

# Feature selection in omics prediction problems using cat scores and false non-discovery rate control

Miika Ahdesmäki <sup>\*†</sup> and Korbinian Strimmer <sup>\*</sup>

11 March 2009; revised 29 July 2009

## Abstract

We revisit the problem of feature selection in linear discriminant analysis (LDA), i.e. when features are correlated. First, we introduce a pooled centroids formulation of the multi-class LDA predictor function, in which the relative weights of Mahalanobis-transformed predictors are given by correlation-adjusted  $t$ -scores (cat scores). Second, for feature selection we propose thresholding cat scores by controlling false non-discovery rates (FNDR). Third, training of the classifier is based on James-Stein shrinkage estimates of correlations and variances, where regularization parameters are chosen analytically without resampling. Overall, this results in an effective and computationally inexpensive framework for high-dimensional prediction with natural feature selection. The proposed shrinkage discriminant procedures are implemented in the R package “sda” available from the R repository CRAN.

---

<sup>\*</sup>Institute for Medical Informatics, Statistics and Epidemiology, University of Leipzig, Härtelstr. 16–18, D-04107 Leipzig, Germany

<sup>†</sup>Department of Signal Processing, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland

# 1 Introduction

Class prediction of biological samples based on their genetic or proteomic profile is now a routine task in genomic studies. Accordingly, many classification methods have been developed to address the specific statistical challenges presented by these data – see, e.g., Schwender et al. (2008) and Slawski et al. (2008) for recent reviews. In particular, the small sample size  $n$  renders difficult the training of the classifier, and the large number of variables  $p$  makes it hard to select suitable features for prediction.

Perhaps surprisingly, despite the many recent innovations in the field of classification methodology, including the introduction of sophisticated algorithms for support vector machines and the proposal of ensemble methods such as random forests, the conceptually simple approach of linear discriminant analysis (LDA) and its sibling, diagonal discriminant analysis (DDA), remain among the most effective procedures also in the domain of high-dimensional prediction (Efron, 2008a; Hand, 2006; Efron, 1975).

In order to be applicable for high-dimensional analysis, it has been recognized early that regularization is essential (Friedman, 1989). Specifically, when training the classifier, i.e. when estimating the parameters of the discriminant function from training data, particular care needs to be taken to accurately infer the (inverse) covariance matrix. A rather radical, yet highly effective way to regularize covariance estimation in high dimensions is to set all correlations equal to zero (Bickel and Levina, 2004). Employing a diagonal covariance matrix reduces LDA to the special case of diagonal discriminant analysis (DDA), also known in the machine learning community as “naive Bayes” classification.

In addition to facilitating high-dimensional estimation of the prediction function, DDA has one further key advantage: it is straightforward to conduct feature selection. In the DDA setting with two classes ( $K = 2$ ) it can be shown that the optimal criterion for ordering features relevant for prediction are the  $t$ -scores between the two group means (e.g., Fan and Fan, 2008), or in the multi-class setting, the  $t$ -scores between group means and the overall centroid.

The nearest shrunken centroids (NSC) algorithm (Tibshirani et al., 2002, 2003), commonly known by the name of “PAM” after its software implementation, is a regularized version of DDA with multi-class feature selection. The fact that PAM has established itself as one of the most popular methods for classification of gene expression data is ample proof that DDA-type procedures are indeed very effective for large-scale prediction problems - see also Bickel and Levina (2004) and Efron (2008a).

However, there are now many omics data sets where correlation among predictors is an essential feature of the data and hence cannot easily be ignored. For example, this includes proteomics, imaging, and metabolomics data where correlation among biomarkers is commonplace and induced by spatial dependencies and by chemical similarities, respectively. Furthermore, in many transcriptome measurements there are correlations among genes within a functional group or pathway (Ackermann and Strimmer, 2009).

Consequently, there have been several suggestions to generalize PAM to account for

correlation. This includes the SCRDA (Guo et al., 2007), Clanc (Dabney and Storey, 2007) and MLDA (Xu et al., 2009) approaches. All these methods are regularized versions of LDA, and hence offer automatic provisions for gene-wise correlations. However, in contrast to PAM and DDA, they lack an efficient and elegant feature selection scheme, due to problems with multiple optima in the choice of regularization parameters (SCRDA) and the large search space for optimal feature subsets (Clanc).

In this paper, we present a framework for efficient high-dimensional LDA analysis. This is based on three cornerstones. First, we employ James-Stein shrinkage rules for training the classifier. All regularization parameters are estimated from the data in an analytic fashion without resorting to computationally expensive resampling. Second, we use correlation-adjusted  $t$ -scores (cat scores) for feature selection. These scores emerge from a restructured version of the LDA equations and enable simple and effective ranking of genes even in the presence of correlation. Third, we employ false non-discovery rate thresholding for selecting features for inclusion in the prediction rule. As we will show below, this is a highly effective method with similar performance to the recently proposed “higher criticism” approach.

The remainder of the paper is organized as follows. In Sections 2–5 we detail our framework for shrinkage discriminant analysis and variable selection. Subsequently, we demonstrate the effectiveness of our approach by application to a number of high-dimensional genomic data sets. We conclude with a discussion and comparison to closely related approaches.

## 2 Linear discriminant analysis revisited

### 2.1 Standard formulation

LDA starts by assuming a mixture model for the  $p$ -dimensional data  $x$

$$f(x) = \sum_{j=1}^K \pi_j f(x|j),$$

where each of the  $K$  classes is represented by a multivariate normal density

$$f(x|k) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \times \exp\left\{-\frac{1}{2}(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)\right\}$$

with group-specific centroids  $\mu_k$  and a common covariance matrix  $\Sigma$ . The probability of group  $k$  given  $x$  is computed from the a priori mixing weights  $\pi_j$  by application of Bayes’ theorem,

$$\Pr(k|x) = \frac{\pi_k f(x|k)}{f(x)}.$$

We define here the LDA discriminant score as the log posterior  $d_k(\mathbf{x}) = \log\{\Pr(k|\mathbf{x})\}$ , which after dropping terms constant across groups becomes

$$d_k^{\text{LDA}}(\mathbf{x}) = \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log(\pi_k). \quad (1)$$

Due to the common covariance,  $d_k^{\text{LDA}}(\mathbf{x})$  is linear in  $\mathbf{x}$ . Prediction in LDA works by evaluating the discriminant function at the given test sample  $\mathbf{x}$  for all possible  $k$ , choosing the class maximizing the posterior probability (and hence  $d_k^{\text{LDA}}$ ).

## 2.2 Pooled centroid formulation

We now rewrite the standard form of the LDA predictor function (Eq. 1) with the aim to elucidate the influence of each individual variable in prediction. Specifically, we simply add a class-independent constant to the discriminant function – note that this does not change in any way the prediction. We compute the pooled mean

$$\boldsymbol{\mu}_{\text{pool}} = \sum_{j=1}^K \frac{n_j}{n} \boldsymbol{\mu}_j,$$

representing the overall centroid ( $n_j$  is the sample size in group  $j$  and  $n = \sum_{j=1}^K n_j$  the total number of observations) and the corresponding discriminant score

$$d_{\text{pool}}^{\text{LDA}}(\mathbf{x}) = \boldsymbol{\mu}_{\text{pool}}^T \boldsymbol{\Sigma}^{-1} \mathbf{x} - \frac{1}{2} \boldsymbol{\mu}_{\text{pool}}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_{\text{pool}}.$$

The centered score

$$\Delta_k^{\text{LDA}}(\mathbf{x}) = d_k^{\text{LDA}}(\mathbf{x}) - d_{\text{pool}}^{\text{LDA}}(\mathbf{x})$$

can be interpreted as log posterior ratio and is in terms of prediction completely equivalent to the original  $d_k^{\text{LDA}}(\mathbf{x})$ . After some careful algebra it simplifies to

$$\Delta_k^{\text{LDA}}(\mathbf{x}) = \boldsymbol{\omega}_k^T \boldsymbol{\delta}_k(\mathbf{x}) + \log(\pi_k) \quad (2)$$

with feature weight vector

$$\boldsymbol{\omega}_k = \mathbf{P}^{-1/2} \mathbf{V}^{-1/2} (\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\text{pool}}) \quad (3)$$

and Mahalanobis-transformed predictors

$$\boldsymbol{\delta}_k(\mathbf{x}) = \mathbf{P}^{-1/2} \mathbf{V}^{-1/2} \left( \mathbf{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_{\text{pool}}}{2} \right). \quad (4)$$

Here, we have made use of the variance-correlation decomposition of the covariance matrix  $\boldsymbol{\Sigma} = \mathbf{V}^{1/2} \mathbf{P} \mathbf{V}^{1/2}$ , where  $\mathbf{V} = \text{diag}\{\sigma_1^2, \dots, \sigma_p^2\}$  is a diagonal matrix containing the variances and  $\mathbf{P} = (\rho_{ij})$  is the correlation matrix.

A remarkable property of the above restructuring (Eq. 2–Eq. 4) of the LDA discriminant function (Eq. 1) is that both  $\boldsymbol{\omega}_k$  and  $\boldsymbol{\delta}_k(\mathbf{x})$  are *vectors* and not matrices. Furthermore, note that  $\boldsymbol{\omega}_k$  is not a function of the test data  $\mathbf{x}$  and that its components control how much each individual variable contributes to the score  $\Delta_k^{\text{LDA}}$  of group  $k$ .

### 2.3 James-Stein shrinkage rules for learning the LDA predictor

In order to train the LDA discriminant function (Eq. 1 or Eq. 2) we estimate group centroids  $\mu_k$  by their empirical means, and otherwise rely on three different James-Stein-type shrinkage rules. Specifically, we employ

1. for the correlations  $\mathbf{P}$  the ridge-type estimator from Schäfer and Strimmer (2005),
2. for the variances  $\mathbf{V}$  the shrinkage estimator from Opgen-Rhein and Strimmer (2007), and
3. for the proportions  $\pi_k$  the frequency estimator from Hausser and Strimmer (2009).

All three James-Stein-type estimators are constructed by shrinking towards suitable targets and analytically minimizing the mean squared error. The precise formulas are given in *Appendix A*. For the statistical background we refer to the above mentioned references.

We remark that the advantages of using James-Stein rules for data analysis have recently become (again) more appreciated in the literature, especially in the “small  $n$ , large  $p$ ” setting, where James-Stein-type estimators are very efficient both in a statistical as well as in a computational sense. In training of the LDA predictor function by James-Stein shrinkage we follow Dabney and Storey (2007) and Xu et al. (2009), who give a comprehensive comparison with competing approaches such as support vector machines. Slawski et al. (2008) also implement a shrinkage version of LDA.

## 3 Feature selection

### 3.1 A natural variable selection score for LDA

Following Zuber and Strimmer (2009) we define the vector  $\tau_k^{adj}$  of correlation-adjusted  $t$ -scores (cat scores) to be a scaled version of the feature weight vector  $\omega_k$ :

$$\begin{aligned} \tau_k^{adj} &\equiv \left(\frac{1}{n_k} - \frac{1}{n}\right)^{-1/2} \omega_k \\ &= \mathbf{P}^{-1/2} \times \left\{ \left(\frac{1}{n_k} - \frac{1}{n}\right) \mathbf{V} \right\}^{-1/2} (\mu_k - \mu_{\text{pool}}) \\ &= \mathbf{P}^{-1/2} \tau_k . \end{aligned} \tag{5}$$

The vector  $\tau_k$  contains the gene-wise gene-specific  $t$ -scores between the mean of group  $k$  and the pooled mean. Thus, the correlation-adjusted  $t$ -scores ( $\tau_k^{adj}$ ) are decorrelated  $t$ -scores ( $\tau_k$ ). If there is no correlation  $\tau_k^{adj}$  reduces to  $\tau_k$ . The factor  $\left(\frac{1}{n_k} - \frac{1}{n}\right)^{-1/2}$  in Eq. 5 standardizes the error of  $\hat{\mu}_k - \hat{\mu}_{\text{pool}}$  and is the same as in PAM (Tibshirani et al., 2003). Note the minus sign, which is due to correlation  $\sqrt{n_k/n}$  between  $\hat{\mu}_k$  and  $\hat{\mu}_{\text{pool}}$ <sup>1</sup>.

<sup>1</sup>The plus sign in the original PAM paper (Tibshirani et al., 2002, p. 6567) is a typographic error.

In DDA approaches, such as PAM, regularized estimates of the  $t$ -scores  $\tau_k$  are employed for feature selection. From Eqs. 2–4 it follows directly that the cat scores  $\tau_k^{adj}$  provide the most natural generalization in the LDA setting (see also Remark A).

As a summary score to measure the total impact of feature  $i \in \{1, \dots, p\}$  we use

$$S_i = \sum_{j=1}^K (\tau_{i,j}^{adj})^2, \quad (6)$$

i.e. the squared  $i$ -th component of the cat score vector  $\tau_k^{adj} = (\tau_{1,k}^{adj}, \dots, \tau_{p,k}^{adj})^T$  summed across the  $K$  groups. For comparison, PAM uses the criterion

$$S'_i = \max_{j=1, \dots, K} (|\tau_{i,j}|). \quad (7)$$

Using the squared sum of the group-specific cat scores in  $S_i$  rather than taking the maximum over the absolute values as in  $S'_i$  has two distinct advantages. First, the sample distribution of the estimated  $S_i$  is more tractable, being approximately  $\chi^2$ . Second, if a feature is discriminative with regard to more than one group this additional information is not disregarded.

### 3.2 Feature selection by controlling the false non-discovery rate

When constructing an efficient classifier it is desirable to eliminate features that provide no useful information for discriminating among classes. The conventional but computationally tedious approach is to choose the optimal threshold by estimating the prediction error via cross-validation along a grid of possible threshold values. Faster alternative thresholding procedures include “higher criticism” (Donoho and Jin, 2008), “FAIR” (Fan and Fan, 2008), and “Ebay” (Efron, 2008b). The latter two methods are primarily developed with the correlation-free setting and  $t$ -scores in mind (however, “Ebay” also offers correlation corrections for prediction errors).

Here, we advocate using the false discovery rate (FDR) framework to select features for classification. We emphasize, however, that in the problem of constructing classifiers the FDR approach can *not* be applied in the same fashion as in differential expression. In the latter case, the aim is to compile a set of genes one has confidence in to be differentially expressed. This is controlled by the FDR criterion. In contrast, when furnishing classifiers one aims at identifying with confidence the set of null features that are not informative with regard to group separation, in order to eliminate them from the classifier. This is controlled by the false *non*-discovery rate, FNDR. For a discussion of the relation between FDR and FNDR see, e.g., Strimmer (2008).

This subtle but important distinction is best illustrated by referring to Fig. 1, which plots the local FDR  $\text{fdr}(S_i) = \text{Prob}(\text{"null"}|S_i)$  computed for (and from) the statistic  $S_i$  of feature  $i$ . In a list of differentially expressed genes we decide to include, say, genes  $i$  with  $\text{fdr}(S_i) < 0.2$ . A similar constraint on the local false non-discovery rate,  $\text{fnDR}(S_i) < 0.2$ , gives a confidence set of the null genes. The local false discovery and local false non-discovery rates add up to one,  $\text{fnDR}(S_i) = 1 - \text{fdr}(S_i)$ . Hence, the set of features to be

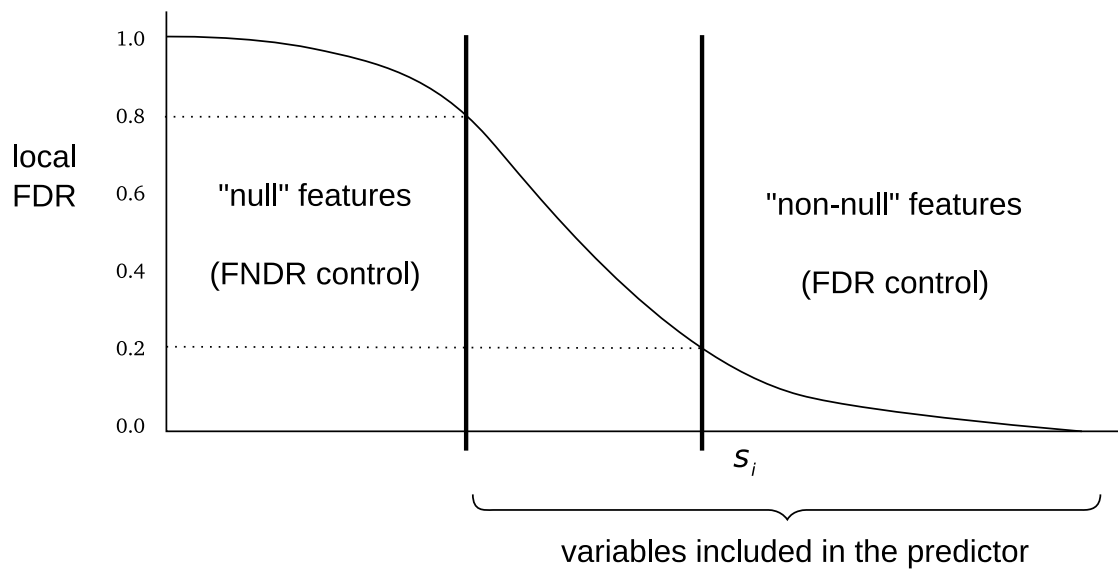


Figure 1: Local false discovery rates as a function of the summary score  $S_i$ . There are three distinct areas: an acceptance and a rejection zone, which are separated by a “buffer zone” in the middle. Note that the features to be included in the classifier by FNDR control of the null genes form a superset of the differentially expressed genes controlled by FDR.

retained in the classifier have local false discovery rates smaller than 0.8 – instead of 0.2. Thus, the features included in the predictor form a superset of the differentially expressed variables. A similar argument applies when using distribution-based Fdr ( $q$ -values) and Fndr values.

In short, our proposal is to identify the null features by controlling (local) FNDR, and subsequently using all features except the identified null set in prediction. For estimating FDR quantities we use the semiparametric approach outlined in Strimmer (2008). Note that this and other FDR procedures assume that there are enough null features so that the null model can be properly estimated (Efron, 2004).

Table 1: The general feature selection score  $S_i$  and special cases thereof.

$S_i =$	$K$ arbitrary	$K = 2$
correlation present:	$\sum_{j=1}^K (\tau_{i,j}^{adj})^2$	$2(\tau_i^{adj})^2$
no correlation ( $\mathbf{P} = \mathbf{I}$ ):	$\sum_{j=1}^K \tau_{i,j}^2$	$2\tau_i^2$

## 4 Special cases

### 4.1 Two groups

For  $K = 2$  the cat score  $\tau_k^{adj}$  between the group centroid and the pooled mean reduces to the cat score between the two means:

$$\begin{aligned}\tau_1^{adj} &= \mathbf{P}^{-1/2} \times \left\{ \left( \frac{1}{n_1} - \frac{1}{n_1 + n_2} \right) \mathbf{V} \right\}^{-1/2} (\boldsymbol{\mu}_1 - \left( \frac{n_1}{n} \boldsymbol{\mu}_1 + \frac{n_2}{n} \boldsymbol{\mu}_2 \right)) \\ &= \mathbf{P}^{-1/2} \times \left\{ \left( \frac{1}{n_1} + \frac{1}{n_2} \right) \mathbf{V} \right\}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2).\end{aligned}$$

Note that  $\tau_1^{adj} = -\tau_2^{adj}$ . The feature selection score  $S_i$  reduces to the squared cat score between the two means, cf. Tab. 1. Likewise, for  $K = 2$  the difference  $\Delta_1^{\text{LDA}}(\mathbf{x}) - \Delta_2^{\text{LDA}}(\mathbf{x})$  reduces to  $\boldsymbol{\omega}^T \boldsymbol{\delta}(\mathbf{x}) + \log\left(\frac{\pi_1}{\pi_2}\right)$  with  $\boldsymbol{\omega} = \mathbf{P}^{-1/2} \mathbf{V}^{-1/2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$  and  $\boldsymbol{\delta}(\mathbf{x}) = \mathbf{P}^{-1/2} \mathbf{V}^{-1/2} (\mathbf{x} - \frac{\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2}{2})$ .

For extensive discussion of the two group case, including comparison of gene rankings with many other test statistics, we refer to Zuber and Strimmer (2009).

### 4.2 Vanishing correlation

In case of no correlation ( $\mathbf{P} = \mathbf{I}$ ) the cat scores reduce (by construction) to standard  $t$ -scores between the two centroids of interest, either between the group and the pooled mean (general  $K$ ) or between the two groups ( $K = 2$ ). The gene summary  $S_i$  reduces to the sum of the respective squared  $t$ -scores (Tab. 1). The discriminant function reduces to the standard form of diagonal discriminant analysis. The pooled centroids formulation of LDA reduces to that of PAM (except for the shrinkage of the means present in PAM but not in our approach).

## 5 Remarks

### Remark A: Definition of feature weights

The definition of feature weights according to Eq. 3 is most natural. Other ways of splitting up the product  $\omega_k^T \delta_k(x)$  lead to various inconsistencies. For example, instead of using  $\omega_k = \Sigma^{-1/2}(\mu_k - \mu_{\text{pool}})$  it has been suggested to consider  $\Sigma^{-1}(\mu_k - \mu_{\text{pool}})$ , for example in Witten and Tibshirani (2009), page 627. However, this choice implies that for the case of no correlation variable selection would be based on  $V^{-1}(\mu_k - \mu_{\text{pool}})$  rather than on  $t$ -scores.

Furthermore, dividing the inverse correlation  $P^{-1}$  equally between Eq. 3 and Eq. 4 greatly simplifies interpretation: feature selection takes place on the level of centered, scaled as well as decorrelated predictors  $\delta_k(x)$ . Note that this interpretation is not hampered by the fact the decorrelation involves all features, because typically there is no substantial correlation between non-null and null features, so that the overall correlation matrix decomposes into correlation within null and within non-null variables.

### Remark B: Grouping of features

Using cat scores for feature selection also greatly facilitates the grouping of features. Specifically, adding the squared cat scores of each feature contained in a given set (e.g. gene sets specified by biochemical pathways) leads to Hotelling's  $T^2$ , see Zuber and Strimmer (2009). Note that if another decomposition than that of Eq. 3 and Eq. 4 was used the connection of cat scores with Hotelling's  $T^2$  would be lost.

### Remark C: FDR methods for feature selection

The usefulness of false discovery rates for feature selection in prediction is disputed, e.g. in Donoho and Jin (2008). What we show here is that the unfavorable performance is due to naive application of FDR, leading to the elimination of too many predictors. If instead FNDR is controlled to determine the null-features to be excluded from the discriminant function, then much more efficient prediction rules are obtained.

### Remark D: Fast computation of matrix square root

The inverse square root of the correlation matrix, required in Eqs. 1 and 5, can be computed very efficiently for the James-Stein shrinkage estimator, see Zuber and Strimmer (2009) for details.

### Remark E: Normalizing the null model

Estimating false discovery rates using summary scores  $S_i$  (Eq. 6) assumes as null model a  $\chi^2$ -distribution with unknown parameters. To employ standard FDR software we apply the cube-root transformation, which provides a normalizing transform for the  $\chi^2$ -distribution (Wilson and Hilferty, 1931).

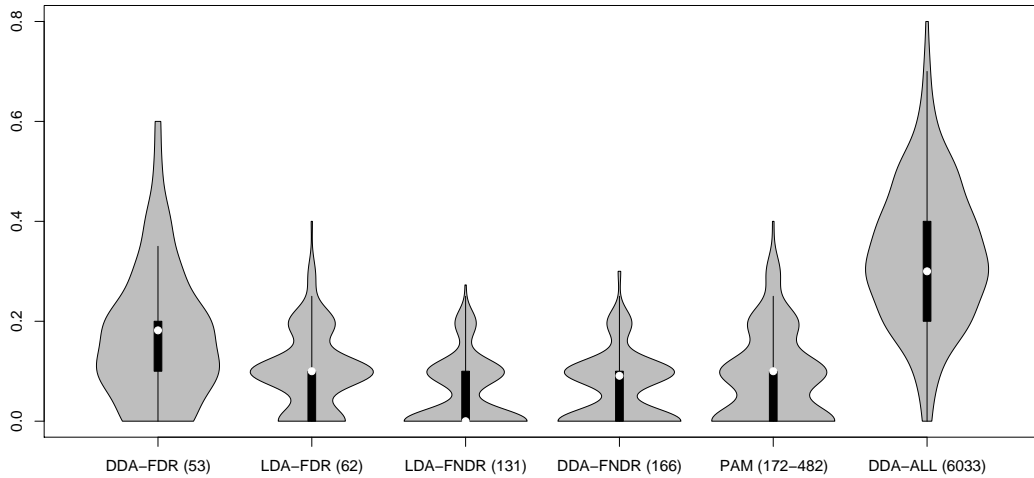


Figure 2: Violin plots of prediction error rates of various classification methods for the Singh *et al.* (2002) data. The violin plot is a generalization of the box plot, showing the median and upper and lower quartiles, as well as the density. Underlying each plot are 200 estimates of prediction error computed from the 200 splits arising from balanced 10-fold cross-validation with 20 repetitions. The number in round brackets indicates the number of selected features. See also Tab. 2.

## 6 Results

We now illustrate our shrinkage DDA and LDA approaches with variable selection using cat scores and FNDR control by analyzing a number of reference data examples, and compare our results with that of competing approaches. We also investigate the performance of FNDR feature selection in comparison with that of “higher criticism” (Donoho and Jin, 2008).

### 6.1 Singh et al. (2002) gene expression data

First, we investigated the prostate cancer data set of Singh et al. (2002). This consists of gene expression measurements of  $p = 6033$  genes for  $n = 102$  patients, of which 52 are cancer patients and 50 are healthy (thus  $K = 2$ ). To facilitate cross-comparison we analyzed the data exactly in form as used in Efron (2008a). Our results are summarized in Tab. 2, and corresponding violin plots (Hintze and Nelson, 1998) are shown in Fig. 2.

Initially, we assumed zero correlation and applied the shrinkage DDA method. By controlling the local FNDR to be smaller or equal than 0.2 we determined that 5867

Table 2: Prediction errors and number of selected features for Singh *et al.* (2000) gene expression data. The number in the round brackets is the estimated standard error.

Method	Prediction Error	Features
Ebay	0.092	51
DDA-FDR	0.1682 (0.0093)	53
LDA-FDR	0.0989 (0.0056)	62
LDA-FNDR	<b>0.0550</b> (0.0048)	131
DDA-FNDR	0.0640 (0.0049)	166
PAM	0.0859 (0.0063)	172–482
DDA-ALL	0.3327 (0.0099)	6033

The prediction error of Ebay is taken from Efron (2008a).

genes were null genes, hence that 166 genes needed to be included in the prediction rule. For comparison, a local FDR cutoff on the same level yielded only 53 genes, lacking the 103 genes in the “buffer zone” between the two cutoffs (cf. Fig. 1). Note that we recommend using the larger FNDR-based feature set, not just the 53 genes considered to be differentially expressed.

We estimated the prediction error of the resulting classification rule using balanced 10-fold cross-validation with 20 repetitions. For each of the in total 200 splits we trained a new prediction rule and estimated new feature rankings and FDR statistics, thereby including the selection process in the error estimate to avoid overoptimistic results (Ambroise and McLachlan, 2002). The number of selected features shown in Tab. 2 is based on the complete data. However, for estimation of prediction error for each of the splits a new set of features was determined.

For the FNDR-based cutoff with 166 included features we obtained an estimate of the prediction error of 0.0640 whereas for the naive FDR cutoff resulting in 53 predictors the error is much higher (0.1682). For comparison, we also computed the error using all 6033 features, yielding a massive 0.3327. The PAM program selected between 172 and 482 genes for inclusion in its predictor with error rate 0.0859 (note that the number of selected features by the PAM algorithm is highly variable and differs from run to run even for the same data set). According to Efron (2008a) the Ebay approach used 51 genes for prediction with error rate 0.092.

If correlation was taken into account, i.e. if the order of ranking was determined by cat rather than *t*-scores, interestingly both the number of differentially expressed genes and of the null genes increases, implying that the “buffer zone” shown in Fig. 1 becomes smaller. Thus, the LDA classifier with FNDR cutoff contained for this data fewer predictors (131) but at the same time nevertheless achieved the smallest overall prediction error (Fig. 2).

Table 3: Estimated prediction errors for several multi-class reference data sets.

Data	Method	Prediction Error	Features	DE
Lymphoma ( $K = 3, n = 62,$ $p = 4026$ )	DDA-FNDR	0.0517 (0.0062)	162	0
	LDA-FNDR	<b>0.0036</b> (0.0018)	392	55
	PAM	0.0254 (0.0045)	2796–3201	
SRBCT ( $K = 4, n = 63,$ $p = 2308$ )	DDA-FNDR	0.0007 (0.0007)	90	62
	LDA-FNDR	<b>0.0000</b> (0.0000)	89	76
	PAM	0.0145 (0.0034)	39–87	
Brain ( $K = 5, n = 42,$ $p = 5597$ )	DDA-FNDR	0.1892 (0.0146)	33	8
	LDA-FNDR	<b>0.1525</b> (0.0120)	102	23
	PAM	0.1939 (0.0112)	197–5597	

The last column (DE) shows the number of differentially expressed genes, which equals the number of significant features if FDR rather than FNDR is used as criterion.

## 6.2 Performance for multi-class reference data sets

For extended comparison we applied our approach to a number of further reference data sets. In particular we analyzed gene expression data for lymphoma (Alizadeh et al., 2000), small round blue cell tumors (SRBCT) (Khan et al., 2001) and brain cancer (Pomeroy et al., 2002). The data sets have in common that all contain more than two classes, thus allowing to study the multi-class summary statistic (Eq. 6). A summary of the results obtained by shrinkage LDA/DDA and FNDR feature selection and by PAM is given in Tab. 3.

The Khan et al. (2001) data are very easy to classify. All methods performed equally well on this data, with no substantial difference between the LDA and DDA approaches.

For the lymphoma data set the PAM approach failed to identify a compact set of predictive features. In contrast, the FNDR approach selects a comparatively small number of genes both in the LDA and DDA case. Intriguingly, for this data there were no differentially expressed genes, if correlation is ignored, yet the FNDR criterion yielded 162 non-null features.

The brain data set is the largest and most difficult data set. Again, the PAM approach failed to determine a stable set of features, whereas FNDR control yielded a compact set of informative predictors. Here, as well as for the lymphoma data, the LDA approach clearly outperformed the DDA approaches in terms of prediction error.

## 6.3 Comparison with “higher criticism” feature selection

Using the data examples above we demonstrated that feature selection based on simple FDR cutoffs is not sufficient for prediction. In particular, if features are weak and sparse it may easily happen that no predictor has sufficiently small false discovery rate to be called significant (cf. the lymphoma data).

Table 4: Estimated prediction errors employing higher criticism as feature selection criterion.

Data	Method	Prediction Error	Features	local FDR
Prostate	DDA-HC	0.0707 (0.0055)	129	0.69
	LDA-HC	0.0497 (0.0045)	116	0.73
Lymphoma	DDA-HC	0.0185 (0.0038)	179	1.00
	LDA-HC	0.0000 (0.0000)	345	0.78
SRBCT	DDA-HC	0.0035 (0.0016)	138	1.00
	LDA-HC	0.0007 (0.0007)	174	1.00
Brain	DDA-HC	0.1572 (0.0118)	33	0.77
	LDA-HC	0.1417 (0.0108)	131	1.00

The last column (local FDR) shows the local FDR of the least significant feature.

In such a setting Donoho and Jin (2008) suggest as alternative to FDR-based thresholding the “higher criticism” (HC) approach. The HC criterion is based on  $p$ -values. For each feature, the  $p$ -value is centered and standardized using the estimated mean and variance of the corresponding order statistic. The optimal threshold is determined as the maximum of the absolute HC scores within a fraction (say 10%) of the top ranking features (Donoho and Jin, 2008).

Our feature selection approach based on FNDR control shares with HC that we aim to overcome the limitations resulting from naive application of FDR-based feature selection. For this reason, it is instructive to investigate our shrinkage prediction rule in combination with the HC thresholding procedure. The  $p$ -values underlying the HC objective function were obtained by fitting a two-component mixture model, so that the same empirical null model was used as in the FDR analysis.

The results are given in Tab. 4. Again, in all cases the LDA approach using cat scores for feature selection leads to smaller prediction error than employing DDA and  $t$ -scores. Remarkably, the performance of FNDR and HC approach are on an equal level, implying that efficient feature selection is indeed possible *within* the FDR framework. The set of features selected by HC is on average a bit smaller than that chosen by FNDR, and larger than the FDR-based set, which indicates that the HC threshold is typically situated in the “buffer zone” of Fig. 1.

## 7 Discussion

### 7.1 Shrinkage discriminant analysis and feature selection

In this paper we have revisited high-dimensional shrinkage discriminant analysis and presented a very efficient procedure for prediction. Our approach contains three distinct elements:

- the use of James-Stein shrinkage for training the predictor,
- feature ranking based on cat scores, and
- feature selection based on FNDR thresholding.

Employing James-Stein shrinkage estimators is efficient both from a statistical as well as from a computational perspective. Note that shrinkage is used here only as a means to improve the estimated parameters, but not for model selection as in the approaches by Tibshirani et al. (2002) and Guo et al. (2007).

The correlation-adjusted  $t$ -score (cat score) emerges as a natural gene ranking criterion in the presence of correlation among predictors (Zuber and Strimmer, 2009). Here we have shown how to employ cat scores in the multi-class LDA setting and demonstrated on high-dimensional data that using cat scores rather than  $t$ -scores leads to more effective choice of predictors. We note that the order of ranking induced by the cat and  $t$ -scores, respectively, may differ substantially. Hence univariate thresholding procedures to select interesting features will differ, even if the testing procedures account for dependencies.

Finally, we propose feature selection by controlling FNDR rather FDR and show that this is as efficient in terms of predictive accuracy as when “higher criticism” is employed. Moreover, we explain why variable selection based on FDR leads to inferior prediction rules.

### 7.2 Recommendations

For extremely high-dimensional data estimating correlation is very difficult, hence in this instance we recommend to conduct diagonal discriminant analysis (see also Bickel and Levina (2004)). From our analysis it is clear the shrinkage DDA as proposed here, combined with variable selection by control of FNDR or HC, is most effective. In contrast to the PAM approach no randomization procedures are involved and hence the prediction rule and the number of selected features is stable.

In all other cases we recommend a full shrinkage LDA analysis, with feature selection based on cat scores. While this approach is computationally more expensive than the shrinkage DDA approach, it has a significant impact on predictive accuracy. Typically, in comparison with DDA taking account of correlation either leads to more compact feature sets or improved prediction error, or both. Furthermore, relative to competing full LDA approaches, such as Guo et al. (2007), our procedure is computationally fast, due to the avoidance of computer-intensive procedures such as resampling.

## Acknowledgments

We thank Verena Zuber for critical comments and helpful discussion. M.A. is grateful to the Alexander von Humboldt Foundation for a postdoctoral research fellowship.

## Appendix A: James-Stein shrinkage estimators for training the LDA predictor

For “small  $n$ , large  $p$ ” inference of the LDA predictor function (Eqs. 1 and 2) and the cat score (Eq. 5) we rely on three different James-Stein-type estimators.

The correlation matrix is estimated by shrinking empirical correlations  $r_{ij}$  towards zero,

$$r_{ij}^{\text{shrink}} = (1 - \hat{\lambda}_1)r_{ij} ,$$

with estimated intensity

$$\hat{\lambda}_1 = \min\left(1, \frac{\sum_{i \neq j} \widehat{\text{Var}}(r_{ij})}{\sum_{i \neq j} r_{ij}^2}\right)$$

(Schäfer and Strimmer, 2005).

The variances are estimated by shrinking the empirical estimates  $v_i$  towards their median,

$$v_i^{\text{shrink}} = \hat{\lambda}_2 v_{\text{median}} + (1 - \hat{\lambda}_1)v_i ,$$

using

$$\hat{\lambda}_2 = \min\left(1, \frac{\sum_{i=1}^p \widehat{\text{Var}}(v_i)}{\sum_{i=1}^p (v_i - v_{\text{median}})^2}\right)$$

(Opgen-Rhein and Strimmer, 2007).

The class frequencies are estimated following Hausser and Strimmer (2009) by

$$\hat{\pi}_j^{\text{shrink}} = \hat{\lambda}_3 \frac{1}{K} + (1 - \hat{\lambda}_3) \frac{n_j}{n} ,$$

using

$$\hat{\lambda}_3 = \frac{1 - \sum_{j=1}^K \left(\frac{n_j}{n}\right)^2}{(n-1) \sum_{j=1}^K \left(\frac{1}{K} - \frac{n_j}{n}\right)^2} .$$

## Appendix B: Relationship to other DDA and LDA approaches

Our proposed shrinkage discriminant approach is closely linked to a number of recently proposed methods.

## **NSC**

The NSC / PAM classification rule was first presented in Tibshirani et al. (2002) and later discussed in more statistical detail in Tibshirani et al. (2003). PAM is a DDA approach, so no gene-wise correlations are taken into account. Genes are ranked according to Eq. 7, and feature selection is determined by soft-thresholding, using prediction error estimated by crossvalidation as optimality criterion.

## **Ebay**

The “Ebay” approach of Efron (2008a) is also a DDA approach. Feature selection is based on an empirical Bayes model that links prediction error with false discovery rates. Thus, it is very similar to PAM but computationally and statistically more efficient. In addition, the “Ebay” algorithm provides correlation corrections of prediction errors, see Section 5 in Efron (2008b).

## **Clanc and MLDA**

The “Clanc” algorithm is described in Dabney and Storey (2007) and the “modified LDA” (MLDA) in Xu et al. (2009). Both methods are based on the LDA framework, and both use James-Stein shrinkage to estimate the pooled covariance matrix. MLDA uses standard  $t$ -scores for feature selection, whereas Clanc employs a greedy algorithm search to find optimal subsets of features based on a multivariate criterion.

## **SCRDA**

The “shrunk centroids regularized discriminant analysis” (SCRDA) procedure is described in Guo et al. (2007) and uses a similar soft-thresholding procedure for variable selection as PAM. The covariance matrix is estimated by a ridge estimator. Regularization and feature selection parameters are simultaneously determined by cross-validation. The main issues with SCRDA are the computational expense and problems in finding unique parameters (Guo et al., 2007).

## **Appendix C: Computer implementation**

We have implemented the proposed shrinkage discriminant procedures (both DDA and LDA) and the associated FNDR and higher criticism variable selection in the R package “sda”, which is freely available under the terms of the GNU General Public License (version 3 or later) from CRAN (<http://cran.r-project.org/web/packages/sda/>).

## References

- Ackermann, M. and Strimmer, K. (2009). A general modular framework for gene set enrichment. *BMC Bioinformatics*, 10:47.
- Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T., Hudson, J., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511.
- Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. USA*, 99:6562–6566.
- Bickel, P. J. and 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*, 10:989–1010.
- Dabney, A. R. and Storey, J. D. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS ONE*, 2:e1002.
- Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA*, 105:14790–15795.
- Efron, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *J. Amer. Statist. Assoc.*, 70:892–896.
- Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.*, 99:96–104.
- Efron, B. (2008a). Empirical Bayes estimates for large-scale prediction problems. *Technical Report*, Dept. of Statistics, Stanford University.
- Efron, B. (2008b). Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.*, 23:1–22.
- Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, 36:2605–2637.
- Friedman, J. H. (1989). Regularized discriminant analysis. *J. Amer. Statist. Assoc.*, 84:165–175.
- Guo, Y., Hastie, T., and Tibshirani, T. (2007). Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100.
- Hand, D. J. (2006). Classifier technology and the illusion of progress. *Statistical Science*, 21:1–14.

- Hausser, J. and Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, 10:in press.
- Hintze, J. L. and Nelson, R. D. (1998). Violin plots: a box plot-density trace synergism. *The American Statistician*, 52:181–184.
- Khan, J., Wei, J. S., Ringner, M., Saal, L. H., Ladanyi, M., Westermann, F., Berthold, F., Schwab, M., Antonescu, C. R., Peterson, C., and Meltzer, P. S. (2001). Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.*, 7:673–679.
- Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6:9.
- Pomeroy, S. L., Tamayo, P., Gaasenbeek, M., Sturla, L. M., Angelo, M., McLaughlin, M. E., Kim, J. Y. H., Goumnerova, L. C., Black, P. M., Lau, C., Allen, J. C., Zagzag, D., Olson, J. M., Curran, T., Wetmore, C., Biegel, J. A., Poggio, T., Mukherjee, S., Rifkin, R., Califano, A., Stolovitzky, G., Louis, D. N., Mesirov, J. P., Lander, E. S., and Golub, T. R. (2002). Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442.
- Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4:32.
- Schwender, H., Ickstadt, K., and Rahnenführer, J. (2008). Classification with high-dimensional genetic data: assigning patients and genetic features to known classes. *Biometr. J.*, 50:911–926.
- Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209.
- Slawski, M., Daumer, M., and Boulesteix, A.-L. (2008). CMA - a comprehensive Bio-conductor package for supervised classification with high dimensional data. *BMC Bioinformatics*, 9:439.
- Strimmer, K. (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303.
- Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer type by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, 99:6567–6572.
- Tibshirani, R., Hastie, T., Narsimhan, B., and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, 18:104–117.

- Wilson, E. and Hilferty, M. (1931). The distribution of chi-square. *Proc. Nat. Acad. Sci.*, 17:684–688.
- Witten, D. M. and Tibshirani, R. (2009). Covariance-regularized regression and classification for high-dimensional problems. *J. R. Statist. Soc. B*, 71:615–636.
- Xu, P., Brock, G. N., and Parrish, R. S. (2009). Modified linear discriminant analysis approaches for classification of high-dimensional microarray data. *Comp. Stat. Data Anal.*, 53:1674–1687.
- Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics*, in press.