THIELE CENTRE
FOR APPLIED MATHEMATICS IN NATURAL SCIENCE

Classification error of the thresholded independence rule

Britta Anker Bak, Morten Fenger-Grøn
and Jens Ledet Jensen

Research Report

No. 05 I December 2012

Classification error of the thresholded independence rule

Britta Anker Bak[1], Morten Fenger-Grøn[2] and Jens Ledet Jensen[1]

[1]Department of mathematics, Aarhus University, Denmark
[2]Research Unit for General Practice, Aarhus University, Denmark

Abstract

We consider classification in the situation of two groups with normally distributed data in the 'large p small n' framework. To counterbalance the high number of variables we consider the thresholded independence rule. An upper bound on the classification error is established which is taylored to a mean value of interest in biological applications.

1 Introduction

Many modern measurement devices are of the high throughput type, whether that be chemometrics measurements (Savorani et al. (2010)), medical imaging problems (Garzon et al. (2011)) or microarray based techniques for cancer classification (Dyrskjøt et al. (2003)). From a statistical point of view the challenge is to handle situations where the number of variables p is much larger than the number of samples n. In this paper we consider classification into two groups based on a p-dimensional vector, and take our inspiration from a cancer setting where the two groups, as an example, can be two subtypes of a cancer and where the measurement comes from a microarray. In the classical setting, with p fixed and $n \to \infty$, the solution to the classification problem is well established, but when the number of variables becomes large compared to the number of observations the situation is much less straightforward.

When the parameters are known the optimal classifier is Bayes rule, see Mardia et al. (1979). However, when $p/n \to \infty$ Bickel and Levina (2004) prove that the estimated version of Bayes rule, known as Fishers rule, asymptotically is no better than a random guess. Intuitively, the estimation of an increasingly large number of covariances makes the generalized inverse of the covariance matrix less and less precise. Avoiding the estimation of the increasing number of covariances naturally leads to the independence rule, also known as naive Bayes, where the covariance matrix in Fishers rule is replaced by its diagonal. Bickel and Levina (2004) discuss this rule and find an upper bound for the classification error when $\log(p)/n \to 0$ and with a restrictive setting for the mean values of the variables.

In this paper we consider a setting, aimed at a microarray experiment, where the number of variables carrying information for discrimination may be increasing with n, although the majority of variables are irrelevant. To get rid of the irrelevant variables the thresholded version of the independence rule is considered, that is, only variables for which the t-statistic is significantly large are included. Fan and Fan (2008) show that in a suitable setting for p and n tending to infinity the t-statistic can, with a probability tending to one, separate the variables with a nonzero mean difference between the two groups and those variables with a zero difference. This points to the relevance of the thresholded independence rule.

We prove in this paper an upper bound for the classification error resembling that of Bickel and Levina (2004), but allowing for a quite different set of conditions on the mean values of the variables. Section 2 contains the setup of the paper and states the main result. The proof is given in section 3 and an appendix collects all the basic inequalities used in the proof.

2 Notation and main result

Based on n_0 observations from group 0 and n_1 observations from group 1 of a p-dimensional vector x, we construct a classifier ξ that maps an observation x to one of the two groups, $\xi(x) \in \{0,1\}$. Let the training data be x_{ij}, $i = 0,1$, $j = 1,\dots,n_i$. Then for an observation x from group 0 the classification error is $W(\xi,\theta) = P_\theta(\xi(x) = 1|\{x_{ij}\})$, where θ parametrize the distributions. Our aim is to control the classification error $W(\xi,\theta)$ uniformly for θ in a chosen set (and at the same time controlling the classification error for an observation from group 1). We consider a setup where an observation x is p-variate normal with mean μ_i dependent on the group $i = 0,1$, and covariance matrix Σ. For the training set we assume that $\kappa_1 \le n_0/n_1 \le \kappa_2$, for some positive constants κ_1 and κ_2.

First we introduce the notation used throughout the paper. The diagonal matrix with variances σ_k^2, $k = 1,\dots,p$, is denoted D, and the correlation matrix is $\Sigma_0 = D^{-1/2}\Sigma D^{-1/2}$. The difference between the means $\Delta_k = \mu_{1k} - \mu_{0k}$, $k = 1,\dots,p$, is called the differential expression and $\delta_k = \Delta_k/\sigma_k$ the scaled differential expression. The average of the kth variable in group i is $\bar{x}_{ik}$, and the observed differential expression is $d_k = \bar{x}_{1k} - \bar{x}_{0k}$. The pooled variance estimate for the kth variable is $s_k^2 \sim \sigma_k^2\chi^2(n)/n$, with $n = n_1 + n_2 - 2$, and $\hat{D}$ is the diagonal matrix with entries s_k^2.

The theoretical optimal classifier when the parameters are known, Bayes rule, is defined as

$$\xi_B(x) = 1(\Delta^{\mathrm{T}}\Sigma^{-1}(x - \tfrac{1}{2}(\mu_0+\mu_1)) \quad \text{with} \quad W(\xi_B,\theta) = \overline{\Phi}(\tfrac{1}{2}(\Delta^{\mathrm{T}}\Sigma^{-1}\Delta)^{1/2}).$$

Here $\overline{\Phi}(x) = 1 - \Phi(x)$ is the tail of the standard normal distribution, and $W(\xi_B,\theta)$ is known as Bayes risk. Replacing Σ by its diagonal we get the theoretical independence rule

$$\xi_{\mathrm{TI}}(x) = 1(\Delta^{\mathrm{T}}D^{-1}(x - \tfrac{1}{2}(\mu_0+\mu_1))$$

width

$$W(\xi_{\mathrm{TI}},\theta) = \overline{\Phi}\Big(\frac{1}{2}\Big(\frac{\Delta^{\mathrm{T}}D^{-1}\Delta}{2(\Delta^{\mathrm{T}}D^{-1}\Sigma D^{-1}\Delta)^{1/2}}\Big)\Big),$$

and where the independence rule ξ_I is obtained on replacing parameters by their estimates. Bickel and Levina (2004) obtain the upper bound $\overline{\Phi}(c\sqrt{K_0}/(1+K_0))$ for $EW(\xi_I, \theta)$ over a subset of $\{\Delta, \Sigma : \ \Delta^{\mathrm{T}}\Sigma^{-1}\Delta \geq c\}$ and where K_0 is an upper bound on $\lambda_{\max}(\Sigma_0)/\lambda_{\min}(\Sigma_0)$ with $\lambda_{\max}$ and $\lambda_{\min}$ the largest and smallest eigenvalue of Σ_0.

The classifier we consider is a thresholded version of the independence rule. For this we define

$$t_k = \frac{d_k}{\sqrt{s_k^2/m}}, \ m = \frac{n_0 n_1}{n_0 + n_1}, \ \text{ and } \ I_k = 1(|t_k| > \sqrt{m}\alpha),$$

$$\hat{\Delta}_k = d_k I_k, \ \text{ and } \ \hat{\mu}_{ik} = \begin{cases} \bar{x}_{ik} & \text{if } I_k = 1, \\ \frac{n_0}{n_0+n_1}\bar{x}_{0k} + \frac{n_1}{n_0+n_1}\bar{x}_{1k} & \text{if } I_k = 0. \end{cases}$$

The classifier is

$$\xi(x) = 1\big(\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}(x - \tfrac{1}{2}(\bar{x}_{1k} + \bar{x}_{0k})) > 0\big). \tag{2.1}$$

The threshold α that appears in the definition depends on n, $\alpha = \alpha_n$, but for notational convenience we hide this dependency.

The model is parametrized by $\theta = (\mu_1, \mu_2, \Sigma)$ and the parameter space we consider is defined in two steps. The first step restricts the covariance matrix Σ and the second step restrict the mean values μ_0 and μ_1. We define

$$\Theta = \big\{\theta : \forall k \ c_1^D \leq \sigma_k^2 \leq c_2^D, \ \lambda_{\max}(\Sigma_0) \leq c_2, \ \theta \in B\big\}, \tag{2.2}$$

where c_1^D, c_2^D, c_2 are positive constants, $\lambda_{\max}$ is the maximal eigenvalue and B is a set putting restrictions on the mean values. For the set B we consider two possibilities. The first covers the case when the number of differentiable expressed variables, with an expression above $\alpha/2$, is of smaller order than n and at least one of the differentiable expressions is not small,

$$B_1 = \big\{\theta : \ \#\{k : |\delta_k| \geq \tfrac{\alpha}{2}\} \leq b_n n, \ \#\{k : |\delta_k| > c_0\} \geq 1\big\}, \tag{2.3}$$

where c_0 is a constant and $b_n \to 0$ as $n \to \infty$. In the second case we do not restrict the number of differentiable expressed variables, instead we require that there is not a disproportionally large number of expressed variables around the threshold α,

$$\begin{aligned} K_n &= \#\{k : |\delta_k| > 2\alpha\} \geq 1 \\ B_2 &= \big\{\theta : \ \#\{k : \tfrac{\alpha}{2} \leq |\delta_k| \leq 2\alpha\} \leq c_1 K_n\big\}, \end{aligned} \tag{2.4}$$

where c_1 is a constant. Note that in the specification of the parameter space the dependecy on n has been hidden. The important point is that the c-constants are independent of n.

To formulate our main result we let $\xrightarrow{P_\Theta}$ denote uniform convergence in probability, that is $X_n \xrightarrow{P_\Theta} 0$ if for all $\epsilon_1 > 0$ and $\epsilon_2 > 0$ there exists $n(\epsilon_1, \epsilon_2)$ such that $P(|X_n| > \epsilon_1) < \epsilon_2$ for $n > n(\epsilon_1, \epsilon_2)$ for all $\theta \in \Theta$. Similarly, $\xrightarrow{P<}$ denotes onesided uniform convergence, that is $|X_n|$ is replaced by X_n in the above statement.

Theorem 1. *Let p tend to infinity with n in such a way that $\log(p)/n = \tau_n \to 0$, and let $\alpha \geq c_\alpha \tau_n^{1/2-\gamma}$ where $c_\alpha > 0$ and $0 < \gamma < \frac{1}{2}$. Consider the parameter space given through* (2.2) *and either* (2.3) *or* (2.4)*. Then*

$$W(\xi, \theta) - \overline{\Phi}\bigg(\frac{1}{2\sqrt{c_2}}\sqrt{\sum_{k:|\delta_k|>2\alpha} \delta_k^2}\bigg) \xrightarrow{P<} 0.$$

Remark 1. By exchanging the group labels it is clear that the upper bound of Theorem 1 applies also to the classification error for a new observation from group 1. Furthermore, the formulation of Theorem 1 allows for a triangular array where means and variances depend on n.

Remark 2. The result in Theorem 1 differs in two ways from the result in Bickel and Levina (2004). Firstly, we only use a restriction on the maximal eigenvalue of Σ_0 whereas both the maximal and the minimal eigenvalue enters the bound of Bickel and Levina (2004). Secondly, where we have the term $\sum_{k:|\delta_k|>2\alpha} \delta_k^2$ the situation in Bickel and Levina (2004) (looking into their proof) is comparable to the sum $\sum_k \delta_k^2$. The difference comes from less assumptions on the mean values in our case, achieved by using the thresholded version of the independence rule. Furthermore, for the setup in Bickel and Levina (2004) we have for any sequence $k_n \to \infty$ that $\sum_k \delta_k^2 - \sum_{k\le k_n} \delta_k^2 \to 0$, and since $|\sum_{k\le k_n} \delta_k^2 - \sum_{k\le k_n:|\delta_k|>2\alpha} \delta_k^2| \le 4k_n\alpha^2$ we can take $k_n = 1/\alpha$ so that when $\alpha \to 0$ the two bounds are equivalent.

The proof of the theorem is given in the next section. We use a number of inequalities for the normal distribution and for the t distribution that we have gathered in an appendix.

3 Proof

We start by stating and proving a fundamental lemma. To this end we define for a p-dimensional vector a and a symmetric $p \times p$ matrix M

$$\Psi_\Sigma(a, M) = \frac{a^{\mathrm{T}} M^{-1} a}{2(a^{\mathrm{T}} M^{-1} \Sigma M^{-1} a)^{1/2}},$$

and let $\omega(\hat{D}) = \max\{\max_k s_k^2/\sigma_k^2, (\min_k s_k^2/\sigma_k^2)^{-1}\}$.

Lemma 2. *Let the covariance matrix Σ satisfy the bounds stated explicitly in* (2.2). *Then $\omega(\hat{D}) \xrightarrow{P_\Theta} 1$ and if*

$$\frac{\sum_{k:\hat{\Delta}_k\neq 0} (\hat{\mu}_{0k} - \mu_{0k})^2/\sigma_k^2}{\sum_{k:\hat{\Delta}_k\neq 0} \hat{\Delta}_k^2/\sigma_k^2} \xrightarrow{P_\Theta} 0, \tag{3.1}$$

we have

$$W(\xi, \theta) - \overline{\Phi}(\Psi_\Sigma(\hat{\Delta}, \hat{D})) \xrightarrow{P_\Theta} 0. \tag{3.2}$$

Furthermore, on Θ we have:

$$2\Psi_\Sigma(\hat{\Delta}, \hat{D}) \ge \frac{1}{\omega(\hat{D})\sqrt{c_2}} |D^{-1/2}\hat{\Delta}|. \tag{3.3}$$

Proof. From the multivariate normal distribution we find that

$$\begin{aligned} W(\xi, \theta) &= \overline{\Phi}\bigg(\Psi_\Sigma(\hat{\Delta}, \hat{D}) + \frac{\hat{\Delta}^{\mathrm{T}} \hat{D}^{-1} (\hat{\mu}_0 - \mu_0)}{2(\hat{\Delta}^{\mathrm{T}} \hat{D}^{-1} \Sigma \hat{D}^{-1} \hat{\Delta})^{1/2}} \bigg) \\ &= \overline{\Phi}\bigg(\Psi_\Sigma(\hat{\Delta}, \hat{D}) \Big(1 + \frac{2\hat{\Delta}^{\mathrm{T}} \hat{D}^{-1} (\hat{\mu}_0 - \mu_0)}{|\hat{D}^{-1/2}\hat{\Delta}|^2}\Big) \bigg). \end{aligned}$$

Using lemma 3 point (ii) we see that we need only show that the last term in the inner parenthesis tends to zero uniformly. Using the Cauchy-Schwarz inequality we find

$$\begin{aligned}
\frac{|\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}(\hat{\mu}_0-\mu_0)|}{|\hat{D}^{-1/2}\hat{\Delta}|^2} &= \frac{\left|\sum_{k:\hat{\Delta}_k\neq 0}\frac{\hat{\Delta}_k}{s_k}\frac{(\hat{\mu}_{0k}-\mu_{0k})}{s_k}\right|}{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/s_k^2} \\
&\leq \frac{\sqrt{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/s_k^2}\sqrt{\sum_{k:\hat{\Delta}_k\neq 0}(\hat{\mu}_{0k}-\mu_{0k})^2/s_k^2}}{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/s_k^2} \\
&\leq \omega(\hat{D})\left\{\frac{\sum_{k:\hat{\Delta}_k\neq 0}(\hat{\mu}_{0k}-\mu_{0k})^2/\sigma_k^2}{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/\sigma_k^2}\right\}^{1/2}.
\end{aligned}$$

By assumption the expression within the curly parenthesis tends to zero uniformly. For $\omega(\hat{D})$ we use that $s_k^2/\sigma_k^2 \sim \chi^2(n)/n$. The Chernoff type bound given in lemma 3 point (iii) together with Boole's inequality gives that

$$P\Big(\max_k \frac{s_k^2}{\sigma_k^2} > 1+\epsilon\Big) \leq \sum_{k=1}^{p} P\Big(\frac{s_k^2}{\sigma_k^2} > 1+\epsilon\Big) \leq e^{-\frac{n}{2}(\epsilon-\log(1+\epsilon)-2\tau_n)},$$

with a similar bound for the minimum being less than $1-\epsilon$. Thus $\omega(\hat{D}) \xrightarrow{P_\Theta} 1$ and (3.2) has been proven.

Finally, (3.3) follows from the inequalities

$$\begin{aligned}
2\Psi_\Sigma(\hat{\Delta},\hat{D}) &= \frac{\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}\hat{\Delta}}{(\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}\Sigma\hat{D}^{-1}\hat{\Delta})^{1/2}} \geq \frac{1}{\sqrt{c_2}}\frac{|\hat{D}^{-1/2}\hat{\Delta}|^2}{|D^{1/2}\hat{D}^{-1}\hat{\Delta}|} \\
&\geq \frac{1}{\sqrt{\omega(\hat{D})c_2}}|\hat{D}^{-1/2}\hat{\Delta}| \geq \frac{1}{\omega(\hat{D})\sqrt{c_2}}|D^{-1/2}\hat{\Delta}|.
\end{aligned}$$

□

Proof of main theorem. We start by proving (3.1) and then obtain the result of the theorem from (3.3). We use the bound $m \geq n\kappa_1/(\kappa_2+1/2) = n\kappa_3$ for $n_1 > 4$ and recall that $\log(p) = n\tau_n$.

To study the denominator of (3.1) we first note that $P(I_k = 1, \forall k : |\delta_k| \geq 2\alpha) \to 1$ since the probability of the complement from lemma 3 point (vi) is bounded by

$$\sum_{k:|\delta_k|\geq 2\alpha} P(|t_k| < \sqrt{m}\alpha) \leq pa_1 e^{-m\alpha^2 a_2} \leq a_1 e^{-n\tau_n(\kappa_3 a_2 c_\alpha^2 \tau_n^{-2\gamma}-1)} \to 0. \qquad (3.4)$$

For case B_1 of the parameter space we have from lemma 3 point (i) that

$$P(|d_k|/\sigma_k > c_0/2)) \to 1 \qquad \text{when } |\delta_k| > c_0.$$

For the parameter space B_2 we have $P(|d_k|/\sigma_k > \alpha, \forall k : |\delta_k| \geq 2\alpha) \to 1$ since the probability of the complement from lemma 3 point (i) is bounded by

$$\sum_{k:|\delta_k|\geq 2\alpha} P(|d_k|/\sigma_k < \alpha) \leq pe^{-m\alpha^2/2} \leq e^{-n\tau_n(\kappa_3 c_\alpha^2 \tau_n^{-2\gamma}/2-1)} \to 0.$$

Thus, with a probability tending to one we have that the denominator in (3.1) is bounded by

$$\sum_{k:\hat{\Delta}_k\neq 0}\frac{\hat{\Delta}_k^2}{\sigma_k^2} \geq \sum_{k:|\delta_k|>2\alpha} I_k\frac{d_k^2}{\sigma_k^2} = \sum_{k:|\delta_k|>2\alpha}\frac{d_k^2}{\sigma_k^2} \geq \begin{cases}\frac{c_0^2}{4} & \text{case } B_1,\\ \alpha^2 K_n & \text{case } B_2.\end{cases} \tag{3.5}$$

For the numerator in (3.1) we introduce the notation $\bar{x}_k = \frac{n_0}{n_0+n_1}\bar{x}_{0k} + \frac{n_1}{n_0+n_1}\bar{x}_{1k}$, $\bar{\mu}_k = \frac{n_0}{n}\mu_{0k} + \frac{n_1}{n}\mu_{1k}$. Note that $\bar{x}_k$ is independent of d_k, $\text{Var}(\bar{x}_k) = \sigma_k^2/(n_0+n_1)$ and $\text{Var}(d_k) = \sigma_k^2/m$. Taking expectation and using lemma 3 point (iv) and (v) we get for the nominator in (3.1)

$$\begin{aligned}
E\Big[\sum_{k:\hat{\Delta}_k\neq 0}\frac{(\hat{\mu}_{0k}-\mu_{0k})^2}{\sigma_k^2}\Big] &= E\Big[\sum_{k=1}^{p} 1(\hat{\Delta}_k\neq 0)\frac{(\bar{x}_{0k}-\mu_k)^2}{\sigma_k^2}\Big]\\
&= E\Big[\sum_{k=1}^{p} 1(\hat{\Delta}_k\neq 0)\frac{(\bar{x}_k-\bar{\mu}_k-\frac{n_1}{n_0+n_1}(d_k-\Delta_k))^2}{\sigma_k^2}\Big]\\
&= \sum_{k=1}^{p}\Big\{E\Big[1(\hat{\Delta}_k\neq 0)\frac{1}{n_0+n_1}\Big] + \frac{n_1^2}{(n_0+n_1)^2}E\Big[1(\hat{\Delta}_k\neq 0)\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\Big\}\\
&\leq \sum_{k:|\delta_k|<\frac{\alpha}{2}}\Big\{P(|t_k|>\alpha\sqrt{m})\frac{1}{n_0+n_1} + \frac{n_1^2}{(n_0+n_1)^2}E\Big[1(|t_k|>\alpha\sqrt{m})\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\Big\}\\
&\quad + \sum_{k:|\delta_k|>\frac{\alpha}{2}}\frac{1}{n_0+n_1} + \frac{n_1^2}{(n_0+n_1)^2}E\Big[\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\\
&\leq \begin{cases} 2pa_1e^{-n\alpha^2a_2} + b_n(1+n/m) & \text{case } B_1,\\ 2pa_1e^{-n\alpha^2a_2} + (c_1+1)K_n\big(\frac{1}{n_0+n_1}+\frac{1}{m}\big) & \text{case } B_2.\end{cases}
\end{aligned} \tag{3.6}$$

Dividing (3.6) by (3.5) we see immediately the convergence to zero for case B_1. For case B_2 the second term of (3.6) is the dominating part and dividing this by (3.5) we get

$$\frac{(c_1+1)K_n\frac{1+1/\kappa_3}{n}}{K_n\alpha^2} = \frac{(c_1+1)(1+1/\kappa_3)}{n\alpha^2} \leq \frac{(c_1+1)(1+1/\kappa_3)}{c_\alpha^2 n^{2\gamma}\log(p)^{1-2\gamma}} \to 0.$$

This ends the proof of (3.1).

We next turn to the use of (3.3) to obtain the result of the theorem. We need to show that $|D^{-1/2}\hat{\Delta}|^2 \geq S_\alpha(1+W_n)$, where $S_\alpha = \sum_{k:|\delta_k|>2\alpha}\delta_k^2$ and where W_n tends to zero in probability. We write

$$|D^{-1/2}\hat{\Delta}|^2 = \sum_{k:|\hat{\Delta}_k|\neq 0} I_k\frac{d_k^2}{\sigma_k^2} \geq \sum_{k:|\delta_k|>2\alpha} I_k\frac{d_k^2}{\sigma_k^2}.$$

From the argument in (3.4) all the indicators I_k in this expression are one with a probability tending to one. Thus, we remove I_k from the expression and write

$d_k/\sigma_k = \delta_k + U_k/\sqrt{m}$, where the U_ks are independent standard normal variables. This gives

$$\sum_{k:|\delta_k|>2\alpha} \frac{d_k^2}{\sigma_k^2} = S_\alpha + \frac{2\sqrt{S_\alpha}}{\sqrt{m}}U + \frac{1}{m}V_n = S_\alpha\left(1 + \frac{2}{\sqrt{mS_\alpha}}U + \frac{1}{mS_\alpha}V_n\right),$$

where $U \sim N(0,1)$ and $V_n \sim \chi^2(K_n)$.

For case B_1 notice that $S_\alpha \geq c_0^2$, so that $1/\sqrt{mS_\alpha} \to 0$, and that $K_n/(mS_\alpha) \leq b_n n/(mc_0) \to 0$. For case B_2 we get $mS_\alpha \geq 4m\alpha^2 K_n \to \infty$ and $K_n/(mS_\alpha) \leq \alpha^2/(4\kappa_3 \log(p)) \to 0$. Thus in both cases we have that $|D^{-1/2}\hat{\Delta}|^2 = S_\alpha(1 + W_n)$ with W_n tending to zero in probability and the result of the theorem is obtained. □

4 Discussion

Theorem 1 extends the result of Bickel and Levina (2004) to a more general structure for the mean values in the two groups by using a thresholded version of the independence rule. A similar approach has been considered in Fan and Fan (2008). In their Theorem 5 the case with a known covariance matrix $\Sigma = I$ is considered, where the thresholding is based on the estimated group differences. The bound given by Fan and Fan (2008) can be compared to the bound in Theorem 1 on taking b_n and a of their paper equal to 2α and $\alpha/2$, respectively. When the set of differential expressed variables, $\{j : \delta_j \neq 0\}$, is finite the two bounds agree. More generally, for the cases considered in this paper the asymptotic upper bound by Fan and Fan (2008) is larger than the bound from Theorem 1. It is possible to construct situations where the upper bound of Fan and Fan (2008) tends to one, whereas the bound of Theorem 1 is strictly less than one (for an example consider $\delta_1 \neq 0$ fixed and all remaining nonzero δ_js between $\alpha/2$ and α).

Looking at the proof of Theorem 1 we find that the assumption $\alpha \geq c_\alpha \tau_n^{1/2-\gamma}$ is used to make sure that the expected number of false positives tends to zero. We can turn this upside down and let α be determined by specifying the expected number of false positives. Thus let $\omega_n = pP(|t| > \alpha\sqrt{m})$ be an upper bound on the expected number of false positives among p variables, where t is t-distributed. We can then select ω_n, tending to zero at a sufficiently fast rate, and determine α from ω_n. At the intuitive level the thresholded independence rule should exclude all false positives and should include some true positives. To illustrate this intuitive background of the classifier, we have in Table 1 chosen the expected number of false positives to be $\omega_n = 0.1$, chosen α accordingly, and then calculated the scaled differential expression needed in order to include a variable in the classififer with a high probability, here taken as 0.9. The interesting aspect of the table is the amount of differential expression needed in order to include a variable in the classifier with some certainty. In particular se see that for moderate values of n, the number of observations, the number of variables p can be quite large still allowing for inclusion of true positives.

n	p	α	δ
40	1000	1.37	1.82
80	1000	0.92	1.22
160	1000	0.64	0.85
160	4000	0.69	0.90
160	20000	0.75	0.96

Table 1: Threshold and differential expression to achieve separation. For each value of the number of observations, $n_0 = n_1 = n/2$, and each value of the number of variables p, the threshold α has been chosen such that the upper bound ω_n on the expected number of false positives among p variables is 0.1. The scaled differential expression δ has been chosen such that an expressed variable is included in the classifier with probability 0.9

Appendix

In this appendix we have put together the bounds used for the normal distribution and the t distribution.

Lemma 3. *Let $U \sim N(0,1)$, $V \sim \chi^2(n)/n$ and $t = \sqrt{m}(\delta + \frac{1}{\sqrt{m}}U)/\sqrt{V}$, where $m \geq \kappa_3 n$ and $n \to \infty$. Then there exists constants a_1 and a_2 such that the following inequalities hold.*

(i) For $x > 0$ we have $\overline{\Phi}(x) \leq \frac{1}{2}e^{-x^2/2}$.

(ii) For $x > 0$ and $|\epsilon| < \frac{1}{2}$ we have $|\overline{\Phi}(x(1+\epsilon)) - \overline{\Phi}(x)| \leq \epsilon/4$.

(iii) For $a > 0$ we have

$$\begin{aligned} P(V > 1 + a) &\leq \exp\{-n(\sqrt{1+2a} - 1)^2/4\}, \\ P(V < 1 - a) &\leq \exp\{-na^2/4\}. \end{aligned}$$

(iv) For $|\delta| < \frac{\alpha}{2}$ we have $P(|t| \geq \sqrt{m}\alpha) \leq a_1 e^{-a_2\alpha^2 n}$.

(v) For $|\delta| < \frac{\alpha}{2}$ we have $E\big[1(|t| > \sqrt{m}\alpha)U^2\big] \leq a_1 e^{-a_2\alpha^2 n}$.

(vi) For $|\delta| > 2\alpha$ we have $P(|t| \leq \alpha\sqrt{m}) \leq a_1 e^{-a_2\alpha^2 n}$.

Proof. Proof of (i). This follows from

$$\overline{\Phi}(x) = \int_x^\infty \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz = e^{-x^2/2} \int_0^\infty \frac{e^{-u^2/2}}{\sqrt{2\pi}} e^{-ux} du \leq \tfrac{1}{2} e^{-x^2/2}.$$

Proof of (ii). This is simply the mean value theorem together with the bound $y\phi(y) < 1/4$, $y > 0$, where ϕ is the standard normal density.

Proof of (iii). The two bounds follows from (4.3) and (4.4) in Laurent and Massart (2000).

Proof of (iv). Let f_V be the density of V and consider δ with $|\delta|/\alpha \leq \omega < 1$. Then we have

$$
\begin{aligned}
P(t > \sqrt{m}\alpha) &= \int_0^\infty \overline{\Phi}\big(\sqrt{m}(\sqrt{v}\alpha - \delta)\big) f_V(v)dv \\
&\leq P(V \leq \omega^2) + \overline{\Phi}\big(m(\omega\alpha - \delta)\big) \\
&\leq e^{-n(1-\omega^2)/4} + \tfrac{1}{2}e^{-\alpha^2 m(\omega-\delta/\alpha)^2/2},
\end{aligned}
$$

where the last inequality follows from the bounds (i) and (iii). For the case $\delta < \frac{\alpha}{2}$ we use $\omega = \frac{3}{4}$ and obtain

$$
P(t > \sqrt{m}\alpha) \leq e^{-7n/64} + \tfrac{1}{2}e^{-\alpha^2 n\kappa_3/32} \leq a_1 e^{-a_2\alpha^2 n},
$$

for suitable values of a_1 and a_2 and α bounded from above. For the lower tail, and with $\delta \leq \frac{\alpha}{2}$, we find

$$
\begin{aligned}
P(t < -\sqrt{m}\alpha) &= \int_0^\infty \Phi(-\sqrt{m}(\sqrt{v}\alpha - \delta)) f_V(v)dv \\
&\leq P\big(V \leq \tfrac{1}{2}\big) + \Phi\big(-\sqrt{m}\alpha(\sqrt{1/2} - \tfrac{\delta}{\alpha})\big) \\
&\leq e^{-n/16} + \tfrac{1}{2}e^{-\alpha^2 m(\sqrt{2}-1)^2/8} \\
&\leq a_1 e^{-a_2\alpha^2 n},
\end{aligned}
$$

for suitable values of a_1 and a_2 and α bounded from above.

Proof of (v). As above we consider δ with $|\delta|/\alpha \leq \omega < 1$. Using partial integration we have $\int_z^\infty u^2\phi(u)du = z\phi(z) + \overline{\Phi}(z)$ so that

$$
\begin{aligned}
E\big[1(t > \sqrt{m}\alpha)U^2\big] &= \int_0^\infty \int_{\sqrt{m}(\alpha\sqrt{v}-\delta)}^\infty u^2\phi(u) f_V(v)dudv \\
&\leq P(V \leq \omega^2) + \int_{\omega^2}^\infty \big\{z(v)\phi(z(v)) + \overline{\Phi}(z(v))\big\} f_V(v)dv, \quad z(v) = \sqrt{m}(\alpha\sqrt{v} - \delta) \\
&\leq e^{-n(1-\omega^2)/4} + e^{-\alpha^2 m(\omega-\delta/\alpha)^2/3} + \tfrac{1}{2}e^{-\alpha^2 m(\omega-\delta/\alpha)^2/2},
\end{aligned}
$$

where we have used $x\phi(x/\sqrt{3}) < \frac{1}{2}$ in the last inequality. As before when $\delta < \frac{\alpha}{2}$ we use $\omega = \frac{3}{4}$ and obtain a bound on the form $a_1\exp(-a_2\alpha^2 n)$. For the lower tail $E[1(t < -\sqrt{m}\alpha)U^2]$ the above argument is combined with the argument in point (iv).

Proof of (vi). For $|\delta| > 2\alpha$ we find

$$
\begin{aligned}
P(|t| \leq \alpha\sqrt{m}) \leq P(t \leq \alpha\sqrt{m}) &= \int_0^\infty \Phi\big(\sqrt{m}(\alpha\sqrt{v} - \delta)\big) f_V(v)dv \\
&\leq P\big(V \geq 2\big) + \Phi\big(\sqrt{m}(\alpha\sqrt{2} - 2)\big) \\
&\leq e^{-n(\sqrt{3}-1)^2/4} + \tfrac{1}{2}e^{-m\alpha^2(2-\sqrt{2})^2/2},
\end{aligned}
$$

where we have used points (i) and (iii). As before we obtain a bound on the form $a_1\exp(-a_2\alpha^2 n)$ for suitable a_1 and a_2. □

References

Bickel, P. J. and E. Levina (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.

Dyrskjøt, L., T. Thykjær, M. Kruhøffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Ørntoft (2003). Classification and characterization of bladder cancer stages using microarrays. stage and grade of bladder cancer defined by gene expression patterns. *Nature Genetics 33*, 90–96.

Fan, J. and Y. Fan (2008). High-dimensional classificaion using features annealed independence rules. *The Annals of Statistics 36*, 2605–2637.

Garzon, B., K. Emblem, K. Mouridsen, B. Nedregaard, P. Due-Tønnesen, T. Nome, J. Hald, A. Bjørnerud, A. Håberg, and Y. Kvinnsland (2011). Multiparametric analysis of magnetic resonance images for glioma grading and patient survival time prediction. *Acta Radiology 52*, 1052–1060.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *The Annals of Statistics 28*, 1302–1328.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate Analysis.* Academic Press.

Savorani, F., M. Kristensen, F. Larsen, A. Astrup, and S. Engelsen (2010). High throughput prediction of chylomicron triglycerides in human plasma by nuclear magnetic resonance and chemometrics. *Metabolism and Nutrition 7*, 43.