Sparse Discriminant Analysis

Clemmensen, Line Katrine Harder; Hastie, Trevor; Ersbøll, Bjarne Kjær

Publication date:
2008

Document Version
Publisher's PDF, also known as Version of record

Sparse Discriminant Analysis

IMM-Technical Report-2008-06

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.

June 10, 2008

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 as microarrays, 1D NMR, and spectroscopy have become everyday tools for measuring thousands of features in samples of interest. Furthermore, the samples are often costly and therefore many such problems have few observations in relation to the number of features. Traditionally such 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 is 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 regards to classification rates and sparseness. The method is extended to mixtures of Gaussians which is useful when e.g. biological clusters are present within each class. 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, LDA provides low-dimensional projections of data onto the most discriminative directions. However, it 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.

The mentioned situations where LDA fails were previously addressed in penalized discriminant analysis (Hastie et al., 1995a) and discriminant analysis by gaussian mixtures (Hastie and Tibshrani, 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 is desired. A low number of nonzero parameters ensures a better interpretation of the model and additionally tends to overfit training data less than nonsparse 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 nonzero 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., 2001; 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; Zou and Hastie, 2005).

The introduction of a sparseness criterion is well known in the regression framework (Tibshirani, 1996; Zou and Hastie, 2005; Zou et al., 2006) and we shall therefore consider LDA by optimal scoring which performs LDA by regression (Hastie et al., 1995a; Ye, 2007). Furthermore, the optimal scoring framework allows for an extension to mixtures of Gaussians (Hastie and Tibshrani, 1996).

The paper is organized as follows. Section two describes the sparse LDA and sparse mixture discriminant analysis algorithms, introducing a modification of the elastic net algorithm to include various penalizing matrices. Section three illustrates experimental results on a small illustrative shape based data set of female and male silhouettes and on three high-dimensional data sets: A microarray data set plus spectral, and chemical identification of fungi. We round off with a discussion in section four.

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 μ_j, $j = 1, ..., k$ and equal dispersion Σ (see e.g. Hastie et al. (2001)). Reduced-rank LDA has the ability to provide low-dimensional views of data of up to at most $k - 1$ dimensions. 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, β_js are the ones which maximize the variance between classes and minimize the variance within classes and are orthogonal to each other. Hence, we maximize the between sums of squares, Σ_B relative to the within sums of squares, Σ_W (the Fisher's criterion)

$$\arg\max_{\beta_j} \beta_j^T \Sigma_B \beta_j \tag{1}$$

under the orthogonality constraint

$$\beta_j^T \Sigma_W \beta_l = \begin{cases} 0 & l = 1, ..., j-1 \\ 1 & l = j \end{cases} , \tag{2}$$

to find the discriminating directions β_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 ℓ_1-norm of the parameters β. The ℓ_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

$$\arg\max_{\beta_j} \beta_j^T \Sigma_B \beta_j - \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \tag{3}$$

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

The elastic net proposed by Zou and Hastie (2005) solves a regression problem regularized by the ℓ_2-norm and the ℓ_1-norm in a fast and effective manner. The elastic net is defined as

$$\beta_j^{en} = \arg\min_{\beta_j}(\|y - X\beta_j\|_2^2 + \lambda_2\|\beta_j\|_2^2 + \lambda_1\|\beta_j\|_1) \quad . \tag{4}$$

As the sparse discriminant criterion is also regularized by an ℓ_2-norm and an ℓ_1-norm penalty it seems advantageous to rewrite the criterion to 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. The idea behind optimal scoring is to turn categorical variables[1] into quantitative variables. Optimal scoring assigns a score, θ_{ji} for each class i and for each parameter vector β_j. The optimal scoring problem is defined as

$$(\hat{\theta}, \hat{\beta})^{os} = \arg\min_{\theta,\beta} n^{-1}\|Y\theta - X\beta\|_2^2 \tag{5}$$

$$s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad , \tag{6}$$

where Y is a matrix of dummy variables representing the k classes.

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

$$(\hat{\theta}, \hat{\beta})^{pos} = \arg\min_{\theta,\beta}(n^{-1}\|Y\theta - X\beta\|_2^2 + \lambda_2\|\Omega^{\frac{1}{2}}\beta\|_2^2) \quad , \tag{7}$$

s.t. (6), where Ω is a symmetric and positive definite matrix. In this paper, a sparseness criterion is added to the penalized optimal scoring criterion in form

[1]The categorical variables will here be encoded as $\{0, 1\}$ *dummy* variables.

of the ℓ_1-norm of the regression parameters β. 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 Tibshrani (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. However, it is possible to run the algorithm on raw data which is comparable to performing principal component analysis on the covariance matrix rather than the correlation matrix.

2.1 Sparse discriminant analysis by optimal scoring

In this section we introduce constraints to the optimal scoring problem in (15) in order to obtain sparseness in the PDA. The score vector θ_j assigns a real number θ_{ji} for each class i, $i = 1, ..., k$. The scored training data $Y\theta$ is an $n \times q$ matrix on which we will regress 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. We define sparse optimal scoring as

$$\begin{aligned}(\theta, \beta)^{sos} &= \arg\min_{\theta,\beta} n^{-1}(\|Y\theta - X\beta\|_2^2 + \lambda_2\|\Omega^{\frac{1}{2}}\beta\|_2^2 + \lambda_1\|\beta\|_1) \qquad (8)\\ &\quad s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad , \qquad (9)\end{aligned}$$

where Ω is a penalization matrix, as introduced in PDA (Hastie et al., 1995a). The ℓ_1-norm introduces sparseness as in lasso or elastic net regularization. In appendix the relation between sparse discriminant analysis (3) and sparse optimal scoring (8) is given.

For fixed θ we obtain:

$$\beta_j^{sos} = \arg\min_{\beta_j} n^{-1}(\|Y\theta_j - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \qquad (10)$$

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

$$\begin{aligned}\theta^{os} &= \arg\min_{\theta} n^{-1}\|Y\theta - X\beta\|_2^2 \qquad (11)\\ &\quad s.t. \quad n^{-1}\|Y\theta\|_2^2 = 1 \quad .\end{aligned}$$

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

$$\begin{aligned} \hat{\theta}^* &= \arg\min_{\theta^*} n^{-1}\|YD_\pi^{-\frac{1}{2}}\theta^* - \hat{Y}\|_2^2 && (12) \\ & s.t. \quad \|\theta^*\|_2^2 = 1 \quad , && (13) \end{aligned}$$

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

$$\begin{aligned} \hat{\theta}^* &= UV^T \Leftrightarrow && (14) \\ \hat{\theta} &= D_\pi^{-\frac{1}{2}}UV^T \quad . && (15) \end{aligned}$$

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 $\beta^T x$, and a *nearest class mean* rule, where "nearest" is measured using Σ_{W_p}, is applied in the $q < k-1$ reduced-dimensional discriminant subspace to obtain class labels.

2.2 Modified elastic net

For generalization, we modify the elastic net algorithm to include an arbitrary penalty matrix Ω rather than the identity. The modified naive elastic net solution becomes

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|y - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \quad . \qquad (16)$$

We can transform the naive elastic net problem into an equivalent Lasso problem on the augmented data (Zou and Hastie, 2005, Lemma 1).

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2\mathbf{\Omega}} \end{bmatrix} \quad , \quad y^* = \begin{bmatrix} y \\ 0_p \end{bmatrix} \quad . \qquad (17)$$

The normal equations, yielding the OLS solution, to this augmented problem are

$$\begin{aligned} \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix}^T \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix} \hat{\beta}^* &= \begin{bmatrix} \mathbf{X} \\ \sqrt{\lambda_2 \mathbf{\Omega}} \end{bmatrix}^T \begin{bmatrix} y \\ 0_p \end{bmatrix} \Leftrightarrow \\ \left(\mathbf{X}^T\mathbf{X} + \lambda_2 \mathbf{\Omega}\right) \hat{\beta}^* &= \mathbf{X}^T y \quad . \end{aligned} \tag{18}$$

We see that β^* is the Ω-penalized regression estimate with weight λ_2. Hence, performing Lasso on this augmented problem yields a modified elastic net solution. Since Ω is symmetric and positive definite, $\sqrt{\Omega}$ always exists. For examples of various penalty matrices Ω and their applications we refer to Hastie et al. (1995a).

2.3 Sparse Discriminant Algorithm

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

Algorithm 1 Sparse Discriminant Analysis:

1. Initialize $\theta = (k \sum_{j=1}^{k} D_{\pi,\{jj\}})^{-1} I_{1:k-1}$.
2. For $j = 1, ..., q$ solve the modified elastic net problem with fixed θ

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|Y\theta_j - X\beta_j\|_2^2 + \lambda_2 \beta_j^T \Omega \beta_j + \lambda_1 \|\beta_j\|_1) \tag{19}$$

3. For fixed β and $Y^T\hat{Y} = USV^T$ compute the optimal scores from (15).
4. Repeat step 2 and 3 until convergence.
5. Update β for fixed θ using (19), the 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 nonzero coefficients. This can be massive if p and m are large. However, in general few nonzero coordinates are desired in the mentioned applications, and the algorithm therefore becomes very effective. Additionally, the number of iterations needed is generally small.

2.4 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 μ_{jr} and common covariance matrix Σ. Since the single prototype problem is formulated as an optimal scoring problem it is straight forward to extend it to mixtures of Gaussians in line with Hastie and Tibshrani (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_{jr} for each observation. Let π_{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_{jr} and the mixing probabilities, π_{jr} are

$$z_{ir} = \frac{\pi_{jr} \exp\{-\frac{(X-\mu_{jr})\Sigma^{-1}(X-\mu_{jr})}{2}\}}{\sum_{r=1}^{R_j} \pi_{jr} \exp\{-\frac{(X-\mu_{jr})\Sigma^{-1}(X-\mu_{jr})}{2}\}} \tag{20}$$

$$\pi_{jr} = \sum_{i\in g_i} z_{ir}, \quad \sum_{r=1}^{R_j} \pi_{jr} = 1 \tag{21}$$

with the *maximization* steps

$$\mu_{jr} = \frac{\sum_{i\in g_i} x_i z_{ir}}{\sum_{i\in g_i} z_{ir}} \tag{22}$$

$$\Sigma = n^{-1} \sum_{j=1}^{k} \sum_{i\in g_i} \sum_{r=1}^{R_j} z_{ir}(x_i - \mu_{jr})(x_i - \mu_{jr})^T \quad . \tag{23}$$

We now write the SMDA algorithm by computing $Q \leq R-1$ sparse directions for the subclasses in the mixture of Gaussians model as described in algorithm 2.

Algorithm 2 Sparse Mixture Discriminant Analysis:

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

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

$$\beta_j = \arg\min_{\beta_j} n^{-1}(\|Z\theta_j - X\beta_j\|_2^2 + \lambda_2\beta_j^T\Omega\beta_j + \lambda_1\|\beta_j\|_1) \tag{24}$$

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

$$\theta = D_p^{-\frac{1}{2}}UV^T \quad , \tag{25}$$

where D_p is a diagonal matrix of subclass probabilities, π_{jr}. π_{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 (20) and (21).

6. Repeat step 2-5 until convergence.

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

$$\theta = D_p^{-\frac{1}{2}}UV_{1:m}^T \quad , \tag{26}$$

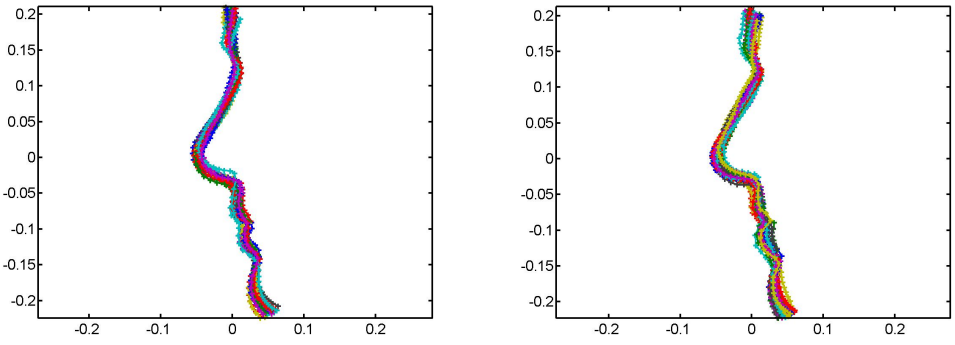
For $j = 1, ..., m$ solve the modified elastic net problem with fixed θ using (24) to obtain the m nontrivial discriminant directions.

3 Experimental results

This section illustrates results on a small data set of shapes from female and male silhouettes and on three different high-dimensional data sets: A benchmark high-dimensional microarray data set, a data set based on spectral imaging of *Penicillium* fungi for classification to the species level, and a data set with 1D NMRs of three fungal genera for classification to the genus level. The number of iterations the algorithms used in the following applications were less than 30 in all cases. The parameters for the elastic net were chosen using leave-one-out cross validation on the training data. Data was normalized and the penalty matrix $\Omega = I$ unless otherwise mentioned.

3.1 Female and male silhouettes

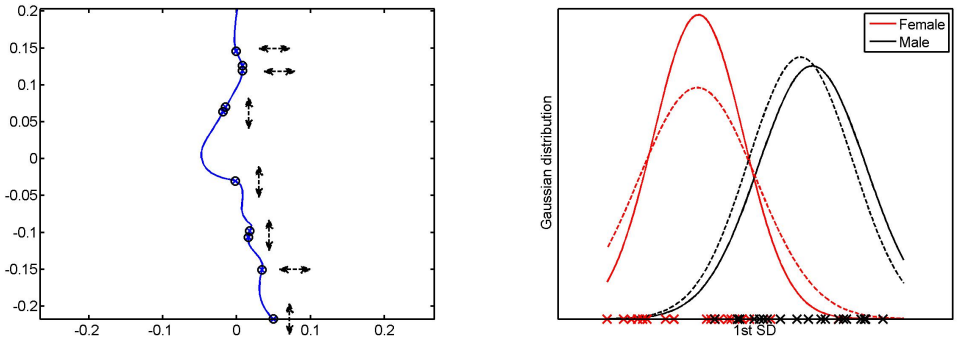
To illustrate the sparse representation of the discriminant directions from SDA we considered a shape based data set consisting of 20 male and 19 female silhouettes from adults. A minimum description length (MDL) approach to annotate the silhouettes were used as in Thodberg and Ólafsdóttir (2003), and Procrustes alignment was performed on the resulting 65 MDL marks of (x, y)-coordinates. For training the model we 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.



(a) Female (b) Male

Figure 1: The silhouettes and the 65 markers for the two groups: Female and male subjects.

Performing leave-one-out cross validation on the training data we selected 10 nonzero features and $\lambda_2 = 10^{-2}$ as parameters for SDA. The SDA results are illustrated in figure 2. Note, how the few markers 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, 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 a 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 differences between the two genders are.



(a) Model (b) SD

Figure 2: Results from SDA on the silhouette data. (a) The mean shape of the silhouettes and the model with the 10 nonzero loadings illustrating which markers 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.

3.2 Leukemia-subtype microarray

This section considers a high-dimensional benchmark data set from the Kent Ridge Biomedical Data Set Repository [2], namely the leukemia-subtype data set published in Yeoh and et. al (2002). The study aimed at classifying subtypes of pediatric acute lymphoblastic leukemia (ALL). Cancer diseases require fast and correct diagnosis and 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 diagnostic groups in data were: BCR-ABL, E2A-PBX1, Hyperdiploid>50 chromosomes, MLL rearrrangement, T-ALL and TEL-AML1. Originally, in Yeoh and et. al (2002), data was analyzed in two steps: A feature selection step and a classification step. Furthermore, data was 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 nonzero features, compared to 40 in Yeoh and et. al (2002), in each of the 5 discriminant directions good classification rates were obtained. The results are summarized in table 1 and are on non-normalized data for comparison with the original analysis of data. There were 2 misclassified observations in the training set and 3 misclassified observations in the test set. In the latter case all the misclassified observations belonged to the BCR_ABL group but were classified as Hyperdiploid>50.

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. Expression profiles of the selected genes for each sparse direction can be found in appendix.

3.3 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

[2] http://sdmc.i2r.a-star.edu.sg/rp/

Table 1: Subgroup predictions using SDA with 25 nonzero features in each of the 5 discriminant directions. The ridge weight, $\lambda_2 = 10^{-1}$ as well as the number of nonzero loadings were chosen using leave-one-out cross validation on the training set.

Group	Training set	Test set
All groups	99%	96%
BCR-ABL	89%	50%
E2A-PBX1	100%	100%
Hyperdiploid>50	98%	100%
T-ALL	100%	100%
TEL-AML1	100%	100%
MLL	100%	100%

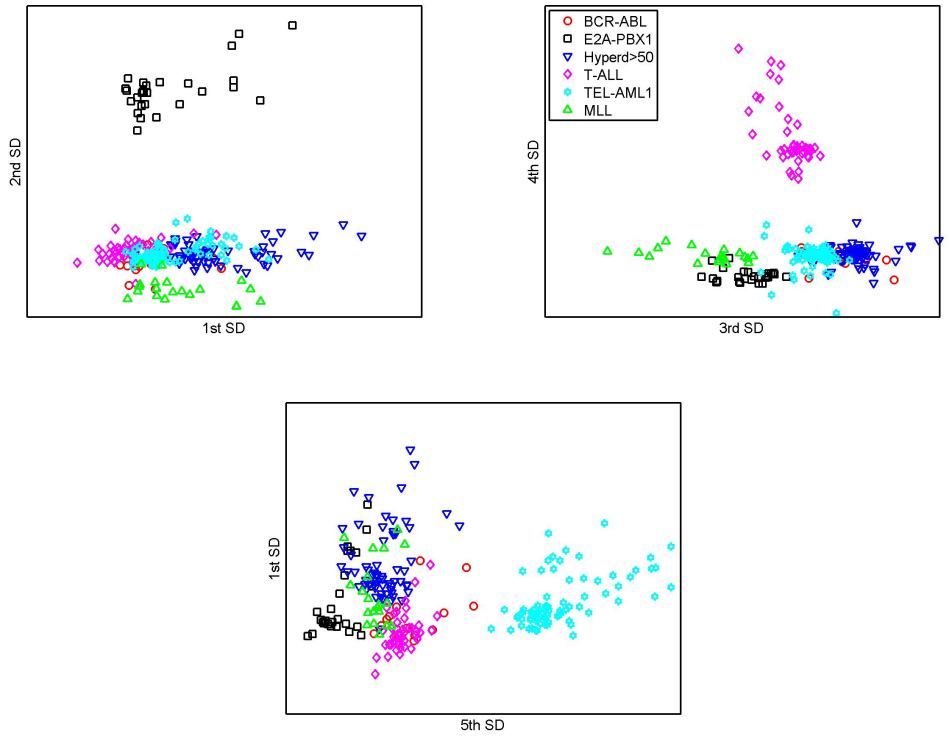


Figure 3: Sparse discriminant variables in SDA of the Leukemia-subtype data set.

and accurately classify known species and identify unknown ones. Here, we consider the three *Penicillium* species: *Melanoconodium, 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 in general faster than chemical or genetic methods for classification. Using image analysis to perform the classification additionally gives an objective and accurate method which can be reproduced in various laboratories.

For each of the three species, four strains were inoculated on 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 8 in the near infra red 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, and SDA. The data was partitioned into 2/3 which was the training data and 1/3 which was 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 and FS. Furthermore, the computational cost of SDA was smaller than for FS based on Wilk's Λ. The CPU-time was more than doubled which for just two nonzero loading doesn't seem alarming but as the number of nonzero loadings grow, the computational effort likewise grows. On top of that, the two methods: FS and SDA had one of the selected variables in common. Figure 4 illustrates the sparse discriminant directions in SDA. It is not surprising that the three groups are completely discriminated as they differ in their conidium color which range from green to blue, see Clemmensen et al. (2007). The selected features are thus also percentiles in differences of blue and green spectral bands.

Table 2: Classification rates from PDA, SDA and forward selection based on Wilk's Λ (FS) combined with LDA on the *Penicillium* data. The Ridge penalty weight was 10^{-6} for PDA and SDA, chosen using leave-one-out cross-validation on the training set. Likewise the number of nonzero loadings was chosen using cross-validation. The covariance matrix in the reduced-rank PDA was ridge regularized since $p >> n$. Note, that the computational complexity for forward selection was much larger than for SDA.

Method	Train	Test	Nonzero loadings	CPU-time
PDA	100%	100%	7084	384.3s
FS	100%	100%	2	0.4s
SDA	100%	100%	2	0.1s

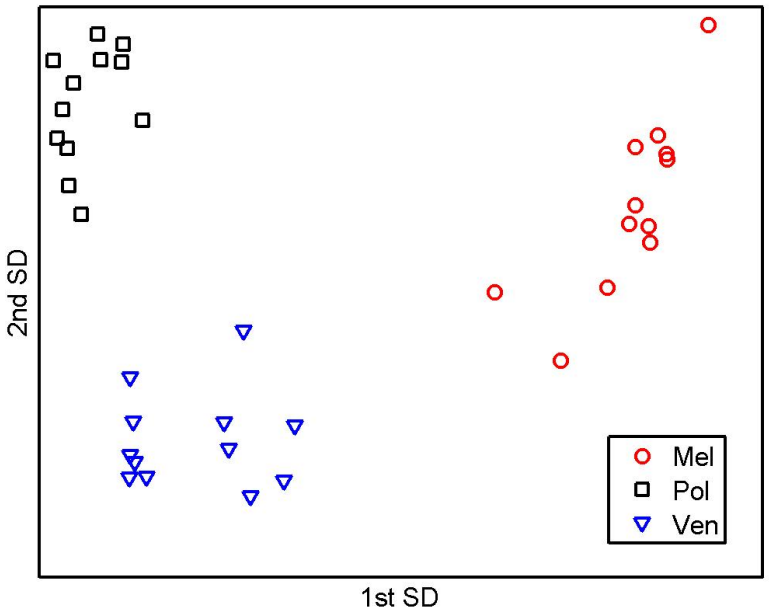


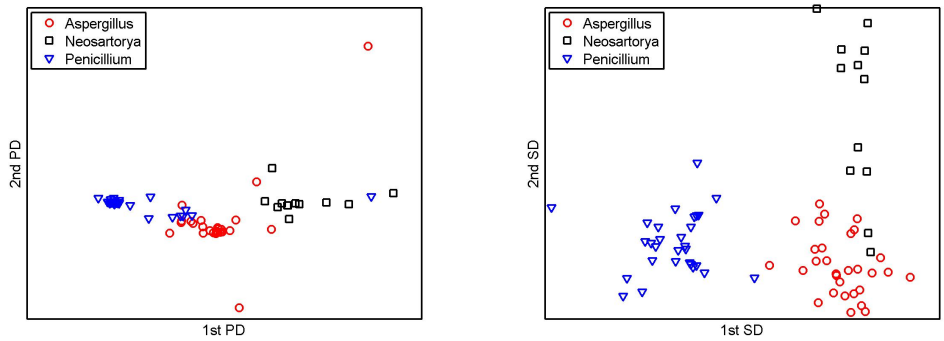
Figure 4: The *Penicillium* data set projected onto the sparse discriminant directions in SDA.

3.4 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 form of 1D NMR of fungi for classification to 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 logarithmicly 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. Table 3 summarizes the results using PDA, SDA and SMDA on the 1D NMR data. In addition to improved classification rates the sparse methods provided insight in which chemical features that 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 PDA model. Figure 5 and 6 illustrate the (sparse) discriminative directions for PDA, SDA, and SMDA. 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.

Table 3: Errors from PDA, SDA and SMDA on the 1D NMR data. With few nonzero loadings in SDA and SMDA the test classification rates are improved. The Ridge penalty weight is in $[10^{-3}, 10^{-1}]$ for the three methods and was as well as the number of nonzero loadings chosen using leave-one-out cross validation on the training set. The covariance matrix in the reduced-rank PDA eas ridge regularized since $p >> n$.

Method	Train	Test	Nonzero loadings
PDA	100%	76%	1900
SDA	97%	91%	10
SMDA	100%	94%	44



(a) PDA (b) SDA

Figure 5: Discriminant directions in PDA and SDA of the 1D NMR data set. In particular for *Aspergillus* and *Neosartorya* there seem to be subclusters within the genera.

4 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 discriminant analysis and mixture discriminant analysis. The methods are especially useful when the number of observations is small in relation to the number of variables ($n \ll p$), and in general 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 and on visual and chemical data for classification of the fungi to the species or the genus level. Sparse mixture discriminant analysis was illustrated on the chemical data for classification of fungi to the genus level. The methods are faster than methods first performing feature selection and then subsequently classification. Furthermore, the classification results are comparable or better than for such methods. 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`.

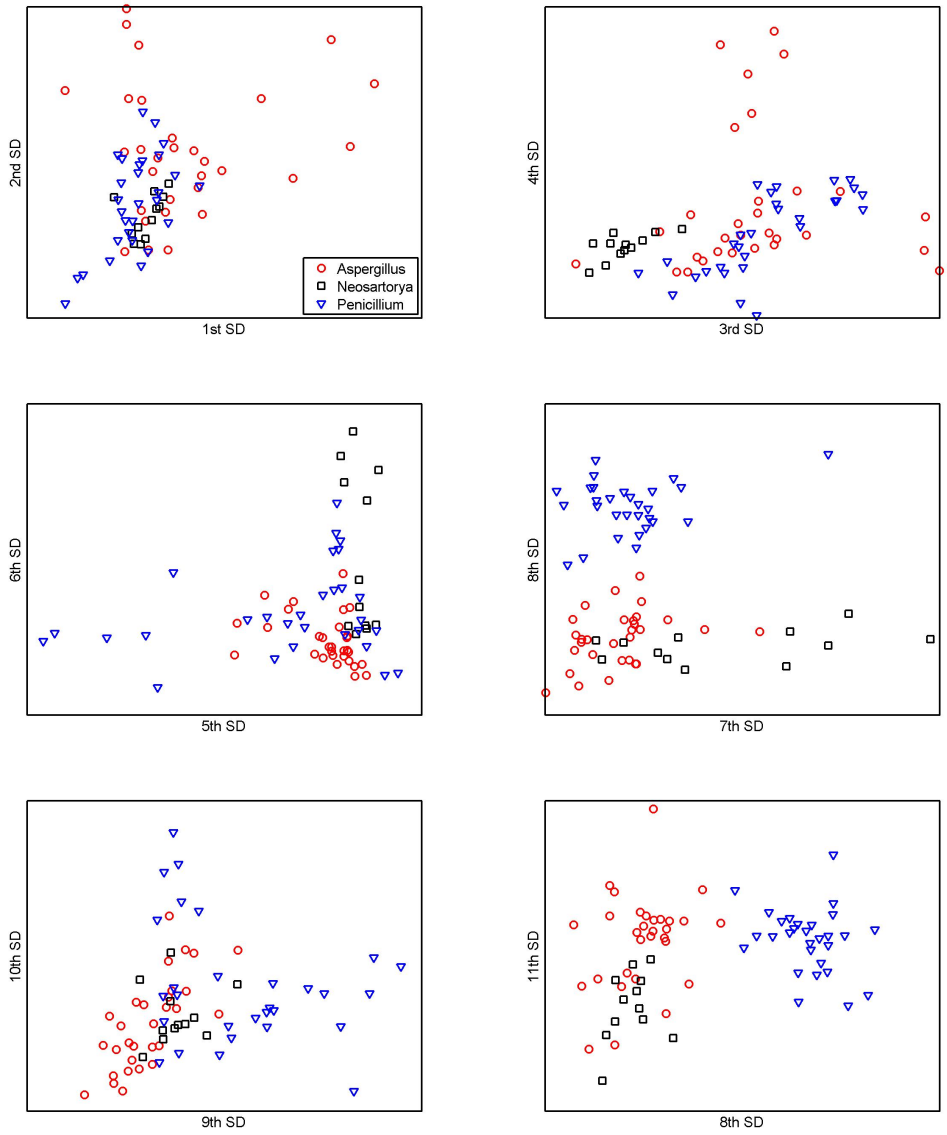


Figure 6: Sparse discriminant directions in SMDA of the 1D NMR data set. Note how the distribution of each group has changed due to the underlying mixture of Gaussians model. Here, each sparse direction aims at separating one sub group from the remaining.

Acknowledgements

The authors would like to thank Gritt Rasmussen, Thomas Ostenfeld Larsen, Charlotte Held Gotfredsen and Michael E. Hansen at BioCentrum, The Technical University of Denmark for making the 1D NMR data available. Also thanks to Hildur Ólafsdóttir for making the silhouette data available, and Karl Sjöstrand for valuable comments.

References

Bickel, P., Levina, E., 2004. Some theory for fisher's linear discriminant function, 'naive bayes', and some alternatives when there are many more variables than observations. Bernoulli 6, 989–1010.

Clemmensen, L., Hansen, M., Ersbøll, B., Frisvad, J., jan 2007. A method for comparison of growth media in objective identification of penicillium based on multi-spectral imaging. Journal of Microbiological Methods 69, 249–255.

Donoho, D. L., 2000. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture on August 8, to the American Mathematical Society 'Math Challenges of the 21st Century'. Avialable from http://www-stat.stanford.edu/∼donoho.

Duda, R. O., Hart, P. E., Stork, D. G., 2001. Pattern Classification. John Wiley & Sons.

Elden, L., Park, H., 1999. A procrustes problem on the stiefel manifold. Numerische Mathematik.

Hand, D. J., 2006. Classifier technology and the illusion of progress. Statistical Science 21 (1), 115.

Hastie, T., Buja, A., Tibshrani, R., 1995a. Penalized discriminant analysis. The Annals of Statistics.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The Elements of Statistical Learning. Springer.

Hastie, T., Tibshrani, R., 1996. Discriminant analysis by gaussian mixtures. J. R. Statist. Soc. B 58, 158–176.

Hastie, T., Tibshrani, R., Buja, A., 1995b. Flexible discriminant and mixture models. In: Neural Networks and Statistics conference, Edinburgh. J. Kay and D. Titterington, Eds. Oxford University Press.

Hawksworth, D. L., 2001. The magnitude of fungal diversity: the 1.5 million species estimate revisited. Mycol. Res. 105, 11422–1432.

Rasmussen, G., 2006. Hr-mas nmr data acquisition and chemometric analysis of fungal extracts. Master's thesis, E06, BioCentrum, Technical University of Denmark.

Thodberg, H. H., Ólafsdóttir, H., sep 2003. Adding curvature to minimum description length shape models. In: British Machine Vision Conference, BMVC.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Statist. Soc. B 58 (No. 1), 267–288.

Ye, J., 2007. Least squares linear discriminant analysis. In: Proc. of the 24th Int. Conf. on Machine Learning. pp. 1087 – 1093.

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

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Statist. Soc. B 67 (Part 2), 301–320.

Zou, H., Hastie, T., Tibshirani, R., June 2006. Sparse principal component analysis. Journal of Computational and Graphical Statistics 15, 265–286.

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 (θ, β), but implicitly it is a sequence of solutions (θ_j, β_j) with orthogonality given by the inner product $n^{-1} < Y\theta_j, Y\theta_l >= \delta_{jl}$ implied in the constraint (9). The sparse optimal scoring criterion can be rewritten to

$$ASR(\theta_j, \beta_j) = \theta_j^T \Sigma_{11} \theta_j - 2\theta_j^T \Sigma_{12} \beta_j + \beta_j^T \Sigma_{22} \beta_j + \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \quad , \tag{27}$$

which is to be minimized under the constraint

$$\theta_j^T \Sigma_{11} \theta_j = 1 \quad , \tag{28}$$

and where

$$\begin{aligned} \Sigma_{11} &= n^{-1} Y^T Y & (29) \\ \Sigma_{22} &= n^{-1}(X^T X + \lambda_2 \Omega) & (30) \\ \Sigma_{12} &= n^{-1} Y^T X \quad ; \quad \Sigma_{21} = \Sigma_{12}^T \quad . & (31) \end{aligned}$$

A.1.2 Sparse canonical correlation analysis

The sparse canonical correlation problem is defined by the criterion (which apart from the ℓ_1-term is the same as the penalized correlation problem, Hastie et al. (1995a))

$$COR_{\ell_1}(\theta_j, \beta_j) = \theta_j^T \Sigma_{12} \beta_j - \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \quad , \tag{32}$$

which is to be maximized under the constraints

$$\theta_j^T \Sigma_{11} \theta_j = 1 \quad \text{and} \quad \beta_j^T \Sigma_{22} \beta_j = 1 \quad . \tag{33}$$

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 θ the parameters in the optimal scoring problem β_{os} is, up to a scalar, the same as the parameters for the canonical correlation problem:

$$\beta_{j,cca} = \beta_{j,os} / \sqrt{\beta_{j,os}^T \Sigma_{22} \beta_{j,os}} \quad , \tag{34}$$

and the ℓ_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 β have:

$$\theta_{cca} = \theta_{os} = \Sigma_{11}^{-1/2} U V^T \quad , \tag{35}$$

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

A.1.3 Sparse discriminant analysis

The sparse discriminant analysis is defined as in (3)

$$BVAR_{\ell_1}(\beta_j) = \beta_j^T \Sigma_B \beta_j - \lambda_1 \sum_{i=1}^{p} |\beta_{ji}| \quad , \tag{36}$$

which is to be maximized under the constraint

$$WVAR(\beta_j) = \beta_j^T \Sigma_{W_p} \beta_j = 1 \quad , \tag{37}$$

and where

$$\Sigma_B = \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \tag{38}$$

$$\Sigma_{W_p} = \Sigma_W + \lambda_2 n^{-1}\Omega = \Sigma_{22} - \Sigma_B \quad . \tag{39}$$

Recall from penalized discriminant analysis (Hastie et al. (1995a)) that without the ℓ_1-penalization then the penalized discriminant analysis and penalized canonical correlation analysis coordinates are related as

$$\beta_{j,lda} = \beta_{j,cca} / \sqrt{\beta_{j,cca}^T \Sigma_{W_p} \beta_{j,cca}} \quad . \tag{40}$$

Comparing $BVAR_{\ell_1}$ (36) and COR_{ℓ_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 ℓ_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:

$$\beta_{lda} = \beta_{os} / \sqrt{\beta_{os}^T \Sigma_{W_p} \beta_{os}} \quad . \tag{41}$$

Furthermore, the ℓ_1-weights are related as $\lambda_{1,lda} = \lambda_{1,os}/2$.

A.2 Expression profiles for the sparse directions

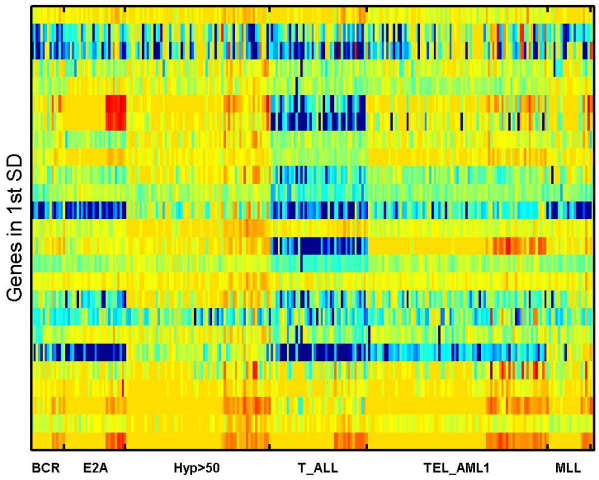


Figure 7: Expression profile of the 6 leukemia subgroups for the 25 selected genes in the first sparse direction of SDA.

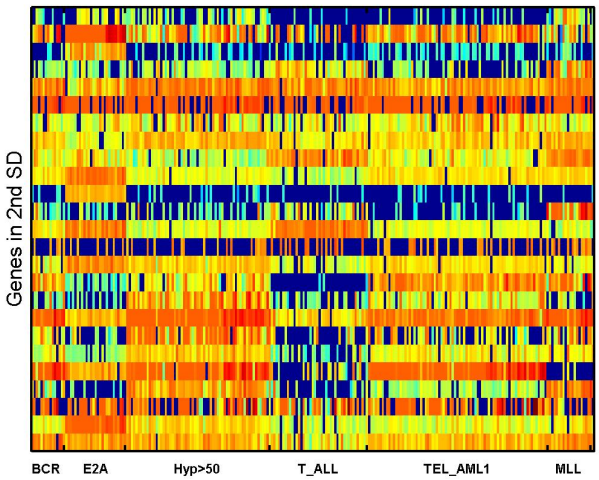


Figure 8: Expression profile of the 6 leukemia subgroups for the 25 selected genes in the second sparse direction of SDA.

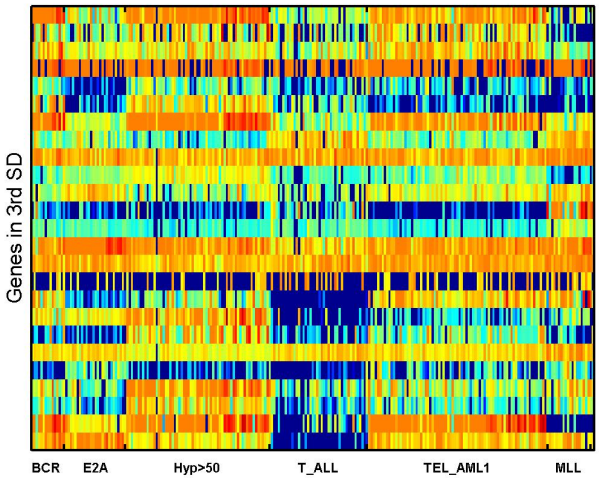


Figure 9: Expression profile of the 6 leukemia subgroups for the 25 selected genes in the third sparse direction of SDA.

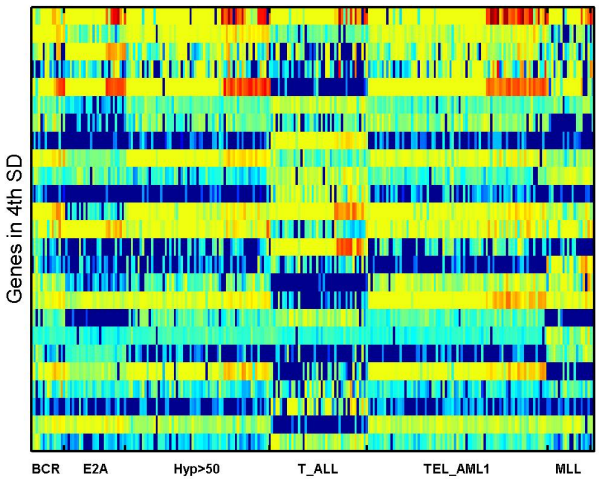


Figure 10: Expression profile of the 6 leukemia subgroups for the 25 selected genes in the fourth sparse direction of SDA.

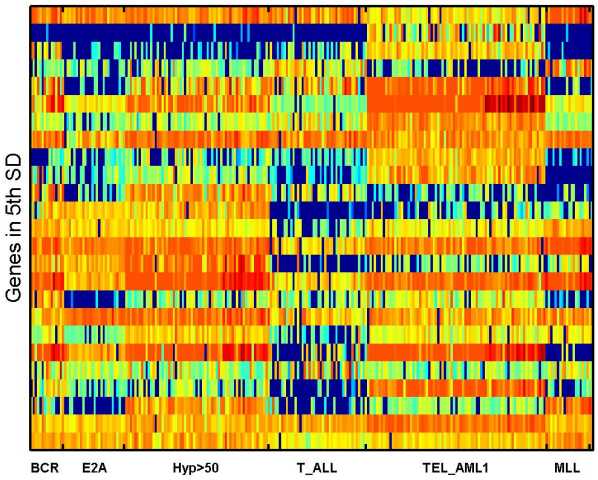


Figure 11: Expression profile of the 6 leukemia subgroups for the 25 selected genes in the fifth sparse direction of SDA.