Sparse Discriminant Analysis

Line Clemmensen*  Trevor Hastie**  Bjarne Ersbøll*

*Department of Informatics and Mathematical Modelling
Technical University of Denmark
Kgs. Lyngby, Denmark

**Statistics Department, Stanford University, California, U.S.A.

March 3, 2010

Abstract

Classification in high-dimensional feature spaces where interpretation and dimension reduction are of great importance is common in biological and medical applications. For these applications standard methods such as microarrays, 1D NMR, and spectroscopy have become everyday tools for measuring thousands of features in samples of interest. The samples are often costly and therefore many problems have few observations in relation to the number of features to be measured. Traditionally data are analyzed by first performing a feature selection before classification. We propose a method which performs
linear discriminant analysis with a sparseness criterion imposed such that the classification, feature selection and dimension reduction are merged into one analysis. The sparse discriminant analysis is faster than traditional feature selection methods based on computationally heavy criteria such as Wilk’s lambda, and the results are better with regard to classification rates and sparseness. The method is extended to mixtures of Gaussians and this is in general useful when boundaries between classes are non-linear and in particular when clusters are present within each class, e.g. biological subgroups. Finally, the methods proposed provide low-dimensional views of the discriminative directions.

1 Introduction

Linear discriminant analysis (LDA) is a favored tool for supervised classification in many applications, due to its simplicity and robustness. Comparison studies show that a large percentage (typically more than 90%) of the achievable improvement in predictive accuracy, over the simple baseline model, is achieved by LDA (Hand, 2006). Furthermore, the computations leading towards LDA provide low-dimensional projections of data onto the most discriminative directions. However, LDA fails in some situations:

- When the number of predictor variables is high in relation to the number of observations \( p \gg n \).
When a single prototype per class is insufficient.

- When linear boundaries are insufficient in separating the classes.

These situations where LDA fails were previously addressed in penalized discriminant analysis (Hastie et al., 1995a) and discriminant analysis by Gaussian mixtures (Hastie and Tibshirani, 1996), see also flexible discriminant and mixture models (Hastie et al., 1995b). However, in some cases, where \( p \gg n \), these methods are not adequate since both sparseness and feature selection are desired. A low number of non-zero parameters ensures a better interpretation of the model and additionally tends to overfit training data less than non-sparse methods, as illustrated with the elastic net and sparse principal components (Zou and Hastie, 2005; Zou et al., 2006).

It is often desirable to perform feature selection in biological or medical applications such as microarrays. In these applications it is essential to identify important features for the problem at hand for interpretation issues and to improve speed by using models with few non-zero loadings as well as fast algorithms.

During the past decade, problems in which the number of features is much larger than the number of observations have received much attention (Donoho, 2000; Hastie et al., 2009; Duda et al., 2001). Here we consider classification problems and propose a method for performing robust discriminant analysis. Previously this issue has been addressed by ignoring correlations between features and assuming independence in the multivariate Gaussian...
model (naive Bayes) (Bickel and Levina, 2004). We will focus on imposing sparseness in the model (Donoho, 2000) in line with models such as lasso and the elastic net (Tibshirani, 1996; Efron et al., 2004; Zou and Hastie, 2005). Sparseness is imposed by an $\ell_1$-norm on the parameters, like in Efron et al. (2004); Tibshirani (1996); Trendafilov and Jolliffe (2007). However, also imposing an $\ell_2$-norm like in the elastic net (Zou and Hastie, 2005) ensures that good solutions are also obtained when the number of selected features exceeds the number of variables.

The introduction of a sparseness criterion is well known in the regression framework (Tibshirani, 1996; Efron et al., 2004; Zou and Hastie, 2005; Zou et al., 2006) and we shall therefore consider computing the projections for LDA by optimal scoring which computes the projections (discriminant directions) by regression (Hastie et al., 1995a; Ye, 2007; Grosenick et al., 2008; Leng, 2008). Furthermore, the optimal scoring framework allows for an extension to mixtures of Gaussians (Hastie and Tibshirani, 1996).

The paper is organized as follows. Section two describes the sparse LDA and sparse mixture discriminant analysis algorithms. Section three briefly describes elastic net regression of dummy variables, shrunken centroids regularized discriminant analysis and sparse partial least squares, which are used for comparison. Section four illustrates experimental results on a small illustrative shape-based data set of female and male silhouettes and on four high-dimensional data sets: a microarray data set, spectral and chemical identification of fungi as well as classification of fish species based on shape
and texture features. We round off with a discussion in section five.

2 Methodology

Linear discriminant analysis (LDA) is a classification method which assumes that the variables in each of the $k$ classes are normally distributed with means $\mu_j$, $j = 1, ..., k$ and equal dispersion $\Sigma$ (see e.g. Hastie et al. (2009)). Reduced-rank LDA has the ability to provide low-dimensional views of data of up to at most $k-1$ dimensions (Hastie et al., 2009). These views, also called discriminant directions, are furthermore sorted such that the direction discriminating the classes most is first and so forth. The at most $k-1$ directions, $\beta_j$s, are $p$-dimensional vectors, where $p$ is the number of predictor variables. They are chosen to maximize the variance between classes and minimize the variance within classes subject to being orthogonal to each other. Hence, we maximize the between-groups sums of squares, $\Sigma_B = \sum_{j=1}^k (\mu_j - \mu)(\mu_j - \mu)^T$ (where $\mu$ is the mean of all groups) relative to the within-groups sums of squares matrix, $\Sigma_W = \sum_{j=1}^k \sum_{i=1}^{n_j} (X_{ij} - \mu_j)(X_{ij} - \mu_j)^T$ (Fisher’s criterion)

$$\arg \max_{\beta_j} \beta_j^T \Sigma_B \beta_j$$ (1)

under the orthogonality constraint

$$\beta_j^T \Sigma_W \beta_l = \begin{cases} 0 & l = 1, ..., j - 1 \\ 1 & l = j \end{cases},$$ (2)
to find the discriminating directions $\beta_j$, $j = 1, ..., k - 1$.

The methodology section is written following the notation of Penalized Discriminant Analysis (PDA) in Hastie et al. (1995a). PDA replaces the within sums of squares matrix in (2) with the penalized term $\Sigma_W + \lambda_2 \Omega$. In order to obtain sparseness in the solution we introduce an extra term which controls the $\ell_1$-norm of the parameters $\beta$. The $\ell_1$-norm has previously proved to be an effective regularization term for obtaining sparseness; see methods such as lasso, elastic net and sparse principal component analysis (Tibshirani, 1996; Zou and Hastie, 2005; Zou et al., 2006). The sparse discriminant criterion then becomes

$$\arg \max_{\beta_j} \beta_j^T \Sigma_B \beta_j - \lambda_1 \sum_{i=1}^p |\beta_{ji}|$$

under the constraint (2) with the penalized within sums of squares matrix $\Sigma_{W_p} = \Sigma_W + \lambda_2 \Omega$ replacing $\Sigma_W$.

The elastic net proposed by Zou and Hastie (2005) solves a regression problem regularized by the $\ell_2$-norm and the $\ell_1$-norm in a fast and effective manner. The elastic net is defined as

$$\beta_j^{en} = \arg \min_{\beta_j} (\|y - X\beta_j\|_2^2 + \lambda_2 \|\beta_j\|_2^2 + \lambda_1 \|\beta_j\|_1) .$$

As the sparse discriminant criterion is also regularized by an $\ell_2$-norm and an $\ell_1$-norm penalty, it seems advantageous to rewrite the criterion as a regression type problem in order to use the elastic net algorithm for solving SDA.
LDA was rewritten in Hastie et al. (1995a) as a regression type problem using optimal scoring; see also the relation between optimal scoring, canonical discriminant analysis and linear discriminant analysis in the appendix A.1. The idea behind optimal scoring is to turn categorical variables into quantitative variables using a linear operator (the categorical variables will here be encoded as \{0, 1\} dummy variables). Optimal scoring assigns a score, \(\theta_{ji}\) for each class \(i\) and for each parameter vector \(\beta_j\). The optimal scoring problem is defined as

\[
(\hat{\theta}, \hat{\beta})_{\text{os}} = \arg \min_{\theta, \beta} n^{-1} \|Y \theta - X \beta\|_2^2 \\
\text{s.t. } n^{-1} \|Y \theta\|_2^2 = 1 ,
\]

where \(Y\) is a matrix of dummy variables representing the \(k\) classes, \(\theta\) is the \(k \times q\) matrix of scores, and \(\beta\) is a \(p \times q\) matrix where the columns are the \(\beta_j\)s. The optimal scoring constructs \(q\) quantitative variables given by \(Y \theta\). These give a continuous ordering of the samples in the \(k\) classes such that we can regress these quantitative variables directly on the predictors.

PDA adds a penalty of \(\beta_j^T \Omega \beta_j\) to the optimal scoring problem such that the penalized optimal scoring criterion becomes

\[
(\hat{\theta}, \hat{\beta})_{\text{pos}} = \arg \min_{\theta, \beta} n^{-1} \|Y \theta - X \beta\|_2^2 + \lambda_2 \|\Omega^{1/2} \beta\|_2^2 \\
\text{s.t. } (6),\text{ where } \Omega\text{ is a symmetric and positive definite matrix. In this paper,}
\]
a sparseness criterion is added to the penalized optimal scoring criterion via the $\ell_1$-norm of the regression parameters $\beta$. The normal equations can thus no longer be applied and it is not possible to solve the sparse discriminant analysis (SDA) problem in one regression and one eigenvalue decomposition step as is the case for PDA. We propose an iterative algorithm for solving SDA. Extending the method to mixtures of Gaussians is straightforward in line with Hastie and Tibshirani (1996).

Since the elastic net (Zou and Hastie, 2005) is used in the algorithm, we will assume that data are normalized, i.e. the features are transformed to have zero mean and length one. The elastic net algorithm uses the correlation between the dependent variable and the predictors to decide which variable to activate in each iteration. Running the algorithm on raw data is comparable to performing principal component analysis on the covariance matrix rather than the correlation matrix. It is therefore recommended that data should always first be normalized.

2.1 Sparse discriminant analysis by optimal scoring

In this section we introduce constraints to the optimal scoring problem in (5) in order to obtain sparseness in the PDA. As before, the score vector $\theta_j$ assigns a real number $\theta_{ji}$ for each class $i$, $i = 1, ..., k$. The scored training data $Y\theta$ is an $n \times q$ matrix which we will regress on the matrix of predictors $X_{n \times p}$ to obtain the parameters or directions $\beta_{p \times q}$. This leads to $q$ components of sparse discriminative directions, where $q \leq k$ since there are up to $k$ non-
trivial directions in the sparse optimal scoring problem. We define sparse optimal scoring as

\[(\theta, \beta)^{sos} = \arg\min_{\theta, \beta} n^{-1} (\|Y\theta - X\beta\|_2^2 + \lambda_2\|\Omega^{1/2}\beta\|_2^2 + \lambda_1\|\beta\|_1) \]
\[s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1, \]

(8)

where \(\Omega\) is a penalization matrix, as introduced in PDA (Hastie et al., 1995a). The \(\ell_1\)-norm of \(\beta\) introduces sparseness as in lasso or elastic net regularization. In appendix A.1 the relation between sparse discriminant analysis (3) and sparse optimal scoring (8) is given.

For fixed \(\theta\) we obtain:

\[\beta_j^{sos} = \arg\min_{\beta_j} n^{-1} (\|Y\theta_j - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \]

(10)

which for \(\Omega = I\) is an elastic net problem. We will later rewrite the elastic net for more general penalty matrices. For fixed \(\beta\) the optimal scores are

\[\theta^{os} = \arg\min_{\theta} n^{-1}\|Y\theta - X\beta\|_2^2 \]
\[s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1. \]

(11)

Set \(D = n^{-1}Y^TY\) which is a diagonal matrix of the class proportions. Then the constraint (9) can be written as \(\theta^T D \theta = I\) and setting \(\theta^* = D^{1/2} \theta\) we can
solve the following problem instead.

\[
\hat{\theta}^* = \arg\min_{\theta^*} n^{-1} \| Y D^\frac{1}{2}_\pi \theta^* - \hat{Y} \|_2^2
\]

s.t. \( \| \theta^* \|_2^2 = 1 \),

where \( \hat{Y} = X\beta \). This is a balanced Procrustes problem when \( Y \) and \( \hat{Y} \) have the same dimensions (for \( q = k \)). If \( q < k \) we pad \( \hat{Y} \) with zeros, so that \( \hat{Y} = [X\beta \ 0] \). The problem can then be solved by taking the svd of \( D^{-\frac{1}{2}}_\pi Y^T \hat{Y} \), as described in Elden and Park (1999). As we only need to estimate \( U \) and \( V \) of the svd in order to obtain a solution, and \( D^{-\frac{1}{2}}_\pi \) is a diagonal matrix, taking the svd of \( Y^T \hat{Y} = USV^T \) suffices, and the solution becomes

\[
\hat{\theta}^* = UV^T \leftrightarrow \\
\hat{\theta} = D^{-\frac{1}{2}}_\pi UV^T .
\]

By analogy with the PDA case, we use heuristics from suitable normal assumptions as guidelines for producing posterior probabilities and a classifier. As a graphical projection of a predictor vector \( x \) we use the set of fits \( \beta^T x \). A nearest class mean rule, where ”nearest” is measured using \( \Sigma W \), is applied in the \( q \leq k \) reduced-dimensional discriminant subspace to obtain class labels.

The supplementary material introduces a modification of the elastic net algorithm to include various penalizing matrices.
2.2 Sparse Discriminant Algorithm

The SDA algorithm using optimal scores and the modified elastic net is described in Algorithm 1.

Algorithm 1 Sparse Discriminant Analysis:

1. Initialize \( \theta = (k \sum_{j=1}^{k} D_{\pi,ij})^{-1} I_{1q} \).
2. For \( j = 1, \ldots, q \) solve the modified elastic net problem with fixed \( \theta \)
   \[
   \beta_j = \arg \min_{\beta_j} n^{-1}(\|Y \theta_j - X \beta_j\|_2^2 + \lambda_2 \beta_j^T \Omega \beta_j + \lambda_1 \|\beta_j\|_1)
   \] (16)
3. For fixed \( \beta \) and \( Y^T \hat{Y} = USV^T \) compute the optimal scores from \( \hat{\theta} = D_{\pi}^{-\frac{1}{2}} UV^T \).
4. Repeat steps 2 and 3 until convergence or a maximum number of iterations is reached.
5. Remove the last \( k - m \) trivial directions, where the \( (m + 1)^{th} \) singular value \( S_{m+1} < \epsilon \) (\( \epsilon \) is some small threshold value). For \( j = 1, \ldots, m \) solve the modified elastic net problem with fixed \( \theta \), the \( m \) sparse discriminant directions are now ordered according to the singular values and thereby degree of discrimination.

The sparse discriminant analysis algorithm has a computational effort similar to that of sparse principal component analysis (Zou et al., 2006). It likewise performs an elastic net step and an SVD in each iteration. The elastic net step for \( p \gg n \) has the highest computational cost, which is in the order of \( qO(pnm + m^3) \), where \( m \) is the number of non-zero coefficients. This can be massive if \( p \) and \( m \) are large. However, in general few non-zero coordinates are desired in the applications mentioned, and the algorithm therefore becomes
very effective. Additionally, the number of iterations needed to obtain a good solution is generally small.

Due to the sparseness the convergence can in some cases be slow as the algorithm may alternate between various sets of active variables which give equally good solutions. However, as either set of variables give good solutions the algorithm can be stopped without full convergence with good results.

2.3 Sparse mixture of Gaussians

Instead of representing each class by a single prototype, we now represent each class by a mixture of Gaussians. We divide each class \( j \) into \( R_j \) subclasses and define the total number of subclasses \( R = \sum_{j=1}^{k} R_j \). To limit the number of parameters we consider a Gaussian mixture model where each subclass has its own mean \( \mu_{jr} \) and common covariance matrix \( \Sigma \). Since the single prototype problem is formulated as an optimal scoring problem it is straightforward to extend it to mixtures of Gaussians in line with Hastie and Tibshirani (1996). Instead of using an indicator response matrix \( Y \), we use a blurred response matrix \( Z_{n \times R} \), which consists of the subclass probabilities, \( z_{ijr} \) for the \( i^{th} \) observation, the \( j^{th} \) class and the \( r^{th} \) subclass. Let \( \pi_{jr} \) be the mixing probability within the \( r^{th} \) subclass within the \( j^{th} \) class, and \( \sum_{r=1}^{R_j} \pi_{jr} = 1 \). Recall the EM steps of using Bayes’ theorem to model Gaussian mixtures. The estimation steps of the subclass probabilities, \( z_{ir} \) and the
mixing probabilities, $\pi_{jr}$ are

$$ z_{ijr} = \frac{\pi_{jr} \exp\{-\frac{(X_i - \mu_{jr})\Sigma^{-1}(X_i - \mu_{jr})^T}{2}\}}{\sum_{r=1}^{R_j} \pi_{jr} \exp\{-\frac{(X_i - \mu_{jr})\Sigma^{-1}(X_i - \mu_{jr})^T}{2}\} }$$

$$(17)$$

$$ \pi_{jr} = \sum_{i \in g_{jr}} z_{ijr}, \quad \sum_{r=1}^{R_j} \pi_{jr} = 1 $$

$$(18)$$

where $g_{jr}$ is the subset of observations in the $r^{th}$ subclass of the $j^{th}$ class.

The maximization steps of the mean and the covariance matrix are

$$ \mu_{jr} = \frac{\sum_{i \in g_{jr}} x_i z_{ijr}}{\sum_{i \in g_{jr}} z_{ijr}} $$

$$(19)$$

$$ \Sigma = \eta^{-1} \sum_{j=1}^{k} \sum_{i \in g_{jr}} \sum_{r=1}^{R_j} z_{ijr} (x_i - \mu_{jr})(x_i - \mu_{jr})^T $$

$$(20)$$

We now write the SMADA algorithm by computing $Q \leq R$ sparse directions for the subclasses in the mixture of Gaussians model as described in algorithm 2.

The graphical projections are again given by the sets of fits $x^T\beta$, and a decision rule is given by summing the subclass probabilities calculated using (17) for each observation.

### 3 Methods for comparison

Apart from PDA mentioned in the previous section, forward selection (FS; Hastie et al. (2009)) combined with LDA, as well as shrunken centroids reg-
Algorithm 2 Sparse Mixture Discriminant Analysis:

1. Initialize the blurred response matrix $Z$ with the subclass probabilities. As in Hastie and Tibshirani (1996) the subclass probabilities can be derived from Learning Vector Quantization or K-means preprocessing, or from a priori knowledge of data. Initialize $\theta = (R \sum_{j=1}^{k} \sum_{r=1}^{R_j} \pi_{jr})^{-1} I_{1:Q}$.

2. For $j = 1, \ldots, Q$, $Q \leq R$ solve the modified elastic net problem with fixed $\theta$

$$\beta_j = \arg \min_{\beta_j} n^{-1}(\|Z\theta_j - X\beta_j\|_2^2 + \lambda_2 \beta_j^T \Omega \beta_j + \lambda_1 \|\beta_j\|_1) \quad (21)$$

3. For fixed $\beta$ and $Y^T \hat{Y} = UV^T$ compute the optimal scores

$$\theta = D_p^{-\frac{1}{2}} UV^T \quad (22)$$

where $D_p$ is a diagonal matrix of subclass probabilities, $\pi_{jr}$. $\pi_{jr}$ is the sum of the elements in the $r^{th}$ column in $Z$ divided by the number of samples $n$.

5. Update the subclass probabilities in $Z$ and the mixing probabilities in $D_p$ using the estimation steps (17) and (18).

6. Repeat step 2-5 until convergence or a maximum number of iterations is reached.

7. Remove the last $R - M$ trivial directions, where the $(M+1)^{th}$ singular value $S_{M+1} < \epsilon$ ($\epsilon$ is some small threshold value):

$$\theta = D_p^{-\frac{1}{2}} UV_{1:M}^T \quad (23)$$

For $j = 1, \ldots, M$ solve the modified elastic net problem with fixed $\theta$ using (21) to obtain the $M$ nontrivial discriminant directions. Update the Gaussian mixture parameters using (17) and (18).

Regularized discriminant analysis (RDA; Guo et al. (2007)) and sparse partial least squares regression (SPLS; Chun and Keles (2009)) are used for compar-
isons. Finally, elastic net (EN) regression of dummy variables is also used for comparison.

3.1 Shrunk centroids regularized discriminant analysis

Shrunk centroids regularized discriminant analysis (RDA) is based on the same underlying model as LDA, i.e. normally distributed data with equal dispersion; Guo et al. (2007). The method regularizes the covariance matrix in LDA, similar to what happens for PDA. The regularization of the covariance matrix $\Sigma$ in RDA is

$$\tilde{\Sigma} = \alpha \hat{\Sigma} + (1 - \alpha) I_p$$  \hspace{1cm} (24)

for some $\alpha$, $0 \leq \alpha \leq 1$. The singular value decomposition trick is used to compute the matrix inversion for the LDA solution, thereby reducing the computational cost from the order of $O(p^3)$ to $O(pn^2)$. The regularization and reduced computational cost make it suitable for $p \gg n$ problems.

Instead of shrinking the centroids directly, it was proposed to shrink $\bar{x}^* = \tilde{\Sigma} \bar{x}$ with the shrinkage parameter $\Delta > 0$, where $\bar{x}$ is the centroid, i.e.

$$\bar{x}' = \text{sgn}(\bar{x}^*) ([\bar{x}^*] - \Delta)_+$$  \hspace{1cm} (25)

which possesses a feature elimination property. However, the false-positive rate for the selected features is high and it is therefore considered a conser-
vative feature selection method.

RDA is based on the same underlying model but uses a different algorithmic approach than SDA to overcome the \( p \gg n \) problem. An R package called rda is available from CRAN (2009).

### 3.2 Sparse partial least squares

Sparse partial least squares (SPLS) builds on the partial least squares model (PLS) which assumes that \( X \) and \( Y \) can be rewritten using basic latent decompositions, thereby obtaining a dimension reduction; Chun and Keles (2009). PLS is widely used in chemometrics for obtaining dimension reduction in \( p \gg n \) problems. SPLS promotes the lasso zero property onto a surrogate direction vector \( c \) instead of the original latent direction vector \( \alpha \), while keeping \( \alpha \) and \( c \) close.

\[
\min_{\alpha, c} -\kappa \alpha^T M \alpha + (1-\kappa)(c-\alpha)^T M (c-\alpha) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2 \quad \text{s.t.} \quad \alpha^T \alpha = 1 \quad (26)
\]

where \( M = X^T Y Y^T X \), \( 0 \leq \kappa \leq 1 \), and \( \lambda_2 \) and \( \lambda_1 \) are the weights on the ridge and lasso constraints, respectively. By using dummy variables to represent the groups with 0s and 1s as \( Y \), PLS and SPLS can be used for classification.

The algorithm uses alternating steps of SVD and elastic net, keeping \( c \) and \( \alpha \) fixed in the respective steps. The algorithmic approach is thus similar to that of SDA but the underlying model differs. An R package called spls is available from CRAN (2009).
3.3 Elastic net regression of dummy variables

One may argue that it is faster and easier to perform regression of dummy variables of the classes \( Y \) on the feature set \( X \) instead of using optimal scoring. The dummy class variables consists of zeros and ones, indicating the classes, with one variable for each class. Additionally, regression of centered dummy class variables is in theory equal to performing linear discriminant analysis. However, for numerical reasons this may not be true in practice, in particular with feature selection which is based on correlations with the response, or the residual of the response. We illustrate that the two methods are not equivalent (although similar) in practice as the selected sets of variables may differ, and hence, likewise the error rates. However, EN is faster than SDA.

4 Experimental results

This section illustrates results on a small data set of shapes from female and male face-silhouettes and on three different high-dimensional data sets: a benchmark high-dimensional microarray data set, a data set based on spectral imaging of \textit{Penicillium} fungi for classification to the species level, a data set with 1D NMRs of three fungal genera for classification to the genus level, and a data set with shape and texture features of three fish species. The number of iterations the algorithms used to obtain good, stable solutions in the following applications were less than 30 in all cases. The parameters for
the elastic net, as well as for the methods included for comparisons, were chosen using leave-one-out cross validation on the training data (Hastie et al., 2009). Five- or ten-fold cross validation could also be used, but as we have very few observations in some of the examples we chose to use leave-one-out cross validation. Subsequently, the models with the chosen parameters were tested using the test data. Data was normalized and the penalty matrix $\Omega = I$ utilized unless otherwise mentioned.

4.1 Female and male silhouettes

To illustrate the sparse representation of the discriminant directions from SDA we considered a shape-based data set consisting of 20 male and 19 female face-silhouettes from adults. A minimum description length (MDL) approach to annotate the silhouettes was used as in Thodberg and Ólafsdóttir (2003), and Procrustes’ alignment was performed on the resulting 65 MDL landmarks of $(x, y)$-coordinates. For training, 22 of the silhouettes were used (11 female and 11 male), which left 17 silhouettes for testing (8 female and 9 male). Figure 1 illustrates the two classes of silhouettes.

Performing leave-one-out cross validation on the training data we selected 10 non-zero features and $\lambda_2 = 10^{-2}$ as parameters for SDA. The SDA results are illustrated in figure 2. There was only one non-trivial sparse direction in this case. Note how the few landmarks included in the model were placed near high curvature points in the silhouettes. The training and test classification rates were both 82%. In the original paper (Thodberg and Ólafsdóttir,
Figure 1: The silhouettes and the 65 landmarks for the two groups: female and male subjects.

2003) a logistic regression was performed on a subset of PCA scores, where the subset was determined by backwards elimination using a classical statistical test for significance. Results were only stated for leave-one-out cross validation on the entire data set, which gave an 85% classification rate, see Thodberg and Ólafsdóttir (2003). The SDA model in figure 2 is easy to interpret compared to a model based on 2-4 principal components, each with contributions from all 65 MDL marks. The SDA model points out exactly where the main differences between the two genders are.

4.2 Leukemia-subtype microarray

This section considers a high-dimensional benchmark data set from the Kent Ridge Biomedical Data Set Repository (http://sdmc.i2r.a-star.edu.sg/rp/), namely the leukemia-subtype data set published in Yeoh et al. (2002). The
Figure 2: Results from SDA on the silhouette data. (a) The mean shape of the silhouettes and the model with the 10 non-zero loadings illustrating which landmarks differ from female to male subjects. The arrows illustrate the directions of the differences. (b) The sparse direction discriminating the classes. The crosses illustrate the observations, the solid curves illustrate the estimated Gaussian distributions of the classes from the training set, and the dashed curves illustrate the estimated Gaussian of the classes from the training and the test set.
study was a genetic analysis of the cancerous cells and aimed at classifying subtypes of pediatric acute lymphoblastic leukemia (ALL). Cancer diseases require fast and correct diagnosis and in order to do that it is of importance to understand genetic factors in the various variants of for example pediatric ALL. One way to facilitate this is by microarray analysis. The microarray data set considered here consisted of 12558 genes, 6 subtypes of cancer, 163 training samples and 85 test samples. The six major cytogenetic diagnostic groups in the data were: BCR-ABL, E2A-PBX1, Hyperdiploid (>50 chromosomes), MLL rearrangement, T-ALL and TEL-AML1. Originally, in Yeoh et al. (2002), the data were analyzed in two steps: a feature selection step and a classification step. Furthermore, the data were analyzed in a decision tree structure such that one group was separated using an SVM at each tree node. Here, we illustrate the strengths of SDA, which performs feature selection, dimension reduction and classification in one step. With only 25 non-zero features in each of the six discriminant directions, compared to 40 in Yeoh et al. (2002), classification rates comparable to or slightly better than those in Yeoh et al. (2002) were obtained. The results are summarized in table 1 and are on non-normalized data for comparison with the original analysis of data. One of the BCR-ABL patients in the training set was misclassified as Hyperdiploid, but there were no misclassifications in the test set.

Figure 3 illustrates scatter plots of the six groups projected onto the sparse directions obtained by SDA. Note that each sparse direction separates different groups. This leads to knowledge not only of the separation of all
groups, but also of which genes have a different expression level for one subtype of cancer compared to the others, similar to the decision tree structure in the original analysis.

Table 1: Subgroup predictions using SDA with 25 non-zero features in each of the six discriminant directions. The ridge weight, $\lambda_2 = 10^{-1}$ as well as the number of non-zero loadings were chosen using leave-one-out cross validation on the training set.

<table>
<thead>
<tr>
<th>Group</th>
<th>Train SDA</th>
<th>Test SDA</th>
</tr>
</thead>
<tbody>
<tr>
<td>All groups</td>
<td>99%</td>
<td>100%</td>
</tr>
<tr>
<td>BCR-ABL</td>
<td>89%</td>
<td>100%</td>
</tr>
<tr>
<td>E2A-PBX1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Hyperdiploid</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>T-ALL</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>TEL-AML1</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>MLL</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 3: Sparse discriminant variables in SDA of the Leukemia-subtype data set.

In comparison, an elastic net regression on dummy variables gave overall classifications of 98% for the training set, and 98% for the test set with
selected parameters of $\lambda = 10^{-2}$ and 20 non-zero elements in each of the six discriminative directions.

4.3 Spectral id of fungal species

This section analyzes another high-dimensional data set which considers multi-spectral imaging for objective classification of fungi. Few of the world’s fungal species are known today (Hawksworth, 2001) and due to the various useful and toxic mycotoxins they can produce, it is of great interest to quickly and accurately classify known species and identify unknown ones. Here, we consider the three *Penicillium* species: *Melanoconidium*, *polonicum*, and *venetum*. The three species all have green/blue conidia (the spores of the fungi) and are therefore visually difficult to distinguish. It is desirable to have accurate and objective classification of the fungi species as they produce different mycotoxins. Some are very useful to us, such as penicillin, while others can be harmful. A visual classification is based on the phenotypes of the species and is generally faster and cheaper than chemical or genetic methods for classification. Using image analysis to perform the classification also gives an objective and accurate method which can be reproduced in laboratories.

For each of the three species, four strains were injected onto yeast extract sucrose (YES) agar in three replica, in total 36 samples. The data set consisted of 3542 variables extracted from multi-spectral images (Clemmensen et al., 2007) with 18 spectral bands (10 in the visual range, and eight in
the near infrared range). The variables were summary statistics taken from histograms of the fungal colonies in each spectral band, and in each pairwise difference and pairwise multiplication between spectral bands. Table 2 summarizes the results from reduced-rank PDA, forward selection (FS) based on Wilk’s Lambda, EN, and SDA. The data was partitioned into 2/3 as the training data and 1/3 as the test data where one of the three repetitions of each strain was left out for testing. This gave 28 training samples and 12 test samples. In this case the classification rates were not improved, but the complexity of the models was reduced by both SDA, EN, and FS. Furthermore, the computational cost of FS based on Wilk’s A was larger than for SDA. The CPU time for FS was more than double the CPU time for SDA. For just two non-zero loadings this is not very excessive but as the number of non-zero loadings increases, the computational cost also increases. EN had the shortest CPU time. Moreover, the three methods - FS, EN, and SDA - had one of the selected variables in common. Figure 4 illustrates the sparse discriminant directions in SDA. There were only two nontrivial directions in this case and the SDA algorithm converged. It is not surprising that the three groups are completely discriminated as they differ in their conidium color, which ranges from green to blue, see Clemmensen et al. (2007). The selected features are thus also percentiles in differences of blue and green spectral bands.
Figure 4: The *Penicillium* data set projected onto the sparse discriminant directions in SDA.
Table 2: Classification rates from PDA, EN, SDA and forward selection based on Wilk’s $\Lambda$ (FS) combined with LDA on the Penicillium data. The Ridge penalty weight was $10^{-6}$ for PDA, EN and SDA, chosen using leave-one-out cross-validation on the training set. Likewise the number of non-zero loadings was chosen using cross-validation. The covariance matrix in the reduced-rank PDA was ridge-regularized since $p >> n$. The chosen parameters for RDA were $\alpha = 0.1$ and $\Delta = 0.2$, and for SPLS there were five latent variables with 762 nonzero variables and $\eta = 0.8$. Note, that the computational complexity for forward selection was much larger than for SDA.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>Test</th>
<th>Nonzero loadings</th>
<th>CPU-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDA</td>
<td>100%</td>
<td>100%</td>
<td>7084</td>
<td>384.3s</td>
</tr>
<tr>
<td>FS</td>
<td>100%</td>
<td>100%</td>
<td>2</td>
<td>0.4s</td>
</tr>
<tr>
<td>RDA</td>
<td>100%</td>
<td>100%</td>
<td>3502</td>
<td>0.0s</td>
</tr>
<tr>
<td>SPLS</td>
<td>100%</td>
<td>100%</td>
<td>3810</td>
<td>0.7s</td>
</tr>
<tr>
<td>EN</td>
<td>100%</td>
<td>100%</td>
<td>3</td>
<td>0.06s</td>
</tr>
<tr>
<td>SDA</td>
<td>100%</td>
<td>100%</td>
<td>2</td>
<td>0.1s</td>
</tr>
</tbody>
</table>

4.4 Chemical id of fungal genera

In the previous section we used visual information to classify fungi to the species level. Here we will use chemical information in the form of 1D NMR of fungi for classification at the genus level (Rasmussen, 2006). Three genera of fungi were considered: Aspergillus, Neosartorya, and Penicillium. For each genus there were 5, 2, and 5 species, respectively. There were 71 observations with 4-8 samples of each species. Information from the 950 highest peaks in the NMR data were used as features. Data were logarithmically transformed, as differences in peaks with lower intensities seemed to have influence. As the biology gave a hierarchy of subgroups within each genus, it seemed reasonable to model each genus as a mixture of Gaussians, i.e. a mixture of species, and therefore we tested the SMDA on this data. Ta-
ble 3 summarizes the results using PDA, EN, SDA and SMDA on the 1D NMR data. RDA and SMDA gave the best test classification rates. In addition to improved classification rates the sparse methods provided insight into which chemical features distinguish the fungal genera. Furthermore, the sparse methods gave models with smaller complexity and thereby smaller variance. Consequently, the sparse methods tended to overfit less than the more complex models such as PDA. Figures 5 and 6 illustrate the (sparse) discriminative directions for PDA, EN, SDA, and SMDA. There were only two nontrivial sparse directions for SDA. Note, that due to the underlying mixture of Gaussians model, the sparse directions in the SMDA provided knowledge of the separation between genera, not only at the genus level but also at the species level.

![](image)

(a) PDA  (b) EN  (c) SDA

Figure 5: Discriminant directions in PDA, EN and SDA of the 1D NMR data set. In particular for *Aspergillus* and *Neosartorya* there seem to be subclusters within the genera.
Table 3: Errors from PDA, RDA, SPLS, EN, SDA and SMDA on the 1D NMR data. With few non-zero loadings in SDA and SMDA, the test classification rates are improved. The Ridge penalty weight is in the range $[10^{-1}, 10^3]$ for the three methods PDA, EN, and SDA, and, together with the number of non-zero loadings, this was chosen using leave-one-out cross validation on the training set. For SMDA the Ridge penalty was $10^8$. The covariance matrix in the reduced-rank PDA was ridge-regularized since $p >> n$.

For RDA on normalized data (n) the chosen parameters were $\alpha = 0.99$ and $\Delta = 0.6$, and on the original data (u) they were $\alpha = 0.9$ and $\Delta = 0.5$. For SPLS, eight latent variables including 79 non-zero variables were chosen and $\eta = 0.8$.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>Test</th>
<th>Nonzero loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>PDA</td>
<td>100%</td>
<td>76%</td>
<td>1900</td>
</tr>
<tr>
<td>RDA(n)</td>
<td>100%</td>
<td>85%</td>
<td>888</td>
</tr>
<tr>
<td>RDA(u)</td>
<td>100%</td>
<td>97%</td>
<td>648</td>
</tr>
<tr>
<td>SPLS</td>
<td>100%</td>
<td>88%</td>
<td>632</td>
</tr>
<tr>
<td>EN</td>
<td>100%</td>
<td>91%</td>
<td>45</td>
</tr>
<tr>
<td>SDA</td>
<td>100%</td>
<td>91%</td>
<td>50</td>
</tr>
<tr>
<td>SMDA</td>
<td>97%</td>
<td>97%</td>
<td>324</td>
</tr>
</tbody>
</table>
Figure 6: Sparse discriminant directions in SMDA of the 1D NMR data set. Note how the distribution of each class has changed due to the underlying mixture of Gaussians model. Here, each sparse direction aims at separating one subclass from the remaining.
4.5 Classification of fish species based on shape and texture

Here we consider classification of three fish species: cod, haddock, and whiting. The classification was performed on the basis of shape and texture features. The data taken from Larsen et al. (2009) consisted of 108 fish: 20 cod, 58 haddock, and 30 whiting. The shape of the fish was represented with landmarks based on MDL, as in the example with the male and female silhouettes, see Figure 7. There were 700 points for the contour of the fish, 300 for the mid line, and one for the eye. The shapes were Procrustes aligned to have full correspondence. The texture features were simply the R, G, and B intensity values from color images taken with a standard camera under a standardized white light illumination and digitized. They were annotated to the shapes using a Delauney triangulation approach; examples of the texture features are illustrated in Figure 7. There was a total of 103,348 shape and texture features in the data set. In Larsen et al. (2009), a principal component analysis followed by a linear discriminant analysis was performed, resulting in a 76% leave-one-out classification rate. Here, we split the data in two: 76 fish for training, and 32 fish for testing. The results are listed in table 4. In this case, SDA gives the sparsest solution and the best test classification rate. Only one of the whiting was misclassified as haddock.

The sparse discriminant directions are illustrated in Figure 8, there were only two nontrivial directions. The 1st SD is mainly dominated by blue
Figure 7: Illustration of the shape features to the left and the texture features to the right. The shape features are marked as white crosses on a gray scale image. The texture features are the intensities in the red, green and blue bands of the color images.

Table 4: Errors from EN, SDA, RDA, and SPLS on the shape and texture features from the images of fish. For SDA the Ridge penalty weight was $10^{-3}$ and 30 non-zero loadings were included for each of the two discriminant directions. For EN the Ridge penalty was $10^{-4}$ and there were 30 non-zero loading for each of the three discriminative directions. The parameters were chosen using leave-one-out cross validation on the training set. For RDA on normalized data (n), the chosen parameters were $\alpha = 0.99$ and $\Delta = 0.2$, and on the original data (u) they were $\alpha = 0.99$ and $\Delta = 0.1$. For SPLS, three latent variables including 105 non-zero variables were chosen and $\eta = 0.9$. The SPLS package implemented in R could not handle the size of the fish data, and therefore SPLS was also used preliminarily on subsets of the data to reduce the number of variables for the problem. Note, that RDA performed much better when data was not normalized.

<table>
<thead>
<tr>
<th>Method</th>
<th>Train</th>
<th>Test</th>
<th>Nonzero loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td>RDA(n)</td>
<td>100%</td>
<td>41%</td>
<td>103084</td>
</tr>
<tr>
<td>RDA(u)</td>
<td>100%</td>
<td>94%</td>
<td>103348</td>
</tr>
<tr>
<td>SPLS</td>
<td>100%</td>
<td>81%</td>
<td>315</td>
</tr>
<tr>
<td>EN</td>
<td>100%</td>
<td>94%</td>
<td>90</td>
</tr>
<tr>
<td>SDA</td>
<td>100%</td>
<td>97%</td>
<td>60</td>
</tr>
</tbody>
</table>
intensities, and shows that cod are in general less blue than haddock and whiting around the mid line and mid fin. This is in line with specialist knowledge that there is in fact a thin dark line present around the mid line in haddock and whiting, whereas this is absent in cod; Larsen et al. (2009). The 2nd SD implies that haddock in general is more blue around the head and tail, less green around the mid line, more red around the tail and less red around the eye, the lower part and the mid line than cod and whiting.

![Figure 8: Sparse discriminant directions in SDA. To the left, the clusters of three fish species. To the right, the selected texture features are marked on the fish mask. The first SD is mainly dominated by blue intensities whereas the 2nd SD consists of both red, green, and blue intensities. Only texture features were selected in the SDA.](image)

5 Discussion

Linear discriminant analysis and classification by mixtures of Gaussians are widely used methods for dealing with supervised classification. In this paper we have proposed algorithms for computing sparse versions of linear discrimi-
nant analysis and mixture discriminant analysis. The methods are especially useful when the number of observations is small in relation to the number of variables \( n \ll p \). Also they are generally useful when it is important to gain knowledge of a subset of features which separates two or more groups in high-dimensional problems. Sparse discriminant analysis has been illustrated on a small shape-based data set of female and male silhouettes, a benchmark microarray data set for classification of leukemia subtypes, on visual and chemical data for classification of fungi to species or the genus level, and on shape and texture features for classification of fish species. Sparse mixture discriminant analysis was illustrated on the microarray data and the chemical data for classification of fungi to the genus level. The methods are generally faster than the methods which perform feature selection followed by a classification analysis, but they are not as fast as the shrunken regularized discriminant analysis algorithm. However, performing an elastic net (EN) regression of class dummy variables on the feature matrix is the fastest. Furthermore, the classification results of EN, SDA and SMDA are comparable or even better than those obtained from the standard methods as well as for sparse partial least squares (SPLS) and shrunken centroids regularized discriminant analysis (RDA). In general, the methods out-performed SPLS in classification rates, whereas they gave comparable classification results to those for RDA. However, RDA is conservative with regards to feature selection and therefore the methods proposed here provide sparser models and additionally provide low dimensional views of the data. It should also be
noted that RDA is sensitive to normalization of data in terms of classification rates. EN and SDA are very similar. Although SDA in some cases gave slightly better results, EN is to be prefered if computational speed is crucial. Finally, the mixture of Gaussians models are useful for modelling data where biological subgroups exist, such as classification of biological data to the species or the genus level. Matlab and R versions of SDA and SMDA are available from: www.imm.dtu.dk/~lhc.

Acknowledgements

The authors would like to thank Gritt Rasmussen, Thomas Ostenfeld Larsen, Charlotte Held Gotfredsen and Michael E. Hansen at BioCentrum, Technical University of Denmark for making the 1D NMR data available. Also thanks to Hildur Ólafrsdóttir and Rasmus Larsen at Informatics and Mathematical Modelling, Technical University of Denmark for making the silhouette and fish data available, and to Karl Sjöstrand for valuable comments. Finally, the authors would like to thank the editor, an associate editor and two referees for valuable comments.

References

Bickel, P., Levina, E., 2004. Some theory for fisher’s linear discriminant function, ‘naïve bayes’, and some alternatives when there are many more vari-
ables than observations. Bernoulli 6, 989–1010.


CRAN, 2009. The comprehensive r archive network.
URL http://cran.r-project.org/


Larsen, R., Olafsdottir, H., Ersbøll, B., 2009. Shape and texture based classi-


Yeoh, E.-J., , et. al, March 2002. Classification, subtype discovery, and pre-
diction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell 1, 133–143.


A Appendix

A.1 The relation between optimal scoring and discriminant analysis

It is convenient to make the relation between the sparse optimal scoring criterion (8) and the sparse discriminant criterion (3) via canonical correlation analysis (CCA).

A.1.1 Sparse optimal scoring

The sparse optimal criterion in (8) is stated in terms of a single solution $(\theta, \beta)$, but implicitly it is a sequence of solutions $(\theta_j, \beta_j)$ with orthogonality given by the inner product $n^{-1} < Y \theta_j, Y \theta_l > = \delta_{jl}$ implied in the constraint
The sparse optimal scoring criterion can be rewritten to

\[ ASR(\theta_j, \beta_j) = \theta_j^T \Sigma_{11} \theta_j - 2\theta_j^T \Sigma_{12} \beta_j + \beta_j^T \Sigma_{22} \beta_j + \lambda \sum_{i=1}^{p} |\beta_{ji}| , \quad (27) \]

which is to be minimized under the constraint

\[ \theta_j^T \Sigma_{11} \theta_j = 1 , \quad (28) \]

and where

\[ \Sigma_{11} = n^{-1}Y^T Y \quad (29) \]
\[ \Sigma_{22} = n^{-1}(X^T X + \lambda_2 \Omega) \quad (30) \]
\[ \Sigma_{12} = n^{-1}Y^T X ; \quad \Sigma_{21} = \Sigma_{12}^T . \quad (31) \]

### A.1.2 Sparse canonical correlation analysis

The sparse canonical correlation problem is defined by the criterion (which apart from the \( \ell_1 \)-term is the same as the penalized correlation problem, Hastie et al. (1995a))

\[ COR_{\ell_1}(\theta_j, \beta_j) = \theta_j^T \Sigma_{12} \beta_j - \lambda \sum_{i=1}^{p} |\beta_{ji}| , \quad (32) \]

which is to be maximized under the constraints

\[ \theta_j^T \Sigma_{11} \theta_j = 1 \quad \text{and} \quad \beta_j^T \Sigma_{22} \beta_j = 1 . \quad (33) \]
Under the CCA constraints we obtain $ASR = 2 - 2COR_{\ell_1}$, and the problems only differ in the additional constraint $\beta^T \Sigma_{22} \beta = 1$. Hence, for fixed $\theta$ the parameters in the optimal scoring problem $\beta_{os}$ is, up to a scalar, the same as the parameters for the canonical correlation problem:

$$\beta_{j,cca} = \frac{\beta_{j,os}}{\sqrt{\beta_{j,os}^{T} \Sigma_{22} \beta_{j,os}}} ,$$

(34)

and the $\ell_1$-weights are related as $\lambda_{1,cca} = \lambda_{1,os}/2$. Finally, we see that the optimal scores are the same for the two problems as we for fixed $\beta$ have:

$$\theta_{cca} = \theta_{os} = \Sigma_{11}^{-1/2} U V^T ,$$

(35)

where $\Sigma_{11}^{-1} \Sigma_{12} \beta_{os} = U S_{os} V^T$ or $\Sigma_{12} \beta_{cca} = U S_{cca} V^T$.

### A.1.3 Sparse discriminant analysis

The sparse discriminant analysis is defined as in (3)

$$BVAR_{\ell_1}(\beta_j) = 2 \beta^T \Sigma_B \beta_j - \lambda_1 \sum_{i=1}^{p} |\beta_{j_i}| ,$$

(36)

which is to be maximized under the constraint

$$WVAR(\beta_j) = \beta^T \Sigma_W \beta_j = 1 ,$$

(37)
and where

\[
\Sigma_B = \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12}
\]

(38)

\[
\Sigma_{W_p} = \Sigma_W + \lambda_2 n^{-1} \Omega = \Sigma_{22} - \Sigma_B
\]

(39)

Recall from penalized discriminant analysis (Hastie et al. (1995a)) that without the \( \ell_1 \)-penalization then the penalized discriminant analysis and penalized canonical correlation analysis coordinates are related as

\[
\beta_{j,lda} = \beta_{j,cca} \frac{\sqrt{\beta_{j,cca}^T \Sigma_{W_p} \beta_{j,cca}}}{\beta_{j,os}^T \Sigma_{W_p} \beta_{j,os}}
\]

(40)

Comparing \( BVAR_{\ell_1} \) (36) and \( COR_{\ell_1} \) (32) and keeping in mind that the constraints are the same as under PDA it is easy to see that the relation still holds, and that the \( \ell_1 \)-weights are related as \( \lambda_{1,lda} = \lambda_{1,cca} \).

### A.1.4 Optimal scoring and discriminant analysis

Finally, we have the relation between sparse discriminant analysis and sparse optimal scoring given via their relations to CCA:

\[
\beta_{lda} = \beta_{os} / \sqrt{\beta_{os}^T \Sigma_{W_p} \beta_{os}}
\]

(41)

Furthermore, the \( \ell_1 \)-weights are related as \( \lambda_{1,lda} = \lambda_{1,os}/2 \).