows a lower overall uncertainty than in the previous case.
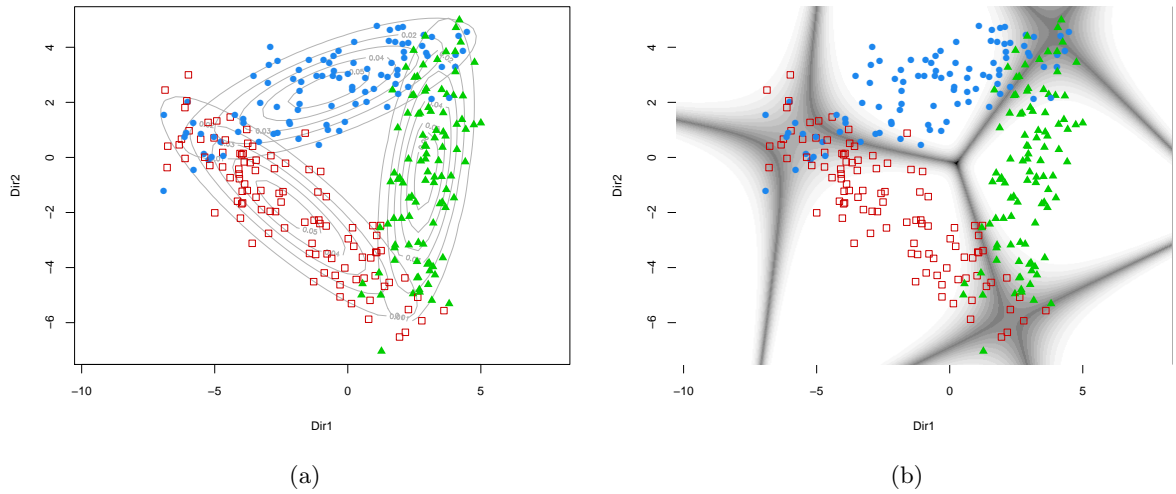

Figure 2: Plot of waveform data projected onto the first two directions for the EDDA model with different class covariances. Panel (a) shows the class density contours, while panel (b) the decision boundary with corresponding uncertainty.

Finally, Figure 3 shows the data projected onto the first two directions estimated from the selected MclustDA mixture model. This model fitted a mixture of three class-specific Gaussian mixtures where the class-specific mixtures had $G_1 = 3$, $G_2 = 4$, and $G_3 = 3$ spherical Gaussian distributions as components. These characteristics are clearly visible on panel (a) of Figure 3. The resulting decision boundaries are shown on Figure 3b; these appear to be highly nonlinear with a relative larger uncertainty where the classes overlap. Finally, note that, on the basis of the corresponding eigenvalues, the first two directions account for 96% of the overall information available in the 21 estimated directions.
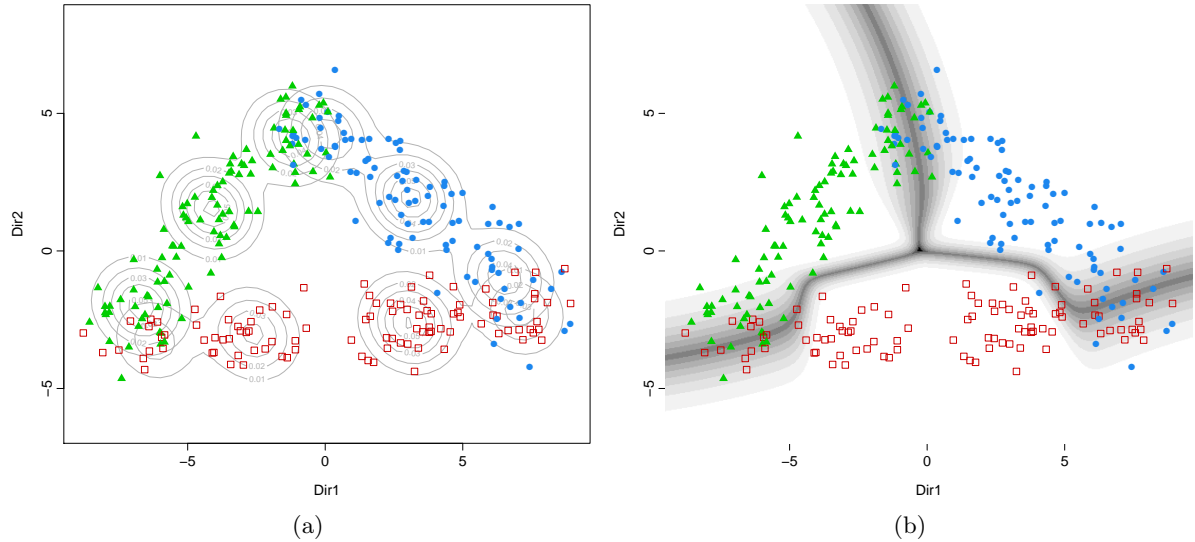

Figure 3: Plot of waveform data projected onto the first two directions for the selected MclustDA model. Panel (a) shows the class density contours, while panel (b) the decision boundary with corresponding uncertainty.

5.2 Swiss banknotes

Flury and Riedwyl (1988, Table 1.1 and 1.2) presented a dataset containing six physical measurements made on a sample of 1000 Swiss Franc bills. 100 observations were classified as genuines, and 100 as counterfeits.

The EDDA mixture model selected by BIC is a EEV model, which assume class covariance structures with different orientations but same volume and shape (see Table 1). Figure 4a shows the data projected onto the first two GMMDRC directions with the corresponding density contours; there appears a clear separation between classes with a larger variability for the group of counterfeit banknotes. The corresponding classification boundary is quadratic with an outlying genuine note classified as counterfeit (see panel (b) of Figure 4). The estimated subspace is quite similar to that obtained by SAVE, which we recall is equivalent to the one we would have obtained by fitting an EDDA mixture model with VVV covariance structure.

Fitting a MclustDA mixture model we obtain the graphs in panels (c) and (d) of Figure (4). The model selected by BIC uses a three components mixture with common covariance structure for the group of counterfeits, and a single component mixture model for the group of genuine notes. The latter appears as an homogeneous group, whereas the counterfeits are more heterogeneous with a clear separated cluster of observations (see panel (c)). Finally, panel (d) of Figure (4) shows the classification boundaries which are clearly nonlinear in this case and classify correctly all the observed banknotes.
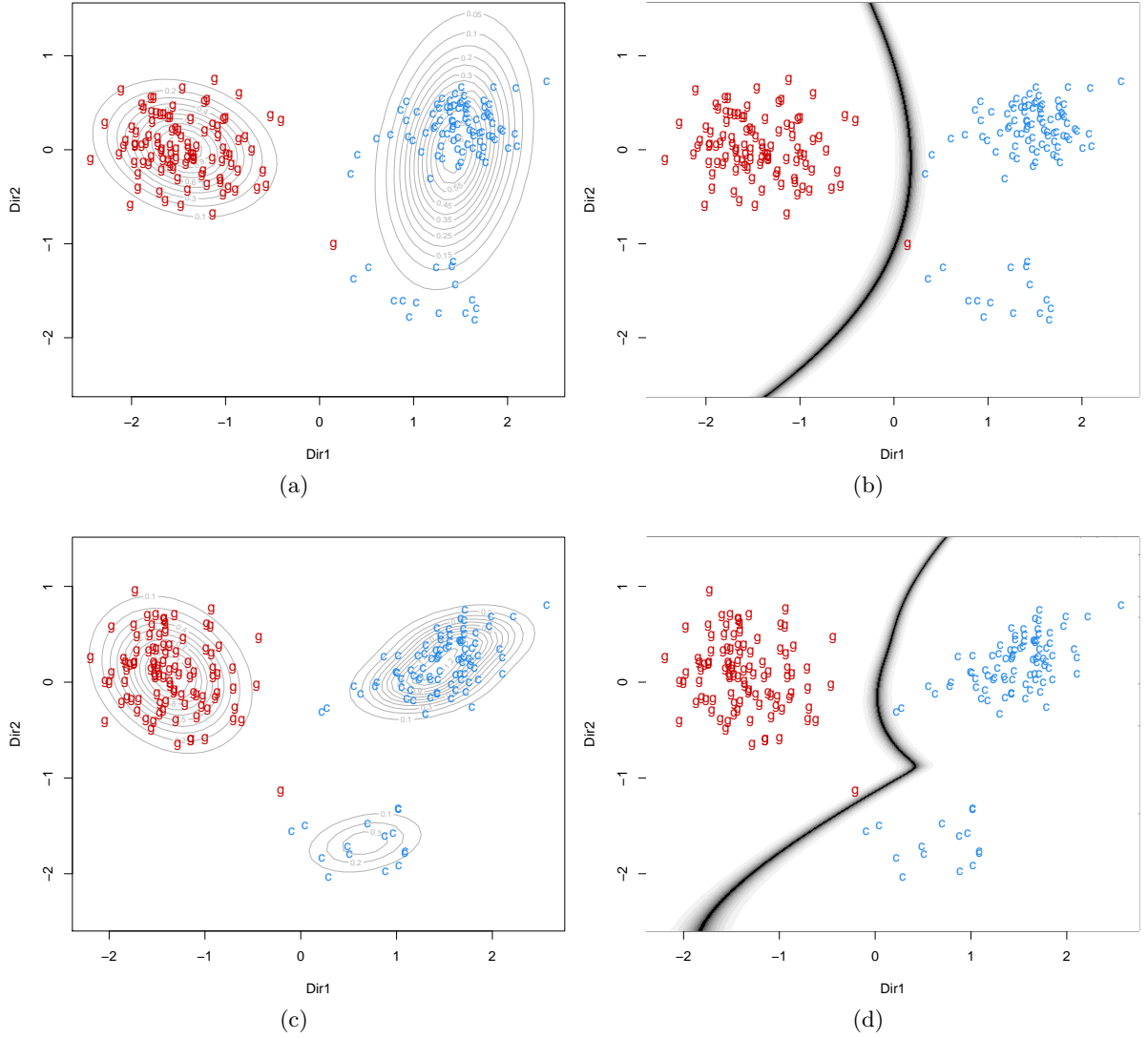
Figure 4: Plot of genuine (g) and counterfeit (c) Swiss banknotes projected onto the first two directions estimated for the "best" EDDA mixture model (top panels) and the "best" MclustDA mixture model (bottom panels). Panels (a) and (c) show the class density contours, while panels (b) and (d) plot the decision boundaries with corresponding uncertainty.

5.3 Simulated data with irrelevant and redundant features

GMMDRC directions are able to identify those variables which contain information about the classification structure. The estimated basis of the subspace is thus formed by linear combinations of the original features. However, when irrelevant and/or redundant correlated variables are present, the corresponding estimated coefficients have negligible values.

Consider the synthetic data example described in Maugis et al (2009, Sec. 6, scenario 5). A sample of size $n = 200$ is generated for a 10-dimensional ifeature vector. The first two variables are drawn from a mixture of four Gaussian distributions $\boldsymbol{x}_{[1:2]} \sim N(\boldsymbol{\mu}_k, \boldsymbol{I}_2)$ with $\boldsymbol{\mu}_1 = (-2, -2)$, $\boldsymbol{\mu}_2 = (-2, 2)$, $\boldsymbol{\mu}_3 = -\boldsymbol{\mu}_2$, $\boldsymbol{\mu}_4 = -\boldsymbol{\mu}_1$, and mixing probabilities $\pi = (0.3, 0.2, 0.3, 0.2)$. The remaining eight variables are simulated according to the model $\boldsymbol{x}_{[3:10]} = \boldsymbol{\beta}^\top \boldsymbol{x}_{[1:2]} + \epsilon$, where $\boldsymbol{\beta} = \begin{pmatrix} 0.5 & 0 & 2 & 0 \\ 0 & 1 & 0 & 3 \end{pmatrix} \boldsymbol{0}_4$, $\epsilon \sim N(\boldsymbol{0}_{10}, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} = \text{diag}(\boldsymbol{I}_2, 0.5\boldsymbol{I}_2, \boldsymbol{I}_4)$, and $\boldsymbol{0}_p$ is the $p \times p$ matrix of zeroes. For this scenario only the first two variables contain relevant information for

classification; the following four variables are correlated with the first two and therefore are redundant for classification purposes, whereas the remaining variables are independent both from the previous variables and from the classification.

The plot of the sample data projected onto the subspace spanned by the first two GMMDRC directions is presented in Figure 5, which also contains the table of coefficients defining the basis of the estimated subspace. The first two GMMDRC directions contain all the information pertaining to the partition of the classes, with the last direction clearly negligible based on the value of the corresponding eigenvalue. Furthermore, the coefficients defining the first two directions are very close to zero for all the variables except the first two, those which are really needed for classification.

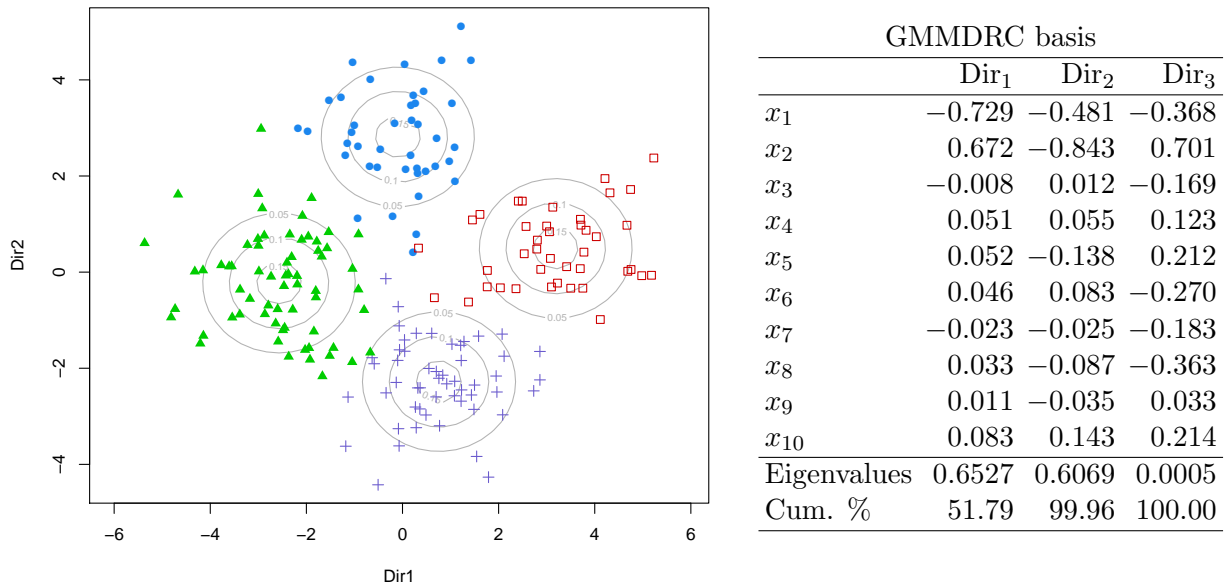|	GMMDRC basis		
	Dir_1	Dir_2	Dir_3
x_1	-0.729	-0.481	-0.368
x_2	0.672	-0.843	0.701
x_3	-0.008	0.012	-0.169
x_4	0.051	0.055	0.123
x_5	0.052	-0.138	0.212
x_6	0.046	0.083	-0.270
x_7	-0.023	-0.025	-0.183
x_8	0.033	-0.087	-0.363
x_9	0.011	-0.035	0.033
x_{10}	0.083	0.143	0.214
Eigenvalues	0.6527	0.6069	0.0005
Cum. %	51.79	99.96	100.00

Figure 5: Plot of simulated data, generated with irrelevant and redundant variables, projected onto the subspace spanned by the first two GMMDRC directions. The table at right shows the coefficients of the linear combinations that define the estimated directions with the corresponding eigenvalues.

5.4 High-dimensional data example

High-dimensional data represents a very challenging problem for many statistical methods, particularly when the number of available observations is small compared to the number of variables. Finite mixture models may be highly parameterized, thus fitting Gaussian mixtures to high-dimensional data requires some form of dimension reduction and/or some form of regularization. For a recent review see Bouveyron and Brunet-Saumard (2013).

Microarray data are an extreme case of high-dimensional data, for the reason that the number of variables (genes) is usually much larger than the number of observations (samples). For instance, consider the famous gene expression cancer dataset from Golub et al (1999). The data contain information on gene expressions in samples from human acute myeloid (AML) and acute lymphoblastic (ALL) leukemias obtained from high-density Affymetrix oligonucleotide arrays. There are 3571 genes and 38 samples: 27 in class ALL, and 11 in class AML. The samples in class ALL could be further split into B-cell and T-cell types. A preliminary filtering of genes, based on t-tests with p-values adjusted for multiple comparisons using the Benjamini and Hochberg (1995) method, selected a subset of 731 genes differentially expressed. Then, an MclustDA model was fitted on the selected subset assuming a common spherical covariance

matrix (EII) for each component within any class. From this model the matrices $\boldsymbol{M}_\mathsf{I}$ and $\boldsymbol{M}_\mathsf{II}$ can be estimated as discussed in Section 4.3. However, to apply the eigen-decomposition (3) we need a regularized estimate of the marginal covariance matrix. Several approaches could be adopted, but here for simplicity we used $\widehat{\boldsymbol{\Sigma}}_X = \mathrm{diag}[s_i^2]_{i=1,\ldots,p}$, where s_i^2 is the sample variance of the i-th gene. Such estimate ignores correlations between genes, which is not biologically realistic, but it has been shown to have no effect on classification accuracy (Dudoit et al, 2002).

Figure 6a shows a boxplot of AML and ALL samples projected along the first GMMDRC direction, which accounts for about 94% of total variation. A single direction is clearly able to separate the two types of cancer. However, the inclusion of the second direction, which accounts for another 4%, allows us to highlight an interesting feature previously not evident. Looking at Figure 6b we see that the group of ALL samples can be further divided into B-cell and T-cell tumour types along the second GMMDRC direction, except for one unusual B-cell which is very close to the group of T-cell samples.
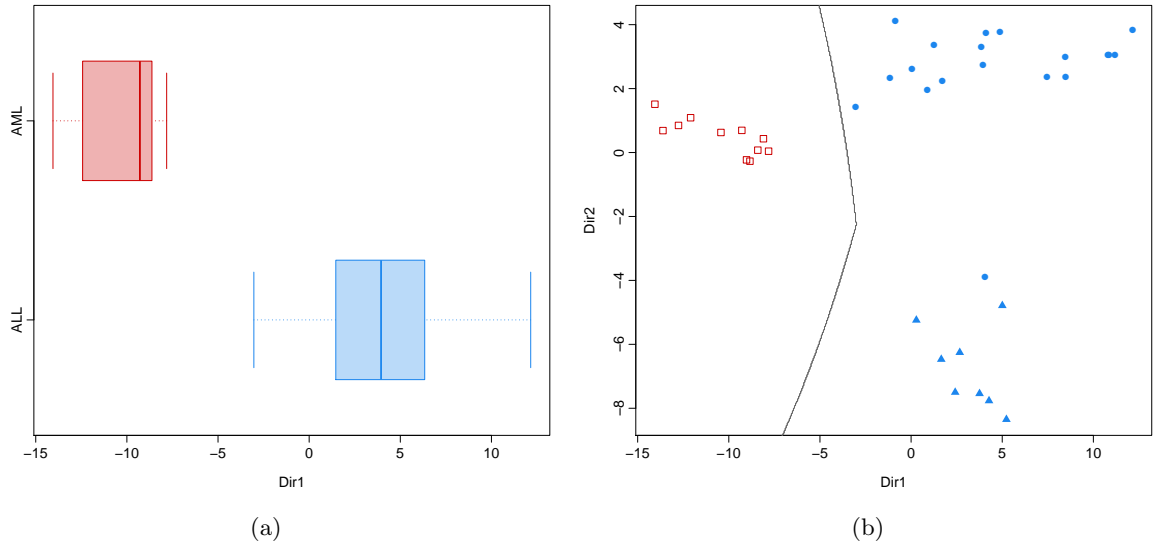

| (a) | (b) |

Figure 6: Plots of Golub data projected along the first GMMDRC direction (a) and the first two GMMDRC directions (b). Points are marked as □ for AML samples, as ● for ALL B-cell and ▲ for ALL T-cell samples.

6 Finding the most discriminant directions

The methodology introduced in Section 4 allows to visualize on a reduced subspace the underlying characteristics of class densities. However, if groups differ not only on location but also on dispersion, this second type of information may be dominant, and the classes would not appear clearly separated along the main directions.

If class separation is the goal, an appropriate modification of the kernel matrix (2) should be adopted, for instance, using the following convex linear combination

$$\boldsymbol{M} = \lambda \boldsymbol{M}_\mathsf{I} \boldsymbol{\Sigma}^{-1} \boldsymbol{M}_\mathsf{I} + (1 - \lambda) \boldsymbol{M}_\mathsf{II}, \tag{5}$$

where $0 \leq \lambda \leq 1$ is a tuning parameter. By choosing a large λ the estimated directions will focus more on differences on location. For $\lambda = 0.5$ we give equal weight to the two types of information, while for $\lambda = 1$ differences in class covariances are completely ignored. More generally, we could decide to optimize a measure of class separation, or minimize the uncertainty in classification.

Recently, Zhu and Hastie (2003) proposed a generalized log-likelihood-ratio (LR) statistic criterion to find the relevant directions for classification. They compare their proposal with SAVE on a simple bi-dimensional dataset with two groups. Figure 7a shows a data sample generated from this setting (for details see the mentioned paper). Zhu and Hastie (2003) argue that the relevant direction for discriminating the two groups corresponds to the first variable X_1, where there are differences in means. This turns out to be the direction selected by the LR criterion they proposed. On the contrary, the first direction selected by SAVE corresponds to X_2, a direction which contains only differences in variances, so the two groups do not appear well separated.
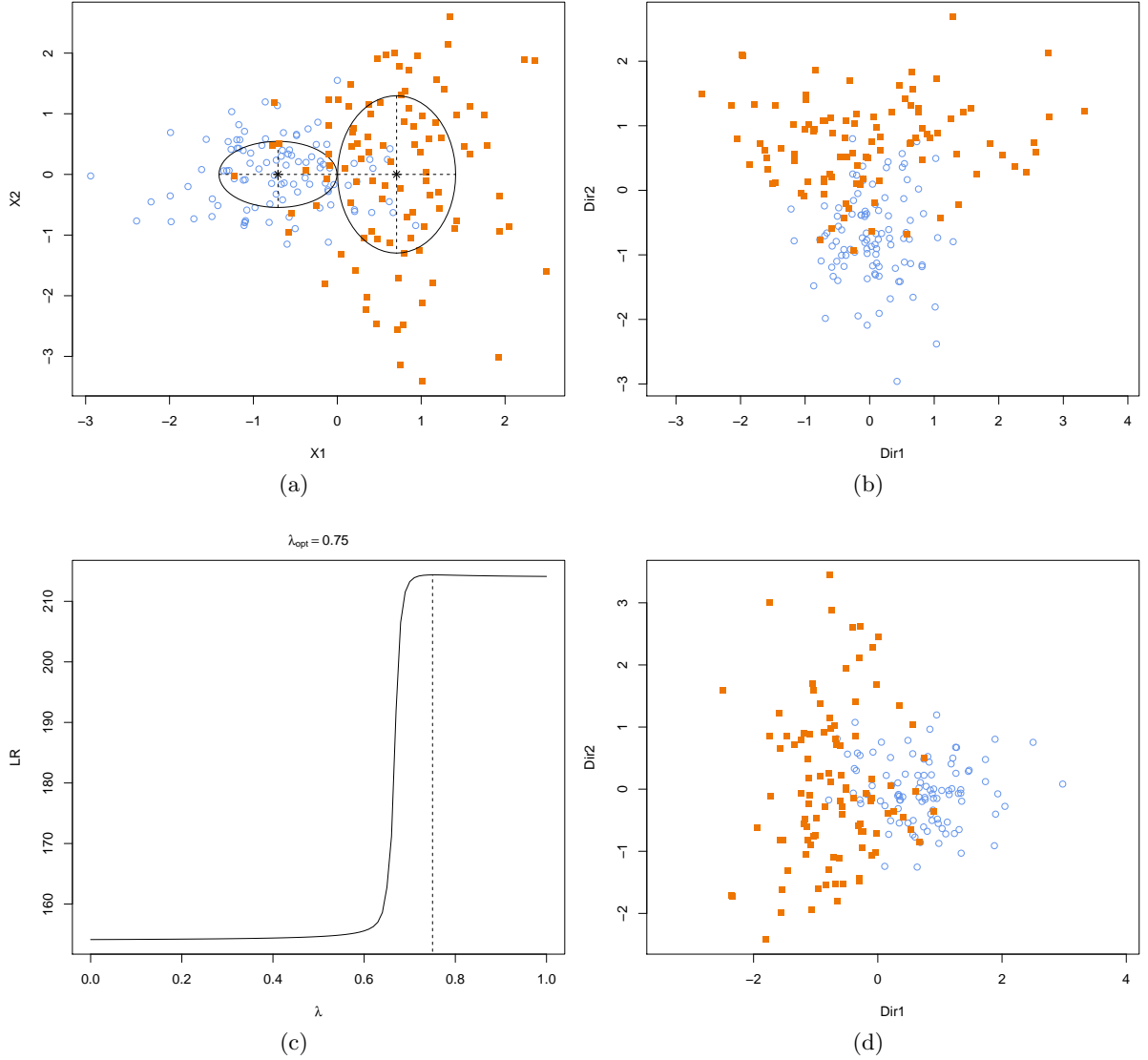

Figure 7: Sample data from Zhu and Hastie (2003) simulation scheme. Panel (a) contains the scatterplot of the two classification variables. Panel (b) and (d) show the data projected along the first two estimated GMMDRC directions with, respectively, $\lambda = 0.5$ (default) and $\lambda = 0.75$. The latter value has been selected on the basis of the LR criterion, whose trace is shown in panel (c).

We fit a Gaussian mixture model to such dataset after adding eight noise variables generated from independent standard normals. The first two directions appears to be needed with associated eigenvalues $(0.54923, 0.28002)$, which accounts for a total contribution of about 83%.

Figure 7b shows the data projected along such directions: the first direction correspond essentially to X_2, while the second direction to X_1. As for SAVE, the information coming from difference in variances is overwhelming that coming from difference in means, and this is what the plot shows. However, if our goal is to look for the most separating directions we can adopt the LR criterion for selecting the value of λ in equation (5). Figure 7c shows the trace of LR over a regular grid for λ: the optimal value is obtained for $\lambda = 0.75$ (or greater), which yields the projection shown in Figure 7d. Now, the first estimated direction is essentially equivalent to X_1, the most discriminating variable, while the second direction is equivalent to X_2.

Instead of optimizing the LR criterion as discussed above, we could adopt a different perspective based on dynamic graphics. For example, one could imagine to build a dynamic graph that, using a slider to manipulate the λ parameter, allows us to change interactively the data projection. In this way, a user would be able to appreciate the transition between the focus directed to differences in location to differences in the dispersion. Furthermore, depending on the purpose of the analysis, by tuning the λ parameter we could decide to highlight the structure of the classes and their characteristics, or to favor the separation of classes.

6.1 Ionosphere data

The ionosphere data were collected by 16 high-frequency antennas in Goose Bay, Labrador, Canada, and contain information about radar signals returned from the ionosphere. "Good" samples are those showing evidence of some type of structure in the ionosphere, while "bad" returns are those whose signals pass directly through the ionosphere and show no structure. A total of 351 signals were received, 225 were "good" returns and the remaining 126 were "bad" returns. The signals were processed using a function of 2 attributes for each of 17 pulse numbers that describe the complex electromagnetic signal. There are 34 continuous-valued feature variables, although one is a constant of all zeroes. The dataset is taken from the UCI Machine Learning Repository and it is available in the R package `mlbench`.

Figure 8a shows the data projected onto the first two GMMDRC directions using the default $\lambda = 0.5$ for a MclustDA mixture model having covariance structure VII with 4 components for the "bad" returns, and covariance VVV with 2 components for the "good" returns. On the basis of the graph it can be said that the two groups of signals differ mainly in the dispersion, with the "bad" returns showing a larger variance and "good" signals which are concentrated around the center of the graph.

These findings are similar to those obtained with SAVE by Pardoe et al (2007). To improve groups separation we selected the tuning parameter λ using the LR criterion discussed previously. The graph in Figure 8b shows the projection onto the first two GMMDRC directions estimated using the optimal value $\lambda = 1$. Here the separation between the two types of signal is clearly shown along the second direction, while the group of "good" signals appears to be composed of two separable sub-groups along the first direction. The latter is an interesting feature not previously recognized.

7 Final comments

The paper discussed a dimension reduction method for visualizing the classification structure and the geometric characteristics induced by a Gaussian mixture model. The methodology can also be easily adapted in order to recover the directions showing the maximum separation between the classes.

Although in the article we have used two-dimensional projections, the proposed method can, in principle, be easily extended to subspaces of higher dimensions. However, graphical representations on spaces of dimension greater than 3 can be quite difficult. Preliminary results for implementing a guided tour in 2-dimensions seem to be promising.
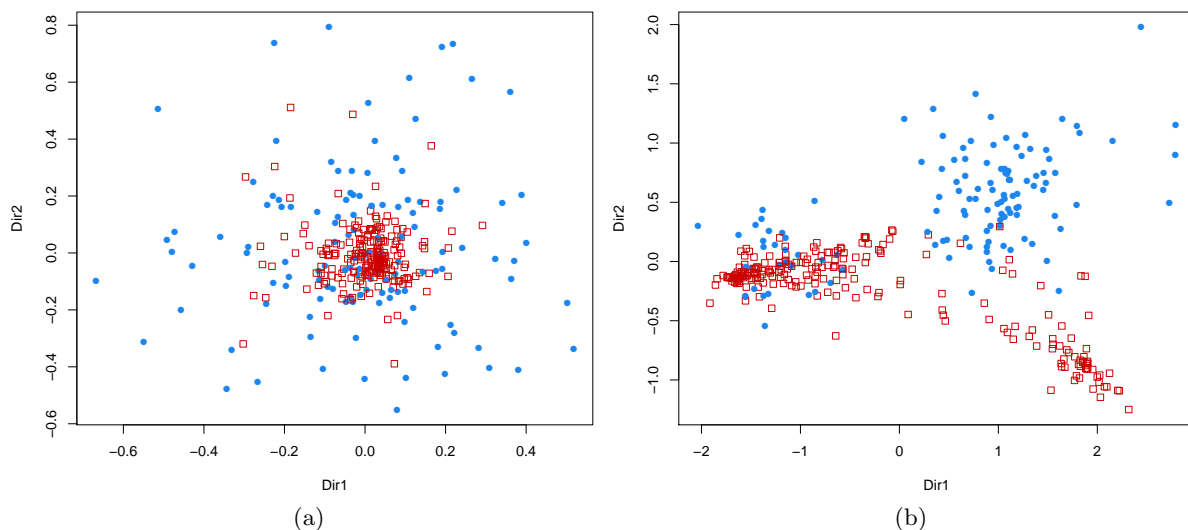
Figure 8: Plot of ionosphere data projected onto two different estimated subspaces: (a) using the default $\lambda = 0.5$; (b) using the optimal $\lambda = 1$ for groups separation. Points marked as $\square$ refer to "good" signals, those marked as • to "bad" signals.

The methodology and corresponding plots discussed in this paper are available in the `MclustDR` function of the R package `mclust` (Fraley et al, 2012).

References

Banfield J, Raftery AE (1993) Model-based Gaussian and non-Gaussian clustering. Biometrics 49:803–821

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B (Methodological) 57:289–300

Bensmail H, Celeux G (1996) Regularized Gaussian discriminant analysis through eigenvalue decomposition. Journal of the American Statistical Association 91:1743–1748

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE Trans Pattern Analysis and Machine Intelligence 22(7):719–725

Bouveyron C, Brunet-Saumard C (2013) Model-based clustering of high-dimensional data: A review. Computational Statistics & Data Analysis DOI 10.1016/j.csda.2012.12.008

Breiman L, Friedman J, Olshen R, C S (1984) Classification and Regression Trees. Wadsworth, New York

Celeux G, Govaert G (1995) Gaussian parsimonious clustering models. Pattern Recognition 28:781–793

Chen CH, Li KC (2001) Generalization of fisher's linear discriminant analysis via the approach of sliced inverse regression. Journal of the Korean Statistical Society 30:193–217

Cook DR, Forzani L (2009) Likelihood-based sufficient dimension reduction. Journal of the American Statistical Association 104(485):197–208

Cook RD, Weisberg S (1991) Discussion of Li (1991). Journal of the American Statistical Association 86:328–332

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm (with discussion). Journal of the Royal Statistical Society, Series B: Statistical Methodology 39:1–38

Dudoit S, Fridlyand J, Speed T (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97:77–87

Fisher RA (1936) The use of multiple measurements in taxonomic problems. Annals of Eugenics 7:179–188

Flury B, Riedwyl H (1988) Multivariate Statistics: a Practical Approach. Chapman & Hall Ltd

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. Journal of the American Statistical Association 97(458):611–631

Fraley C, Raftery AE, Murphy TB, Scrucca L (2012) MCLUST version 4 for R: Normal mixture modeling for model-based clustering, classification, and density estimation. Technical Report 597, Department of Statistics, University of Washington

Friedman JH (1989) Regularized discriminant analysis. Journal of the American Statistical Association 84:165–175

Golub T, Slonim D, Tamayo P, Huard C, Gaasenbeek M, Mesirov J, Coller H, Loh M, Downing JR, Caligiuri M, Bloomfield C, Lander E (1999) Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science 286:531–537

Hastie T, Tibshirani R (1996) Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58(1):155–176

Hastie T, Tibshirani R, Buja A (1994) Flexible discriminant analysis by optimal scoring. Journal of the American Statistical Association 89:1255–1270

Hastie T, Buja A, Tibshirani R (1995) Penalized discriminant analysis. Annals of Statistics 23:73–102

Hennig C (2004) Asymmetric linear dimension reduction for classification. Journal of Computational and Graphical Statistics 13(4):930–945

Kent JT (1991) Discussion of Li (1991). Journal of the American Statistical Association 86:336–337

Li KC (1991) Sliced inverse regression for dimension reduction (with discussion). Journal of the American Statistical Association 86:316–342

Mardia K, Kent J, Bibby J (1979) Multivariate Analysis. Academic Press, London

Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with gaussian mixture models. Biometrics 65(3):701–709

Pardoe I, Yin X, Cook R (2007) Graphical tools for quadratic discriminant analysis. Technometrics 49(2):172–183

R Core Team (2013) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/

Schwartz G (1978) Estimating the dimension of a model. Annals of Statistics 6:31–38

Scrucca L (2010) Dimension reduction for model-based clustering. Statistics and Computing 20(4):471–484, DOI 10.1007/s11222-009-9138-7

Velilla S (2008) A method for dimension reduction in quadratic classification problems. Journal of Computational and Graphical Statistics 17(3):572–589

Velilla S (2010) On the structure of the quadratic subspace in discriminant analysis. Journal of Multivariate Analysis 101(5):1239–1251

Zhu M, Hastie TJ (2003) Feature extraction for nonparametric discriminant analysis. Journal of Computational and Graphical Statistics 12(1):101–120

Proof of proposition 1

Assume an EDDA mixture model with common full class covariance matrix. The last condition implies that the matrix $\boldsymbol{M}_{\mathsf{II}}$ in equation (2) cancels out, so the kernel matrix simplifies to

$$\boldsymbol{M} = \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_{\mathsf{I}},$$

where $\boldsymbol{M}_{\mathsf{I}} = \sum_{k=1}^{K} \pi_k (\boldsymbol{\mu}_k - \boldsymbol{\mu})(\boldsymbol{\mu}_k - \boldsymbol{\mu})^\top = \boldsymbol{\Sigma}_B$, the between-class covariance matrix. The basis of the subspace $\mathcal{S}(\boldsymbol{\beta})$ provided by GMMDRC is obtained as the solution of the following problem

$$\boldsymbol{M}\boldsymbol{\beta}_j = l_j \boldsymbol{\Sigma}_X \boldsymbol{\beta}_j,$$

with $l_1 \geq \ldots \geq l_d$, and $d = \min(p, K-1)$. Thus, $\boldsymbol{\beta}_j$ is the jth eigenvector associated to the jth largest eigenvalue l_j $(j = 1, \ldots, d)$ of the $(p \times p)$ matrix

$$\begin{aligned}
\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{M} \boldsymbol{\Sigma}_X^{-1/2} &= \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1/2} \\
&= (\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2})^\top (\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2}).
\end{aligned}$$

The subspace estimated by SIR is obtained as the solution of

$$\boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} = l_j^{\mathsf{SIR}} \boldsymbol{\Sigma}_X \boldsymbol{\beta}_j^{\mathsf{SIR}} \tag{6}$$

which is given by the eigen-decomposition of $\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2}$. It is easily seen that $\boldsymbol{\beta}_j = \boldsymbol{\beta}_j^{\mathsf{SIR}}$ and $l_j = (l_j^{\mathsf{SIR}})^2$, for $j = 1, \ldots, d$. Thus, the basis of the subspace provided by GMMDRC under model EDDA with full common class covariance matrix is equivalent to the basis estimated by SIR.

We now consider the relation of GMMDRC with LDA canonical variates. From (6), we may subtract $l_j^* \boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}}$ from both side and, recalling the decomposition of the total variance, $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$, we may write

$$\begin{aligned}
\boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} - l_j^{\mathsf{SIR}} \boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} &= l_j^{\mathsf{SIR}} \boldsymbol{\Sigma}_X \boldsymbol{\beta}_j^{\mathsf{SIR}} - l_j^{\mathsf{SIR}} \boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} \\
(1 - l_j^{\mathsf{SIR}}) \boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} &= l_j^{\mathsf{SIR}} (\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_B) \boldsymbol{\beta}_j^{\mathsf{SIR}} \\
\boldsymbol{\Sigma}_B \boldsymbol{\beta}_j^{\mathsf{SIR}} &= l_j^{\mathsf{SIR}} / (1 - l_j^{\mathsf{SIR}}) \boldsymbol{\Sigma}_W \boldsymbol{\beta}_j^{\mathsf{SIR}}.
\end{aligned}$$

It is clear that $l_j^{\mathsf{SIR}} / (1 - l_j^{\mathsf{SIR}})$ and $\boldsymbol{\beta}_j^{\mathsf{SIR}}$ are, respectively, the jth eigenvalue and the associated eigenvector of $\boldsymbol{\Sigma}_W^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_W^{-1/2}$, the decomposition solving the Rayleigh quotient used to derive canonical variates in LDA. Thus, the basis of the subspace $\mathcal{S}(\boldsymbol{\beta}^{\mathsf{LDA}})$ is equivalent to $\mathcal{S}(\boldsymbol{\beta}^{\mathsf{SIR}})$, which in turn is equivalent to that provided by GMMDRC under the specific model assumption.

Proof of proposition 2

The kernel matrix of SAVE can be written in the original scale of the variables as

$$M_{\mathsf{SAVE}} = \sum_{k=1}^{K} \omega_k \left(\boldsymbol{I}_p - \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_k \boldsymbol{\Sigma}_X^{-1/2} \right)^2.$$

Recalling that $\boldsymbol{\Sigma}_X = \boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W$, we may write the expression within parenthesis as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_X^{-1/2} (\boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_X^{-1/2} &= \boldsymbol{\Sigma}_X^{-1/2} (\boldsymbol{\Sigma}_B + \boldsymbol{\Sigma}_W - \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_X^{-1/2} \\
&= \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2} + \boldsymbol{\Sigma}_X^{-1/2} (\boldsymbol{\Sigma}_W - \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_X^{-1/2}.
\end{aligned}$$

Then,

$$\begin{aligned}
M_{\mathsf{SAVE}} &= \sum_{k=1}^{K} \omega_k \left(\boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2} + \boldsymbol{\Sigma}_X^{-1/2} (\boldsymbol{\Sigma}_W - \boldsymbol{\Sigma}_k) \boldsymbol{\Sigma}_X^{-1/2} \right)^2 \\
&= \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1} \boldsymbol{\Sigma}_B \boldsymbol{\Sigma}_X^{-1/2} + \\
&\quad \boldsymbol{\Sigma}_X^{-1/2} \left(\sum_{k=1}^{K} w_k (\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_W) \boldsymbol{\Sigma}_X^{-1} (\boldsymbol{\Sigma}_k - \boldsymbol{\Sigma}_W)^\top \right) \boldsymbol{\Sigma}_X^{-1/2} \\
&= \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1/2} + \boldsymbol{\Sigma}_X^{-1/2} \boldsymbol{M}_{\mathsf{II}} \boldsymbol{\Sigma}_X^{-1/2} \\
&= \boldsymbol{\Sigma}_X^{-1/2} (\boldsymbol{M}_{\mathsf{I}} \boldsymbol{\Sigma}_X^{-1} \boldsymbol{M}_{\mathsf{I}} + \boldsymbol{M}_{\mathsf{II}}) \boldsymbol{\Sigma}_X^{-1/2},
\end{aligned}$$

where $\boldsymbol{M}_{\mathsf{I}}$ and $\boldsymbol{M}_{\mathsf{II}}$ are those obtained from an EDDA Gaussian mixture model with a single component for each class and different class covariance matrices (VVV).

Proof of proposition 3

The proof is analogous to that provided for Prop. 2 in Scrucca (2010) and it is not replicated here.