Classification Method Performance in High Dimensions

Claus Weihs and Tobias Kassner

Abstract We discuss standard classification methods for high-dimensional data and a small number of observations. By means of designed simulations illustrating the practical relevance of theoretical results we show that in the 2-class case the following rules of thumb should be followed in such a situation to avoid the worst error rate, namely the probability π_1 of the smaller class:

Avoid "complicated" classifiers: The independence rule (*ir*) might be adequate, the support vector machine (*svm*) should only be considered as an expensive alternative, which is additionally sensitive to noise factors. From the outset, look for stochastically independent dimensions and balanced classes. Only take into account features which influence class separation sufficiently. Variable selection might help, though filters might be too rough. Compare your result with the result of the data independent rule "*Always predict the larger class*".

Claus Weihs · Tobias Kassner
Department of Statistics, TU Dortmund
Vogelpothsweg 87, 44227 Dortmund
✉ claus.weihs@tu-dortmund.de
✉ tobias.kassner@tu-dortmund.de

Archives of Data Science, Series A
(Online First)
KIT Scientific Publishing
Vol. 3, No. 1, 2018

DOI 10.5445/KSP/1000083488/03
ISSN 2363-9881
CC BY-SA

1 Introduction

In this paper we discuss typical classification methods in the context of the analysis of high-dimensional data. This means that we assume many more features p than observations n, i.e. $p \gg n$ (cp., e.g., Weihs (2016)). Examples for this situation can be found in high throughput biotechnology like in data acquisition platforms as micro arrays, SNP chips, and mass spectrometers (cp., e.g., Kiiveri (2008)). As possible consequences, specialized classification methods are proposed (cp., e.g., Mai (2013), Tan et al (2014)) or theoretical results concerning the performance of well-known classifiers are derived (Bickel and Levina (2004), Fan et al (2010)). In this paper, we discuss implications of this theory for "classical methods", originally developed for low dimensional situations, in high-dimensional data.

In Section 2 we will consider theoretical results for standard classification methods if $p \gg n$. In Section 3 we define our research questions. In Section 4 we construct an experimental design for simulations to investigate the effects of different factors on the error rate. We vary the classifiers, the prior class probabilities, the true error rates, the correlations of features, as well as the asymptotic behavior of the Bayes error (constant vs. decreasing for $p \to \infty$). In Section 5 the corresponding simulation results are discussed, in particular the convergence of error rates for $p \to \infty$. Noise features are ignored until Section 6 where we briefly discuss the influence of noise on classifier performance. In Section 7 we conclude.

2 Theory

2.1 Linear Discriminant Analysis

We start with a strong warning concerning the application of linear discriminant analysis (*lda*) in high dimensions derived by Bickel and Levina (2004). For the data structure, Gaussians $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$, $\mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$ are assumed in two classes, equal prior probabilities $\pi_1 = \pi_2 = 0.5$, and $n_1 = n_2$ observations. Linear discriminant analysis (*lda*) optimally fits this structure. The performance of *lda* in the case of $p \gg n$ is discussed by Bickel and Levina (2004), stating:

Let positive constants k_1, k_2, c be given. Consider feature distributions with

- *true covariance matrix $\boldsymbol{\Sigma}$ not ill-conditioned, i.e.* $0 < k_1 \leq \lambda_{\min}(\boldsymbol{\Sigma}) \leq \lambda_{\max}(\boldsymbol{\Sigma}) \leq k_2 < \infty$ *for $\lambda_{\max}$ and $\lambda_{\min}$ the maximal and minimal eigenvalues,*
- *Mahalanobis class distance* $\Delta = \sqrt{(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)} > c > 0$*, and*
- *$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ in a compact set.*

Then, if $p/n \to \infty$, the worst case error rate of lda converges to $\pi_1 = 0.5$, i.e. asymptotical class assignment might be no better than random guessing.

Note that since $p > n$, the inverse of the estimated pooled covariance matrix does not exist, therefore the Moore-Penrose generalized inverse is used instead in *lda*. Also note that this statement is about the worst case error rate over all $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}$ with the mentioned properties. For applications, though, this asymptotical behavior should be assumed if there is no indication for a special case with better asymptotical error rates (for an example cp. Section 2.3).

2.2 Independence Rule

Noise accumulation is suspected to be one reason for the above adverse property of *lda*. Therefore, a diagonal covariance matrix is often tried. An asymptotic result for the corresponding so-called *independence rule* (*ir*, linear discriminant analysis with diagonal covariance matrix) is again given in Bickel and Levina (2004), under the same distributional assumptions as for the asymptotic property of *lda*:

If $p/e^n \to 0$, i.e. p grows slower than e^n, then the error rate of ir is bounded by $\Phi\Big(-\frac{\sqrt{K_0}}{1+K_0}\Delta\Big) \leq 0.5$ for Φ the cumulative standard normal distribution function, $K_0 = \max_{\boldsymbol{\Sigma}_0} \frac{\lambda_{\max}(\boldsymbol{\Sigma}_0)}{\lambda_{\min}(\boldsymbol{\Sigma}_0)}$, $\boldsymbol{\Sigma}_0 :=$ correlation matrix.

Note that this statement, again, refers to worst-case behavior since K_0 is a maximum over all possible correlation structures $\boldsymbol{\Sigma}_0$. What matters, though, is that this statement leads to a possible superiority of *ir* over full *lda* for $p \gg n$. For a graphic on the behavior of the error rate bound for different K_0 see Bickel and Levina (2004).

If $\mathbf{\Sigma}_0 = \boldsymbol{I}$, then $K_0 = 1$ and *ir* is Bayes optimal (as expected) since the Bayes error is $\Phi(-\Delta/2)$. If $\mathbf{\Sigma}_0$ has eigenvalues $\to 0$ or $\to \infty$, then $K_0 \to \infty$ and the above bound is $\Phi(0) = 0.5$, as for full *lda*. For normal distributions, *ir* is equivalent to *Naive Bayes*. Since we also assume normal distributions for *Naive Bayes* (*NB*), results different from *ir* may only originate from the different realization procedures.

2.3 Distance-Based Classifiers

Naturally, classification quality depends on class distance. Fortunately, perfect class prediction is possible for so-called distance-based classifiers (Fan et al (2010)). A distance-based classifier g is defined by two properties:

(a) g assigns an observation $\boldsymbol{x}$ to class 1 if it is closer to each observation in class 1 than to any observation in class 2.

(b) If g assigns $\boldsymbol{x}$ to class 1, then $\boldsymbol{x}$ is closer to at least one observation in class 1 than to the most distant observation in class 2.

For such classifiers, the following property is shown:

Let the data structure be $\boldsymbol{X}_i = \boldsymbol{\mu}_i + \boldsymbol{\epsilon}$, $i = 1, 2$ *the class index,* $\boldsymbol{\epsilon} := (\epsilon_j)$, j *= feature index,* ϵ_j *i.i.d. with expectation* 0. (*Note that this way the correlation matrix is assumed to be* $\mathbf{\Sigma}_0 = \boldsymbol{I}$.) *Then,* g *achieves the error rate* 0 *asymptotically for* $p \to \infty$ *iff* $D := ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1||$ *grows faster than* $p^{1/4}$. (*Note that* $D = \Delta$ *for* $\mathbf{\Sigma} = \boldsymbol{I}$.)

This result is independent of sample size n.

Note that the property "$D := ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1||$ grows faster than $p^{1/4}$" can be interpreted as "all involved dimensions contribute sufficiently to class separation".
Examples for distance-based classifiers are the k-Nearest-Neighbor classifiers (*kNN*), the linear support vector machine (*svm*), as well as *lda* and *ir* for $\pi_1 = \pi_2 = 0.5$ (cp. Hall et al (2008)). Therefore, the error rates not only of *ir*, but also of *lda* may converge to 0 if $\mathbf{\Sigma}_0 = \boldsymbol{I}$ and $\pi_1 = \pi_2 = 0.5$ (cp. with Section 2.1).

3 Research Questions

The main idea of this paper is to study the behavior of many relevant classifiers in situations where theoretical results were developed for only some of these classifiers. Besides the above mentioned classifiers *lda*, *ir*, *NB*, 1*NN* (as a special case of *kNN*), and *svm*, we additionally included the decision tree (*tree*) into the study as a representative of classification rules explicitly using only series of univariate rules, i.e. no linear combinations of features.

From the theoretical results in Section 2 we derived the following four research questions for the practical application of classification methods: *Parameter Dependence*, *Convergence*, *Classifier Ranking*, and *Noisy Performance*.

Parameter Dependence: How will the performance of the classifiers depend on the parameters p, π_1, and others?

Convergence: How do the classifiers behave for $p \to \infty$, e.g.:

1. Will error rates of the classifiers converge to $\pi_1 < 0.5$ for $p \to \infty$ if the Bayes error is the same for each number of dimensions p?
2. Will error rates of the classifiers converge to 0 for $p \to \infty$ for distance-based classifiers if $D := ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1||$ grows faster than $p^{1/4}$, but the involved features are dependent?

Classifier Ranking: How do the classifiers compare concerning Bayes error approximation?

All the above three research questions will be mainly discussed for the situation where all observed features influence classes. Finally, we will discuss the behavior of the classifiers in the case of noise features:

Noisy Performance: How will the performance of the classifiers react to noise factors, i.e. to factors not contributing to class separation?

4 Simulation Design

In this section we will develop the experimental design for our simulation.

4.1 General Design and two Cases

For the data structure, we choose the ideal situation for *lda*, i.e. 2 classes with influential features i.i.$\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ distributed, $i = 1, 2, \boldsymbol{\mu}_1 \neq \boldsymbol{\mu}_2$, class 1 with probability $\pi_1 \leq 0.5$. We distinguish two very different cases of Bayes error development (see Section 4.5 for details):

A. The Bayes error, i.e. the classification difficulty, decreases for $p \to \infty$, i.e. classification gets simpler for $p \to \infty$. This is realized by including more and more independent blocks of features with the same contribution to class separation.

B. The Bayes error is constant $\forall p$. Note that in this case the contribution of individual features to class separation is decreasing for $p \to \infty$.

4.2 Correlation Setting

For the $p \times p$ covariance matrix we assume a special structure, namely (see, e.g., Bickel and Levina (2004)):

$$\begin{aligned}\boldsymbol{\Sigma} := \boldsymbol{R}_{\kappa;p} &:= \begin{pmatrix} 1 & \kappa & \kappa & \cdots & \kappa \\ \kappa & 1 & \kappa & \cdots & \kappa \\ \kappa & \kappa & 1 & & \kappa \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \kappa & \kappa & \cdots & & 1 \end{pmatrix} \\ &= (1-\kappa)\,\boldsymbol{I}_p + \kappa \boldsymbol{v}_1 \boldsymbol{v}_1^T, \quad 0 < \kappa < 1, \\ \boldsymbol{v}_1^T &:= (1 \ldots 1). \end{aligned} \tag{1}$$

This leads to $\boldsymbol{\Sigma}^{-1} = \boldsymbol{R}_{\kappa;p}^{-1} = \frac{1}{1-\kappa}\boldsymbol{I}_p - \frac{\kappa}{1-\kappa}\frac{1}{1+\kappa(p-1)}\boldsymbol{v}_1\boldsymbol{v}_1^T$ (using the Sherman-Morrison formula). The eigenvalues of $\boldsymbol{\Sigma}^{-1}$ are $\frac{1}{1-\kappa}$ ($(p-1)$-times) and $\frac{1}{1+\kappa(p-1)} \xrightarrow[p\to\infty]{} 0$ (once). The eigenvalues of $\boldsymbol{\Sigma}$ are $\lambda_i = (1-\kappa), i \neq 1$, and $\lambda_1 = 1 + (p-1)\kappa \xrightarrow[p\to\infty]{} \infty$. Therefore, $K_0 \to \infty$ in the error bound of *ir* and *ir* is not theoretically superior to *lda* (see Section 2.2).

The structure of the covariance matrix can be generalized by *blocking*. For that, we introduce a block-diagonal covariance matrix $\boldsymbol{\Sigma}$ with p/b diagonal blocks $\boldsymbol{R}_{\kappa;b}$ of block size b.

Example: $p = 6,\ \kappa = 0.5,\ b = 3$: $\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0.5 & 1 & 0.5 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.5 & 0.5 \\ 0 & 0 & 0 & 0.5 & 1 & 0.5 \\ 0 & 0 & 0 & 0.5 & 0.5 & 1 \end{pmatrix}$.

For eigenvalue 2, there are eigenvectors $(1\ 1\ 1\ 0\ 0\ 0)^T$, $(0\ 0\ 0\ 1\ 1\ 1)^T$, and for eigenvalue 0.5, there are eigenvectors $(1\ 0\ -1\ 0\ 0\ 0)^T$, $(0\ 1\ -1\ 0\ 0\ 0)^T$, $(0\ 0\ 0\ 1\ 0\ -1)^T$, $(0\ 0\ 0\ 0\ 1\ -1)^T$.

In the general eigenstructure we have p/b eigenvalues $1 + (b-1)\,\kappa$ and $p - p/b$ eigenvalues $1 - \kappa$ (cp. **Case B1** below). Eigenvectors can be constructed to lie in subspaces of dimension b. In the case of no blocking, we have $b = p$.

Note that the sign of κ may be changed in one dimension $q \in \{1, \ldots, p\}$ leaving eigenvalues λ_i unchanged (see Section 4.4). However, this will not change the Bayes rule for our choice of mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ as we will prove in Section 4.4. Therefore, we will ignore this generalization.

4.3 Class Means

In our design, we would like to pre-specify the Bayes error rate f and the probability π_1 of class 1 at the same time. To achieve this, we fix w.l.o.g. the mean of class 1 as $\boldsymbol{\mu}_1 := \mathbf{0}$, and the mean of class 2 $\boldsymbol{\mu}_{2i}$ so that the Bayes error rate is $f \in (0, 0.5)$, where i in the index of the mean corresponds to the chosen discriminant direction, i.e. of the ith normalized eigenvector $\boldsymbol{e}_i$.

Let the projections on $\boldsymbol{e}_i$ be $m_1 := \boldsymbol{e}_i^T \boldsymbol{\mu}_1 = 0 \leq \boldsymbol{e}_i^T \boldsymbol{\mu}_{2i} =: m_{2i}$ (w.l.o.g.). The class variance in direction i is $\sigma_i^2 = \boldsymbol{e}_i^T \boldsymbol{\Sigma} \boldsymbol{e}_i = \lambda_i = i$th eigenvalue of $\boldsymbol{\Sigma}$, i.e., $\sigma_i^2 = 1 - \kappa$ for $i > p/b$ and $\sigma_i^2 = 1 + (p-1)\,\kappa \xrightarrow[p\to\infty]{} \infty$ for $i \leq p/b$. For our study, we only use $i > p/b$. Then, for the Bayes error the following equation holds:

$$f = \pi_1 \Big(1 - \Phi\Big(\frac{\tau_i - m_1}{\sigma_i}\Big)\Big) + (1-\pi_1)\,\Phi\Big(\frac{\tau_i - m_{2i}}{\sigma_i}\Big) \quad \text{for}$$

$$\tau_i := \frac{m_1 + m_{2i}}{2} + \frac{\sigma_i^2 \log\left(\frac{\pi_1}{1-\pi_1}\right)}{m_{2i} - m_1} = \frac{m_{2i}}{2} + \frac{(1-\kappa)\log\left(\frac{\pi_1}{1-\pi_1}\right)}{m_{2i}}, \tag{2}$$

since $m_1 = 0$, $\sigma_i^2 = 1 - \kappa$ and with τ representing the location where the densities of the two distributions intersect (cp. Figueiredo (2004)). From this formula, we derived numerical solutions for all relevant combinations of factors $f,\ \kappa,\ \pi_1$ with $0 < f < \pi_1 \leq 0.5$ (see **Case A** below). Estimation of the linear model for m_{2i} on the features $u_{1-f} := (1-f)-$quantile of the standard normal, $\sigma = \sqrt{1-\kappa}$, and $\tilde{\pi}_1 := 1 - \left(1/2 \log\left(\frac{\pi_1}{1-\pi_1}\right)\right)^2$, all their 2-factor interactions, and the one 3-factor interaction [1] in R (cp. R Core Team (2017)) leads to

$$\begin{aligned} m_{2i} \approx &- 2.13952 \cdot \sigma + 2.91430 \cdot u_{1-f} \cdot \sigma \\ &+ 2.12119 \cdot \sigma \cdot \tilde{\pi}_1 - 0.89714 \cdot u_{1-f} \cdot \sigma \cdot \tilde{\pi}_1 \end{aligned} \tag{3}$$

after elimination of non-significant features and interactions, leading to (using coefficients rounded to integers except the optimized 2.2722):

$$\begin{aligned} 0 < m_{2i}(\kappa, f, \pi_1) &\approx 2\sqrt{1-\kappa}\,u_{1-f} - \sqrt{1-\kappa}\,(2.2722 - u_{1-f})\left(1/2 \log\Big(\frac{\pi_1}{1-\pi_1}\Big)\right)^2 \\ &= m_{2i}(\kappa, f, \pi_1 = 0.5) - \sqrt{1-\kappa}\,(2.2722 - u_{1-f})\left(1/2 \log\Big(\frac{\pi_1}{1-\pi_1}\Big)\right)^2 \\ &\leq m_{2i}(\kappa, f, \pi_1 = 0.5) \quad \text{if} \quad f \geq 0.012. \end{aligned} \tag{4}$$

Note that the estimated model is nearly exact ($R^2 > 0.999$), so that the above argument not only leads to a nearly exact general formula for $m_{2i}(\kappa, f, \pi_1)$, but

[1] $m_{2i} \sim u_{1-f} * \sigma * \tilde{\pi}_1$ in R-notation

also to a proof that $m_{2i}\,(\kappa, f, \pi_1) \leq m_{2i}\,(\kappa, f, \pi_1 = 0.5)$ for relevant error rates f.
For blocking, $\boldsymbol{\mu}_2$ is set constant for all p/b blocks of size b. Let $b = 2 \cdot p^{\eta}$, $0 \leq \eta < 1$. Then, $D := ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|| = ||m_{2ib}\boldsymbol{e}_{ib}||\sqrt{p/b} = \Theta(\sqrt{p/b}) = \Theta(p^{0.5\,(1-\eta)})$. We choose b so that $0.5(1-\eta) \geq 1/4$ (because of Section 2.3) using $\eta = 0,\ 1/3,\ 1/2$, i.e., $0.5\,(1-\eta) = 1/2, 1/3, 1/4$.

4.4 Theoretical consequences

We now derive theoretical consequences of our settings in Sections 4.2 and 4.3.

Generalization of Bickel and Levina (2004)

For the Bayes error f, we have seen relation (2). With $\pi_1 = 0.5$ this leads to

$$f = 0.5\,(1 - \Phi\Big(\frac{m_{2i}}{2\sigma_i}\Big) + \Phi\Big(-\frac{m_{2i}}{2\sigma_i}\Big)\Big) = 1 - \Phi\Big(\frac{m_{2i}}{2\sigma_i}\Big). \tag{5}$$

For *lda*, Bickel and Levina (2004, p. 995) show that the argument of Φ, i.e. $\frac{m_{2i}}{2\hat{\sigma}_i} \xrightarrow{\mathrm{P}} 0$ for $p/n \to \infty$, leading to $f \xrightarrow{\mathrm{P}} 0.5 = \pi_1$. Since this result does not depend on π_1, we can show that $\frac{\hat{\tau}_i}{\hat{\sigma}_i} = \frac{m_{2i}}{2\hat{\sigma}_i} + \frac{\hat{\sigma}_i \log\left(\frac{\pi_1}{1-\pi_1}\right)}{m_{2i}} = \frac{m_{2i}}{2\,\hat{\sigma}_i} + \frac{\log\left(\frac{\pi_1}{1-\pi_1}\right)}{m_{2i}/\hat{\sigma}_i} \xrightarrow{\mathrm{P}} 0 - \infty = -\infty$, i.e. $\Phi(\frac{\hat{\tau}_i}{\hat{\sigma}_i}) \xrightarrow{\mathrm{P}} 0$ for $0 < \pi_1 < 1/2$. Therefore, with formula (2) the asymptotic behavior of the estimated error rate $\hat{f}$ can be characterized as $\hat{f} \xrightarrow{\mathrm{P}} \pi_1\,(1-0) + (1-\pi_1)\,0 = \pi_1$. This generalizes the result of Bickel and Levina (2004) for *lda* in that we have shown that the "worst-case error rate $\to \pi_1$ for $0 < \pi_1 \leq 0.5$, $p/n \to \infty$". Thus for *lda* the asymptotic result might not be better than the data independent rule: *Always predict the larger class*. This partly answers the research question *Convergence* in Section 3.

Sign of κ

One can show that the sign of κ may be changed in one dimension $q \in \{1, \ldots, p\}$ leaving eigenvalues λ_i unchanged. Changing the sign of κ in dimension q of $\boldsymbol{R}_{\kappa;p}$ means changing the sign of all entries κ in line and column q.

Example: Let $p = 6, \kappa = 0.5, q = 2$: $\Sigma = \begin{pmatrix} 1 & -0.5 & 0.5 & 0.5 & 0.5 & 0.5 \\ -0.5 & 1 & -0.5 & -0.5 & -0.5 & -0.5 \\ 0.5 & -0.5 & 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & 1 & 0.5 & 0.5 \\ 0.5 & -0.5 & 0.5 & 0.5 & 1 & 0.5 \\ 0.5 & -0.5 & 0.5 & 0.5 & 0.5 & 1 \end{pmatrix}$.

Then the 1st eigenvector = $(1\ -1\ 1\ 1\ 1\ 1)^T$ has eigenvalue $1+(p-1)\kappa = 3.5$ and the eigenvectors $(1\ 0\ 0\ 0\ 0\ -1)^T$, $(0\ 1\ 0\ 0\ 0\ 1)^T$, $(0\ 0\ 1\ 0\ 0\ -1)^T$, $(0\ 0\ 0\ 1\ 0\ -1)^T$, and $(0\ 0\ 0\ 0\ 1\ -1)^T$ have eigenvalue $1-\kappa = 0.5$. Thus, only the 1st eigenvector with eigenvalue $1+(p-1)\,\kappa = 3.5$ is not the same as in the case with all signs equal (cp. one block of the example in Section 4.2).

At the same time, a sign change in the correlation will not change the Bayes rule either, as we will prove now. In Section 4.3 we have shown that $m_{2i} \approx 2\sqrt{1-\kappa}\,u_{1-f} - \sqrt{1-\kappa}\,(2.2722 - u_{1-f})\,(1/2 \log(\frac{\pi_1}{1-\pi_1}))^2$ which is independent of $\boldsymbol{e}_i$ and, thus, independent of q. The decision hyperplane of the *Bayes rule* is given by $h_1(\boldsymbol{x}) = h_2(\boldsymbol{x})$, where $h_k(\boldsymbol{x}) := (\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k)^T\boldsymbol{x} - 0.5\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k)$, $k = 1, 2$, i.e.

$$\begin{aligned}
\log(\pi_1) &= h_1(\boldsymbol{x}) = h_2(\boldsymbol{x}) \\
&= m_{2i}\boldsymbol{x}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_i - 0.5m_{2i}^2\boldsymbol{e}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_i + \log(1-\pi_1) \\
&= m_{2i}\cdot(\sum_j \alpha_j\boldsymbol{e}_j)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{e}_i - 0.5m_{2i}^2/\lambda_i + \log(1-\pi_1) \text{ for adequate } \alpha_j \in \mathbb{R}, \\
&= m_{2i}\cdot\alpha_i/\lambda_i - 0.5m_{2i}^2/\lambda_i + \log(1-\pi_1), \text{ i.e.} \\
\alpha_i &= 0.5m_{2i} + \lambda_i\cdot\log(\pi_1/(1-\pi_1))/m_{2i} \qquad (6)
\end{aligned}$$

is independent of q, the α_j can be arbitrary, $j \neq i$, and the decision hyperplanes of the *Bayes rule* are independent of q. Therefore, the parameter q is ignored, i.e. we choose the same correlation sign for all dimensions.

Choice of $\boldsymbol{e}_i$

As discrimination directions we only choose eigenvectors $\boldsymbol{e}_i$ with the same variance $\sigma_i^2 = (1-\kappa)$. Therefore, with the same argument as in the previous paragraph on the sign of κ we can show that the Bayes rule is independent of the choice of $\boldsymbol{e}_i$ so that $\boldsymbol{e}_i$ can be fixed deliberately guaranteeing that $\sigma_i^2 = (1-\kappa)$.

Therefore, the parameter i is ignored also in the simulation design, i.e. we fix this parameter according to the rules in Section 4.5.

4.5 Implemented Design

In order to study the dependence on the number of dimensions, we choose $p = 12, 120, 480, 1020, 1980$ features. We use training samples with $n = 12$ observations and test samples with 2000 observations. For each covariance matrix $\mathbf{\Sigma}$ we use 200 repetitions, i.e. different random samples, to minimize variation in results. Notice that the training samples are very small in absolute numbers as well as related to the highest numbers of features. The maximum ratio of the number of features to the number of observations is 165. The parameter design for all parameters is fixed as follows:

Vary p, κ, b, π_1, f_b on a grid so that

- $p = 12,\ 120,\ 480,\ 1020,\ 1980,$
- $\kappa = 0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9,$
- $b = 1,\ 2,\ 2p^{1/3},\ 2p^{1/2}$ (exact numbers for $b = 2p^{1/3},\ 2p^{1/2}$ see below),
- $\pi_1 = 0.1,\ 0.2,\ 0.3,\ 0.4,\ 0.5,$
- $f_b = 0.05,\ 0.15,\ 0.25,\ 0.35,\ 0.45 \quad \text{if} \quad f_b < \pi_1.$

Note that for training we approximate $\pi_1 = 0.1, 0.2, 0.3, 0.4, 0.5$ by using $1, 2, 4, 5, 6$ observations in class 1. In the test sample, the theoretical π_1 is realized. The block size is generally set to $b = 2p^{\eta}$ so that $D = ||\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1|| = \Theta(p^{0.5\,(1-\eta)})$, $0.5\,(1-\eta) \geq 1/4$. Because of the restrictions $b = 2p^{\eta} \in \mathbb{N}$, $p/b = p/(2p^{\eta}) \in \mathbb{N}$, and $b/2 \in \mathbb{N}$ (see Section 4.3), for $\eta = \{1/3, 1/2\}$ we use $b = \{4, 6\}, \{10, 20\}, \{16, 40\}, \{20, 60\}, \{22, 90\}$ for $p = 12, 120, 480, 1020, 1980$, correspondingly. Additionally, we study $b = 1$ (diagonal covariance matrix) fixing $\kappa = 0$ (complete independence). The class means are chosen $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2$ individually for each case below. The following simulation cases mainly differ in the generation of eigenvectors and corresponding samples.

Case A: Decreasing Bayes error for increasing dimension p

Here, eigenvectors $\boldsymbol{e}_i$ are constructed in b dimensions. Globally, we fix $i = b$. The class mean $\boldsymbol{\mu}_2$ is chosen identical in the p/b blocks as derived in Section 4.3 (with prefixed f_b, π_1, κ). The samples are independently drawn for each block. Then, the Bayes error is $f_b^{p/b} = f_b^{0.5p^{1-\eta}} \to 0$ for $p \to \infty$ since $(1-\eta) \geq 1/2$. Note that even in the worst case $f_b = 0.45, p = 1980, b = 90$ the expression $f_b^{p/b}$ is as small as 2.3e-08. Overall, we have $5 \cdot (1 + 5 \cdot 3) \cdot (5 + 4 + 3 + 2 + 1) \cdot 200 = 1200 \cdot 200 = 240,000$ simulation runs per classification method.

Case B1: Constant Bayes error for all p (version 1)

Here, the eigenvectors $\boldsymbol{e}_i$ are determined according to the full block-diagonal covariance matrix, but in subspaces of dimension b (setting all other entries to 0). These eigenvectors are used for the construction of $\boldsymbol{\mu}_2$ in Section 4.3. Thus, the Bayes error is the prefixed f_b for all block sizes b and all numbers of dimensions p. We additionally allow for $b = p$ and globally fix $i = p$. Note that for the full covariance matrix only the first p/b eigenvalues are $\neq (1-\kappa)$. Overall, we have $5 \cdot (1 + 5 \cdot 4) \cdot (5 + 4 + 3 + 2 + 1) \cdot 200 = 1575 \cdot 200 = 315,000$ simulation runs per classification method.

Case B2: Constant Bayes error for all p (version 2)

Here, b is chosen as in the general parameter design and $i = b$. The eigenvectors $\boldsymbol{e}_i$ are built blockwise, but identical for all blocks. Afterwards, normalization is realized over the combined eigenvectors by means of $\boldsymbol{e}_i := (\boldsymbol{e}_{ib} \ldots \boldsymbol{e}_{ib})^T / ||(\boldsymbol{e}_{ib} \ldots \boldsymbol{e}_{ib})^T||$ so that $||\boldsymbol{e}_i|| = 1$. Eigenvectors are not restricted to subspaces of dimension b, leading to the most general eigenvector structure. As in **Case B1**, the Bayes error is f_b for all block sizes b and all numbers of dimensions p. The number of runs is the same as in **Case A**.

All simulations were carried out by means of the software R (R Core Team (2017)) on the Linux-HPC-Cluster at TU Dortmund (LiDOng).[2]

[2] For the classifiers, standard implementations in the package mlr (Bischl et al (2016)) are used except in the specified cases: "classif.lda" for *lda*, "classif.sda" with options diagonal = TRUE, lambda = 0, lambda.var = 0, lambda.freqs = 0 for *ir*, "classif.naiveBayes" for *NB*, "classif.knn" with k=1 for *1NN*, "classif.svm" with kernel = linear and the cost parameter tuned on the grid $\{2^{-4}, 2^{-3}, \ldots, 2^3, 2^4\}$ for *svm*, and "classif.rpart" with minsplit=4, minbucket=2 for *tree*.

5 Results

5.1 Parameter Dependence

Let us first discuss the research question *Parameter Dependence* of Section 3, i.e. the effects of the parameters p, κ, b, π_1, f_b and their 2-parameter interactions on the mean error rates over the 200 replications. Also, the contrasts of the classifier effects with the 'basic classifier' *lda* and the corresponding interactions with the parameters are reported. Please note that the high number of replications leads to very small variation in the mean error rates so that we expect the parameter effects not to be blurred by noise, i.e. we expect relevant effects to be very highly significant.

With this reservation, in **Case A** from the regression analysis we see that (cp. Table 1) the main effects on mean error rates are positive for p (dimension), b (block size), as well as f_b, π_1 (Bayes error, class 1 probability). The correlation coefficient κ is only indirectly significant (on the 5 %-level) via interactions with the other parameters. The probability π_1 of class 1 is not very highly significant (p-value = 0.2 %) since for every π_1 there is also convergence to $a = 0$ (cp. Table 4). All classifiers except *NB* are significantly different from *lda* (basic classifier) and all corresponding 2-parameter interactions are significant except of κ, b with *tree*. Also, the significance of the interactions of κ with p, b is only around 5 %.

Defining the *contribution* of a parameter to the variation in mean error rate as the difference between the product of its estimated coefficient with the maximum and minimum parameter value, the contribution 0.36 of the Bayes error f_b appears to be most relevant, followed by the contribution 0.13 of the dimension p. Note that only the coefficients of these two parameters are very highly significant. The fit of the regression model is not optimal ($R^2 = 0.79$), i.e. there should be influences of even higher-order terms.

In **Case B1**, main effects on mean error rates are negative for p (dimension) and positive for κ, b (correlation, block size) and f_b, π_1 (Bayes error rate, class 1 prob.) (see Table 2). Many interactions are not highly significant and classifier *ir* differs the least significant from the basic classifier *lda*. The *contributions* 0.41 of the Bayes error f_b and 0.30 of the probability π_1 of class 1 appear to be most relevant. Note that here all main effects except of the constant are very highly significant. The model fit is distinctly better than in **Case A** ($R^2 = 0.85$).

Table 1: Case A: Parameter Estimates (p-value) from Linear Regression ($R^2 = 0.79$).

param	main [1]	:κ [2]	:b	:π_1	:f_b	:*ir*	:*NB*	:*1NN*	:*svm*	:*tree*
const.	3.5e-2 (0.001)									
p	6.6e-5 *** [3]	-8.7e-6 (0.070)	-1.2e-6 ***	-5.7e-5 (3e-5)	-1.0e-4 (6e-15)	-7.5e-5 ***	-4.7e-5 ***	-8.1e-5 ***	-6.4e-5 ***	-1.4e-5 (0.006)
κ	-3.9e-3 (0.755)		-2.9e-4 (0.046)	5.5e-1 ***	-3.7e-1 ***	1.2e-1 ***	4.9e-2 (6e-6)	1.1e-1 ***	6.9e-2 (3e-10)	9.7e-3 (0.370)
b	8.6e-4 (2e-5)			8.3e-3 ***	-4.2e-3 ***	2.5e-3 ***	1.3e-3 ***	2.8e-3 ***	2.7e-3 ***	-8.8e-5 (0.556)
π_1	8.9e-2 (0.002)				-2.8e-1 (0.001)	-4.2e-1 ***	-3.9e-1 ***	-3.3e-1 ***	-4.4e-1 ***	-1.3e-1 (1e-5)
f_b	0.905 ***					3.6e-1 ***	2.8e-1 ***	3.8e-1 ***	3.8e-1 ***	1.7e-1 (1e-8)
					classifier	*ir*	*NB*	*1NN*	*svm*	*tree*
					contrast to *lda*	-7.7e-2 (4e-11)	1.6e-2 (0.176)	-9.6e-2 ***	-7.0e-2 (6e-8)	3.7e-2 (0.002)

Table 2: Case B1: Parameter Estimates (p-value) from Linear Regression ($R^2 = 0.85$).

param	main [1]	:κ [2]	:b	:π_1	:f	:*ir*	:*NB*	:*1NN*	:*svm*	:*tree*
const.	-3.6e-3 (0.498)									
p	-2.9e-5 *** [3]	-2.3e-5 ***	-2.7e-8 ***	2.1e-4 ***	-2.0e-4 ***	9.8e-6 (4e-5)	5.6e-6 (0.019)	3.1e-5 ***	2.2e-5 ***	2.1e-5 ***
κ	4.9e-2 (3e-15)		6.8e-6 (0.070)	1.3e-1 ***	-3.0e-1 ***	7.2e-2 ***	2.6e-2 (1e-6)	1.7e-2 (0.001)	1.6e-2 (0.004)	5.6e-2 ***
b	7.1e-5 ***			-5.2e-5 (3e-7)	-9.2e-7 (0.926)	2.9e-5 (7e-15)	9.4e-6 (0.012)	-7.9e-6 (0.035)	-7.0e-6 (0.065)	-3.1e-6 (0.407)
π_1	0.739 ***				-1.37 ***	-1.3e-1 ***	5.8e-2 (8e-5)	-2.1e-1 ***	-7.3e-2 (6e-6)	-2.7e-1 ***
f_b	1.02 ***					9.7e-2 (6e-11)	5.8e-3 (0.692)	8.4e-2 (1e-8)	1.0e-1 (1e-11)	2.3e-1 ***
					classifier	*ir*	*NB*	*1NN*	*svm*	*tree*
					contrast to *lda*	-1.0e-2 (0.085)	-4.7e-2 (1e-15)	6.1e-2 ***	-2.0e-2 (0.002)	3.3e-2 (1e-8)

(1) main stands for the direct effect of par,

(2) :κ stands for the interaction of par with κ, e.g. $p : \kappa$ for the interaction between p and κ (other interactions analogously),

(3) *** stands for (< 2e-16), i.e. p-value numerically 0.

In **Case B2**, main parameter effects on the mean error rates are similar to **Case B1** (see Table 3), except that now the effect of classifier *ir* is also highly significantly different from the basic classifier *lda*. Again, most interactions do not appear to be highly significant, and only the *contributions* of the Bayes error f_b and the probability π_1 of class 1 appear to be relevant. Moreover, note that entries in Table 3 marked in bold face have signs different than the corresponding entries in Table 2, but at most one of the corresponding entries in Tables 2 and 3 is significant. The model fit is best among the regressions ($R^2 = 0.87$).

Table 3: Case B2: Parameter Estimates (p-value) from Linear Regression ($R^2 = 0.87$).

par	main (1)	:κ (2)	:b	:π_1	:f	:*ir*	:*NB*	:*1NN*	:*svm*	:*tree*
const.	-1.9e-3 (0.732)									
p	-2.0e-5 (3e-10)	-3.3e-5 *** (3)	-5.1e-7 ***	1.9e-4 ***	-1.9e-4 ***	1.1e-5 (3e-5)	1.0e-5 (1e-4)	3.1e-5 ***	1.8e-5 (9e-12)	2.1e-5 (1e-15)
κ	4.0e-2 (1e-9)		4.5e-4 (4e-9)	1.3e-1 (2e-16)	-2.4e-1 ***	5.9e-2 ***	1.2e-2 (0.029)	1.9e-2 (0.001)	3.7e-2 (1e-10)	**-8.7e-3 (0.125)**
b	7.6e-4 (7e-13)			**4.6e-4 (0.030)**	-1.0e-3 (2e-6)	2.9e-4 (3e-4)	**-7.5e-5 (0.335)**	-4.0e-5 (0.607)	**3.4e-4 (3e-5)**	-1.2e-4 (0.118)
π_1	0.729 ***				-1.29 ***	-8.3e-2 (1e-7)	5.8e-2 (2e-4)	-2.1e-1 ***	-1.1e-1 (6e-11)	-1.1e-1 (5e-13)
f_b	0.966 ***					1.1e-1 (1e-12)	1.5e-2 (0.327)	8.1e-2 (2e-7)	1.4e-1 ***	**-2.3e-2 (0.145)**
					classifier	*ir*	*NB*	*1NN*	*svm*	*tree*
					contrast to *lda*	-3.3e-2 (9e-8)	-4.6e-2 (8e-14)	6.3e-2 ***	-3.0e-2 (1e-5)	8.3e-2 ***

(1) main stands for the direct effect of par,

(2) :κ stands for the interaction of par with κ, e.g. $p : \kappa$ for the interaction between p and κ (other interactions analogously),

(3) *** stands for (< 2e-16), i.e. p-value numerically 0.

5.2 Convergence

Let us now discuss the research questions *Convergence*. We assume convergence to a if $|e_{1980} - a| < 0.025$, where $0 \leq e_{1980} :=$ mean estimated error rate for $p = 1980$.

In **Case A**, convergence of error rates to $a = 0$ is observed in the case of complete independence ($b = 1, \kappa = 0$) (cp. Figure 1) for *ir*, *1NN*, and *svm* in 100 % of the cases except for $\pi_1 = 0.5$, $f_b = 0.45$, for *NB* except for f_b near π_1, and for *lda* only if f_b small and $\pi_1 = 0.5$.

For dependent features ($\kappa > 0$), convergence to $a = 0$ again appears most often for *ir, 1NN, svm* and somewhat less often for *NB* (cp. Figure 2 for an example). Note that convergence to $a = 0$ appears in 14 - 23 % of the cases with higher percentages for higher π_1 (cp. Table 4). Also note that *svm* does not work for $\pi_1 = 0.1$ because there is only one observation in class 1 for training.

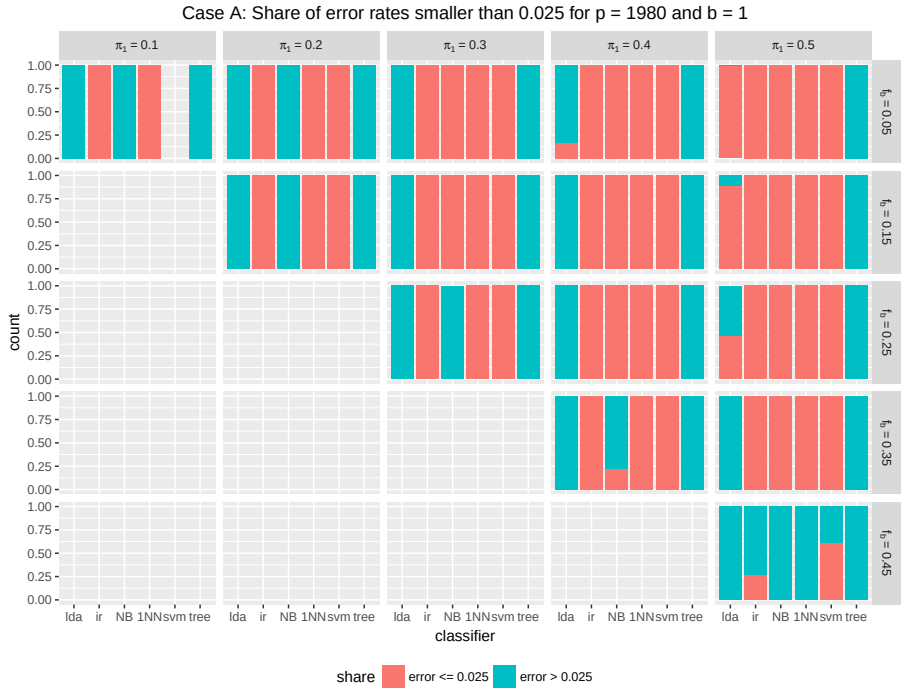

Figure 1: Case A: Individual error rates: Convergence to 0 for $b = 1$ ($\kappa = 0$), on the y-axis "count" gives the percentage of error rates converged to 0.

Convergence is also observed to limits $a \neq 0$. However, the share of the asymptotic Bayes error rate 0 is with 21 % (= (11 + 32 + 60 + 86 + 110) / 1424, cp. Table 4) of all cases bigger than the share of convergence to any probability π_1 of class 1 which is maximum in $\pi_1 = 0.5$ with 12 % (= 167 / 1424). Nevertheless, in 40 % of the cases convergence to π_1 is realized so that error rate convergence to the worst rate π_1 is not unusual. Moreover, note that there is even convergence to unacceptable rates distinctly $> \pi_1$, especially for classifier *tree*.

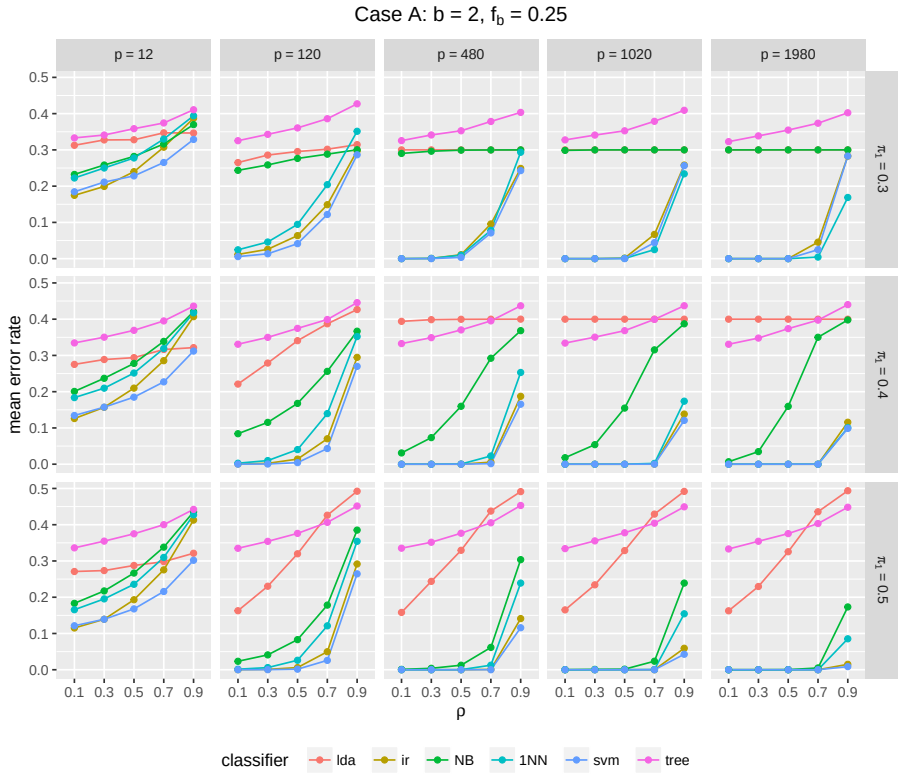

Figure 2: Case A: Dependence on κ for $b = 2$.

Table 4: Case A: Number of Cases: Convergence to lim = 0, 0.1, 0.2, 0.3, 0.4, 0.5 for different π_1. In the first block of rows, the column *max* denotes the maximum number of replicates for the corresponding row. Since $f < \pi_1$ by construction, there are more replicates for higher π_1. Green numbers relate to convergence to 0. For individual classifiers, diagonal red numbers equal the maximum number possible. Violet numbers indicate convergence to a value $> \pi_1$. Note that only the right block in the first row represents percentages.

All classifiers:

π_1\lim	0	0.1	0.2	0.3	0.4	0.5	*max*
0.1	11	62	0	0	0	0	*80*
0.2	32	3	108	20	0	0	*192*
0.3	60	3	12	112	29	0	*288*
0.4	86	8	11	11	122	20	*384*
0.5	110	10	15	14	35	167	*480*

% row-wise

	0	0.1	0.2	0.3	0.4	0.5
0.1	14	78	0	0	0	0
0.2	17	2	56	10	0	0
0.3	21	1	4	39	10	0
0.4	22	2	3	3	32	5
0.5	23	2	3	3	7	35

LDA:

	0	0.1	0.2	0.3	0.4	0.5
0.1	0	16	0	0	0	0
0.2	0	0	32	0	0	0
0.3	0	0	0	45	1	0
0.4	0	0	1	0	55	0
0.5	3	2	2	1	3	51

Independence Rule:

	0	0.1	0.2	0.3	0.4	0.5
0.1	4	11	0	0	0	0
0.2	9	0	18	0	0	0
0.3	17	1	4	10	4	0
0.4	25	2	1	3	7	2
0.5	30	2	1	3	4	18

Naïve Bayes:

	0	0.1	0.2	0.3	0.4	0.5
0.1	0	16	0	0	0	0
0.2	0	0	32	0	0	0
0.3	5	0	1	37	0	0
0.4	12	1	0	1	37	0
0.5	23	1	4	1	7	24

1 Nearest Neighbour:

	0	0.1	0.2	0.3	0.4	0.5
0.1	7	3	0	0	0	0
0.2	13	1	2	6	0	0
0.3	20	1	2	1	8	0
0.4	25	3	1	0	5	6
0.5	25	3	1	1	5	24

SVM:

	0	0.1	0.2	0.3	0.4	0.5
0.1	-	- (does not run) -				
0.2	10	0	19	0	0	0
0.3	18	1	2	12	4	0
0.4	24	2	3	3	7	2
0.5	29	2	3	2	4	18

Decision Tree:

	0	0.1	0.2	0.3	0.4	0.5
0.1	0	16	0	0	0	0
0.2	0	1	5	14	0	0
0.3	0	0	3	7	13	0
0.4	0	0	5	4	11	10
0.5	0	0	4	6	12	32

In order to characterize situations leading to convergence to $a = 0$, we defined two new classes, class 0 with all examples with $e_{1980} < 0.025$ and class 1 with all other examples. This defines a 2^{nd}-stage classification problem with influential features b, κ, π_1, f, and *classifier*. Applying the above *tree* classifier with priors 0.20 for class 0 and 0.80 for class 1 to this problem, leads to the decision tree in Figure 3 with acceptable 4.1 % training error rate, 7.0 % balanced training error rate (taking the mean of the error rates for class 0 and class 1), as well as 5.8 % cross-validated error rate. Note that the priors are motivated by the fact that class 0 appears in 299 examples and class 1 in 1225 of our examples. The decision tree clearly indicates that higher block sizes $b > 2$ are more likely not leading to convergence to 0 though the theoretical convergence condition is valid for $b = 22$ and $p = 1980$ (cp. Sections 2.3, 4.5). Moreover, convergence to 0 is not

restricted to the distance-based classifiers *1NN, svm*, as well as {*lda, ir*} for $\pi_1 = 0.5$ (cp. Section 2.3). Indeed, convergence of *ir* depends on the influential features in the same way as *1NN, svm*, and convergence to 0 is also appearing for *NB*. Obviously, convergence to asymptotic Bayes error 0 can be expected for *ir, 1NN, svm* if $b \leq 2$ and $\kappa \leq 0.7$, $f_b \leq 0.35$ or $\kappa > 0.7$, $f_b \leq 0.15$ and if $b = 22$ (i.e. $b > 2 \wedge b \leq 22$) and $\kappa \leq 0.3$, $f_b \leq 0.05, \pi_1 \geq 0.2$, as well as for *NB* if $b \leq 2$ and $\pi_1 > 0.3$, $f_b \leq 0.25$. Such dependence on π_1 might be related to the fact that *lda, ir* are only distance-based for $\pi_1 = 0.5$. The dependence on f_b shows that the dimension-wise or block-wise Bayes error should not be too high to allow for convergence to 0, and the dependence on κ might reflect that the theoretical result in Section 2.3 is only valid for $\kappa = 0$. Note that not too unbalanced classes should be preferred for the standard implementations of the classifiers. In summary, for *ir, 1NN*, and *svm* convergence to 0 can be expected if block size b and f_b, κ are reasonably small, but for *lda, tree* and for *NB* with $\pi_1 \leq 0.3$ or $f_b > 0.25$ convergence to 0 should not to be expected (cp. also Table 4).

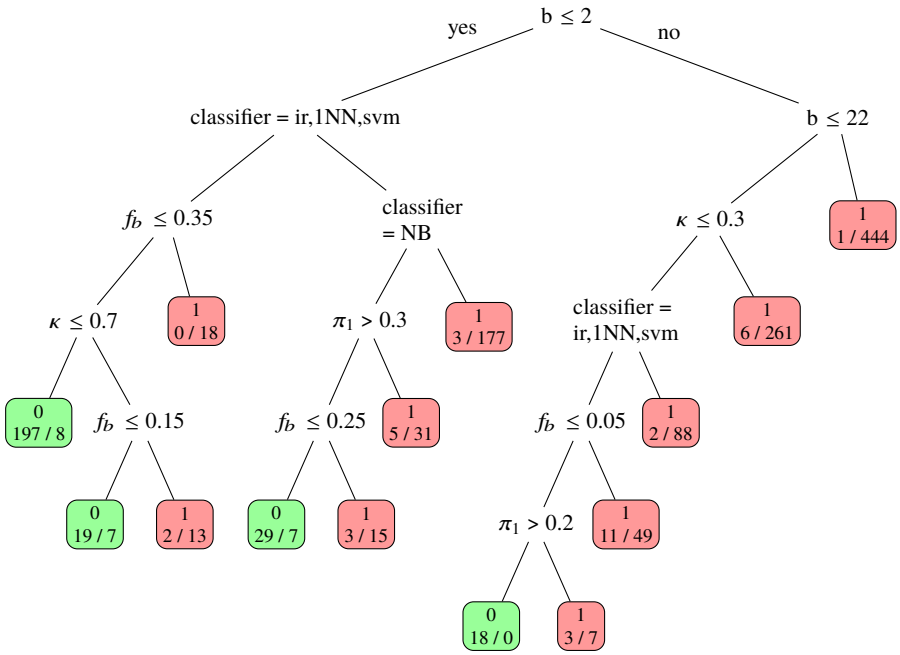

Figure 3: Case A: Characterization of convergence to $a = 0$ (class 0, green) vs. $a \neq 0$ (class 1, red) by a classification tree. In the non-terminal nodes the split is characterized. Note that in the tree the implicit "no"-alternatives to the classifiers *ir, 1NN, svm* are *NB, lda, tree*, and to *NB* these alternatives are *lda, tree*. In the terminal nodes the predicted class is denoted, and the number of cases in the two classes.

Figure 4 shows an example for general convergence behavior with the small fixed Bayes error $f_b = 0.05$. Note that for *lda* convergence to π_1 is obvious, except for $\pi_1 = 0.5$. For *ir, 1NN,* and *svm* convergence to 0 is always realized, and for *NB* for $\pi_1 > 0.2$.

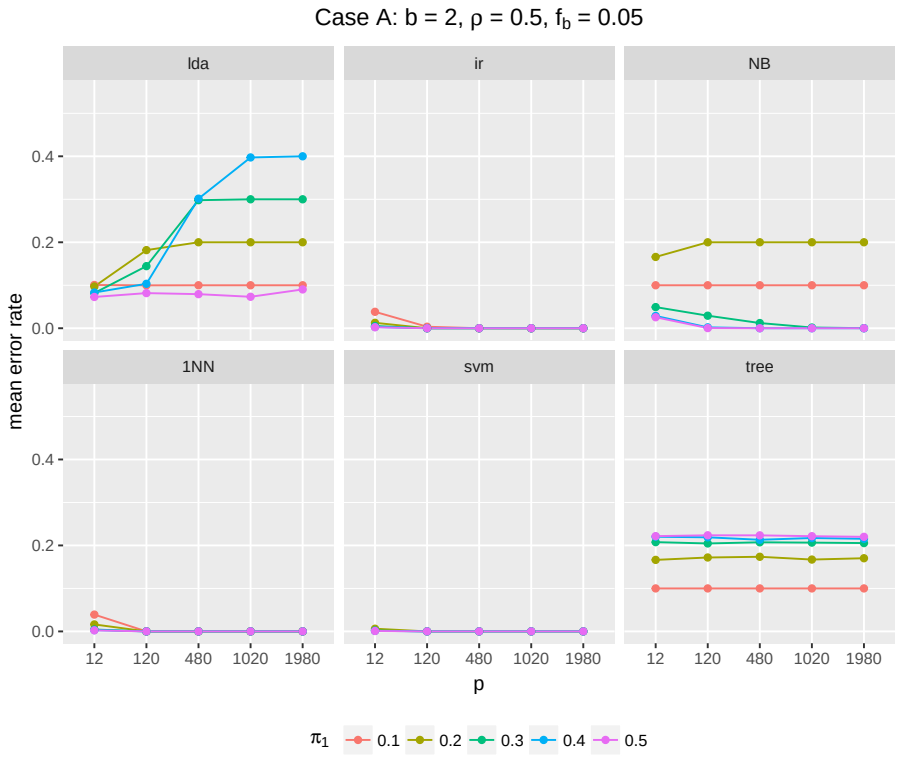

Figure 4: Case A: Example: Convergence to 0 or π_1?

In **Case B1** (considering Table 5), we observe convergence to $\pi_1 = 0.5$ in 92 % of the cases, but less often for $\pi_1 = 0.1, 0.2, 0.3, 0.4$ (cp. the main diagonal of Table 5). For *lda* and *NB* convergence to π_1 is observed in around 88 % (= $(21+40+51+63+103)/(21+42+63+84+105)$) of the cases, for 1*NN* only for $\pi_1 = 0.5$. Overall, convergence to π_1 appeared in 63 % (= $(80+153+168+201+581)/1869$) of the cases.

Table 5: Case B1: Number of Cases: "Convergence" to lim = 0.1, 0.2, 0.3, 0.4, 0.5 for different π_1. For individual classifiers, diagonal bold red numbers equal the maximum number possible and diagonal numbers in brackets indicate the corresponding unmatched maximum. Violet numbers indicate convergence to a value $> \pi_1$.

All classifiers:

π_1\lim	0.1	0.2	0.3	0.4	0.5	max
0.1	80	**1**	0	0	0	*105*
0.2	1	153	**72**	**2**	0	*252*
0.3	0	0	168	**49**	**2**	*378*
0.4	0	0	2	201	**148**	*504*
0.5	0	0	4	3	581	630

% row-wise

	0.1	0.2	0.3	0.4	0.5
0.1	76	**1**	0	0	0
0.2	0	61	**29**	**1**	0
0.3	0	0	46	**13**	**1**
0.4	0	0	0	40	**29**
0.5	0	0	1	0	**92**

LDA:

	0.1	0.2	0.3	0.4	0.5
0.1	**21**	0	0	0	0
0.2	0	41(42)	0	0	0
0.3	0	0	49(63)	**3**	0
0.4	0	0	0	60(84)	0
0.5	0	0	0	0	**105**

Independence Rule:

	0.1	0.2	0.3	0.4	0.5
0.1	17	**1**	0	0	0
0.2	0	30	**1**	**2**	0
0.3	0	0	31	**7**	**2**
0.4	0	0	0	36	**19**
0.5	0	0	0	0	95

Naïve Bayes:

	0.1	0.2	0.3	0.4	0.5
0.1	**21**	0	0	0	0
0.2	0	40	0	0	0
0.3	0	0	51	**3**	0
0.4	0	0	0	63	**5**
0.5	0	0	0	0	103

1 Nearest Neighbour:

	0.1	0.2	0.3	0.4	0.5
0.1	0	0	0	0	0
0.2	0	0	**42**	0	0
0.3	0	0	0	**22**	0
0.4	0	0	0	0	**63**
0.5	0	0	0	0	99

SVM:

	0.1	0.2	0.3	0.4	0.5
0.1	- (does not run) -				
0.2	0	38	0	0	0
0.3	0	0	36	**9**	0
0.4	0	0	0	37	**5**
0.5	0	0	0	0	97

Decision Tree:

	0.1	0.2	0.3	0.4	0.5
0.1	**21**	0	0	0	0
0.2	1	4	**29**	0	0
0.3	0	0	1	**12**	0
0.4	0	0	2	5	**56**
0.5	0	0	4	3	82

Note, however, that the asymptotic error is very seldom better than π_1, only for *tree* some asymptotic error rates are $< \pi_1$. Moreover, many asymptotic error rates are $> \pi_1$, especially for classifier *1NN*. Also note that all estimated error rates are bigger than the pre-fixed Bayes error f_b.

In **Case B2**, convergence to $\pi_1 = 0.5$ is observed in 95 % of the cases, for $\pi_1 = 0.1, 0.2, 0.3, 0.4$ such convergence is less systematically realized (cp. Table 6). For *lda* and *NB* convergence to π_1 is observed in 95 - 97 % of the cases, for 1*NN* convergence to π_1 is only realized for $\pi_1 = 0.5$. Overall, convergence to π_1 is realized in 69 % of the cases. Note that the asymptotic error rate is never better than π_1 and that, again, especially classifier *1NN* converges to a rate $> \pi_1$ very often.

5.3 Classifier Ranking

For the research question *Classifier Ranking*, in **Case A** we look at ranking via mean absolute distance of estimated error rates to Bayes error $f_b^{p/b}$, getting 0.20 for {*svm, ir*}, 0.21 for 1*NN*, 0.25 for *NB*, 0.32 for *lda*, and 0.33 for *tree*. Thus, *svm*, *ir*, and 1*NN* appear to be most adequate in **Case A**. Note that the mean distance would be 0.36 if all error rates would be π_1 for each classifier except 0.38 for *svm*.

In **Case B1**, classifier ranking is realized via mean absolute distance of estimated error rates to Bayes error f_b, getting 0.185 for *NB*, 0.19 for *lda, svm*, 0.20 for *ir*, 0.21 for *tree*, and 0.22 for *1NN*. The mean absolute distance would be 0.183 if all error rates would be π_1 for each classifier except 0.193 for *svm*. Therefore, all classifiers sometimes produce error rates $> \pi_1$ (see Table 5).

Table 6: Case B2: Number of Cases: "Convergence" to lim = 0.1, 0.2, 0.3, 0.4, 0.5 for different π_1. For individual classifiers, diagonal bold red numbers equal the maximum number possible and diagonal numbers in brackets indicate the corresponding unmatched maximum. Violet numbers indicate convergence to a value $> \pi_1$.

All classifiers:

π_1\lim	0.1	0.2	0.3	0.4	0.5	max
0.1	64	**2**	0	0	0	*80*
0.2	0	120	**60**	0	0	*192*
0.3	0	0	155	**19**	0	*288*
0.4	0	0	0	186	**126**	*384*
0.5	0	0	0	0	456	*480*

% row-wise

π_1\lim	0.1	0.2	0.3	0.4	0.5
0.1	80	**3**	0	0	0
0.2	0	63	**31**	0	0
0.3	0	0	54	**7**	0
0.4	0	0	0	48	**33**
0.5	0	0	0	0	**95**

LDA:

	0.1	0.2	0.3	0.4	0.5
0.1	**16**	0	0	0	0
0.2	0	**32**	0	0	0
0.3	0	0	45(48)	**1**	0
0.4	0	0	0	56(64)	0
0.5	0	0	0	0	**80**

Independence Rule:

	0.1	0.2	0.3	0.4	0.5
0.1	**16**	0	0	0	0
0.2	0	28	0	0	0
0.3	0	0	30	**7**	0
0.4	0	0	0	36	**11**
0.5	0	0	0	0	71

Naïve Bayes:

	0.1	0.2	0.3	0.4	0.5
0.1	**16**	0	0	0	0
0.2	0	**32**	0	0	0
0.3	0	0	47	0	0
0.4	0	0	0	58	0
0.5	0	0	0	0	79

1 Nearest Neighbour:

	0.1	0.2	0.3	0.4	0.5
0.1	0	**2**	0	0	0
0.2	0	0	**31**	0	0
0.3	0	0	0	**12**	0
0.4	0	0	0	0	**49**
0.5	0	0	0	0	75

SVM:

	0.1	0.2	0.3	0.4	0.5
0.1	- (does not run) -				
0.2	0	28	0	0	0
0.3	0	0	33	**7**	0
0.4	0	0	0	36	**6**
0.5	0	0	0	0	71

Decision Tree:

	0.1	0.2	0.3	0.4	0.5
0.1	**16**	0	0	0	0
0.2	0	0	**29**	0	0
0.3	**0**	**0**	**0**	**0**	**0**
0.4	0	0	0	0	**64**
0.5	0	0	0	0	**80**

In **Case B2**, classifier ranking is again realized via mean absolute distance of estimated error rates to Bayes error f_b, getting 0.180 for *NB*, 0.19 for *ir, svm, lda*, 0.22 for 1*NN*, and 0.24 for *tree* with, again, a mean distance of 0.183 if all error rates would be π_1 for each classifier except 0.193 for *svm*. Note that this time classifiers *NB* and *svm* produce mean distances smaller than the mean distance corresponding to π_1 errors (cp. Table 6 for cases converged to some π_1).

6 Noisy Performance

For the research question *Noisy performance* we study the influence of noise on the error rate.

6.1 Noisy Performance: Simulation design

We compare the above situations with $p = 12$ and $p = 120$ influential features with a situation with 120 features where only 12 features influence class separation and 108 features are just independent noise. This is called the (12 + 108)-situation in the following. For this, we first generate mean vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ as well as the covariance matrix $\boldsymbol{\Sigma}$ for $p = 12$ as described in Section 4.5. Then, we elongate $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ by 108 zeros. As the new covariance matrix we take the 120×120 identity matrix where the left upper 12×12 block is replaced by the 12×12 covariance matrix generated before. All parameters except p are varied as indicated in Section 4.5.

In order to identify the relevant features, we apply feature selection in the (12 + 108)-situation. For selection, we used the RELIEF criterion in the package mlr (Bischl et al (2016)) of the software R. RELIEF estimates the quality of attributes according to how well their values distinguish between instances of different classes that are near to each other (Kira and Rendell (1992)). Then, we compare the behavior of the classifiers for $p = 12$ and $p = 120$ influential features without feature selection with the (12 + 108)-situation with selection of the most important 12 or 18 features and we test the null-hypothesis:

H_0: By the inclusion of noisy features the mean error rates of the different classifiers are smaller than or equal to the mean error rates in situations without noise and feature selection, i.e. $p = 12$ or $p = 120$ in Section 4.5.

The idea behind this hypothesis is that not all relevant features are identified and that, therefore, the relevant dimension is lower than in the non-noisy case (see results).

6.2 Noisy Performance: Results

First, we report the number of correct identifications of influential features in the (12 + 108)-situation (cp. Table 7). Obviously, the mean number of identified influential features over the 200 replications is small in all cases. To characterize the range of realized correct identifications we used the statistic "mean + 3·std.dev.". In the best case (18 features selected in **Case A**), mean + 3·std.dev. = 11.4 is still relatively close to the pursued number 12. In the worst case (12 features selected in **Case B1**), however, the value of this statistic is only 5.53, i.e. much smaller than 12. Obviously, **Case A** is easier for correct feature selection and the **Cases B1** and **B2** behave similar. Let us see how this poor feature selection affects the error rates.

Table 7: Mean (Std.dev.) of the Number of Identified Influential Features when 12 or 18 features are selected out of 12 influential and 108 noise factors.

Case	12 features selected	18 features selected
A	2.82 (2.32)	3.65 (2.58)
B1	1.57 (1.32)	2.21 (1.59)
B2	1.68 (1.36)	2.36 (1.59)

Corresponding to hypothesis $\mathbf{H_0}$, the number of significant results of the Welch-test at the 1 %-level is given in Table 8 for the different classifiers. We distinguish **Case A** and **Case B**, summarizing the **Cases B1** and **B2**, and we test the mean error rates in the (12 + 108)-situations with 12, 18, and all 120 selected features against the corresponding mean error rates for $p = 12$ and $p = 120$ influential features. Note that the classifiers are sometimes producing error rates exactly $= \pi_1$ in all repetitions, e.g. classifiers *NB* and *tree* for $\pi_1 = 0.1$ and all different b and classifier *NB* sometimes also for $\pi_1 = 0.2$ (cp. Figure 5). In these cases,

the test could not be carried out. Therefore, in Table 8 the reported number of tests differ in the different situations.

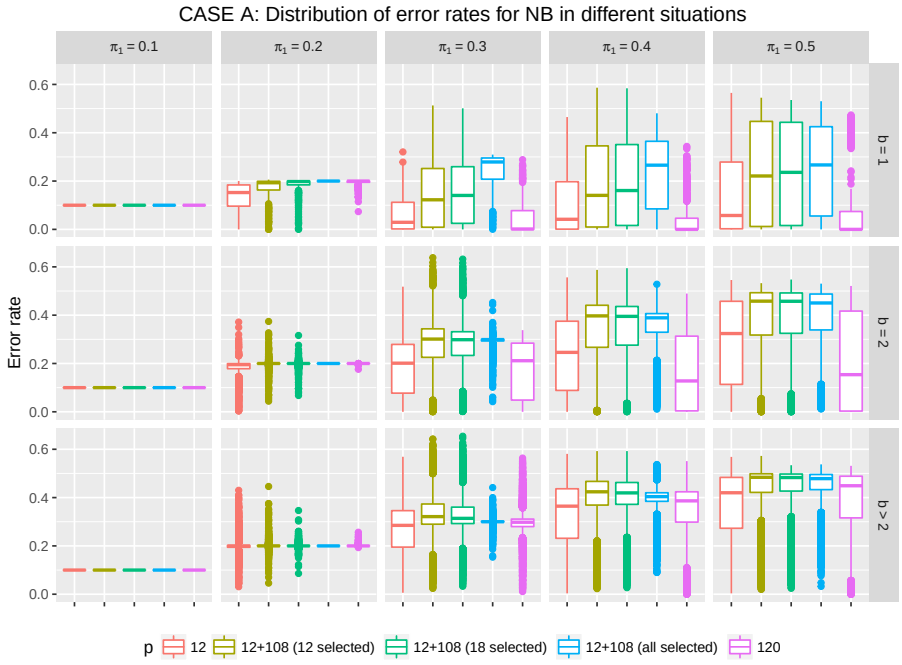

Figure 5: Case A: Behavior of *NB* in noisy and non-noisy situations.

Table 8 shows, e.g., that on the one hand in **Case A** and *svm*, hypothesis $\mathbf{H_0}$ is always rejected for $p = 12$. On the other hand, in **Case B**, hypothesis $\mathbf{H_0}$ nearly always cannot be rejected for classifier 1*NN* when $p = 120$ (cp. bold numbers in Table 8). Note that *svm* appears to react the most negative to noise among the classifiers.

Overall, classification with additional noise ((12 + 108)-situations) is seldom better than without noise ($p = 12$), but frequently better than with more influential features ($p = 120$), in particular in **Case B**. Moreover, in **Case B** the classifiers appear to react more positive to noise than in **Case A**. This might reflect the fact that in **Case B** more influential factors increase the error rate

up to π_1 and that in (12 + 108)-situations the number of influential factors is smaller than for $p = 12$ and particularly $p = 120$.

Table 8: Number of Significant and Non-Significant Results of the Welch-Test at 1 %-Level.

	Case A				**Case B**			
	$p = 12$		$p = 120$		$p = 12$		$p = 120$	
Classifier	signif.	non-sig.	signif.	non-sig.	signif.	non-sig.	signif.	non-sig.
lda	199	41	156	83	315	240	256	276
ir	207	32	173	37	371	184	141	414
NB	194	30	175	29	312	206	162	304
1*NN*	219	18	169	39	370	185	14	**541**
svm	222	**0**	171	18	474	44	197	321
tree	179	45	145	79	426	92	35	483

7 Discussion

We studied standard classification methods for dimensions $p \gg n$. We developed an experimental design to discuss the effects of certain factors on the error rate in **Case A** with decreasing Bayes error for increasing p and in **Case B** with constant Bayes error. The design factors are:

- p = number of features;
- κ = covariance between features in special covariance structure;
- b = block size in covariance matrix;
- π_1 = probability of class 1;
- f_b = true error rate (in blocks).

We saw significance of (nearly) all varied factors and corresponding interactions (research question *Parameter Dependence*). Moreover, convergence of error rates for increasing dimension p (research question *Convergence*) is observed

- to asymptotic Bayes error 0 (**Case A**) most often for independent dimensions ($b = 1, \kappa = 0$), but also for dependent dimensions, especially

for the classifiers *ir, 1NN, svm*. For *lda* and *tree*, convergence to π_1 is often observed, for classifier *NB* if π_1 is small or f_b large. Overall, stochastically independent dimensions should be preferred.

- In the case of constant pre-fixed Bayes error f_b for all numbers of dimensions p (**Case B**), convergence to π_1 is systematically observed for $\pi_1 = 0.5$ as well as for *lda* and *NB* in general. Also, asymptotical rules are often worse than the data independent rule *"Always predict the larger class"*, especially for classifiers *1NN, tree*.

Concerning the research question *Classifier Ranking*, the classifiers *ir* and *svm* show the smallest mean absolute distance to the Bayes error in **Case A**. In **Case B**, however, no classifier is really recommendable.

Concerning the research question *Noisy Performance*, classification with **additional noise factors** ($p = 12 + 108$ situation) is seldom better than without noise ($p = 12$), but frequently better than with more relevant influential features ($p = 120$), in particular in **Case B**. In **Case B** the classifiers appear to react more positive to noise than in **Case A**. This might reflect the fact that in **Case B** more influential factors increase the error rate up to π_1 and that noise factors reduce the number of influential factors. Classifier *svm* appears to react the most negative to noise among the classifiers. As a consequence, adequate *Rules of Thumb* to avoid the worst error rate π_1 for high dimensions and small numbers of training observations may be:

- Avoid "complicated" classifiers: *ir* might be adequate, *svm* should only be considered as an expensive alternative which is additionally sensitive to noise factors.
- From the outset, look for stochastically independent dimensions and not too unbalanced classes.
- Only take into account features which influence class separation sufficiently. Variable selection might help, though filters might be too rough.
- Compare your result with the result of the data independent rule *"Always predict the larger class"*.

Let us compare the results in this paper with our former results in Weihs (2016). In that paper, simulations were performed only for $\pi_1 = 0.5$ and a covariance matrix which was on the one hand somewhat more general in that different correlations between features were used, but on the other hand more restrictive

in that the matrices were nearly diagonal, i.e. had much higher values at the diagonal than in the other entries. The individual error rate f was not controlled, but implicitly pre-fixed once by the class distance in each dimension. **Case A** is represented by the class distance $md = 2.5$ and **Case B** by $md = 20/\sqrt{p}$. Then, the results showed convergence to 0 in **Case A** except for *tree*, and to $\pi_1 = 0.5$ in **Case B** for all methods. Moreover, if class distance sufficiently increases in **Case A** in a specific way for higher dimensions, then error rates were decreasing, even though the theoretical condition that the covariance matrix is diagonal (cp. Section 2.3) is only approximately fulfilled. Finally, if not all features influence class separation, convergence to 0 was slower in **Case A**. In such cases, feature selection choosing the number of selected features somewhat too high appeared to be better than choosing it too low.

Obviously, results in Weihs (2016), are generalized in this paper allowing for pre-specification of a general π_1 and the error rate f_b (corresponding to a certain class distance). Moreover, we study distinctly non-diagonal covariance matrices, albeit of a special structure. We show that our former results for *lda* and $\pi_1 = 0.5$ are somewhat special in **Case A**. Overall, the problems of *lda* with approximating the Bayes error in high dimensions are much clearer now. Finally, in the case of noise, we have seen that feature selection is identifying more relevant features if the number of selected features is higher than the number of relevant features. This, in a way, explains our results in Weihs (2016), concerning feature selection.

However, also in this paper settings are special, particularly the covariance structure. We use normal distributions with special invertible covariance matrices and identical contributions to class choice by all feature blocks. As possible extensions you may want to use other data distributions than normals, vary contributions of feature blocks to class separation, or use other covariance structures. Most easily, you may want to choose different error rates f_b and different correlations in the different blocks. Also, you may want to include other classification methods in the comparison such as methods with nonlinear decision borders like radial basis *svm* and ensemble methods like bagged trees (as in Weihs (2016)). Another extension would be to study evaluation by *leave-one-out*, possibly with more observations per class in order to make the results more reliable.

Acknowledgements The authors would like to thank Daniel Horn and Sarah Schnackenberg for critical discussions as well as Rosa Pink for coding a basic version of Figure 3. We also thank two reviewers for their very helpful remarks.

References

Bickel PJ, 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(6):989–1010. DOI: 10.3150/bj/1106314847.

Bischl B, Lang M, Kotthoff L, Schiffner J, Richter J, Studerus E, Casalicchio G, Jones ZM (2016) mlr: Machine Learning in R. Journal of Machine Learning Research 17(170):1–5. URL: http://jmlr.org/papers/v17/15-066.html.

Fan J, Fan Y, Wu Y (2010) High-dimensional classification., World Scientific, New Jersey, pp. 3–37. DOI: 10.1142/9789814324861_0001.

Figueiredo MA (2004) Lecture Notes on Bayesian Estimation and Classification. Tech. Rep., Instituto de Telecomunicacoes, and Instituto Superior Tecnico, Lisboa, Portugal, p. 38. URL: http://www.lx.it.pt/~mtf/learning/Bayes_lecture_notes.pdf.

Hall P, Pittelkow Y, Gosh M (2008) Theoretical measures of relative performance of classifiers for high dimensional data with small sample sizes. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1):159–173. DOI: 10.1111/j.1467-9868.2007.00631.x.

Kiiveri HT (2008) A general approach to simultaneous model fitting and variable elimination in response models for biological data with many more variables than observations. BMC Bioinformatics 9:195. DOI: 10.1186/1471-2105-9-195.

Kira K, Rendell LA (1992) A practical approach to feature selection. In: Proceedings of International Conference on Machine Learning, Sleeman D, Edwards P (eds), Morgan Kaufmann, pp. 249–256. DOI: 10.1016/B978-1-55860-247-2.50037-1.

Mai Q (2013) A review of discriminant analysis in high dimensions. WIREs Computational Statistics 5:190–197. DOI: 10.1002/wics.1257.

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Tan M, Tsang MI, Wang L (2014) Towards Ultrahigh Dimensional Feature Selection for Big Data. Journal of Machine Learning Research 15:1371–1429.

Weihs C (2016) Big Data Classification - Aspects on Many Features, Lecture Notes in Computer Science, vol. 9580, Springer Berlin Heidelberg, pp. 139–147. ISBN: 978-3-319417-06-6, DOI: 10.1007/978-3-319-41706-6_6.