Journal of Data Science **11**(2013), 537-558

Effect Size Estimation and Misclassification Rate Based Variable Selection in Linear Discriminant Analysis

Bernd Klaus
European Molecular Biology Laboratory (EMBL)

Abstract:

Supervised classifying of biological samples based on genetic information, (e.g., gene expression profiles) is an important problem in biostatistics. In order to find both accurate and interpretable classification rules variable selection is indispensable.

This article explores how an assessment of the individual importance of variables (effect size estimation) can be used to perform variable selection. I review recent effect size estimation approaches in the context of linear discriminant analysis (LDA) and propose a new conceptually simple effect size estimation method which is at the same time computationally efficient.

I then show how to use effect sizes to perform variable selection based on the misclassification rate, which is the data independent expectation of the prediction error. Simulation studies and real data analyses illustrate that the proposed effect size estimation and variable selection methods are competitive. Particularly, they lead to both compact and interpretable feature sets.

Program files to be used with the statistical software `R` implementing the variable selection approaches presented in this article are available from my homepage: http://b-klaus.de.

Key words: Correlation-adjusted t-score, effect size estimation, linear discriminant analysis, misclassification rate, variable selection.

1. Introduction

Modern medical research has been revolutionized by the possibility of characterizing diseases at a molecular level using microarrays. Classification of biological samples based on their gene expression continues to be a field of active research. See e.g., Cao *et al.* (2011), Pang *et al.* (2009), Xiaosheng and Simon (2011) and Shao *et al.* (2011). Current reviews of the subject can be found in Schwender *et al.* (2008), Slawski *et al.* (2008) as well as in Kim and Simon (2011).

In order to develop classifiers which are potentially useful for molecular diagnostics, it is important to construct them based on a selection of genes (variables) strongly associated with the respective class labels (e.g., cancer and healthy tissue). These genes have a large effect size which is generally measured by standardized differences.

Three distinct but closely related objectives need to be achieved to identify a group of genes with high effect sizes (Ahdesmäki and Strimmer, 2010, Matsui and Noma, 2011):

(i) to establish a reliable variable ranking,

(ii) to provide a reasonable estimate of the effect size for each gene, and

(iii) to find a suitable cutoff point that allows to disregard (the usually large) number of noise-features.

Problems (ii) and (iii) are the main concerns of the current chapter. For the ranking problem (obj. (i)), I will rely on correlation adjusted t–scores (a.k.a. "cat" – scores) introduced by Zuber and Strimmer (2009). The cat–score is a t–type statistic which takes correlation into account and has been shown to induce a reliable variable ranking even in the presence of correlation among the variables. I therefore am going to use cat–scores to obtain effect size estimates (obj. (ii)). Based on these estimates, a nominal prediction error is computed. It is dependent on the number of variables included. Variable selection is then performed (ob. (iii)) by determining the number of variables necessary to achieve a certain nominal error level.

The approach presented here is similar to that of Efron (2009) and Dabney and Storey (2007). However, in contrast to Efron (2009), my method applies to any number of classes and allows empirical null modeling. In contrast to Dabney and Storey (2007), it does not need a computationally expensive greedy algorithm to select variables due to the variable ranking performed beforehand.

The article is organized as follows: I will present basic theory on LDA in Chapter 2, then I obtain effect size estimates based on cat–scores and compare them to other effect size estimation approaches in Section 3. Notably, the methods of Efron (2009) and Matsui and Noma (2011) are presented in a unifying way using cat–scores, which sheds new light on their similarities. Section 4 shows how to perform variable ranking and selection using different methods based on a variable ranking. Results of variable selection methods on simulated and real data are then presented in Section 5.

2. Linear Discriminant Analysis (LDA) and Its Misclassification Rate

2.1 Linear Discriminant Analysis (LDA) and Effect Sizes

LDA forms the basis of most classification algorithms currently employed, e.g., Nearest Shrunken Centroids commonly abbreviated as NSC, and also known as Prediction Analysis for Microarrays (PAM), see Tibshirani *et al.* (2003), Shrinkage Discriminant Analysis – SDA, Ahdesmäki and Strimmer (2010) – and many more. It starts by assuming a mixture model for the d-dimensional data $\boldsymbol{x}$

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}|k),$$

where each class k is represented by a multivariate normal density

$$f(\boldsymbol{x}|k) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\right\},$$

with group–specific centroids $\boldsymbol{\mu}_k$ and a common covariance matrix $\boldsymbol{\Sigma}$. A sample $\boldsymbol{x}$ is assigned to the class yielding the highest LDA discriminant score defined as the log posterior probability $d_k^{\mathrm{LDA}}(\boldsymbol{x}) = \log\{P(k|\boldsymbol{x})\}$. This score can be written as

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

The standard form of the LDA predictor function shown in (1) can be transformed into a scalar product which is given by

$$\Delta_k^{\mathrm{LDA}}(\boldsymbol{x}) = \left(\boldsymbol{\omega}^{(k,\mathrm{pool})}\right)^T \boldsymbol{\delta}_k(\boldsymbol{x}) + \log(\pi_k). \tag{2}$$

See Ahdesmäki and Strimmer (2010) for details. In (2), we have an inner product of Mahalanobis transformed variables (commonly called features) $\boldsymbol{\delta}(\boldsymbol{x})$ and a corresponding feature weight vector $\boldsymbol{\omega}^{(k,\mathrm{pool})}$ given by

$$\boldsymbol{\delta}_k(\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_{\mathrm{pool}}}{2}\right), \tag{3}$$

and

$$\boldsymbol{\omega}^{(k,\mathrm{pool})} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\mathrm{pool}}), \tag{4}$$

respectively. In this equation the pooled mean is calculated as $\boldsymbol{\mu}_{\mathrm{pool}} = \sum_{k=1}^{K} \frac{n_k}{n}\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}$ is decomposed as: $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$, with a diagonal matrix containing the variances $\boldsymbol{V} = \mathrm{diag}\{\sigma_1^2, \cdots, \sigma_d^2\}$ and the correlation matrix $\boldsymbol{P} = (\rho_{ij})$. Remarkably, both $\boldsymbol{\omega}^{(k,\mathrm{pool})}$ and $\boldsymbol{\delta}_k(\boldsymbol{x})$ are vectors and not matrices.

The decomposition in (2) shows that $\boldsymbol{\omega}^{(k,\mathrm{pool})}$ gives the influence of the transformed variables $\boldsymbol{\delta}(\boldsymbol{x})$ in prediction. Zuber and Strimmer (2009) have shown

that this Mahalanobis–transformation leads to an improved ranking of the original variables since it removes the effect of correlation. Thus, as in Ahdesmäki and Strimmer (2010), the feature weights $\boldsymbol{\omega}$ will serve as a measure of variable importance and the terms variables and features will be used interchangeably in the following sections.

Additionally, from (4) it can be seen that the components of $\boldsymbol{\omega}^{(k,\text{pool})}$ are decorrelated and standardized differences (i.e., effect sizes) between the class k and the "pooled class" (Matsui and Noma, 2011). This is readily generalized. The effect size vector $\boldsymbol{\omega}^{(k,l)}$ between any two classes k and l is defined as the difference between the two respective feature weight vectors $\boldsymbol{\omega}^{(k,\text{pool})}$ and $\boldsymbol{\omega}^{(l,\text{pool})}$

$$\boldsymbol{\omega}^{(k,l)} := \boldsymbol{\omega}^{(k,\text{pool})} - \boldsymbol{\omega}^{(l,\text{pool})} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l). \tag{5}$$

Note that $\boldsymbol{\omega}^{(k,l)}$ is up to the scale factor $(1/n_k + 1/n_l)^{-1/2}$ equivalent to the cat–score vector between the classes k and l on the population level, i.e., assuming known model parameters (Zuber and Strimmer, 2009). Hence there is a close relationship between test statistics and effect sizes: The effect size is simply a sample size independent version of the test statistic. The statistic is denoted by a "cat" subscript in this article, i.e.,

$$\boldsymbol{\omega}^{(k,l)}_{\text{cat}} = (1/n_k + 1/n_l)^{-1/2}\boldsymbol{\omega}^{(k,l)}.$$

2.2 The Misclassification Rate of Linear Discriminant Analysis

In this section, I am going to look at an unconditional (i.e., not depending on the data) misclassification error of LDA on the population level. This quantity is called (unconditional) misclassification rate in the literature (Dabney and Storey, 2007, Shao *et al.*, 2011).

Let $\boldsymbol{x}^{(k)}$ be a sample vector drawn from the multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ associated with class k. In the LDA algorithm, it is assigned to the class yielding the highest score (1). Using the scalar product of (2) a misclassification (on the population level) of $\boldsymbol{x}^{(k)}$ occurs if $[\boldsymbol{\omega}^{(k,\text{pool})}]^T\boldsymbol{\delta}_k(\boldsymbol{x}^{(k)}) + \log(\pi_k) < \max_l[\boldsymbol{\omega}^{(l,\text{pool})}]^T\boldsymbol{\delta}_l(\boldsymbol{x}^{(k)}) + \log(\pi_l)$. It is easily verified that this is equivalent to the condition

$$\min_{l\neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x}^{(k)} - \frac{\boldsymbol{\mu}_k+\boldsymbol{\mu}_l}{2}\right)] + \log\left(\frac{\pi_k}{\pi_l}\right)}{\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}]}} < 0\,.$$

Since $\boldsymbol{x}^{(k)} \sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ holds for all $k \in \{1, \cdots, K\}$, the unconditional (i.e., expected) probability of misclassifying a sample from class k into a wrong class

$j \neq k$ can be deduced from the above formula as:

$$P(j \neq k|k) = \Phi\left(-\min_{l \neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}] + 2\log\left(\frac{\pi_k}{\pi_l}\right)}{2\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}]}} \right).$$

This results in a misclassification rate (total error probability) of

$$\begin{aligned} P(\text{error}) &= \sum_{k=1}^{K} P(j \neq k|k) \times P(k) \\ &= \sum_{k=1}^{K} \Phi\left(-\min_{l \neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}] + 2\log\left(\frac{\pi_k}{\pi_l}\right)}{2\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}]}} \right) \times \pi_k. \end{aligned} \tag{6}$$

Observe that (6) is the result of applying an expectation operator twice, once with regard to the model parameters $\boldsymbol{\omega}^{(k,l)}$ and once with regard to the transformed data $\boldsymbol{\delta}_k(\boldsymbol{x}^{(k)}) - \boldsymbol{\delta}_l(\boldsymbol{x}^{(k)}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{x}^{(k)} - (\boldsymbol{\mu}_k + \boldsymbol{\mu}_l)/2)$. The first application leads to the population version of the statistical model, with $\hat{\boldsymbol{\omega}}^{(k,l)}$ replaced by $\boldsymbol{\omega}^{(k,l)}$, the second results in an unconditional (not dependent on the data) error rate.

3. Effect Size Estimation

For two given classes k and l, a feature i with a large corresponding effect size $\omega_i^{(k,l)}$ is most influential in differentiating between class k and l. However, a "naive" estimation of $\omega_i^{(k,l)}$ (e.g., estimation by plug-in estimates) suffers from the so–called "selection bias": Estimates of $\omega_i^{(k,l)}$ are biased upwards in general. For example, an estimated effect size of 1.5 based on t–scores might correspond to a true effect size of 0.7, see Figure 1. Therefore, reliable estimates of $\omega_i^{(k,l)}$ are needed in order to furnish a good estimate of (6).

3.1 Three Empirical Bayes Approaches

Bayesian approaches are "immune" to selection effects (Dawid, 1994, Senn, 2008). Thus, Efron (2009) as well as Matsui and Noma (2011) employ empirical Bayes estimates to tackle the estimation of effect sizes.

I am going to present their ideas in a unified way using cat–scores. This will show similarities between the two methods that are not readily apparent from studying the two original papers. Therefore, both methods are presented in considerable detail in order to clearly demonstrate the conceptual overlap between them. This will also help to indicate their respective weaknesses.

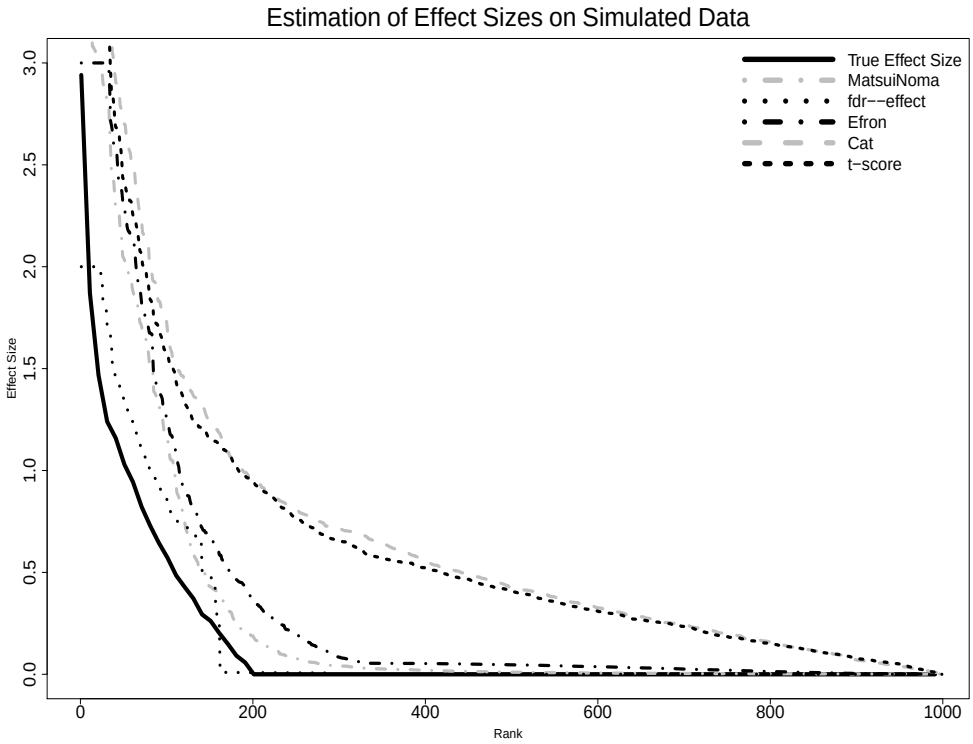

Figure 1: Comparison of effect size estimates on simulated data following the Smyth (2004) model.

Furthermore, the current section can be read as a concise and yet comprehensive review of both methods, which can be of great help to the interested reader. The empirical Bayes estimator presented in Section 3.1.3 is an attempt to combine the strengths of both approaches while addressing their shortcomings.

Let k and l be any two classes. For the sake of simplicity, the feature index i ($i \in \{1, \cdots, d\}$) will be dropped in the upcoming subsections.

3.1.1 Efron's Method

Efron (Efron, 2009) begins by transforming the statistics $\omega_{\text{cat}}^{(k,l)}$ into z–scores via a t–distribution with $n_l + n_k - 2$ degrees of freedom:

$$z = \Phi^{-1}\left(F_{n_l+n_k-2}(\omega_{\text{cat}}^{(k,l)})\right),$$

where $F_{n_l+n_k-2}$ denotes the distribution function of a t–distribution with $n_l + n_k - 2$ degrees of freedom. He then assumes a prior density g on $\omega_{\text{cat}}^{(k,l)}$ given by the mixture

$$g(\omega_{\text{cat}}^{(k,l)}) = \eta_0 I_0(\omega_{\text{cat}}^{(k,l)}) + (1 - \eta_0) g_A(\omega_{\text{cat}}^{(k,l)}), \tag{7}$$

where I_0 is a delta-function at 0 and η_0 the proportion of genes having a true effect size of zero. The alternative group, i.e., the nonzero effect sizes are represented by g_A. In the following, I will in general abbreviate conditioning on the alternative group with an "A" subscript. The statistic z is assumed to be distributed as

$$z|\omega_{\text{cat}}^{(k,l)} \sim \mathcal{N}(\omega_{\text{cat}}^{(k,l)}, 1).$$

Together with (7), this results in the following mixture model for z

$$f(z) = \eta_0\varphi(z) + (1-\eta_0)f_A(z), \tag{8}$$

where $\varphi(z)$ is the normal distribution density and f_A is a mixture of the densities $\varphi(z - \omega_{\text{cat}}^{(k,l)})$:

$$f_A(z) = \int_{-\infty}^{\infty} \varphi(z - \omega_{\text{cat}}^{(k,l)}) g_A(\omega_{\text{cat}}^{(k,l)})\, d\omega_{\text{cat}}^{(k,l)}.$$

(8) is a typical case of two–groups mixture model. It consists of a theoretical (i.e., no additional parameters) "null" model $f_0 = \varphi$ and an alternative component f_A from which the "interesting" cases are assumed to be drawn (Efron, 2008). In order to present the ideas of both Matsui and Noma (2011) and Efron (2009) in a unified fashion, I will start with computing the posterior density conditioned on the alternative i.e., $f(\omega_{\text{cat}}^{(k,l)}|z, z \in \text{"alternative"}) = f(\omega_{\text{cat}}^{(k,l)}|z, \omega_{\text{cat}}^{(k,l)} \neq 0)$. As introduced above, the "A" subscript indicates conditioning on the alternative so that $f_A(\omega_{\text{cat}}^{(k,l)}|z) = f(\omega_{\text{cat}}^{(k,l)}|z, z \in \text{"alternative"})$. Finally, using Bayes' rule this density can be computed as

$$\begin{aligned} f_A(\omega_{\text{cat}}^{(k,l)}|z) &= \frac{f_A(z|\omega_{\text{cat}}^{(k,l)}) \cdot g_A(\omega_{\text{cat}}^{(k,l)})}{f_A(z)} \\ &= \exp(\omega_{\text{cat}}^{(k,l)} z - \log\{f_A(z)/\varphi(z)\})[\exp\{-(\omega_{\text{cat}}^{(k,l)})^2/2)\}] g_A(\omega_{\text{cat}}^{(k,l)}). \end{aligned}$$

It has the form of a natural exponential family with natural parameter $\omega_{\text{cat}}^{(k,l)}$, sufficient statistic z and cumulant generating function $\log\{f_A(z)/\varphi(z)\} = \log\{[(1 - \text{fdr}(z))/\text{fdr}(z)]\} \cdot \eta_0(1-\eta_0)\}$, where

$$\text{fdr}(z) = P(\text{"null"}|z) = \eta_0 \frac{\varphi(z)}{f(z)} = \eta_0 \frac{f_0(z)}{f(z)} \tag{9}$$

is the local false discovery rate (Efron, 2008). Conditional on the alternative component, this leads to an effect size estimate of the simple form

$$E_A\left(\omega^{(k,l)}|z\right) = -(1/n_l + 1/n_k)^{1/2} \frac{d}{dz} \log\left(\frac{1-\text{fdr}(z)}{\text{fdr}(z)} \frac{\eta_0}{1-\eta_0}\right). \tag{10}$$

Since by (9) the relationship $P(\text{“alternative”}|z) = 1 - P(\text{“null”}|z) = 1 - \text{fdr}(z)$ holds, the unconditional effect size estimate is:

$$E\left(\omega^{(k,l)}|z\right) = E_A\{\omega^{(k,l)}|z\}\{1 - \text{fdr}(z)\}$$
$$= -(1/n_l + 1/n_k)^{1/2}\frac{d}{dz}\log\left\{\frac{1-\text{fdr}(z)}{\text{fdr}(z)}\frac{\eta_0}{1-\eta_0}\right\}\{1-\text{fdr}(z)\}, \quad (11)$$

which after some further calculations becomes

$$E\left(\omega^{(k,l)}|z\right) = -(1/n_l + 1/n_k)^{1/2}\frac{d}{dz}\log\{\text{fdr}(z)\}. \quad (12)$$

Note that if one used an empirical null $\mathcal{N}(0, \sigma^2)$ with estimated σ as null density f_0, the connection to the natural exponential family would be lost. Then both the natural parameter and the sufficient statistic would depend on σ.

Unfortunately, in this case the elegant formula (12) no longer holds. This basically is the only downside of Efron's approach: It is conceptually simple and computationally efficient but it is not possible to include an additional variance parameter in the null model without “destroying” (12).

3.1.2 Matsui and Noma's Method

Matsui and Noma (2011) introduce empirical null modeling into the approach of Efron (2009) via an empirical Bayes method. They start with a similar z–score transform. However, as a starting point absolute values are used:

$$z = \Phi^{-1}\left[1 - 2\cdot\left\{1 - F_{n_l+n_k-2}\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)\right\}\right].$$

Additionally, only a prior on the absolute non-null effect sizes $g_A(|\omega_{\text{cat}}^{(k,l)}|)$ is assumed. The non–null z have the conditional density

$$f_A\left(z|\left|\omega_{\text{cat}}^{(k,l)}\right|\right) = \varphi\left(\frac{\left|\omega_{\text{cat}}^{(k,l)}\right| - z}{V\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)}\right).$$

The variance function V and the prior g_A are estimated from the data. As in Efron (2009), they also assume a two-group mixture model for the z–scores:

$$f(z) = \eta_0\varphi\left(\frac{z-\mu_0}{\sigma_0}\right) + (1-\eta_0)f_A(z).$$

The null density is (in contrast to Efron) an empirical null, i.e., mean and variance are estimated from the data: $f_0(z) = \varphi\left((z-\mu_0)/\sigma_0\right)$. The alternative density

f_A is computed as:

$$f_A(z) = \int_0^\infty f_A\left(z|\left|\omega_{\text{cat}}^{(k,l)}\right|\right) g_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right) d\left|\omega_{\text{cat}}^{(k,l)}\right|$$
$$= \int_0^\infty \varphi\left(\frac{\left|\omega_{\text{cat}}^{(k,l)}\right| - z}{\sqrt{V\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)}}\right) g_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right) d\left|\omega_{\text{cat}}^{(k,l)}\right|.$$

The application of Bayes' rule gives a posterior expectation of $|\omega_{\text{cat}}^{(k,l)}|$ which is unfortunately not as simple as (10):

$$E_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right| |z\right) = \int_0^\infty \left|\omega_{\text{cat}}^{(k,l)}\right| \frac{f_A\left(z|\left|\omega_{\text{cat}}^{(k,l)}\right|\right) g_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)}{f_A(z)} d\left|\omega_{\text{cat}}^{(k,l)}\right|$$
$$= \int_0^\infty \left|\omega_{\text{cat}}^{(k,l)}\right| \frac{\varphi\left(\frac{\left|\omega_{\text{cat}}^{(k,l)}\right| - z}{\sqrt{V\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)}}\right) g_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right|\right)}{f_A(z)} d\left|\omega_{\text{cat}}^{(k,l)}\right|.$$

The statistic $|\omega_{\text{cat}}^{(k,l)}|$ is then transformed back into an absolute value effect size:

$$E_A\left(\left|\omega^{(k,l)}\right| |z\right) = (1/n_l + 1/n_k)^{1/2} F_{n_l+n_k-2}^{-1}\left(1 - \frac{1}{2}\left[1 - \Phi\left\{E_A\left(\left|\omega_{\text{cat}}^{(k,l)}\right| |z\right)\right\}\right]\right).$$

As in (12), the final effect size estimate is:

$$E\left(\left|\omega^{(k,l)}\right| |z\right) = E_A\left(\left|\omega^{(k,l)}\right| |z\right)(1 - \text{fdr}(z)). \tag{13}$$

In contrast to Efron's method, the approach of Matsui and Noma (2011) allows empirical null modeling and thus leads to better effect size estimates in general, as Matsui and Noma (2011) convincingly show in their article.

However, this increased accuracy comes at a price. The estimation of variance function V can take up to two hours. Furthermore, it has to be estimated for every number of class samples n_k and n_l separately. This makes cross-validation based assessment of predictive accuracy extremely time consuming. Additionally, even if V has been computed for fixed n_k and n_l , the estimation of the final effect size will take up to several minutes.

In summary, while Matsui and Noma (2011) provide a method that is superior to Efron's method in terms of bias, it is at the same time computationally very demanding.

3.1.3 A Simple Empirical – Bayes Approach

In this section I will derive another more heuristic approach to the reliable

estimation of effect sizes that tries to combine the advantages of Matsui and Noma's (2011) as well as Efron's (2009) methods. Empirical null modeling will be included, it will be computationally tractable and provide sufficient accuracy.

Observe that in non–empirical Bayes frameworks, reliable estimation of effect sizes is generally achieved by shrinking initial estimates of statistics playing the same role as $\omega_{\text{cat}}^{(k,l)}$. For example, in the popular PAM algorithm (Tibshirani *et al.*, 2003), the estimated t–scores are shrunk using a parameter λ estimated by cross validation.

Therefore, an appropriate adaptive shrinkage of the original test-statistic should provide us with reasonable effect size estimates. As it turns out, this adaptive shrinkage can easily be achieved by employing false discovery rates.

The first step in my heuristic approach to achieve a shrinkage of $\omega^{(k,l)}$ is the assumption of a two–component mixture model on the effect sizes:

$$f(\omega_{\text{cat}}^{(k,l)}) = \eta_0 f_0(\omega_{\text{cat}}^{(k,l)}) + (1-\eta_0) f_A(\omega_{\text{cat}}^{(k,l)}), \tag{14}$$

leading to corresponding fdr estimates of (9). Assuming a centered null distribution, we can now make use of the "naive" estimates $E_A\left(\omega^{(k,l)}\right) = \omega^{(k,l)}$ and correspondingly $E_0\left(\omega^{(k,l)}\right) = 0$ (since f_0 is centered). The 0 subscript indicates a conditioning on the null distribution, $E_0\left(\omega^{(k,l)}\right) = E\left(\omega^{(k,l)} \,|\omega^{(k,l)} \in \text{“null”}\right)$. It now holds by the law of total probability and (9) that the effect size is given by

$$\begin{aligned}
E\left(\omega^{(k,l)}\right) &= (1/n_l + 1/n_k)^{1/2} \Bigg\{ E_0\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot P\left(\omega_{\text{cat}}^{(k,l)} \in \text{“null”} | \omega_{\text{cat}}^{(k,l)}\right) \\
&\quad + E_A\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot P\left(\omega_{\text{cat}}^{(k,l)} \in \text{“alternative”} | \omega_{\text{cat}}^{(k,l)}\right) \Bigg\} \\
&= (1/n_l + 1/n_k)^{1/2} E_A\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot P\left(\omega_{\text{cat}}^{(k,l)} \in \text{“alternative”} | \omega_{\text{cat}}^{(k,l)}\right) \\
&= E_A\left(\omega^{(k,l)}\right) \cdot \left(1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})\right) \\
&= \omega^{(k,l)} \left(1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})\right).
\end{aligned} \tag{15}$$

(15) is very similar to (13) and (11), however, no full Bayesian posterior is computed. Instead, simple non–Bayesian estimates for the expectations in the two–groups model (14) are employed. This makes the implementation of (15) computationally efficient.

There is an obvious downside though: Large (with respect to their absolute value) statistics usually have a high fdr value close to 0. Therefore, they are hardly shrunk at all although their effect size is usually grossly overestimated. Thus, it is necessary to impose a minimum shrinkage. From the results of the real data analysis in Table 1 of Matsui and Noma (2011), it can easily be seen that

the empirical Bayes method that these authors apply imposes a shrinkage of at least 50% on the top 5 test statistics. I therefore also set the minimum shrinkage to 50% leading to the formula

$$\boldsymbol{\omega}_{\text{fdr}}^{(k,l)} = \boldsymbol{\omega}^{(k,l)} \cdot \min\left\{0.5;\, [1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})]\right\}. \tag{16}$$

I call this fdr–effect size estimation (fdr–effect) and abbreviate $\omega^{(k,l)}(1-\text{fdr}(\omega_{\text{cat}}^{(k,l)}))$ by $\omega_{\text{fdr}}^{(k,l)}$. Note that a fdr cutoff of 50% is conceptually very close to Higher Criticism Thresholding, see Klaus and Strimmer (2013).

Table 1: Prediction errors and number of selected features for simulation setup 1, the number in the round brackets is the estimated standard error over 25 runs. The true number of differentially expressed features is 100

Method	Prediction Error	Features
DDA–MR	0.1077 (0.0177)	156.48 (64.70)
DDA–FNDR	0.2482 (0.1272)	39.24 (23.72)
DDA–HC	0.1880 (0.0626)	152.32 (193.48)
PAM	0.0923 (0.0163)	253.6 (116.26)
DDA–ALL	0.1555 (0.0180)	500

Perhaps surprisingly, in next section it will be shown that it is competitive with regard to the attained accuracy, even though no sophisticated posterior estimates are used. The adaptive shrinkage performed in (16) can be interpreted as being in between the full empirical Bayes approaches of Efron (2009) or Matsui and Noma (2011) and soft thresholding using a single shrinkage parameter for all statistics as in Tibshirani *et al.* (2003).

3.2 Evaluation of Effect Size Estimation Methods on Real and Simulated Data

A comparison of effect size estimation methods using simulated data is shown in Figure 1. Specifically, I will compare the effect size estimation using "naive" approaches (simple cat and t–scores) and the more sophisticated ones described in the previous section abbreviated as MatsuiNoma, Efron and fdr–effect, respectively. For the methods MatsuiNoma and Efron, I use the implementations offered by the authors, for fdr–effect, I perform cat–score and fdr estimation using the `R`–packages (R Development Core Team, 2012) `st` and `fdrtool` (Strimmer, 2008a). In the real data analysis displayed in Figure 2, the package `locfdr` (Efron, 2004, 2007, 2008) is applied since this allows a straightforward use of an theoretical null as it has been suggested in Matsui and Noma (2011) and Efron (2004) for this data set.

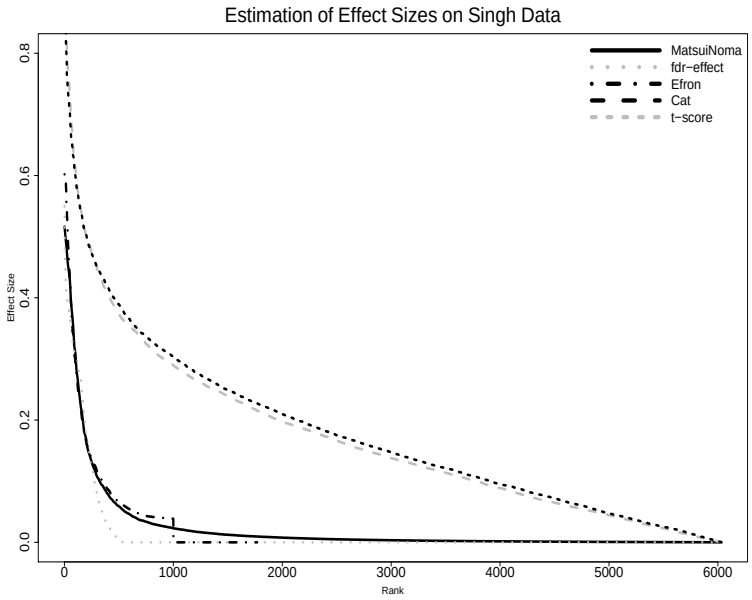

Figure 2: Comparison of effect size estimates for the Singh *et al.* (2002) data

I am going to follow closely the setup used in Opgen-Rhein and Strimmer (2007), Smyth (2004) and Zuber and Strimmer (2009) to simulate gene expression data. The parameters are chosen in such a way that effect sizes between 1 and 3 are obtained, which roughly corresponds to the range considered in the simulation studies of Matsui and Noma (2011).

The number of statistics was fixed at $d = 1000$ with 200 statistics designated to be differentially expressed. The variances across genes were drawn from a scale–inverse–chi–square distribution Scale–inv–$\chi^2(d_0, s_0^2)$ with $s_0^2 = 1$ and $d_0 = 1$, i.e., the variances vary moderately from gene to gene. Furthermore, the difference of means for the differentially expressed genes (1–200) were drawn from a normal distribution with mean zero and the gene-specific variance multiplied with a scale factor set to 0.3. For the non–differentially expressed genes (201–1000), the difference was set to zero. The data were generated by drawing from group–specific multivariate normal distributions with the given variances and means employing a block diagonal correlation structure intended to mimic gene expression data. This structure was generated as in Guo *et al.* (2007) with block size 100 and block entries equal to $0.9^{|i-j|}$. Furthermore, the sample sizes n_1 and n_2 are equal with $n_1 = n_2 = 8$.

The effect size estimates are plotted in Figure 1 according to their rank. It is important to note that this does not tell us whether the respective ranking is correct. Thus, even though the effect size estimates of the cat–score and an ordinary t–score are very similar, this does not mean that their induced ranking is

comparable. Efron's and Matsui and Noma's method will also change the ranking of the supplied cat–scores at least slightly.

It can be seen that fdr–effect and MastsuiNoma yield good results, while Efron's method has a higher bias for effect sizes up to 1, a phenomenon already observed by Matsui and Noma (2011). The "naive" approaches (cat–scores and t–scores) are far off for effect sizes up to 1.5. However, all methods overestimate large effect sizes. It follows that variable selection methods relying on effect size estimates will generally have a tendency of choosing only a relatively small number of variables in data sets with large effects.

This is in fact a phenomenon already observed by Ahdesmäki and Strimmer (2010) for the Efron algorithm applied to the Singh (Singh *et al.*, 2002) prostate cancer gene expression data. This data consists of gene expression measurements of $d = 6033$ genes for $n = 102$ patients, of which 52 are cancer patients and 50 are healthy. It has already been analyzed in Efron (2009) and Matsui and Noma (2011). Figure 2 shows the analysis results. As in the simulated data, the "naive" approaches are far off, while Efron and MatsuiNoma are quite similar. Note, however, that MatsuiNoma gives significantly lower estimates of large effect sizes than Efron. This is a phenomenon already noted in Matsui and Noma (2011). The fdr–effect method yields similar results to MatsuiNoma for large effect sizes but reaches zero estimates much faster than MatsuiNoma and Efron. In conclusion, all empirical Bayes methods considered seem to give sound results here, while the naive methods are probably grossly overestimating the effect sizes.

4. Variable Selection and Estimation of the Prediction Rule

4.1 Estimation of the Prediction Rule and Local False Discovery Rates

For the estimation of the prediction rule (2) I mostly employ James-Stein-type estimators as in shrinkage discriminant analysis – SDA, Ahdesmäki and Strimmer (2010). The group centroids $\boldsymbol{\mu}_k$ are estimated by the empirical means, for the correlations $\boldsymbol{P}$ the ridge-type estimator from Schäfer and Strimmer (2005) is used and the variances $\boldsymbol{V}$ are estimated by the shrinkage estimator from Opgen-Rhein and Strimmer (2007). Finally the proportions π_k are obtained by using the frequency estimator from Hausser and Strimmer (2009). For SDA I employ the implementation provided by the `R` package `sda`. The local false discovery rates used in the fdr–effect approach are learned by using the Grenander density estimator and truncated maximum likelihood for the empirical null as in Strimmer (2008b). As in Chapter 3 the implementation offered by the `R` package `fdrtool` is employed.

4.2 Variable Ranking and Selection

4.2.1 Variable Ranking

Before being able to select variables, a variable ranking needs to be established (obj. (i)). In the two class case, this is straightforward since the feature weight vector for class one $\boldsymbol{\omega}_1$ is up to a scale factor of n_2/n equal to the effect size vector $\boldsymbol{\omega}^{(1,2)}$, ($\boldsymbol{\omega}_1 = (n_2/n)\boldsymbol{\omega}^{(1,2)}$). Correspondingly, the feature weight vector for class two $\boldsymbol{\omega}_2$ is equal to the effect size vector $-\boldsymbol{\omega}^{(1,2)}$ up to a scale factor of n_1/n ($\boldsymbol{\omega}_2 = (-n_1/n)\boldsymbol{\omega}^{(1,2)}$). Thus, variables can be ranked according to the absolute value of $\boldsymbol{\omega}^{(1,2)}$. In the the case of multiple classes, the situation is more complicated. The feature weight vectors of the different classes need to be summarized in a certain way to obtain the importance of each feature i in class prediction. Here, I am going to use the summary statistic S_i proposed by Ahdesmäki and Strimmer (2010) and given by

$$S_i = \sum_{k=1}^{K} \left(\omega_{\text{cat},i}^{(k,\text{pool})}\right)^2, \tag{17}$$

where $\omega_{\text{cat},i}^{(k,\text{pool})} = (1/n_k - 1/n)^{-1/2}\omega_i^{(k,\text{pool})}$. Since false discovery rates are generally assumed to be monotone, (15) shows that using fdr–effect effect size estimates $\boldsymbol{\omega}_{\text{fdr}}^{(k,\text{pool})}$ would produce the same ranking as the cat–scores if they were used instead of $\boldsymbol{\omega}_{\text{cat}}^{(k,\text{pool})}$ to compute S_i in (17).

4.2.2 Misclassification Rate Based Variable Selection

Having obtained estimates $\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}$ of $\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}$ and $\widehat{\pi_k}$ of π_k, we can now compute an estimate of the misclassification rate using (6). Let $\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)$ be the vector of the t top–ranked variables according to the ranking induced by the vector $\boldsymbol{S}$ of all statistics S_i given by (17). This gives an estimate of the misclassification rate, which depends on t:

$$\widehat{P}(\text{error})(t) = \sum_{k=1}^{K} \Phi\left(- \min_{l \neq k} \frac{[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]^T[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)] + 2\log\left(\frac{\widehat{\pi_k}}{\widehat{\pi_l}}\right)}{2\sqrt{[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]^T[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]}} \right) \times \widehat{\pi_k}. \tag{18}$$

Efron performs feature selection by choosing a level $\alpha = 0.05$ as a target misclassification rate for the estimate in (18). Although one could view α as a tuning parameter, I follow his suggestion in this regard. Experiments with lower α led to very large feature sets showing only a negligible improvement of the classification performance.

After the target error α has been set, a feature threshold t^* is obtained by including as many features as necessary to reach it, i.e., $\widehat{P}(\text{error})(t^*) = \alpha$. Since usually a lot of features are shrunken to zero, it is possible that the target error can not be reached. Then, all the features will be included. This, however, is extremely unlikely to happen in real high dimensional data analysis. Finally, all features fulfilling $S_i \geq S_t^*$ are included in the classifier. I call the approach presented in this section misclassification rate (MR) based variable thresholding (MRT). Figure 3 gives a flowchart detailing the implementation of this method.

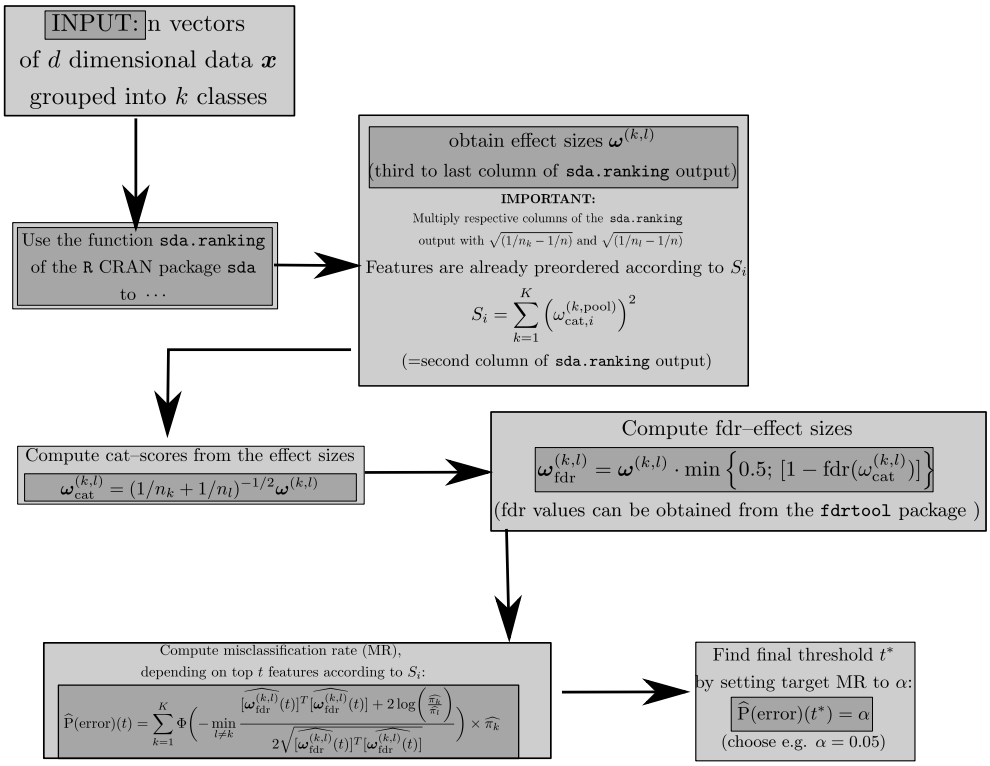

Figure 3: Flowchart describing misclassification rate (MR) based variable thresholding (MRT)

5. Analysis of Real and Simulated Data

5.1 Simulations

In this section, I will compare variable selection based on the misclassification rate (MR) with several other state of the art thresholding variable selection approaches, namely false-non discovery rate (FNDR) thresholding (Ahdesmäki and Strimmer, 2010), HCT (Donoho and Jin, 2008) and the PAM/NSC algorithm

(Tibshirani *et al.*, 2003). All methods are performed using empirical null modeling. As a base line classifier, I also include the results of classification with all features, i.e., performing no variable selection.

The simulations closely follow the setup of Witten and Tibshirani (2011). A training set of size 100 and a test set of 1000 samples are created with a dimension of $d = 500$ variables. In total, 25 runs of each simulation setup are performed.

5.1.1 Simulation Setup 1

In this setup, there are four classes with equal probability (0.25) no correlation and unit variance. In each class 25 features are differentially expressed with an effect size of 0.7, yielding a total number of 100 differentially expressed features. Since there is no correlation, I perform Diagonal Discriminant Analysis (DDA), i.e., LDA with identity covariance $\mathbf{\Sigma} = I_d$. The results are displayed in Table 1.

It can be seen that thresholding the summary statistic S (17) by false-non discovery rates or Higher Criticism yields hardly any significant features in most runs. Consequently, the estimated prediction errors are quite high.

Misclassification rate based feature selection as well as PAM, however, identify features useful for classification. This indicates that "analytical" thresholding methods, which do not rely on the optimization of a tuning parameter, may not work reliably when the effect sizes are small.

5.1.2 Simulation Setup 2

In this simulation, I am going to use a Guo *et al.* (2007) type block correlation with 5 blocks of size 100×100. As in Section 3, each block entry is given by $0.9^{|i-j|}$, thus we have some highly correlated variables within blocks but variables in different blocks are independent.

Note that Witten and Tibshirani (2011) report using an entry size of 0.6. This is probably a misprint since my results obtained for PAM are quite similar to the ones reported in their article, while for 0.6 the error of PAM is only about 5%.

There are two classes with equal probability (0.5) and 200 features are differentially expressed with effect size 0.6, all of them are attributed to class 2. Since there is correlation present in this setting, I will perform LDA.

It can be seen in Table 2 that all feature selection methods except for PAM, which does not take correlation into account, perform quite well here.

Table 2: Prediction errors and number of selected features for simulation setup 2, the number in the round brackets is the estimated standard error over 25 runs. The true number of differentially expressed features is 200

Method	Prediction Error	Features
LDA–MR	0.000 (0.000)	63.16 (7.215)
LDA–FNDR	0.000 (0.000)	60.96 (6.567)
LDA–HC	0.000 (0.000)	85.04 (8.677)
PAM	0.088 (0.018)	294.0 (69.43)
LDA–ALL	0.093 (0.014)	500

5.2 Gene Expression Data

In Ahdesmäki and Strimmer (2010), the relative effectiveness of the FNDR and HC thresholds to select relevant genes in shrinkage discriminant analysis applied to gene expression data has already been compared. I am going follow their setup here and will analyze four clinical gene expression data sets related to prostate cancer (Singh *et al.*, 2002), B-cell lymphoma (Alizadeh *et al.*, 2000), colon cancer (Alon *et al.*, 1999) and brain cancer (Pomeroy *et al.*, 2002).

Specifically, balanced 10–fold cross–validation with 20 repetitions was performed to obtain error estimates and their standard deviations. The number of selected features is inferred by a single run of the respective variable selection method on the whole data set. Only for PAM this was repeated several times since the number of selected variables selected by this algorithm varies considerably between several runs in a row on the same data set.

In Table 3, it can bee seen that the MRT approach has a performance similar to the other approaches. Interestingly, the MRT approach shows a more "adaptive" feature selection, leading to appropriate feature sets for each problem. In the brain data set, a very compact set of features is selected yielding a prediction error which is nonetheless in the range of the other approaches. The same is true for the Lymphoma and Colon data sets. This demonstrates that a variable selection method based on effect sizes leads to compact and yet effective molecular signatures. Furthermore, FNDR and HC thresholding yield very similar results.

6. Discussion

In this paper I reviewed and extended statistical techniques related to effect size estimation in linear classification and showed how to use them for variable selection. The fdr–effect method proposed for effect size estimation has been shown to work as well as competing approaches while being conceptually simple and computationally inexpensive. It therefore successfully unites the strengths of the approaches presented in Efron (2009) and Matsui and Noma (2011).

Table 3: Analysis of four cancer gene expression data sets with shrinkage discriminant analysis. The number of selected features are determined by a single feature selection run on the whole data set

Data / Method	Prediction Error	Selected Variables
Prostate ($d = 6033, n = 102, K = 2$)		
LDA–MR	0.0630 (0.0050)	134
LDA–FNDR	0.0550 (0.0048)	131
LDA–HC	0.0497 (0.0045)	116
PAM	0.0850 (0.0061)	172–377
Lymphoma ($d = 4026, n = 62, K = 3$)		
LDA–MR	0.0211 (0.0039)	34
LDA–FNDR	0.0036 (0.0018)	392
LDA–HC	0.0000 (0.0000)	345
PAM	0.0234 (0.0041)	2796–2383
Colon ($d = 2000, n = 62, K = 2$)		
LDA–MR	0.1291 (0.0093)	28
LDA–FNDR	0.1278 (0.0088)	168
LDA–HC	0.1233 (0.0087)	122
PAM	0.1160 (0.0921)	13–23
Brain ($d = 5597, n = 42, K = 5$)		
LDA–MR	0.1628 (0.0126)	56
LDA–FNDR	0.1525 (0.0120)	102
LDA–HC	0.1417 (0.0108)	131
PAM	0.2023 (0.0118)	42–5587

Additionally, I gave a unified treatment of the effect size estimation approaches presented in these two papers elucidating similarities not apparent when considering the original publications only.

Variable selection by minimizing the misclassification rate has been somewhat neglected in the literature but I showed in accordance with Dabney and Storey (2007), Efron (2009) and Matsui and Noma (2011) that it is indeed very well suited for real world problems. In addition, it is also much more intuitive than selecting a non-interpretable regularization parameter as for example in the PAM algorithm and leads to compact and interpretable feature sets.

In this work I proposed a conceptually simple and competitive variable selection algorithm that gives priority to genes with large effect sizes and is thus easy to interpret. This has been achieved by extending and combining the ideas of Dabney and Storey (2007), Efron (2009) and Matsui and Noma (2011).

High expectations are associated with the promise of a personalized medicine promising tailored treatments based on genetic and other information of the patient. In order to develop molecular diagnostics guiding these treatments, statis-

tical approaches for effective and interpretable classification are indispensable.

The methodology presented in this article provides interpretability and applicability for biological study and medical use. Reliable effect size estimates allow one to identify genes having discriminative power while variable selection based on these effect size estimates allows the selection of the most important genes for the construction of classification algorithms.

Program files to be used with the statistical software R (R Development Core Team, 2012), implementing the variable selection approaches presented in this article are available from my homepage: http://b-klaus.de. They are published under the GNU General Public License 3.0.

Acknowledgements

I would like to thank Shigeyuki Matsui for providing R–Code implementing the method of Matsui and Noma (2011). Furthermore, I thank Stéphane Robin, Tristan Mary-Huard, Marie-Laure Martin-Magniette (all at AgroParisTech) and Korbinian Strimmer, David Petroff as well as Verena Zuber for fruitful discussions of this work. Korbinian Strimmer and Verena Zuber also provided R-Code implementing the simulation setup of Smyth (2004). I also thank Miika Ahdesmäki (Almac Diagnostics) for R-Code performing CV–based prediction error estimation of several classification methods. The suggestion of adding an additional flowchart describing the MR based variable selection method made by an anonymous referee is also gratefully acknowledged.

References

Ahdesmäki, M. and Strimmer, K. (2010). Feature selection in omics prediction problems using cat scores and false nondiscovery rate control. *Annals of Applied Statistics* **4**, 503-519.

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.

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide

arrays. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 6745-6750.

Lê Cao, K. A., Boitart, S. and Besse, P. (2011). Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* **12**, 253.

Dabney, A. R. and Storey, J. D. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS ONE* **2**, e1002.

Dawid, A. P. (1994). Selection paradoxes of Bayesian inference. In *Multivariate Analysis and Its Applications* (*Hong Kong, 1992*) (Edited by T. W. Anderson, K. T. Fang and I. Olkin), Volume 24, 211-220. IMS Lecture Notes - Monograph Series, Institute of Mathematical Statistics, Hayward, California.

Donoho, D. and Jin, J. (2008). Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 14790-14795.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99**, 96-104.

Efron, B. (2007). Size, power and false discovery rates. *Annals of Applied Statistics* **35**, 1351-1377.

Efron, B. (2008). Microarrays, empirical Bayes, and the two-groups model. *Statistical Science* **23**, 1-22.

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. *Journal of the American Statistical Association* **104**, 1015-1028.

Guo, Y., Hastie, T. and Tibshirani, T. (2007). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics* **8**, 86-100.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research* **10**, 1469-1484.

Kim, K. I. and Simon, R. (2011). Probabilistic classifiers with high-dimensional data. *Biostatistics* **12**, 399-412.

Klaus, B. and Strimmer, K. (2013). Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics* **14**, 129-143.

Matsui, S. and Noma, H. (2011). Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics* **12**, 223-233.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statistical Applications in Genetics and Molecular Biology* **6**, 9.

Pang, H., Tong, T. and Zhao, H. (2009). Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics* **65**, 1021-1029.

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.

R Development Core Team. (2012). *R*: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology* **4**, 32.

Schwender, H., Ickstadt, K. and Rahnenfuhrer, J. (2008). Classification with highdimensional genetic data: assigning patients and genetic features to known classes. *Biometrical Journal* **50**, 911-926.

Senn, S. (2008). A note concerning a selection "paradox" of Dawid's. *American Statistician* **62**, 206-210.

Shao, J., Wang, Y., Deng, X. and Wang, S. (2011). Sparse linear discriminant analysis by thresholdig for high dimensional data. *Annals of Statistics* **39** 1241-1265.

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 Bioconductor package for supervised classification with high dimensional data. *BMC Bionformatics* **9**, 439.

Smyth, G. K. (2004). Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* **3**, 3.

Strimmer, K. (2008a). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* **24**, 1461-1462.

Strimmer, K. (2008b). A unified approach to false discovery rate estimation. *BMC Bioinformatics* **9**, 303.

Tibshirani, R., Hastie, T., Narsimhan, B. and Chu, G. (2003). Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statistical Science* **18**, 104-117.

Witten, D. M. and Tibshirani, R. (2011). Penalized classification using Fisher's linear discriminant. *Journal of the Royal Statistical Society, Series B* **73**, 753-772.

Xiaosheng, W. and Simon, R. (2011). Microarray-based cancer prediction using single genes. *BMC Bioinformatics* **12**, 391.

Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation. *Bioinformatics* **25**, 2700-2707.

Received December 19, 2012; accepted March 1, 2013.

Bernd Klaus
European Molecular Biology Laboratory (EMBL)
Meyerhofstraße 1, 69117 Heidelberg, Germany
bernd.klaus@embl.de