High-dimensional Classification

PhD Thesis

Britta Anker Bak

Supervisor: Jens Ledet Jensen

July 31, 2015

Department of Mathematics

Aarhus University

Preface

This thesis consists of the work made during my four years as a PhD student at Department of Mathematics, Aarhus University, Denmark. It considers high-dimensional classification from various points of view, and includes an introduction followed by four self-containing chapters. The notation may differ between the chapters.

- Chapter 1 gives a review of existing methods and problems within the field of high-dimensional classification.
- Chapter 2 is based on joint work with Jens Ledet Jensen and Morten Fenger-Grøn, and has been published in the Scandinavian Journal of Statistics (March, 2015). In Section 2.6 an extension of the result from the paper is described. All content of this chapter was also a part of my progress report in 2013.
- Chapter 3 is a submitted paper, based on joint work with Jens Ledet Jensen.
- Chapter 4 is joint work with Jens Ledet Jensen, and contains a corrected version of an existing theorem. We thank Xin Tong for reviewing this material. The content can also be found on arXiv (`http://arxiv.org/abs/1405.5989`).
- Chapter 5 consists of unpublished joint work with Jens Ledet Jensen.

Britta Anker Bak
Aarhus, July 2015

Acknowledgements

First and foremost, I want to thank my supervisor Jens Ledet Jensen for sharing his high level of competence and flow of ideas with me during the past four years. His scientific curiosity and natural scepticism have been an invaluable inspiration, while his patience and always open door have been essential for the completion of this thesis.

Thanks to Sara van de Geer and the Seminar für Statistik at ETH Zürich for welcoming me during the spring semester 2013.

Thanks to everyone at B.3 for creating such a pleasant atmosphere to work in, with the cake club as an enjoyable and deserved interruption from work. A separate thank goes to my office mate, Ina Trolle Andersen, for being around to share pleasures and frustrations of PhD life, and to previous and current members of the famous Statistical Knitting Club.

Finally, I want to thank family, friends, SF Aarhus, and, last but not least, Benjamin, for persistently reminding me that there is a world outside statistics.

Abstract

The scientific and technological development has given rise to an increasing demand of statistical methods that can handle the “large p, small n” paradigm, where the number of variables is large, while the number of observations remains small. One such example is microarray analysis, where the expressions of a large number of genes are measured on a limited number of individuals, bounded by either cost or a limited number of patients.

In this thesis we focus on the high dimensional classification problem of two normally distributed groups with equal covariance matrices. Inspired by the applications in microarray studies, we are primarily interested in situations where the vector of differences between the two group means is sparse, meaning that most variables do not differ systematically between the two groups.

It is well known that Fishers rule fails to work in a scenario where $p/n \to \infty$, due to noise accumulation. Contrary, the independence rule, which ignores correlations between variables, has been proved to be useful in a very specific, extremely sparse, setting. We extend this result under more general assumptions justifiable for microarray data, and in this way theoretically justify current practice.

While ignoring correlation in classification problems decreases the noise accumulation, it also omits potential information and is thus not optimal. In recent years, this recognition has motivated procedures searching directly for the optimal classification vector, instead of estimating it as a product of a matrix and a vector. Such procedures have good properties theoretically as well as empirically when compared to independence classifiers in correlated settings. We perform a comparison of four such procedures through a large simulation study. We furthermore correct a technical mistake and reformulate a theorem regarding the ROAD classifier.

A problem, which has often been either overlooked or neglected in the statistical field, as well as in broader scientific societies, is the bias in classification when the group sizes in the training data differ. When p is fixed the problem is minor, but as p increases it gets more substantial, and classification to the smaller group is sometimes almost impossible. Existing literature on the topic primarily focuses on intuition and methods that are shown empirically to work. We take a more analytic approach in a simple setting where all variables are independent, and thereby obtain an in-depth understanding of the origin of the problem. This insight leads to the suggestion of two new classifiers with practically no bias. Simulation studies support our theoretical conclusion, and further simulations indicate that our methodology can be of relevance also for the situation of dependent variables. Finally, we see that oversampling, a commonly applied method in imbalanced situations, worsens the bias.

Resume

Den videnskabelige og teknologiske udvikling har medført et voksende behov for metoder, der kan håndtere data for “stort p, lille n”, dvs. situationer hvor antallet af variable er højt, mens antallet af observationer fortsat er lavt. Et område hvor dette har relevans, er microarray analyse, hvor ekspressionsniveauet måles på en lang række gener, men på et begrænset antal individer, da målingerne er dyre og der ofte kun er et begrænset antal patienter til rådighed.

Dennne afhandling betragter højdimensionel klassifikation mellem to normalfordelte grupper med samme kovariansmatrix. Inspireret af microarray analyse er vi primært interesserede i situationer svarende til, at de fleste gener ikke afviger systematisk mellem de to grupper.

Når $p/n \to \infty$ er det velkendt, at Fishers lineære diskriminationsregel bryder sammen asymptotisk som følge af akkumulationen af støj. Uafhængighedsregler, der ignorerer korrelation, kan derimod beviseligt udføre en brugbar klassifikation i restriktive situationer. Vi udvider et eksisterende resultat til antagelser, der synes rimelige for microarray data, og på den måde retfærdiggør vi teoretisk den nuværende praksis.

Udover at udeladelsen af korrelation reducerer mængden af støj, vil det også betyde, at en væsentlig information bliver ignoreret i klassifikationen. I de seneste år har dette givet inspiration til en række metoder, der estimerer den optimale klassifikationsvektor direkte i stedet for som et produkt af en matrix og en vektor. Disse metoder har gode egenskaber teoretisk, og også empirisk når de sammenlignes med klassifikationsregler baseret på en naiv uafhængighedsantagelse. Vi sammenligner fire sådanne klassifikationsregler i et simulationsstudium. Desuden reformulerer og retter vi et teoretisk resultat angående klassifikationsreglen ROAD.

Både i den statistiske verden og i bredere videnskabelige sammenhænge bliver det ofte overset eller ignoreret, at der fremkommer et bias i klassifikationen, når antallet af observationer varierer imellem grupperne. Dette giver en tilbøjelighed til at klassificere til den største gruppe. Når p er lille, er dette bias sjældent afgørende, men i højdimensionale situationer kan det være næsten umuligt at klassificere til den mindste af grupperne. Størstedelen af den eksisterende forskning på området har foreslået nye metoder med baggrund i intuition, og efterfølgende påvist at disse virker godt empirisk. Gennem en analytisk tilgang under en antagelse om uafhængige variable opnår vi en dybere forståelse af årsagen til problemet. Dette leder os til to nye metoder, der praktisk talt ikke har noget bias uanset graden af ubalance. Vores teoretiske udledninger bakkes op af simulationer, og vi ser endvidere, at vores metoder også kan være anvendelige i situationer med afhængige variable. Endelig viser vi, at oversampling, der er en gængs metode til at håndtere ubalance, blot gør biasproblemet værre.

Contents

1

Introduction

The development of statistics has always been driven by the demands of other scientific fields as well as the surrounding society. Classical statistical methods, such as the t-test and analysis of variance, were developed due to the needs from growing industries and agricultural organizations with a desire of a more effective production. In such situations, one considers a small number of variables measured on a limited number of observations. This can be an agricultural experiment where various levels of fertilizer and pesticides are added to different plots and subplots, and the crop is evaluated as an outcome. Another example could be determining whether a patient should be assigned to further inspection for a specific disease, based on the levels of a few factors in a blood sample.

As the advances in science and technology have not stopped, the demand of new statistical methods have not either. Today, computers are able to collect and store large amounts of data automatically, so there is no need to restrict the interest to a few parameters only. Simultaneously, methods justifiable and efficient also in situations with a vast amount of data, where the dimension of each observation might be large, are needed. In Committee on Mathematical Sciences Research for DOE's Computational Biology (2005) a thorough description of situations where the biological field challenges the curiosity and ingenuity of statisticians in this millennium is given.

One example of a high-dimensional situation is in image analysis, where one picture is represented by a large number of pixels. If the aim is to group all pictures on the internet into themes, the amount of data points can roughly be considered as tending to infinity, so lack of data is not an issue, while computational limitations definitely are.

Another high-dimensional example is microarray data sets, where the expression levels of a large number of genes are measured simultaneously. The purpose of a such study can be to differentiate between various cancer types, or to enable population screening for a specific form of cancer. Though the price of extracting DNA is decreasing dramatically in these years, the number of observations in such data sets is usually small due to a low number of patients available. This leads to situations where the number of variables, typically denoted by p, is dramatically bigger than the number of observations, typically denoted by n. A further challenge is that probably only a few of the genes in a microarray study are *differentially expressed*, meaning that their values differ systematically between the considered groups of patients, but it is unknown a priori which ones. Finding these important variables is complicated by the fact that some variables coincidentally appear to differ between the groups without having any causal effect. That this aspect is fundamental to take into account is obvious, since 10 000 hypotheses tested simultaneously at a 0.05 level, are expected to detect 500 false positives.

In the following, we focus on statistical methods in the latter situation, known as the 'large p, small n' regime, with focus towards applications in microarray data. In Section 1.1, we introduce our setting along with Bayes rule, Fishers rule and the independence rule, while in Section 1.2, we focus on problems arising for these classifiers in high-dimensional settings. Section 1.3 gives a description of common variable selection methods for independence classifiers, including nearest shrunken centroids, features annealed independence rule and higher criticism thresholding. Section 1.4 and Section 1.5 describe methods allowing the use of correlation in high-dimensional classification. Section 1.6 relates methods from penalized regression to a classification setting. Section 1.7 describes a bias problem arising in classification when the sample size in the various groups differ, and a range of methods meant to eliminate this bias are presented. Finally, Section 1.8 gives a brief overview over the remaining chapters of this thesis.

1.1 Bayes rule, Fishers rule and the independence rule

In this section, we introduce our setting and some of the most well-known linear classifiers. Mathematically, we consider independent observations from two normally distributed groups with equal covariance, that is $x_{0i} \sim N_p(\mu_0, \Sigma)$ for $i = 1, \ldots, n_0$, and $x_{1i} \sim N_p(\mu_1, \Sigma)$ for $i = 1, \ldots, n_1$. Let $\Delta = \mu_1 - \mu_0$ and $n = n_0 + n_1$. An (approximately) sparse setting means that most entries in Δ are (approximately) zero. Inspired by the microarray terminology, the term *differentially expressed* are used for variables where Δ_j is not (approximately) zero after scaling by the inverse standard deviation. Similarly, we call all other variables non-expressed. Our aim is to construct a classifier $\xi(z)$ with a high probability of assigning a new observation z to the correct group, when p is possible much larger than n. That is, we compare classifiers in terms of their classification error $P(\xi(z) \neq y)$, where y is the group label of z.

It is well known (Mardia et al., 1979) that the optimal classifier, in terms of minimizing the classification error, is Bayes rule, which classifies a new observation to group $\xi_B(z)$ where

$$\xi_B(z) = \mathbf{1}\big\{\Delta^{\mathsf{T}}\Sigma^{-1}\big(z - \tfrac{1}{2}(\mu_0 + \mu_1)\big) > \log(\pi_0/\pi_1)\big\}, \tag{1.1}$$

and π_k is the prior probabilities of the k'th group. Until Section 1.7, we always assume $\pi_0 = \pi_1 = 0.5$, so the decision threshold in (1.1) is simply zero. In this situation, the classification error of both groups is $\overline{\Phi}\big((\Delta^{\mathsf{T}}\Sigma^{-1}\Delta)^{1/2}/2\big)$, also known as Bayes risk, where $\overline{\Phi}$ denotes the upper tail of a standard normal distribution.

Bayes rule is an *oracle* rule, meaning that it involves the true parameters, and thus has limited practical applicability. When plugging in maximum likelihood estimates of each of the parameters in (1.1), a more useful classifier is naturally suggested, which in our situation with two normally distributions equals Fishers rule:

$$\xi_F(z) = \mathbf{1}\big\{\hat{\Delta}^{\mathsf{T}}\hat{\Sigma}^{-1}\big(z - \tfrac{1}{2}(\bar{x}_0 + \bar{x}_1)\big) > 0\big\}. \tag{1.2}$$

Here $\bar{x}_k$ is the k'th group average. When p is fixed, Fishers rule approaches Bayes rule, as n tends to infinity. When p is larger than n, $\hat{\Sigma}$ is singular, and Fishers rule is undefined. This can be repaired by applying alternative estimates of Σ^{-1}, but it does not necessarily gives good results, as described in Section 1.2.

When Fishers rule does not perform well, an obvious explanation is the lack of information to estimate $p(p-1)$ parameters of Σ. One solution is to ignore correlation, and estimate the variances only, which leads to the independence rule

$$\xi_I(z) = \mathbf{1}\{\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}(z - \tfrac{1}{2}(\bar{x}_0 + \bar{x}_1)) > 0\}, \tag{1.3}$$

where $\hat{D} = \mathrm{diag}(\hat{\Sigma})$. When the variables are truly independent, the independence rule can be considered as Fishers rule incorporating this a priori information. When the data is not independent, some information is lost by omitting correlations, but often this loss is less serious than the noise arising from estimating $p(p-2)$ covariances.

In the following, $\lambda_i(A)$ for $i = 1, \ldots, p$ denotes the eigenvalues of a symmetric matrix A, with $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denoting the largest and smallest ones, respectively. Define also $\Sigma_0 = D^{-1/2}\Sigma D^{-1/2}$, where $D = \mathrm{diag}(\Sigma)$.

1.2 Problems in classification when $p/n \to \infty$

Bickel and Levina (2004) consider the classification between two p-variate normal distributions, as described above, in an asymptotic setting with $p/n \to \infty$. We here state their main results which show that while Fishers rule works poorly in high-dimensional settings, the independence rule can result in a useful classifier. We further mention a result of Fan and Fan (2008), stating that the independence rule also fails in some high-dimensional settings though.

The parameter space considered in Bickel and Levina (2004) is

$$\Gamma = \{(\mu_0, \mu_1, \Sigma) : \Delta^{\mathrm{T}}\Sigma^{-1}\Delta > c^2, c_1 \leq \lambda_{\min}(\Sigma) \leq \lambda_{\max}(\Sigma) \leq c_2, \mu_0, \mu_1 \in B\}, \tag{1.4}$$

where c, c_1 and c_2 are strictly positive constants, and

$$B = \Big\{ \sum_{j=1}^{\infty} a_j \mu_j^2 \leq K \Big\}, \tag{1.5}$$

for some constant K and $a_j \to \infty$. In words this can be formulated as the covariance matrix being well-behaved while the group means are small on most coordinates, and the difference between the groups is sufficiently large. Now define the a posteriori classification error of a parameter combination $\theta \in \Gamma$ as

$$W(\xi, \theta) = P_\theta(\xi(z) \neq y \mid x).$$

Classifiers are examined in terms of their worst case classification error over Γ, defined by

$$\overline{W}_\Gamma(\xi) \equiv \max_{\theta \in \Gamma} E[W(\xi, \theta)],$$

where the mean value is with respect to the distribution of the training data.

Bickel and Levina (2004) first consider the behaviour of Fishers rule in (1.2), but instead of the maximum likelihood estimates they use estimates $\hat{\mu}_0$ and $\hat{\mu}_1$ inspired by Pinsker's Theorem (see Johnstone (2002) for details). In particular, $\hat{\mu}_{0j}$ and $\hat{\mu}_{1j}$ are zero for large values of j. Since $p > n$, $\hat{\Sigma}$ is singular, and $\hat{\Sigma}^{-1}$ in (1.2) is replaced by the Moore-Penrose inverse

$$\hat{\Sigma}^{-} = \sum_{i=1}^{f} \frac{1}{\lambda_i(\hat{\Sigma})} v_i v_i^{\mathrm{T}},$$

where only the $n-2$ nonzero eigenvalues of $\hat{\Sigma}$ is included, and the $v_i's$ are the corresponding eigenvectors. When $p/n \to \infty$, it is proved that the worst case classification error tends to $1/2$, the classification error of a random guess, so Fishers rule is of no use. This failure, caused by the estimation of too many parameters compared to the amount of data, encourages to consider instead the independence rule from (1.3). By this simplification, the number of parameters in Σ is reduced from $p(p+1)/2$ to p. For

$$K_0 = \max_{\theta \in \Gamma} \frac{\lambda_{\max}(\Sigma_0)}{\lambda_{\min}(\Sigma_0)},$$

it is proved that as long as $\log(p)/n \to 0$,

$$\limsup_{n\to\infty} \overline{W}_\Gamma(\xi_I) = \overline{\Phi}\left(\frac{\sqrt{K_0}}{1+K_0}c\right). \tag{1.6}$$

Note that the right hand side is strictly less than $1/2$ for any $K_0 < \infty$. If the true covariance is the identity, the independence rule is asymptotically as good as Bayes rule.

In a less restrictive setting than (1.4), avoiding the estimation of all the covariance coefficients is not sufficient to assure an informative classification. Fan and Fan (2008) consider the parameter space

$$\Gamma_2 = \left\{(\Delta,\Sigma) : \Delta^{\mathsf{T}} D^{-1}\Delta > c_p, \lambda_{\max}(R) \le c, \min_{j=1,\dots,p} \sigma_j^2 > 0\right\},$$

where σ_j^2 is the variance of the j'th variable. They prove that even when $\log(p)/n \to 0$, the classification error of the independence rule can tend to one half. This occurs when many variables together contribute only little information, such that noise constitutes the dominating part of the classifier. Therefore, unless most variables have large differential expression, variable selection is useful, and may even be necessary to obtain a good classifier in high-dimensional situations

1.3 Variable selection for independence classifiers

Inspired by the result of Fan and Fan (2008) described above, we give a review of methods performing variable selection through thresholding. When theoretical results on the classifer exist we mention these as well.

Note that the independence rule can be written as

$$\xi_I(z) = \mathbf{1}\left\{\sum_{j=1}^{p} \left((z_j - \bar{x}_{0j})^2 - (z_j - \bar{x}_{1j})^2\right)/s_j^2 > 0\right\}, \tag{1.7}$$

where $s_j^2 = \left(\sum_{i=1}^{n_0}(x_{0ij} - \bar{x}_{0j})^2 + \sum_{i=1}^{n_1}(x_{1ij} - \bar{x}_{1j})^2\right)/(n-2)$.

1.3.1 Hard and soft thresholding

In the classifier, we want to incorporate only variables that appear differentially expressed. The obvious way of doing so is to include variables where the t-statistic

$$t_j = \frac{\bar{x}_{1j} - \bar{x}_{0j}}{\sqrt{s_j^2\left(\frac{1}{n_0} + \frac{1}{n_1}\right)}},$$

is large, and exclude all other variables. This can be done by incorporating a weight function into each term in $\xi_I(z)$:

$$\xi_T(z) = \mathbf{1}\left\{(\sum_{j=1}^{p} \left((z_j - \bar{x}_{0j})^2 - (z_j - \bar{x}_{1j})^2\right)/s_j^2 \cdot w(t_j) > 0\right\}. \tag{1.8}$$

Hard thresholding, as studied in Chapter 2, has $w(t) = \mathbf{1}\{|t| > \alpha\}$, meaning that $\xi_T(z)$ is $\xi_I(z)$ with terms having small t-statistics omitted. Soft thresholding uses a continuous weight function instead, often of the form

$$w(t) = \frac{|t| - \alpha}{\theta + |t|}\mathbf{1}\{|t| > \alpha\}. \tag{1.9}$$

Similar to hard thresholding, soft thresholding omits variables with small t-statistics, but concurrently weights the rest of the variables according to their t-statistics.

When the type of threshold has been selected, one needs to determine a good threshold value α. For simulated data, the exact classification error can be calculated from the true parameter values, for a range of threshold values. From these values, an optimal α can be found. For real data the optimal threshold is often estimated by cross-validation (CV) instead. Section 1.3.3 and Section 1.3.4 suggest two alternative methods for selecting the threshold.

Simulations in Jensen (2006) show that thresholding decreases the mean of the classification error, and at the same time reduces a negative correlation existing between $P\big(\xi_I(z) = 1 \,|\, y = 1\big)$ and $P\big(\xi_I(z) = 0 \,|\, y = 0\big)$. It is seen that the use of a data-dependent threshold chosen by cross-validation often improves classification compared to a threshold fixed a priori. In most situations, soft thresholding decreases the classification error more than hard thresholding, but at the cost of including more variables, and thereby increasing the number of variables to be measured on new observations when using the classifier.

A final note on thresholding is that in situations where certain variables are expected to be strongly correlated, for instance genes within the same pathway in microarray analysis, one may include the whole group of variables as soon as one of them is selected.

1.3.2 Nearest shrunken centroids

Tibshirani et al. (2003) suggest a threshold method called *nearest shrunken centroids* (NSC). The idea is to replace the maximum likelihood estimates of the mean values by adjusted means. Define

$$\tilde{x}_{0j} = \begin{cases} a_j & \text{if } |t_j| < \alpha, \\ \bar{x}_{0j} + \operatorname{sgn}(\bar{x}_{1j} - \bar{x}_{0j})\alpha s_j\sqrt{\frac{1}{n_0} - \frac{1}{n}} & \text{if } |t_j| \geq \alpha. \end{cases} \tag{1.10}$$

$$\tilde{x}_{1j} = \begin{cases} a_j & \text{if } |t_j| < \alpha, \\ \bar{x}_{1j} - \operatorname{sgn}(\bar{x}_{1j} - \bar{x}_{0j})\alpha s_j\sqrt{\frac{1}{n_1} - \frac{1}{n}} & \text{if } |t_j| \geq \alpha. \end{cases}$$

The classifier $\xi_{\mathrm{NSC}}(z)$ is defined as (1.7) with $\bar{x}_{kj}$ replaced by $\tilde{x}_{kj}$. According to Jensen (2006), when $n_0 = n_1$, $\xi_{\mathrm{NSC}}(z)$ equals (1.8) with soft thresholding and $\theta = 0$ in (1.9). In Section 1.7 an extension of NSC is described.

1.3.3 Features annealed independence rule

Fan and Fan (2008) consider two normally distributed groups with $\Sigma = I_p$, and $p \to \infty$ sufficiently slow, as $n \to \infty$. A theoretical analysis of the independence rule leads to the suggestion of a new method to select its threshold. When incorporating this threshold, the resulting classifier is called the *Features Annealed Independence Rule* (FAIR). The background and results of FAIR are described in this section.

First note that with known $\Sigma = I_p$, the independence rule can be written as:

$$\xi_I(z) = \mathbf{1}\{\hat{\Delta}^{\mathrm{T}}(z - \tfrac{1}{2}(\bar{x}_0 + \bar{x}_1)) > 0\}.$$

When assuming that the variables are ordered with respect to their absolute differential expression, an intuitive way of simplifying the independence rule is by including only the first m variables, for some appropriate choice of $m < p$. This leads to the following classifier:

$$\xi_I^m(z) = \mathbf{1}\{(\hat{\Delta}^m)^{\mathrm{T}}(z^m - \tfrac{1}{2}(\bar{x}_0^m + \bar{x}_1^m)) > 0\}, \tag{1.11}$$

where z^m is a vector consisting of the first m coordinates of z only, and $\bar{x}_0^m$, $\bar{x}_1^m$ and $\hat{\Delta}^m$ are defined in a similar way. In the asymptotic setting of Fan and Fan (2008), m is assumed to tend to infinity with n. For specific values of p and n, the optimal value of m is wanted.

Under the assumption $\frac{n}{\sqrt{m}} \sum_{j=1}^m \Delta_j^2 \to \infty$ as $n \to \infty$, it is proved that

$$W(\xi_I^m, \theta) = \overline{\Phi}\left(\frac{(1 + o(1)) \sum_{j=1}^m \Delta_j^2 + m \frac{n_0 - n_1}{n_1 n_0}}{2\sqrt{(1 + o(1)) \sum_{j=1}^m \Delta_j^2 + \frac{nm}{n_1 n_0}}} \right), \tag{1.12}$$

where $o(1)$ denotes a term tending to zero in probability as $n \to \infty$. Ignoring small terms, $W(\xi_I^m, \theta)$ is minimized as a function of m in

$$m_0 = \underset{1 \le m \le p}{\arg\max} \frac{\left(\sum_{j=1}^m \Delta_j^2 + m \frac{n_0 - n_1}{n_0 n_1} \right)^2}{\sum_{j=1}^m \Delta_j^2 + \frac{nm}{n_0 n_1}}. \tag{1.13}$$

If the positions of most importance in Δ are known a priori, but are not the first ones, the classifier

$$\xi_{\mathrm{orc}}^{\alpha}(z) = \mathbf{1}\{\sum_{j=1}^p \hat{\Delta}_j (z_j - \tfrac{1}{2}(\bar{x}_{0j} + \bar{x}_{1j}))\mathbf{1}\{|\Delta_j| > \alpha\} > 0\},$$

can be used instead of the truncated classifier in (1.11). Here α is chosen to be smaller than exactly m of the $|\Delta_j|$'s. The result in (1.12) obviously remains valid if the sum over $\{1, \ldots, m\}$ is replaced by summing over $\{j : |\Delta_j| > \alpha\}$.

In data scenarios, oracle information on the positions of important variables are not available. Therefore, for a threshold α_n, FAIR is defined by substituting $\hat{\Delta}_j$ for Δ_j also in the thresholding process, that is

$$\xi_{\mathrm{FAIR}}^{\alpha_n}(z) = \mathbf{1}\{\sum_{j=1}^p \hat{\Delta}_j (z_j - \tfrac{1}{2}(\bar{x}_{0j} + \bar{x}_{1j}))\mathbf{1}\{|\hat{\Delta}_j| > \alpha\} > 0\}.$$

The defining point of FAIR is now to estimate $\hat{m}_0$ from (1.13) by plugging in estimates of Δ_j, and subsequently selecting a threshold α_n, such that exactly $\hat{m}_0$ terms are included in $\xi_{\mathrm{FAIR}}^{\alpha}(z)$.

Next, the behaviour of FAIR is considered theoretically. Define $\mathcal{A} = \{j : |\Delta_j| > \alpha_n\}$ with $|\mathcal{A}| = m$ and assume:

(i) $\max_{j \in \mathcal{A}^c} |\Delta_j| < \alpha_n$,

(ii) $\frac{\log(p-m)}{n(\alpha_n - \max_{j \in \mathcal{A}^c} |\Delta_j|)^2} \to 0$ for $n \to \infty$,

(iii) $\frac{n}{\sqrt{m}} \sum_{j \in \mathcal{A}} \Delta_j^2 \to \infty$ for $n \to \infty$,

(iv) $\frac{\sum_{j\in A}|\Delta_j|}{\sqrt{n}\sum_{j\in\mathcal{A}}\Delta_j^2} \to 0$ for $n \to \infty$.

Assumption (i) means that variables omitted in the oracle classifier are not allowed to have differential expression above the threshold, and (ii) that these are not allowed to get too close to the threshold either. Furthermore, (ii) restricts the growth rate of p. A lower bound on the sum of differential expressions of the oracle set is given by (iii) and (iv).

Under (i)–(iv), the following upper bound of the classification error is found:

$$W(\xi_{\text{FAIR}}^{\alpha_n}, \theta) \leq \overline{\Phi}\left(\frac{(1+o(p))\sum_{j\in\mathcal{A}}\Delta_j^2 + \frac{mn}{n_1 n_0} - m\alpha_n^2}{2\sqrt{(1+o(p))\sum_{j\in\mathcal{A}}\Delta_j^2 + \frac{nm}{n_1 n_0}}}\right). \tag{1.14}$$

Note (1.14) differs from (1.12) only by the term $m\alpha_n^2$. Not surprisingly, this means that the upper bound of FAIR is increased compared to (1.12), since we now need to estimate $\mathcal{A}$ in addition to the Δ_j's.

FAIR is generalized to a situation with any covariance matrix Σ by using the t-statistics in thresholding:

$$\xi_{\text{FAIR}}^{\alpha_n}(z) = \sum_{j=1}^{p}\hat{\Delta}_j\frac{(z_j - \frac{1}{2}(\bar{x}_{0j} + \bar{x}_{1j}))}{s_j^2}\mathbf{1}\{\sqrt{n/(n_0 n_1)}|t_j| > \alpha_n\}.$$

Fan and Fan (2008) state that after ordering the variables by decreasing values of $\hat{\Delta}_j$, the optimal number of variables to use in classification is:

$$\hat{m}_1 = \underset{1\leq m\leq p}{\arg\max}\frac{1}{\lambda_{\max}(\hat{\Sigma}_0^m)}\frac{n\left(\sum_{j=1}^m t_j^2 + \frac{m(n_0-n_1)}{n}\right)^2}{mn_0n_1 + n_0n_1\sum_{j=1}^m t_j^2}.$$

Here $\lambda_{\max}(\hat{\Sigma}_0^m)$ is the maximum eigenvalue of the estimated correlation matrix for the m included variables. Since $1/\lambda_{\max}(\hat{\Sigma}_0^m)$ is decreasing in m, $\hat{m}_1$ is usually smaller than $\hat{m}_0$. No proof is given that $\hat{m}_1$ is actually the optimal number of variables to include, and no boundary on the classification error is found for a general Σ either.

To justify the application of FAIR when the variances are unknown a priori, Fan and Fan (2008) prove that the differential expressed variables can be separated from the non-expressed ones by the t-statistics, asymptotically. This occurs when some technical assumptions are met, including that Δ is sparse with only the first s_n elements being non-zero. These non-zero values are not allowed to decrease too fast, and the threshold α_n grows slowly. Formally, the result states

$$P\left(\min_{j\leq s}|t_j| > \alpha_n, \max_{j>s}|t_j| < \alpha_n\right) \to 1 \qquad \text{as } n \to \infty.$$

This is proved by showing that the probabilities of the complementary events, that is $\{\min_{j>s}|t_j| > \alpha_n\}$ and $\{\max_{j\leq s}|t_j| < \alpha_n\}$, tend to zero sufficiently fast by using normal tail probability inequalities. The result is closely related to Lemma 2.5 in Chapter 2 of this thesis, though none of the two results directly imply the other.

Simulations show that FAIR has a lower classification error than NSC, and also that FAIR tends to include a more stable number of variables across simulations. The last point leads to a smaller variation of the classification error of FAIR.

1.3.4 Classification by higher criticism thresholding

Higher criticism (HC) was first introduced by Donoho and Jin (2004) as a test for the global null hypothesis $\bigcap_{j=1}^{p}\{\Delta_j = 0\}$. The phrase "higher criticism" emphasizes that a decision is assessed on p test statistics simultaneously, and not by testing each of the hypotheses separately.

In this section, we first state the original, simple formulation of higher criticism, and describe how this method can be applied as a variable selection method in classification problems. Afterwards, we introduce the asymptotic setting of Donoho and Jin (2009), and give a more general definition of the HC threshold. This more flexible definition enables us to define the ideal HC threshold which can be compared to the ideal threshold. We state theoretical results from the paper Donoho and Jin (2009), which examines when a successful classification is possible for the ideal threshold as well as the ideal HC threshold.

Let $p_{(i)}$ denote the i'th ordered P-value when testing each individual hypothesis of $\Delta_j = 0$ with a two-sided alternative. The HC score for variable i is defined as

$$HC(i, p_{(i)}) = \frac{\frac{i}{p} - p_{(i)}}{\sqrt{\frac{i}{p}(1 - \frac{i}{p})}}. \tag{1.15}$$

The motivation behind these scores is that when the global null hypothesis holds, one has $p \cdot p_{(i)} \sim \text{binomial}(p, i/p)$, and thus large values of the HC scores indicate that some variables are differentially expressed. Define further

$$\hat{i} = \underset{1 \leq i \leq \alpha_0 p}{\arg\max}\, HC(i, p_{(i)}), \tag{1.16}$$

where α_0 is some fixed constant between 0 and 1, with 0.1 as a typical choice.

Donoho and Jin (2008) and Donoho and Jin (2009) use HC for variable selection in binary classification, by including only variables with $p_i \leq p_{(\hat{i})}$ in the classifier. Their model consists of two normally distributed groups with known covariance matrix $\Sigma = I_p$. Most coordinates of Δ are zero, apart from a small fraction, ϵ, where the value is instead Δ_0. For simplicity, $n_0 = n_1$ and $(\mu_0 + \mu_1) = 0_p$, such that the latter does not need to be estimated. Define $\tilde{t} = (\bar{x}_1 - \bar{x}_0)/\sqrt{n}$ as a vector of test statistics for the hypotheses $\Delta_j = 0$, for $i = 1, \dots, p$. The two-sided P-values for $H_j : \Delta_j = 0$ are calculated from the $\tilde{t}$-statistics. These P-values are applied for calculation of the HC scores in (1.15), and finally the HC threshold is found as $\hat{\alpha}^{\text{HC}} = |\tilde{t}_{(\hat{i})}|$.

Classifiers of the form $L(z) = \sum_{j=1}^{p} w_{\alpha,j}(\tilde{t}) z_j$, which classifies to group 1 when $L > 0$, and to group 0 otherwise, is of interest. The aim is to find a weight function $w_\alpha(\tilde{t})$ which results in a low classification error. To this end, three thresholding functions, all applied to the threshold $\hat{\alpha}^{\text{HC}}$, are considered:

- hard thresholding, where $w_\alpha(\tilde{t}) = \tilde{t}\mathbf{1}\{|\tilde{t}| > \alpha\}$,
- soft thresholding, where $w_\alpha(\tilde{t}) = \text{sgn}(\tilde{t})(|\tilde{t}| - \alpha)_+$,
- clip thresholding, where $w_\alpha(\tilde{t}) = \text{sgn}(\tilde{t})\mathbf{1}\{|\tilde{t}| > \alpha\}$.

Here x_+ denotes the positive part of x.

We now turn to the asymptotic model of Donoho and Jin (2009), where p is the driving parameter, and $n \sim c(\log(p))^\zeta$ for positive constants c and ζ. The differential expressed variables become rare as p increases, described through the fraction of nonzero variables $\epsilon_p = p^{-\beta}$ for some $\beta \in (0, 1)$. The amount of differential expression is described through $\tau_p \equiv \sqrt{n}\Delta_0 = \sqrt{2r\log(p)}$ for some $r \in (0, 1)$. When $\zeta > 1$, the value of each differentially expressed variable decreases with p.

For notational simplicity, we often omit the subscript p of ϵ and τ. As described in the following, the values of the parameters β and r determine whether a successful classification is possible.

To inspect this setting in detail, a more general definition of the HC than given through (1.15) is fruitful. To this end, define the HC functional

$$A_{\mathrm{HC}}(F) = \arg\max_{\alpha > \alpha_0} \frac{\overline{F}_0(\alpha) - \Phi_0(\alpha)}{\sqrt{F_0(\alpha)\overline{F}_0(\alpha)}}.$$

Here, F is a distribution function not equal to Φ, while $\Phi_0(\alpha) = \Phi(\alpha) - \Phi(-\alpha)$, $F_0(\alpha) = F(\alpha) - F(-\alpha)$, and $\overline{F}_0(\alpha) = 1 - F_0(\alpha)$. The *ideal HC threshold* is defined as $A_{\mathrm{HC}}(F)$ when $F = E[F_p]$, where F_p denotes the empirical distribution function for the $\tilde{t}$-values. For comparison, remark that the HC procedure defined from (1.16) leads to the threshold $A_{\mathrm{HC}}(F_p)$.

The behaviour of the ideal HC threshold is compared to the ideal threshold to be defined next, and the two turn out to have similar properties. Recall that the classification error of a classifier defined by the linear vector w, since $\Sigma = I_p$, is a function of $\gamma = w^{\mathrm{T}}\Delta/\sqrt{w^{\mathrm{T}}w}$. The error is minimal when the vector of weights w is given by Δ. Here, we instead search for the threshold value that minimizes the classification error when applied to a given classification vector w. To this end, an approximation of γ is

$$\tilde{\gamma}(\alpha, \epsilon, \tau) = \frac{A}{\sqrt{B}},$$

where for $U \sim N(0,1)$, A and B are defined as

$$A(\alpha, \epsilon, \tau) = \epsilon\tau E[w_\alpha(\tau + U)],$$
$$B(\alpha, \epsilon, \tau) = \epsilon E[w_\alpha(\tau + U)] + (1-\epsilon)E[w_\alpha^2(W)].$$

The ideal threshold is now defined as

$$A_{\mathrm{ideal}}(\epsilon, \tau) = \arg\max_{\alpha} \tilde{\gamma}(\alpha, \epsilon, \tau).$$

The main result of Donoho and Jin (2009) states that the behaviour of the ideal threshold and the ideal HC threshold when applying either hard, soft or clip thresholding are in many aspects the same. Common properties are described below. On the other hand, it is shown that such similarities do not hold between the ideal threshold and a threshold found through ideal procedures controlling either the familywise error rate or the false discovery rate (FDR), that is, either the probability of declaring a non-expressed variable as differential expressed, or the fraction of declared expressed variables that are truly non-expressed (Benjamini and Hochberg, 1995). Thus HC thresholding appears superior to those methods.

Of particular interest regarding ideal thresholding and ideal HC thresholding, the parameter space with respect to r and β is divided into exactly two phases: A phase of success, where the classification error tends to zero, as p tends to infinity, and a phase of failure, where this never occurs. Actually, Jin (2009) shows for a more general relationship between p and n that the classification error in any point of the phase of failure tends to $1/2$, as p tends to infinity. The phase of success for both ideal thresholding and ideal HC thresholding is determined by $r > \rho(\beta)$ where

$$\rho(\beta) = \begin{cases} 0 & 0 < \beta \leq 1/2, \\ \beta - 1/2 & 1/2 < \beta \leq 3/4, \\ (1 - \sqrt{1-\beta})^2 & 3/4 < \beta < 1. \end{cases}$$

An interesting aspect to consider is the value of the ideal threshold when a successful classification is possible. To this end, divide the phase of success into three regions:

(I) $0 < r \leq \beta/3$ and $1/2 < \beta \leq 3/4$; $r > \rho(\beta)$,

(II) $\beta/3 < r \leq \beta$ and $1/2 < \beta \leq 1$; $r > \rho(\beta)$,

(III) $\beta < r \leq 1$ and $1/2 < \beta \leq 1$; $r > \rho(\beta)$.

Recall that the differentially expressed variables have strength $\sqrt{2r\log(p)}$. The ideal threshold as well as the ideal HC threshold have a related form in the phase of success, namely $A_{\text{HC}}(F) = A_{\text{ideal}}(\epsilon, \tau) = \sqrt{2q\log(p)}$ where

$$q = \begin{cases} 4r & \text{region (I)}, \\ (\beta + r)^2/(4r) & \text{region (II), (III)}. \end{cases}$$

In region (I) and (II), the ideal threshold is thus larger than the signals to be detected, which demonstrates the importance of excluding most irrelevant variables, even though it means a majority of the important variables are excluded as well. This is illustrated by the fact that the FDR, which has similar behaviour for ideal thresholding and ideal HC thresholding, is large in those regions: In region (I), the FDR tends to 1 as $p \to \infty$, while in region (II), it is asymptotically strictly between $1/2$ and 1. In region (III), on the other hand, where the expressions are a bit stronger, the optimal threshold is smaller than the signals. Asymptotically, the FDR is $1/2$ in this region. We note that this setting is fundamentally different from the asymptotic settings of Fan and Fan (2008) and Bak et al. (2015), where complete separation between expressed and non-expressed variables is possible. It is surprising that even though most selected variables are false positives, and most expressed variables remain undetected, an asymptotically perfect classification can be performed in region (I).

Obviously, the ideal HC threshold cannot be calculated in practice, since one lacks knowledge of ϵ and τ. As pointed out by Donoho and Jin (2015), the results for the ideal HC threshold also holds for the HC threshold based on the data, when ϵ and τ are not known. Simulations in Donoho and Jin (2008) illustrate that the asymptotic behaviour of the ideal HC classifier is also reflected in realistic finite sample size situations.

Apart from the attractive asymptotic behaviour, a huge advantage of HC thresholding is that no cross-validation step is needed for selecting the threshold. Furthermore, the variance of the threshold itself is smaller than when it is chosen through cross-validation, according to Donoho and Jin (2008).

1.4 Including correlation in classification

In this section, it is illustrated that including correlation can be fundamental for obtaining a better classification. Therefore, we review procedures that make use of correlation in classification, with a particular focus on methods which estimate the covariance matrix consistently in restricted settings.

As described in Section 1.2, Fishers rule cannot be expected to perform well when $p > n$, and Σ^{-1} is estimated by a generalized inverse. Correlation can be helpful in classification, though: When the true parameters are known, we have

$$W(\xi_B) = \overline{\Phi}\big((\Delta^{\mathrm{T}}\Sigma^{-1}\Delta)^{1/2}/2\big) \leq \overline{\Phi}\bigg(\frac{\Delta^{\mathrm{T}}D^{-1}\Delta}{2(\Delta^{\mathrm{T}}D^{-1}\Sigma D^{-1}\Delta)^{1/2}}\bigg),$$

where the inequality follows from Corollary A.9.2.2 in Mardia et al. (1979), and the right hand side is the classification error of the oracle independence rule. The difference between Bayes rule and the independence rule can be substantial as illustrated through this small example from Fan et al. (2012): Set $p = 2$, and consider

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

for some $\rho \in (0,1)$. Bayes risk can be written as $\overline{\Phi}\big((\Gamma_1(\rho))^{1/2}/2\big)$ where

$$\Gamma_1(\rho) = \frac{1}{1-\rho^2}(\Delta_1^2 + \Delta_2^2 - 2\rho\Delta_1\Delta_2).$$

It is easy to see that $\Gamma_1(\rho) \to \infty$, as $\rho \to 1$. On the other hand, the classification error of the oracle independence rule is $\overline{\Phi}\big((\Gamma_2(\rho))^{1/2}/2\big)$ where

$$\Gamma_2(\rho) = \frac{(\Delta_1^2 + \Delta_2^2)^2}{\Delta_1^2 + \Delta_2^2 + 2\rho\Delta_1\Delta_2}.$$

Note that $\Gamma_2(\rho)$ tends to $(\Delta_1^2 + \Delta_2^2)^2/(\Delta_1 + \Delta_2)^2$, a fixed constant, as $\rho \to 1$, and thus the independence rule is considerably worse than Bayes rule when ρ is large. Thus, if we can find a reasonable, or maybe even consistent, estimate of Σ^{-1} in sample situations, we can hope for a better classification than when applying the independence rule to correlated data.

1.4.1 Nearest shrunken centroids regularized discriminant analysis

Inspired by NSC, Guo et al. (2005) introduce *shrunken centroids regularized discriminant analysis* (SRRDA) as

$$\mathrm{SCRDA}(z) = 1\big((\tilde{x}_1^* - \tilde{x}_0^*)^{\mathrm{T}}\tilde{\Sigma}^{-1}(z - \tfrac{1}{2}(\tilde{x}_0^* + \tilde{x}_1^*)) > 0\big),$$

where

$$\tilde{\Sigma} = \beta\hat{\Sigma} + (1-\beta)I_p \quad \text{and} \quad \tilde{x}_k^* = \mathrm{sgn}(\tilde{\Sigma}^{-1}\bar{x}_k)(|\tilde{\Sigma}^{-1}\bar{x}_k| - \alpha)_+,$$

with $\beta \in (0,1)$ and $\alpha \geq 0$. SCRDA incorporates two innovations compared to the thresholded independence rule. First, SCRDA includes correlation between variables by estimating the covariance matrix as an intermediate between the ones applied in the independence rule and Fishers rule. Second, SCRDA thresholds $\hat{\Sigma}^{-1}(\bar{x}_1 - \bar{x}_0)$ instead of the t-statistic.

One drawback of SCRDA is that it involves two parameters, α and β, to be estimated by cross-validation. If the amount of data is limited, this might cause lack of stability, meaning that small deviations in the data can lead to rather different selected parameters. Furthermore, though a nonsingular estimate of Σ is suggested, we cannot expect neither $\hat{\Sigma}$ nor $\hat{\Sigma}^{-1}$ to be close to Σ and Σ^{-1}, respectively.

1.4.2 Estimation of covariance matrices

Though the correlation matrix is usually unknown, one can have prior knowledge of its structure from the scientific situation at hand. As an example, in time series, and to some extend in gene data, it can be reasonable to assume that variables are only short-range dependent, which means $\Sigma_{ij} = 0$ for $|i-j|$ sufficiently large. After realizing that no well-behaved estimate of Σ exists in general high-dimensional situations, a considerable amount of work on improving estimates of the covariance matrix in such more specific situations has been performed within the last decade. We highlight a few of these suggestions. Replacing $\hat{\Sigma}^{-}$ in Fishers rule by such estimates hopefully improves classification in specific settings. This is indeed the case in Shao et al. (2011), described below.

Besides its relevance in classification, estimation of covariance matrices is also of significant importance in principal component analysis, graphical modelling, and when establishing confidence intervals for linear functions of variables.

Evaluating consistency of an estimator of Σ in terms of coordinate wise convergence is inappropriate when p is increasing with n: Even if each coordinate converges, the full matrix can diverge dramatically. Instead, we consider consistency of a symmetric matrix M through convergence in either operator norm, defined by

$$\|M\| = \sup_{\|v\|_2=1} \|Mv\|_2 = max_{1\leq j\leq p}|\lambda_j(M)|,$$

or in Frobenius norm

$$\|M\|_F^2 = \sum_{i,j} M_{ij}^2 = \mathrm{trace}(MM^{\mathrm{T}}).$$

Convergence in operator norm is linked to convergence of eigenvalues, and is thus particularly relevant when the target of the covariance estimation is principal component analysis.

Bickel and Levina (2004) consider a scenario where $\Sigma_{ij} = \sigma(|i-j|)$, and $\sigma(k)$ takes small values when k is large. This justifies the following estimate of the covariance

$$\hat{\Sigma}_{ij}^{*} = \begin{cases} \hat{\sigma}^{*}(|i-j|) & |i-j| \leq d, \\ 0 & \text{otherwise}, \end{cases} \tag{1.17}$$

where

$$\hat{\sigma}^{*}(k) = \frac{1}{p-k} \sum_{a=1}^{p-k} \hat{\Sigma}_{a,a+k}.$$

Compared to the general case, we have much more information available for the estimation of each entry of Σ. While the number of parameters to be estimated possibly increases with p, the information available for each variable to be estimated does as well. Under further conditions, this classifier has attractive asymptotic properties, see Theorem 2 of Bickel and Levina (2004).

In Bickel and Levina (2008a), a more general estimator of Σ under short range dependence is suggested as

$$\hat{\Sigma}_{ij}^{**} = \hat{\Sigma}_{ij}\mathbf{1}\{|i-j| \leq d\}.$$

This method is called banding. Similarly to (1.17) the known structure of the covariance matrix is applied to reduce the noise from estimating too many correlation coefficients, while we have less information for each entry of $\hat{\Sigma}^{**}$ compared to $\hat{\Sigma}^{*}$. Notice that $\hat{\Sigma}^{**}$ is not automatically invertible, which excludes the possibility of direct insertion in Fishers rule. This problem can be solved by replacing $\mathbf{1}\{|i-j| \leq d\}$ with more cleverly selected r_{ij} fulfilling certain conditions, including that r_{ij} is decreasing as the distance between i and j increases. It is proved that $\hat{\Sigma}^{**}$ based on either $\mathbf{1}\{|i-j|\}$ or r_{ij} converge in operator norm under certain conditions, including the assumption of $\log(p)/n \to 0$. Not surprisingly, a fundamental assumption is that distant variables are almost uncorrelated.

Related procedures and results are presented when regularizing the estimate of the inverse of the covariance matrix based on the Cholesky decomposition of Σ in a similar way. Besides classification, this is of particular usage in graphical modelling, where a zero in the inverse covariance matrix corresponds to no causal effect between the associated variables (Lauritzen, 1996, Proposition 5.2).

Bickel and Levina (2008b) consider a similar setting, where most covariances are close to zero while the positions of the nonzero covariances are unknown. Inspired by thresholding in the estimation of Δ, they suggest:

$$\hat{\Sigma}^{***}_{ij} = \hat{\Sigma}_{ij}\mathbf{1}\{|\hat{\Sigma}_{ij}| > \alpha'\}.$$

It is proved that $\hat{\Sigma}^{***}$ estimate Σ well with respect to both operator and Frobenius norm, when the true Σ is sparse. This thresholding estimator can be applied also when the assumptions in Bickel and Levina (2008a) hold, but since $\hat{\Sigma}^{**}$ takes more advantage of the true covariance structure in such situations, it outperforms $\hat{\Sigma}^{***}$ asymptotically.

Shao et al. (2011) consider applying $\hat{\Sigma}^{***}$ in classification. This leads to the proposal of *sparse linear discriminant analysis* (SLDA), where both $\hat{\Delta}$ and $\hat{\Sigma}$ in Fishers rule are replaced by their thresholding estimates. When the differential expressions are sufficiently large, $\log(p)/n \to 0$, and certain sparsity restrictions hold for Σ and Δ, it is proved that SLDA is asymptotically optimal, in the sense that its classification error converges to Bayes risk with the optimal rate. Simulations show that SLDA is significantly better than SCRDA in terms of classification error.

In a setting related to Bickel and Levina (2008a), Rothman et al. (2008) estimate the inverse of the covariance matrix, Ω, directly through ℓ_1 penalization in the *Sparse Permutation Invariant Covariance Estimator* (SPICE) defined as

$$\hat{\Omega} = \underset{\Omega \succ > 0}{\arg\min}\{\text{trace}(\Omega\hat{\Sigma}) - \log|\Omega| + \lambda|\Omega_-|_1\},$$

where Ω_- is the negative part of Ω. Similar rates of convergence of the Frobenius and operator norm as in Bickel and Levina (2008a) are obtained. An efficient algorithm for the calculation is suggested, based on the Cholesky decomposition of the inverse estimated covariance matrix combined with a coordinate descent approach.

1.5 Correlated classifiers avoiding the estimation of the inverse covariance

Attractive estimates of the covariance matrix in high-dimensional settings only appear to exist under rather tight restrictions, so we cannot expect good results in general when building a classifier by plugging estimates of each parameter directly into (1.1). To imitate Bayes rule, one do not need good estimates of neither Σ^{-1} nor Δ, though, as long as a good estimate of $w_{\text{Bayes}} = \Sigma^{-1}\Delta$ exists. This section reviews methods that use correlation in classification without estimating the inverse covariance matrix.

Due to the results of Fan and Fan (2008), a sparse estimate of w_{Bayes} is wanted. While sparsity in $\Sigma^{-1}\Delta$ is harder to interpret than sparsity in Δ, it does reduce the amount of noise in the classifier. Following Mai et al. (2012), sparsity in neither $\Sigma^{-1}\Delta$ nor Δ imply the other.

We thus consider linear classifiers of the form

$$\xi_w(z) = \mathbf{1}\{w^{\mathrm{T}}(z - \tfrac{1}{2}(\mu_0 + \mu_1)) > 0\},$$

and want to estimate w directly. This line of research has been developed within the last few years. In data situations, μ_0 and μ_1 are estimated by their sample averages.

Note that the classification error of an observation from group 0 for a linear classifier as above is given by

$$W(\delta_w) = \overline{\Phi}\bigg(\frac{w^{\mathrm{T}}\Delta}{2(w^{\mathrm{T}}\Sigma w)^{1/2}}\bigg). \tag{1.18}$$

When considering the sample version of a linear classifier, an extra term from the estimation of Δ is added in this expression.

All procedures below are computational more complex than methods described previously, and their computation often involve convex optimization techniques. Unless the number of variables is extremely large, this should not be a serious concern though. In situations where the computational burden is considered too large, the procedures can be combined with an initial variable screening, as described in Fan et al. (2012).

In the remaining of this chapter we use the notation that $|v|_p^p = \sum_i |v_i|^p$, including $|v|_0 = \#\{i : v_i \neq 0\}$.

1.5.1 Regularized optimal affine discriminant

The method broadly known as *Regularized Optimal Affine Discriminant* (ROAD) has been proposed independently by Fan et al. (2012) and Wu et al. (2011). The starting point in Fan et al. (2012) is to minimize the classification error in (1.18), that is to find

$$\operatorname*{arg\,min}_{\Delta^{\mathrm{T}} w = 1} w^{\mathrm{T}}\Sigma w. \tag{1.19}$$

When inserting estimates of the parameters, the solution to this problem is not unique in high-dimensional situations. To assure uniqueness as well as some kind of sparsity of the solution, an ℓ_1 penalty is added, and the ROAD classification vector is defined as

$$w_{\mathrm{ROAD}}(c) = \operatorname*{arg\,min}_{\Delta^{\mathrm{T}} w = 1, |w|_1 \leq c} w^{\mathrm{T}}\Sigma w. \tag{1.20}$$

Wu et al. (2011) start from Fishers rule, which was originally proposed as the classifier which maximizes

$$\operatorname*{arg\,max}_{w} \frac{w^{\mathrm{T}}\hat{\Sigma}_{\mathrm{between}} w}{w^{\mathrm{T}}\hat{\Sigma}_{\mathrm{within}} w}.$$

When considering classification between only two groups, this reduces to the sample version of (1.19), and the ℓ_1 penalty is further introduced.

Both Fan et al. (2012) and Wu et al. (2011) suggest algorithms to solve (1.20). Wu et al. (2011) solve the problem directly through an algorithm closely related to the LARS-algorithm, also known from estimation of the LASSO (see Section 1.6). By convex theory, Fan et al. (2012) instead reformulate (1.20) to

$$w_{\mathrm{ROAD}}(\lambda) = \operatorname*{arg\,min}_{w^{\mathrm{T}}\Delta = 1} w^{\mathrm{T}}\Sigma w + \lambda |w|_1$$

which is further approximated by

$$w_{\text{ROAD}}(\lambda, \gamma) = \underset{w^{\text{T}}\Delta=1}{\arg\min}\, w^{\text{T}}\Sigma w + \lambda|w|_1 + \tfrac{1}{2}\gamma(w^{\text{T}}\Delta - 1)^2.$$

When $\gamma \to \infty$, it is proved that $w_{\text{ROAD}}(\lambda, \gamma) \to w_{\text{ROAD}}(\lambda)$. Simulations show that the value of γ is not crucial for the performance of the resulting classifier, as long as a corresponding $\lambda(\gamma)$ is selected by cross-validation. Therefore, introducing the extra parameter does not increase the computational burden.

Mai and Zou (2012) prove that, after appropriate scaling and adjustment of the regularization parameters, ROAD is equivalent to *Direct Sparse Discriminant Analysis* (DSDA) and *Sparse Optimal Scoring* (SOS), which are two recently suggested procedures avoiding the estimation of the full covariance matrix. DSDA was first introduced by Mai et al. (2012), and can be formulated as

$$w_{\text{DSDA}}(\lambda) = \arg\min_{w}\{w^{\text{T}}(\Sigma + \tfrac{1}{4}\Delta\Delta^{\text{T}})w - \tfrac{1}{2}w^{\text{T}}\Delta + \lambda|w|_1\}.$$

SOS was proposed in Clemmensen et al. (2011), and is only formulated in a sample version. Let X be the $n \times p$ mean centred data matrix, and $\tilde{Y}$ be a $n \times 2$ matrix with $\tilde{Y}_{i1} = 1$ if i belongs to group 0, $\tilde{Y}_{i2} = 1$, if i belongs to group 1, and all other coordinates equal to zero. SOS consider the optimization problem

$$\min_{w,\theta}\{|\tilde{Y}\theta - Xw|_2^2 + \lambda|w|_1\}$$
$$\text{subject to } \tfrac{1}{n}\theta^{\text{T}}\tilde{Y}^{\text{T}}\tilde{Y}\theta = 1, \theta\tilde{Y}^{\text{T}}\tilde{Y}1_2,$$

and define $w_{\text{SOS}}(\lambda)$ as the minimizing w hereof.

The equivalence of the sample versions of w_{ROAD}, w_{DSDA} and w_{SOS} means that theoretical properties proved for one classifier automatically holds for the others. We now summarize such properties.

Under some sparsity assumption on $w_{\text{ROAD}}(c)$ in (1.20), Fan et al. (2012) prove that the empirical ROAD obtains a classification error close to the oracle ROAD. If w_{Bayes} is sparse, $w_{\text{ROAD}}(c)$ is further proved to approach w_{Bayes} in ℓ_2 norm. See Chapter 4 and Chapter 5 for further details on these issues.

Kolar and Liu (2015) prove that the oracle version of w_{DSDA} detects exactly the true nonzero coefficients of w_{Bayes} under certain conditions, and up to scaling estimates these nonzero coefficients as if the support was known a priori. When n is sufficiently large compared to p and the sparsity $|w_{\text{Bayes}}|_0$, and the nonzero coefficients of w_{Bayes} are not too small, a related selection and estimation result is proved for the sample version of DSDA. It is briefly sketched, that the results can be extended to situations where w_{Bayes} is only approximately sparse, and one requires only detection of sufficiently large coefficients hereof.

1.5.2 Linear programming discriminant

Cai and Liu (2011) find their starting point in $\Sigma w_{\text{Bayes}} = \Delta$, and search for a sparse classification vector which almost fulfills this equality coordinatewise. This leads to the *Linear Programming Discriminant* (LPD) defined by

$$w_{\text{LPD}} = \underset{|\Sigma w - \Delta|_\infty \le \lambda}{\arg\min}\ |w|_1.$$

The name originates from the optimization problem being solved through a linear program, similar to the one used for the Dantzig Selector in Candes and Tao (2007). It is proved that the sample version of LPD, denoted by $\hat{w}_{\text{LPD}}$, obtains a classification error approaching Bayes risk under certain sparsity conditions on w_{Bayes} when $\log(p)/n \to \infty$.

Wang et al. (2013) prove under further restrictions that when $|w_{\text{Bayes}}|_0 = s$, the s largest coefficients of $\hat{w}_{\text{LPD}}$ are with high probability the nonzero coefficients of w_{Bayes}. This leads to the proposal of a two-stage procedure: First, select the important variables as those with the largest coefficients in $\hat{w}_{\text{LPD}}$. Next, calculate Fishers rule based only on the selected variables. Calculating Fishers rule only from a subset of variables is a straightforward approach, but previously only procedures ignoring correlation have been considered in the variable selection step in such application (e.g. by selecting variables with large t-test as in Figure 1 of Fan et al. (2012)).

It is shown that this two-step LPD has classification error tending to Bayes risk with a faster rate than LPD itself, under similar assumptions of sparseness. In practice, $|w_{\text{Bayes}}|_0$ is not known a priori, and must be estimated through cross-validation, as the number s that minimizes the CV-error. Naturally, only situations where $s \leq n$ is of interest, such that the inverse of the estimated covariance matrix is well defined.

1.6 Penalized classifiers

In this section we consider prediction methods frequently used in linear classification. We first compare classification and prediction settings, and afterwards sketch a range of high-dimensional linear regression methods, including ridge regression, LASSO, and the elastic net. The content of this section is not a prerequisite for the rest of this thesis.

Let $x_i = (x_{i1}, \dots, x_{ip})$ for $i = 1, \dots, n$ be mean centred predictors generated from normal distributions. Assume for simplicity that $n_0 = n_1$, and let observations from group 1 have group label $y_i = 1$, while observations from group 0 have $y_i = -1$, such that $\bar{y} = 0$. Let X denote the $n \times p$ matrix of predictors, and y the $n \times 1$ vector of group labels.

We classify a new observation z with outcome y_{new} to group 1 if $zb > 0$, for some p-dimensional vector b, and to group 0 otherwise. Let b_{orc} denote the vector which minimizes $E[(zb - y)^2]$. From Zhang (2004) it is known that the classifier build from b_{orc} have classification error close to Bayes risk, and thus applying the prediction approach in classification is reasonable.

When $p < n$, the classical estimation of b_{orc} is by ordinary least squares, that is

$$\hat{b}_{\text{OLS}} = (X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}y.$$

When $n < p$, $(X^{\mathrm{T}}X)^{-1}$ does not exists. A remedy could be to substitute a generalized inverse into the formula for $\hat{b}_{\text{OLS}}$, which causes problems related to those described in Section 1.2. A more customary option is to adjust $X^{\mathrm{T}}X$ by a diagonal matrix, known as ridge regression

$$\hat{b}_{\text{ridge}} = (X^{\mathrm{T}}X + \beta I_p)^{-1}X^{\mathrm{T}}y, \tag{1.21}$$

for some constant β. While $\hat{b}_{\text{ridge}}$ does not estimate b_{orc} unbiasedly, it is more stable towards small deviations in the data, compared to $\hat{b}_{\text{OLS}}$. Due to this stableness, ridge regression is also widely applied when $p < n$.

The expression in (1.21) is equivalent to

$$\hat{b}_{\text{ridge}} = \arg\min_{b} \left\{ \tfrac{1}{n}|y - Xb|_2^2 + \lambda|b|_2^2 \right\},$$

or

$$\hat{b}_{\text{ridge}} = \arg\min_{b} \tfrac{1}{n}|y - Xb|_2^2 \quad \text{subject to} \quad |b|_2^2 \leq c,$$

with a one-to-one correspondence between β, λ and c. For a suitable choice of regularization parameter $\hat{b}_{\text{ridge}}$ always exists, and due to (1.21) it is easy to compute.

When b_{orc} is sparse, ridge regression do not give a good estimate hereof. As an example, when the data originate from normal distributions, where most variables have no influence in classification, the probability of having any zero coefficients in $\hat{b}_{\text{ridge}}$ is zero, which must be suboptimal. In the regression setting, Tibshirani (1996) introduce the *least absolute shrinkage and selection operator* (LASSO) as

$$\hat{b}_{\text{LASSO}}(\lambda) = \arg\min_{b} \{ \tfrac{1}{n}|y - Xb|_2^2 + \lambda|b|_1 \},$$

which can equivalently be formulated as

$$\hat{b}_{\text{LASSO}} = \arg\min_{b} \tfrac{1}{n}|y - Xb|_2^2 \quad \text{subject to} \quad |b|_1 \leq c,$$

with a one-to-one correspondence between c and λ. When $\lambda > 0$, $\hat{b}_{\text{LASSO}}(\lambda)$ is sparse with no more than $\min(p, n)$ nonzero coordinates.

In the regression setting when considering both the prediction of Xb_{orc} and the estimation of b_{orc}, the LASSO is proved to have nice theoretical properties under sparsity of b_{orc} and further distributional assumptions of the error terms. These results cannot be directly translated to the classification setting, but it indicates that the LASSO is a good procedure more generally.

Contrary to ridge regression, there is no explicit expression for $\hat{b}_{\text{LASSO}}$, but it is easy to calculate through convex optimization techniques. In most practical situations, a priori knowledge of a good value of c or λ is unavailable, and the parameter is found through cross-validation. Efron et al. (2004) suggest the LARS-algorithm, which calculates the full path of $\hat{b}_{\text{LASSO,j}}(\lambda)$ as a function of λ for $j = 1, \ldots, p$, by utilizing that the solution is piecewise linear. The computational complexity of LARS is $O(np \min(n, p))$, and it thus makes the cross-validation approach attractive. A sketch of the LARS-algorithm is found in Section 5.6.4. In some situations, a coordinate descend algorithm is faster than LARS in obtaining the full path of LASSO-coefficients.

The LASSO has experienced much success due to its simple formulation and computational algorithms, as well as its nice theoretical properties. It does have some limitations though. First, it cannot include more than n variables. Secondly, though the LASSO incorporates the covariance structure, it have a tendency to include only one or a few variables from a group of very correlated variables, where one often wants to include all of them. In fact, Zou and Hastie (2005) show that if variable i and j are perfectly correlated, and $\hat{b}_{\text{LASSO},i}$ and $\hat{b}_{\text{LASSO},j}$ are their corresponding LASSO estimates, then all estimates of the form $\hat{b}^*_{\text{LASSO},i} = \beta(\hat{b}_{\text{LASSO},i} + \hat{b}_{\text{LASSO},j})$ and $\hat{b}^*_{\text{LASSO},j} = (1 - \beta)(\hat{b}_{\text{LASSO},i} + \hat{b}_{\text{LASSO},j})$ for $\beta \in [0, 1]$ are LASSO solutions as well. This means that the most sparse LASSO solution only estimates one of those two coefficients as nonzero. On the other hand, when a strictly convex penalization function is applied instead of the ℓ_1 penalty, both coefficients have a unique solution where $\hat{b}_i = \hat{b}_j$, according to Zou and Hastie (2005).

The elastic net was proposed by Zou and Hastie (2005) to solve these problems of the LASSO, while retaining some kind of sparsity. For parameters λ_1 and λ_2 define augmented data as

$$X^* = \frac{1}{1+\lambda_2}\begin{pmatrix} X \\ \sqrt{\lambda_2} I_p \end{pmatrix}, \qquad y^* = \begin{pmatrix} y \\ 0_p \end{pmatrix}.$$

Then the elastic net is $\hat{b}_{\mathrm{EN}} = \sqrt{1+\lambda_2}\hat{b}^*_{\mathrm{EN}}$, where

$$\hat{b}^*_{\mathrm{EN}} = \arg\min_b \left\{ |y^* - X^* b|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} |b|_1 \right\}.$$

That is, the elastic net is a scaled version of the LASSO applied to the augmented data with transformed penalty parameter. Contrary to $\hat{b}_{\mathrm{LASSO}}$, $\hat{b}_{\mathrm{EN}}$ is able to have all coordinates nonzero. It turns out that $\hat{b}_{\mathrm{EN}} = (1+\lambda_2)\hat{b}'_{\mathrm{EN}}$, where

$$\hat{b}'_{\mathrm{EN}} = \arg\min_b \tfrac{1}{n}|y - Xb|_2^2 \quad \text{subject to} \quad \beta|b|_1 + (1-\beta)|b|_2 \leq c,$$

for some β between 0 and 1. Thus, the elastic net is an intermediate between LASSO and ridge regression. The elastic net maintains most positive characteristics of both LASSO and ridge regression, and it often achieves a better precision in prediction when compared to the LASSO. A disadvantage, though, is its higher computational complexity, since it involves cross-validation in two parameters.

Figure 1.1 illustrates the behaviour of the LASSO, ridge regression and elastic net with $\beta = 0.5$, for $p = 2$. For the penalized classifiers, the curves mark the boundary of the feasible set, that is the set of b's fulfilling $|b|_1 \leq c$, $|b|_2^2 \leq c$, and $|b|_1 + |b|_2^2 \leq 2c$, respectively. Note that $\hat{b}_{\mathrm{OLS}}$ exists (for $n \geq 2$), and is the minimizer of $g(b) = \frac{1}{n}|y - Xb|_2^2$. The contours of g are marked by dotted curves. The penalized solutions for $\hat{b}$ are found where these contours first intersects their respective feasible set. One can sense that when considering the LASSO, this often happens in an axis point, which on the other hand very rarely occurs regarding the intersection with the ridge curve. The elastic net is an intermediate of those two, which can be sparse, but this happens less frequent than for the LASSO.

The approach considered in this section have been extended in multiple directions. One option is to consider the class of bridge estimators defined for $\gamma > 0$ as

$$\hat{b}_{\mathrm{bridge}}(\lambda, \gamma) = \arg\min_b \left\{ \tfrac{1}{n}|y - Xb|_2^2 + \lambda \sum_{j=1}^{p} |b_j|^\gamma \right\},$$

which includes both the LASSO and ridge regression. Huang et al. (2008) consider properties of bridge estimators with regard to variable selection when $0 < \gamma < 1$. When $1 < \gamma < 2$, the shape of the restriction of a bridge classifier is very similar to one of an elastic net as in Figure 1.1. Small deviations around the axes are fundamental, though, since the elastic net performs variable selection, whereas the bridge classifier does not.

Finally, an example of a more complicated penalty function is the *Smooth Continuous Absolute Deviation* (SCAD) penalty from Fan and Li (2001). Other possible extensions in penalized regression is to replace $|y - Xb|_2^2$ in the minimization criteria by a more general, but preferably convex, loss functions $L_b(X, y)$. This includes an extension of obvious application in classification, namely to replace the prediction function $f(X) = Xb$ by the prediction function from logistic regression. Due to its confined relevance for this thesis, we refer to Bühlmann and van de Geer (2011) for further information on those last extensions.

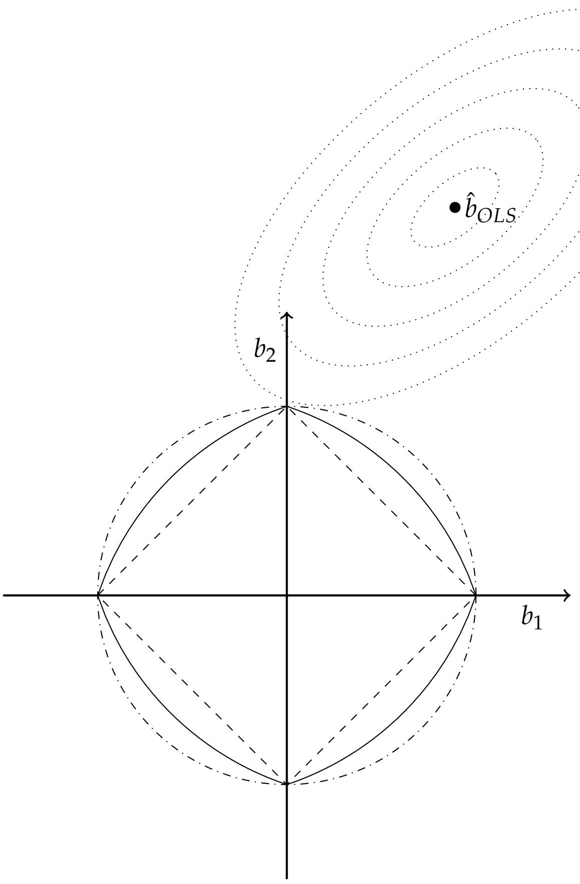

Figure 1.1: Comparison of LASSO (dashed), ridge regression (dashed-dotted), and elastic net with $\beta = 1/2$ (full) in a two-dimensional situation. Contours of the value of $\frac{1}{n}|y - Xb|_2^2$ are marked by dotted lines.

1.7 The imbalance problem

A problem which has been generally overlooked or ignored in the above described work on classification is the imbalance problem, where the number of observations in the groups of the training data differ. Within the last couple of decades, this issue has had some attention, but even today most researchers in the statistical as well as broader scientific fields do not take the issue into account (Shipp et al., 2002; Ramaswamy et al., 2002; Fan et al., 2012). In this section, we describe the background of the problem with the independence rule as a simple, illustrative example. We furthermore review methods for handling the imbalance, including a variety of over- and undersampling methods. Most of these methods have been shown empirically to work well. We refer to Chapter 3 for a deeper theoretical analysis of the imbalance problem.

In real life, imbalanced data sets are the rule, not the exemption. Even when an experiment is designed to have equal sample size in all groups, imbalance often occurs due to missing data points, patients dropping out or dying from external reasons etc. Furthermore, in some groups the amount of data might be limited due to e.g. few patients available with a specific disease, while the control group can be much larger. In such instances the imbalance problem is unavoidable unless one accepts a generally low sample size. Methods to handle imbalance should therefore be of broad demand.

In binary imbalanced settings, the group with most observations is called the majority group while the other is called the minority group. An imbalance can have an intrinsic or extrinsic origin. Intrinsic imbalance occurs when the proportion of two varieties differs in the population, that is $\pi_0 \neq \pi_1$, for example when one compare a group of cancer patients to a control group. Note, though, that in designed experiments, imbalance in the population does not necessarily lead to imbalanced data sets. Extrinsic imbalance occurs due to the way the data are collected, for example when studying the difference in health conditions between genders on a population level by collecting data only among PhD students at the math department at a given university.

It is worth mentioning, that often a classification problem is considered imbalanced only when the difference in sample size among groups is sufficiently large (Blagus and Lusa, 2010; Chawla et al., 2002; He and Garcia, 2009). As illustrated for the independence rule in Section 1.7.1, the problem arises as soon as $n_0 \neq n_1$, though, and in Chapter 3 we see that even small imbalances can have drastic consequences in the high-dimensional case.

In the following, we make use of the group wise probability of correct classification (POCC) defined by

$$POCC_k = P\big(\xi(z) = k \,|\, z \text{ from group } k\big),$$

which is also known as the group specific accuracy. Note that $1 - POCC_k$ equals the classification error of group k.

1.7.1 The imbalance problem for the independence rule

The imbalance problem occurs for most common classifiers, but for simplicity we now address the issue by considering the independence rule in a situation with independent variables. When performing no thresholding, the independence rule is build from the function

$$D(z) = \sum_{i=1}^{p} \frac{(z_j - \bar{x}_{0j})^2 - (z_j - \bar{x}_{1j})^2}{s_j^2}. \tag{1.22}$$

When the imbalance is expected to be intrinsic, imitating Bayes rule, the independence rule classifies to $\mathbf{1}\big\{D(z) > \log(n_0/n_1)\big\}$. Bayes rule puts more weight on the majority group, and is thus not optimal if one wants $POCC_0 = POCC_1$, and we apply instead $\mathbf{1}\{D(z) > 0\}$ irrespective of the imbalance being intrinsic or extrinsic.

The mean value of a term in (1.22), corresponding to a variable with no differential expression, is $\sigma_j^2(1/n_0 - 1/n_1)f/(f-2)$ where $f = n_0 + n_1 - 2$. This gives the independence rule a preference for classifying to the majority group. This bias, due to the varying precisions of the estimates of μ_0 and μ_1, arises whenever $n_0 \neq n_1$, also for small p. However, for large p it gets destructive, and classification to the minority group can be almost impossible even for fairly moderate imbalances. Note that the problem is worse for smaller sample sizes when the imbalance ratio is held constant.

When thresholding the terms in (1.22), the classification bias between the two groups is decreased, but not eliminated. Actually, the irrelevant variables selected through thresholding are the ones that cause the largest bias, as illustrated in Section 3.2. Thus, if the number of variables is decreased to one tenth of the original number through thresholding, it is optimistic to expect the same amount of decline in the bias.

Intuitively, the reason that the independence rule works poorly in imbalanced situations can be explained as follows. Assume without loss of generality that $n_0 > n_1$. Then we expect $\bar{x}_0$ to be a more accurate estimate than $\bar{x}_1$. When $\mu_{0j} = \mu_{1j}$, the j'th coordinate of a new observation is thus on average closer to $\bar{x}_{0j}$ than $\bar{x}_{1j}$. Summarizing this effect over all variables makes classification to the minority group almost impossible in high-dimensional situations.

It is important to realize, that since the bias between groups is inherent in the classification procedure, it cannot be avoided by applying alternative performance measures. In Blagus and Lusa (2013), improved variants of NSC are proposed, where the value of the threshold is selected by considering the G-mean:

$$G\text{-}mean = \sqrt{POCC_0 \cdot POCC_1},$$

While such procedures can decrease the classification bias in some instances, they cannot generally remove it.

In many situations, it is desirable to have different POCCs between groups. For example, when screening a population for a specific cancer type, one typically wants a larger probability of detecting an increased risk of disease, than for correct assignment to the non-risk group. In such situations, the independence rule gives a bias in the opposite direction than desired.

1.7.2 Imbalanced classifiers

In the following, we consider classification between two groups with n_{maj} observations in the majority group, and n_{min} observations in the minority group. Since the bias problem occurs due to different sample sizes in the groups, the most obvious remedy is to adjust the data such that equal sample size is achieved. The simplest way of doing this is by either over- or undersampling (downsizing).

In random oversampling (ROS), one randomly replicates observations from the minority group until it reaches the size of the majority group. Notice that this does not enables the minority group to span a larger fraction of the sample space. Contrary to this, if one obtained more real observations from the minority group, one would hardly expect all of them to lie in the convex hull of the original observations. When considering the independence rule, it is shown in Chapter 3 that the imbalance problem is increased by applying ROS as well as some other oversampling procedures.

In Chawla et al. (2002), an extension of oversampling is proposed under the name Synthetic Minority Oversampling Technique (SMOTE). Instead of direct resampling of the minority observations, one samples random points on the line segments between minority observations. The method is simple: For each observation in the minority group, its h nearest neighbours within the minority class are found, where a reasonable value of h could be 5. For an original observation x and one of its neighbours x', $u \sim U(0,1)$ is generated, and a new observation is defined as $x^* = x + u(x' - x)$. If the imbalance is large, multiple new observations can be generated along each line segment, whereas for almost balanced groups only a random subset of the artificial observations are added to the minority group.

Random undersampling (RUS), removes the bias by using only n_1 observations from the majority group, at the cost of omitting some of the available data, which is clearly not optimal. This can be solved by multiple undersampling, where RUS is repeated a number of times, and a sub-classifier is calculated for each repetition. The conclusion is reached through voting of the sub-classifiers. A slightly different version of multiple undersampling is EasyEnsemble from Liu et al. (2009). Here, q subsets $A_1, \dots, A_q$ are chosen such that $A_i \subseteq \{1, \dots, n_{maj}\}$ and $|A_i| = n_{min}$, and a sub-classifier $\mathbf{1}\{H(z; A_i) - \alpha_i > 0\}$ is calculated, where $H(z; A_i)$ can be built from multiple sub-sub-classifiers. A final ensemble classifier is defined as

$$\mathbf{1}\{\sum_{i=1}^{q} H(z; A_i) - \alpha_i > 0\}. \tag{1.23}$$

Regarding undersampling, a natural question is whether subsampling can be done in a more clever way than simply sampling at random, so that more information is obtained from each sampled observation. We now describe two extensions following this line of thought.

Yang et al. (2014) propose Sample Subset Optimization Technique (SSO) which can be applied to ones preferred classifier. Their suggestion is to split the dataset into a number of cross-validation folds q. When leaving out the observations in one fold as a test set, a good subset of observations from the majority group is searched in the remaining folds in terms of test error. In this way, q subsets $A_1, \dots, A_q$ of suggested optimal majority subsets are obtained. The n_{min} majority observations appearing with highest frequency in these optimal subsets are thus selected for inclusion in the final classifier, while all other majority observations are ignored. Some information is lost when applying SSO, but it is probably less as compared to RUS.

BalanceCascade, suggested by Liu et al. (2009), is an extension of EasyEnsemble which is able to handle larger imbalances due to an iterative method of selecting subsets. In the first step, A_1 is selected at random and used in the first sub-classifier $H_1(z; A_1)$ along with all observations in the minority group. The decision threshold α_1 is adjusted such that the classification error of the majority group, $1 - POCC_{maj}$, is (approximately) $r = \sqrt[q-1]{n_{min}/n_{maj}}$, when classifying by $\mathbf{1}\{H_1(z; A_1) - \alpha_1 > 0\}$. Majority observations which are correctly classified by this classifier are removed from the majority group, and A_2 is randomly selected among the remaining majority observations, leading to $H_2(z; A_2)$. Once again, the threshold α_2 is chosen such that $\mathbf{1}\{H_2(z; A_2) - \alpha_2\}$ fulfils $1 - POCC_{maj} \approx r$. This procedure is repeated q times in total. Finally, all sub-classifiers are combined as in (1.23). The adjustment of the decision threshold makes sure that the majority group includes at least n_{min} observations at the execution of the last step.

Finally, we turn to a method unrelated with sampling strategies. Proposition 5 of Jensen (2006) shows that NSC reduces the bias in imbalanced situations. Tibshirani et al. (2003) further suggest an adjusted version of NSC, which repairs the imbalance problem though it is not its explicit purpose: Calculate $\xi_I(z)$, with $\bar{x}_{0j}$ replaced by $\tilde{x}_{0j}$ as in (1.10), while $\bar{x}_{1j}$ is replaced by

$$\tilde{x}_{1j} = \begin{cases} a_j & \text{if } |t_j| < \theta\alpha, \\ \bar{x}_{1j} - (1 - \frac{1}{\theta})\frac{n_0}{n_0+n_1}(\bar{x}_{1j} - \bar{x}_{0j}) - \operatorname{sgn}(\bar{x}_{1j} - \bar{x}_{0j})\alpha s_j\sqrt{\frac{1}{n_1} - \frac{1}{n}} & \text{if } |t_j| \geq \theta\alpha. \end{cases}$$

For each value of the threshold α, it is now possible to find a value of θ such that the variables with no actual influence do not systematically give any of the groups an advantage over the other, see Jensen (2006) for details.

1.8 Review of the thesis

In this section we give a description of the problems considered in the succeeding chapters of this thesis and state the main conclusions.

1.8.1 Chapter 2: Classification error of the thresholded independence rule

In Chapter 2 the thresholded independence rule is considered in a situation of two p-variate normal distributions when $\log(p)/n \to \infty$. From Bickel and Levina (2004), it is known that a classification superior to random guessing can be conducted under the parameter restrictions in (1.4). When aiming at application in microarray settings, these restrictions are in our opinion inappropriate. We thus consider relaxing them in the discussion below. A more detailed discussion can be found in the master thesis of Morten Fenger-Grøn (Grøn, 2007).

Consider first the covariance matrix, where Bickel and Levina (2004) assume an upper as well as a lower bound on the eigenvalues. This is implied by having both the variances and the eigenvalues of the correlation matrix bounded above and below. The boundedness of the variances appears reasonable, at least after a transformation. The upper bound of the correlation matrix is assured if the sum of the correlations of each variable with all the other variables are upper bounded, which we consider reasonable in microrarray settings. We do not have a justification for the lower bound of the eigenvalues of the correlation matrix Σ_0.

Next, when considering the assumptions of the mean values, (1.5) is extremely restrictive since all coordinates apart from the first few ones are required to be very close to zero. In microarray analysis this rarely happens, but we do expect that the differential expression is close to zero for most genes. Substituting μ by Δ in (1.5) will not fulfil our purpose though, since we do not expect a priori knowledge of the positions of the differentially expressed genes.

Based on the discussion of (1.4) and (1.5) above, we define the following parameter space

$$\Theta = \big\{\theta : \forall j \; c_1^D \leq \sigma_j^2 \leq c_2^D,\ \lambda_{\max}(\Sigma_0) \leq c_2,\ \theta \in B\big\},$$

where c_1^D, c_2^D and c_2 are positive constants, and B is either B_1 or B_2 where

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

for $\delta_j = \Delta_j/\sigma_j$, c_0 a constant, and $b_n \to 0$ as $n \to \infty$, or

$$\begin{aligned} K_n &= \#\{k : |\delta_j| > 2\alpha\} \geq 1 \\ B_2 &= \big\{\theta :\ \#\{j : \tfrac{\alpha}{2} \leq |\delta_j| \leq 2\alpha\} \leq c_1 K_n\big\}. \end{aligned}$$

Note that neither B_1 nor B_2 require most variables to have differential expression equal to zero, but only that the majority of the expressions are small.

Our main result provides an upper asymptotic bound of the classification error of the thresholded independence rule on Θ. Contrary to (1.6), which gives an upper bound of the worst case classification error of the full parameter space, we instead find an asymptotic bound of the classification error for any $\theta \in \Theta$.

Our result uses that perfect separation between expressed and non-expressed variables is obtained asymptotically. This separation rarely occurs in finite sample scenarios. An extension of our theorem shows that the conclusion remains valid when the threshold is chosen such that a fixed number of false detected variables are allowed.

While being an extension of Bickel and Levina (2004), our result is also related to Fan and Fan (2008) and Donoho and Jin (2009). Compared to Fan and Fan (2008), our main improvement is that our result is proved for $\Sigma \neq I_p$, whereas FAIR is only heuristically justified for such Σ. A comparison of the upper bounds of the classification error in our result and Fan and Fan (2008) are given in Chapter 2.

In our setting, the expression levels of the differentiable expressed variables are allowed to vary, whereas Donoho and Jin (2009) require all nonnull variables to be equal. The number of nonzero coefficients in B_1 and B_2 are defined with respect to n, while

Donoho and Jin (2009) use p as the driving parameter. This means that the situations covered differ for the two results. For instance, the situation B_1 with $b_n = \log(p)/n$ is not included in the setting of Donoho and Jin (2009).

FAIR and HC thresholding suggest directly applicable methods for selecting the threshold in data situations, whereas our main result is of a more theoretical nature. Our extension selects a specific threshold fulfilling an upper bound of the expected number of false discoveries, and is thus more directly applicable. No optimality result exists for the resulting value of threshold, though. Actually, in situations included in the setting of Donoho and Jin (2009), we must expect this threshold to be too large to obtain an optimal classification, since it leads to a low FDR when p is large.

1.8.2 Chapter 3: High-dimensional classifiers in the imbalanced case

Chapter 3 considers the imbalance problem introduced in Section 1.7 restricted to a situation with independent variables, such that Bayes rule corresponds to the independence rule. To the best of our knowledge, this situation has not previously been considered from our analytical point of view. We consider theoretically the behaviour of the thresholded independence rule, and clearly see that the bias can be very large in high-dimensional situations, even for rather small imbalances.

The in-depth understanding of the reasons behind the failure of the independence rule leads to the introduction of two new classifiers. The first one is called the *bias adjusted independence* classifier (BAI classifier). The justification for BAI is through subtracting the bias from the independence classifier, as if the variables were non-differentiable expressed, leading to

$$B_0(z) = \sum_{j=1}^{p} \frac{\bar{x}_{0j} - \bar{x}_{1j}}{s_j^2} \Big[z_j - \frac{1}{2}(\bar{x}_{0j} + \bar{x}_{1j}) + \frac{\rho}{2}(\bar{x}_{1j} - \bar{x}_{0j}) \Big] w(t_j),$$

where $\rho = (n_0 - n_1)/(n_0 n_1)$. Regarding the weight function $w(t)$, our main focus is on hard thresholding. In B_0, a bias of unknown size from the differentially expressed variables remains, which is estimated by leave-one-out cross-validation and subsequently subtracted. BAI is thus defined as classifying to group 0 when $B(z) < 0$ and to group 1 otherwise, where

$$B(z) = B_0(z) - \frac{1}{2}\Big[\frac{1}{n_0} \sum_{i=1}^{n_0} B_0(x_{0i}; x_{0i}) + \frac{1}{n_1} \sum_{i=1}^{n_1} B_0(x_{1i}; x_{1i}) \Big],$$

and $B_0(z_1; z_2)$ denotes B_0 calculated from all available training data except for z_2, and afterwards applied to z_1.

To state our second proposed classifier, let $\bar{x}_{kj}(i)$ be the k'th group average with the respective i'th observation left out, and let $s_j^2(x_{ki})$ be the within group variance when x_{ki} is left out. The corresponding t-statistic is denoted $t_j(x_{ki})$. The *leave one out independence* classifier (LOUI classifier) is defined as

$$\begin{aligned} L(z) = \frac{1}{2} \sum_{j=1}^{p} \Big[& \frac{1}{n_0} \sum_{i=1}^{n_0} \frac{\bar{x}_{1j} - \bar{x}_{0j}(i)}{s_j^2(x_{0i})} (z_j - x_{0ij}) w\big(t_j(x_{0i})\big) \\ & + \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{\bar{x}_{1j}(i) - \bar{x}_{0j}}{s_j^2(x_{1i})} (z_j - x_{1ij}) w\big(t_j(x_{1i})\big) \Big]. \end{aligned}$$

The reasoning behind LOUI is that, from a certain point of view, it reestablish independence between $\bar{x}_1 - \bar{x}_0$ and $\bar{x}_1 + \bar{x}_0$.

Both of the suggested classifiers result in a tiny bias of no practical importance in classification. When considering the resulting probabilities of correct classification analytically, the behaviour of BAI and LOUI are very alike. In an empirical study, they are further compared to EasyEnsemble with the thresholded independence rule as sub-classifier. Both BAI and LOUI slightly outperform EasyEnsemble in our scenarios, and BAI furthermore includes fewer variables, which makes it more attractive from an experimentalist and interpretative point of view.

A small simulation study shows that the BAI and LOUI classifiers can be generalized and combined with ROAD, which enables them to handle imbalance also in correlated situations.

In an appendix, a general formulation of oversampling is proved to increase the bias of the independence rule. This general form of oversampling includes ROS and a method familiar to SMOTE. This advices one to be careful when applying oversampling also in more general situations.

1.8.3 Chapter 4: On oracle efficiency of the ROAD classification rule

In Theorem 1 of Fan et al. (2012), the classification errors of the empirical ROAD and the oracle ROAD are concluded to be close under specific restrictions. In Chapter 4, we point out an error in the original proof, and further reformulate and prove the theorem under adjusted assumptions.

The major difference between our result and Theorem 1 is that the sparseness of w_{ROAD} is not included in neither the assumptions nor the result in our corrected theorem. This extends the applicability of ROAD. While Fan et al. (2012) require the eigenvalues of the covariance matrix to be bounded below, our only restriction related to the covariance is that $|\Sigma|_\infty$ should be upper bounded. Following the discussion in Section 1.8.1, this is a realistic scenario in microarray settings. We furthermore include minor extra restrictions on the mean values and mean difference in our theorem, which in our opinion were missing in the original result.

1.8.4 Chapter 5: A numeric comparison of sparse linear classifiers incorporating covariance

In Chapter 5 an extensive simulation study is performed to compare various classifiers in correlated settings. The considered classifiers are ROAD from (1.20) calculated from the algorithms of both Fan et al. (2012) and Wu et al. (2011), LPD from (1.5.2), and a new suggestion, namely the *Linear Lasso Discriminant* (LLD) defined by

$$w_{\text{LLD}} = \operatorname*{arg\,min}_{w:|w|_1 \leq c} \frac{1}{p} |\Sigma w - \Delta|_2^2.$$

The theoretical properties of the various classifiers are summarized, and all of the applied algorithms are described. The simulations are performed in 19 different settings with small, moderate and large correlations, and p varying between 100 and 1000. We fix $n_0 = n_1 = 50$, which is rather low when comparing to related simulation studies, but we consider sample sizes of that order more realistic in microarray settings.

Our main interest in the simulations is to find the classifier that minimizes the classification error, but we consider other aspects as well. The two algorithms for ROAD result in correlated, but different, classifiers. None of the classifiers generally have lower classification error than the others, but ROAD and LPD tend to be the better options in most situations. LLD works poorer than the other classifiers, unless $p = 100$.

In our implementations, ROAD is by far the fastest classifier to calculate, and, contrary to theory its computational complexity appears linear in p. LPD and LLD is less sparse than ROAD. No significant difference in the ability to estimate w_{Bayes} with respect to ℓ_2 error is seen for ROAD and LPD, though ROAD is slightly better at estimating large, nonzero elements of w_{Bayes} precisely.

Bibliography

Bak, B. A., M. F. Grøn, and J. L. Jensen (2015). Classification error of the thresholded independence rule. *Scandinavian Journal of Statistics 42*, 34–42.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 57*(1), 289–300.

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.

Bickel, P. J. and E. Levina (2008a). Covariance regularization by thresholding. *The Annals of Statistics 36*(6), 2577–2604.

Bickel, P. J. and E. Levina (2008b). Regularized estimation of large covariance matrices. *The Annals of Statistics 36*(1), 199–227.

Blagus, R. and L. Lusa (2010). Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics 11*.

Blagus, R. and L. Lusa (2013). Improved shrunken centroid classifiers for high-dimensional class-imbalanced data. *BMC Bioinformatics 14*.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: Synthetic miniority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Clemmensen, L., T. Hastie, D. M. Witten, and B. Ersbøll (2011). Sparse discriminant analysis. *Technometrics 53*, 406–413.

Committee on Mathematical Sciences Research for DOE's Computational Biology, N. R. C. (2005). Mathematics and 21st century biology. Report.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics 32*(3), 962–994.

Donoho, D. and J. Jin (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*(39), 14790–14795.

Donoho, D. and J. Jin (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical transactions of the royal society A 367*, 4449–4470.

Donoho, D. and J. Jin (2015). Higher criticism for high-scale inference, especially for rare and weak effects. *Statistical Science 30*(1), 1–25.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

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

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Grøn, M. F. (2007). Optimal classification of observations from a high dimensional, normal distribution. Master's thesis, Aarhus Universitet.

Guo, Y., T. Hastie, and R. Tibshirani (2005). Regularized linear discriminant analysis and its application in microarrays. *Biostatistics 1*, 1–18.

He, H. and E. A. Garcia (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering 21*(9), 1263–1284.

Huang, J., J. L. Horowitz, and S. Ma (2008). Asymptotic properties of bridge estimators in high-dimensional regression models. *The Annals of Statistics 36*(2), 587–613.

Jensen, J. L. (2006). Maximum likelihood classifiers in microarray studies. Research Report 474, University of Aarhus.

Jin, J. (2009). Impossibility of successful classification when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America 106*(22), 8859–8864.

Johnstone, I. M. (2002). Function estimation and gaussian sequence models. Manuscript.

Kolar, M. and H. Liu (2015). Optimal feature selection in high-dimensional discriminant analysis. *IEEE Transactions on Information Theory 61*, 1063–1083.

Lauritzen, S. (1996). *Graphical Models*. Clarendon Press, Oxford.

Liu, X.-Y., J. Wu, and Z.-H. Zhou (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transations on Sysetems, Man, and Cybernetics Part B: Cybernetics 39*(3), 539–550.

Mai, Q. and H. Zou (2012). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics 55*(2), 243–246.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

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

Ramaswamy, S., K. N. Ross, E. S. Lander, and T. R. Golub (2002). A molecular signature of metastasis in primary solid tumors. *Nature Genetics 33*(1), 49–54.

Rothman, A. J., P. J. Bickel, E. Levina, and J. Zhu (2008). Sparse permutation invariant covariance estimation. *Electronic Journal of Statistics 2*, 494–515.

Shao, J., Y. Wang, X. Deng, and S. Wang (2011). Sparse linear discriminant analysis by thresholding for high dimensional data. *The Annals of Statistics 39*(2), 1241–1265.

Shipp, M. A., K. N. Ross, P. Tamayo, A. P. Weng, J. L. Kutok, R. C. T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. S. Pinkus, T. S. Ray, M. A. Koval, K. W. Last, A. Norton, T. A. Lister, J. Mesirov, D. S. Neuberg, E. S. Lander, J. C. Aster, and T. R. Golub (2002, January). Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature Medicine 8*(1), 68–74.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), pp. 267–288.

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

Wang, C., L. Cao, and B. Miao (2013). Optimal feature selection for sparse linear discriminant analysis and its applications in gene expression data. *Computational Statistics and Data Analysis 66*, 140–149.

Wu, M. C., L. Zhang, and X. Lin (2011). Two-group classfication using sparse linear discriminants analysis. Technical report, Department of Biostatistics, Harvard School of Public Health.

Yang, P., P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya (2014). Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Transations on Cybernetics 44*(3), 445–455.

Zhang, T. (2004). Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics 32*(1), 56–134.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 301–320.

2

Classification error of the thresholded independence rule

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

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.

2.1 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. Often the number of actually relevant variables are much less than p and variable selection is necessary. This is often done by either thresholding or the LASSO (Tibshirani, 1996) and extensions hereof. For a thorough overview of the LASSO see Bühlmann and van de Geer (2011).

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 measurements come 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 2.2 contains the setup of the paper and states the main result. The proof is given in Section 2.3 and the appendix in Section 2.5 collects all the basic inequalities used in the proof. The appendix in Section 2.6 gives an extension of the main theorem where a constant number of expected false positives are allowed.

2.2 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 $W(\xi,\theta) = P_\theta(\xi(x) = 1 \mid \{x_{ij}\})$, where θ parametrizes 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

$$\xi_B(x) = \mathbf{1}\{\Delta^{\mathrm{T}}\Sigma^{-1}\big(x - \tfrac{1}{2}(\mu_0 + \mu_1)\big) > 0\} \qquad \text{with } W(\xi_B,\theta) = \overline{\Phi}\big(\tfrac{1}{2}(\Delta^{\mathrm{T}}\Sigma^{-1}\Delta)^{1/2}\big).$$

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

$$\xi_{\mathrm{TI}}(x) = \mathbf{1}\{\Delta^{\mathrm{T}}D^{-1}\big(x - \tfrac{1}{2}(\mu_0 + \mu_1)\big) > 0\}$$

with

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

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

$$t_k = \frac{d_k}{\sqrt{s_k^2/m}}, \quad m = \frac{n_0 n_1}{n_0 + n_1}, \quad \text{and} \quad I_k = \mathbf{1}\{|t_k| > \sqrt{m}\alpha\},$$
$$\hat{\Delta}_k = d_k I_k, \quad \text{and} \quad \hat{\mu}_{ik} = \begin{cases} \bar{x}_{ik} & \text{if } I_k = 1, \\ \frac{n_0}{n_0+n_1}\bar{x}_{0k} + \frac{n_1}{n_0+n_1}\bar{x}_{1k} & \text{if } I_k = 0. \end{cases}$$

The classifier is

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

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 restricts the mean values μ_0 and μ_1. We define

$$\Theta = \{\theta : \forall k \ c_1^D \leq \sigma_k^2 \leq c_2^D, \ \lambda_{\max}(\Sigma_0) \leq c_2, \ \theta \in B\}, \tag{2.1}$$

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,

$$B_1 = \{\theta : \ \#\{k : |\delta_k| \geq \tfrac{\alpha}{2}\} \leq b_n n, \ \#\{k : |\delta_k| > c_0\} \geq 1\}, \tag{2.2}$$

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,

$$K_n = \#\{k : |\delta_k| > 2\alpha\} \geq 1$$
$$B_2 = \{\theta : \ \#\{k : \tfrac{\alpha}{2} \leq |\delta_k| \leq 2\alpha\} \leq c_1 K_n\}, \tag{2.3}$$

where c_1 is a constant. Note that in the specification of the parameter space the dependency 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 2.1. *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* (2.1) *and either* (2.2) *or* (2.3). *Then*

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

Remark 2.2. *By exchanging the group labels it is clear that the upper bound of Theorem 2.1 applies also to the classification error for a new observation from group 1. Furthermore, the formulation of Theorem 2.1 allows for a triangular array where means and variances depend on n.*

Remark 2.3. *The result in Theorem 2.1 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 enter 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.

2.3 Proof

We start by stating and proving a fundamental lemma. To this end we define for a p-dimensional vector a and a symmetric, nonsingular $p \times p$ 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 2.4. *Let the covariance matrix Σ satisfy the bounds stated explicitly in* (2.1)*. Then $\omega(\hat{D}) \xrightarrow{P_\Theta} 1$ and if*

$$\frac{\sum_{k:\hat{\Delta}_k\neq 0}(\hat{\mu}_{0k}-\mu_{0k})^2/\sigma_k^2}{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/\sigma_k^2} \xrightarrow{P_\Theta} 0, \tag{2.4}$$

we have

$$W(\xi,\theta) - \overline{\Phi}(\Psi_\Sigma(\hat{\Delta},\hat{D})) \xrightarrow{P_\Theta} 0. \tag{2.5}$$

Furthermore, on Θ we have:

$$2\Psi_\Sigma(\hat{\Delta},\hat{D}) \ge \frac{1}{\omega(\hat{D})\sqrt{c_2}}|D^{-1/2}\hat{\Delta}|. \tag{2.6}$$

Proof. From the multivariate normal distribution we find that

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

Using Lemma 2.5(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 2.5(iii) together with Boole's inequality give 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 (2.5) has been proven.

Finally, (2.6) 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 (2.4) and then obtain the result of the theorem from (2.6). 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 (2.4) 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 2.5(vi) is bounded by

$$\sum_{k:|\delta_k|\geq 2\alpha} P(|t_k| < \sqrt{m}\alpha) \leq pa_1e^{-m\alpha^2a_2} \leq a_1e^{-n\tau_n(\kappa_3a_2c_\alpha^2\tau_n^{-2\gamma}-1)} \to 0. \qquad (2.7)$$

For case B_1 of the parameter space we have from Lemma 2.5(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 2.5(i) is bounded by

$$\sum_{k:|\delta_k|\geq 2\alpha} P(|d_k|/\sigma_k < \alpha) \leq pe^{-m\alpha^2/2} \leq e^{-n\tau_n(\kappa_3c_\alpha^2\tau_n^{-2\gamma}/2-1)} \to 0.$$

Thus, with a probability tending to one we have that the denominator in (2.4) is bounded by

$$\sum_{k:\hat{\Delta}_k\neq 0}\frac{\hat{\Delta}_k^2}{\sigma_k^2} \geq \sum_{k:|\delta_k|>2\alpha} I_k\frac{d_k^2}{\sigma_k^2} = \sum_{k:|\delta_k|>2\alpha}\frac{d_k^2}{\sigma_k^2} \geq \begin{cases}\frac{c_0^2}{4} & \text{case } B_1,\\ \alpha^2K_n & \text{case } B_2.\end{cases} \qquad (2.8)$$

For the numerator in (2.4) we introduce the notation $\bar{x}_k = n_0/(n_0+n_1)\bar{x}_{0k} + n_1/(n_0+n_1)\bar{x}_{1k}$, $\bar{\mu}_k = n_0/(n_0+n_1)\mu_{0k} + n_1/(n_0+n_1)\mu_{1k}$. Note that $\bar{x}_k$ is independent of d_k, $\mathrm{Var}(\bar{x}_k) = \sigma_k^2/(n_0+n_1)$ and $\mathrm{Var}(d_k) = \sigma_k^2/m$. Taking expectation and using Lemma 2.5(iv)-(v) we get for the nominator in (2.4)

$$
\begin{aligned}
E\Big[\sum_{k:\hat{\Delta}_k\neq 0}\frac{(\hat{\mu}_{0k}-\mu_{0k})^2}{\sigma_k^2}\Big] &= E\Big[\sum_{k=1}^{p}\mathbf{1}\{\hat{\Delta}_k\neq 0\}\frac{(\bar{x}_{0k}-\mu_k)^2}{\sigma_k^2}\Big]\\
&= E\Big[\sum_{k=1}^{p}\mathbf{1}\{\hat{\Delta}_k\neq 0\}\frac{(\bar{x}_k-\bar{\mu}_k-\frac{n_1}{n_0+n_1}(d_k-\Delta_k))^2}{\sigma_k^2}\Big]\\
&= \sum_{k=1}^{p}\Big\{E\Big[\mathbf{1}\{\hat{\Delta}_k\neq 0\}\frac{1}{n_0+n_1}\Big]+\frac{n_1^2}{(n_0+n_1)^2}E\Big[\mathbf{1}\{\hat{\Delta}_k\neq 0\}\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\Big\}\\
&\le \sum_{k:|\delta_k|<\frac{\alpha}{2}}\Big\{P(|t_k|>\alpha\sqrt{m})\frac{1}{n_0+n_1}+\frac{n_1^2}{(n_0+n_1)^2}E\Big[\mathbf{1}\{|t_k|>\alpha\sqrt{m}\}\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\Big\}\\
&\quad+\sum_{k:|\delta_k|>\frac{\alpha}{2}}\frac{1}{n_0+n_1}+\frac{n_1^2}{(n_0+n_1)^2}E\Big[\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\\
&\le \begin{cases} 2pa_1e^{-n\alpha^2a_2}+b_n(1+n/m) & \text{case } B_1,\\ 2pa_1e^{-n\alpha^2a_2}+(c_1+1)K_n\big(\frac{1}{n_0+n_1}+\frac{1}{m}\big) & \text{case } B_2.\end{cases}
\end{aligned}
\tag{2.9}
$$

Dividing (2.9) by (2.8) we see immediately the convergence to zero for case B_1. For case B_2 the second term of (2.9) is the dominating part and dividing this by (2.8) we get

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

This ends the proof of (2.4).

We next turn to the use of (2.6) 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 (2.7) 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

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

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 $K_n/(mS_\alpha)\leq b_nn/(mc_0)\to 0$. Considering case B_2 we get $mS_\alpha\geq 4m\alpha^2K_n\to\infty$ 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. □

2.4 Discussion

Theorem 2.1 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. Bickel and Levina (2004) require the individual group means μ_{jk} to tend to zero as $k \to \infty$. This is not the situation in an experiment with microarray measurements. Instead, in Theorem 2.1, we have restrictions on the differences $\mu_{1k} - \mu_{0k}$ only. Also, when δ contains m nonzero coefficients, all with differential expression c, these coefficients can not be placed arbitrarily along the sequence $\delta_1, \ldots, \delta_p$ in the setting of Bickel and Levina (2004). In the setting of a microarray experiment we want to consider the possibility of a number of nonzero (small) coefficients spread all over the variables.

Our approach is similar to the one considered in Fan and Fan (2008), where a proof is given for the case $\Sigma = I$. The upper bound on the classification error in their Theorem 5 can be compared to the bound in Theorem 2.1 on taking b_n and a of their paper equal to α 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 2.1. It is possible to construct situations where the upper bound of Fan and Fan (2008) tends to one, whereas the bound of Theorem 2.1 is strictly less than one, as an example consider $\delta_1 \neq 0$ fixed and all remaining nonzero δ_js between $\alpha/2$ and α.

Theorem 2.1 is an asymptotic result for $n \to \infty$. For finite n it is of interest to investigate whether one is close to the asymptotic situation. Looking at the proof of Theorem 2.1 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 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, as in the asymptotic case in Theorem 2.1, and include all true positives with large differential expression. To illustrate this intuitive background, we have in Table 2.1 chosen the expected number of false positives as $\omega_n = 0.1$, chosen α accordingly, and then calculated δ, the scaled differential expression needed in order to include a variable in the classifier with a high probability, here taken as 0.9. With p in the order of thousands, requiring the number of false positives to be below 0.1 imply that only variables with a fairly high level of differential expression are detected. This hardly gives an optimal classifier in terms of classification error. In Table 2.1 we have therefore also included the case where the expected number of false positives is $\omega_n = 5$. We leave it to future work to analyze theoretically the impact of including such non-expressed variables in the classifier.

The interesting aspect of the table is the amount of differential expression needed in order to include an expressed variable in the classifier with some certainty. In particular we see that for moderate values of the number of observations n, the number of variables p can be quite large still allowing for inclusion of true positives both when $\omega_n = 0.1$ and $\omega_n = 5$.

We now take a brief look on the bladder cancer data from Dyrskjøt et al. (2003). After prefiltering there are $p = 3032$ variables and two groups of patients, a group with $n_0 = 15$ patients having no recurrence of the cancer and a group with $n_1 = 16$ having recurrence. In Table 2.2 the number of positive variables and the number of expected positives under the hypothesis of no differential expression, obtained from the t-test, is calculated for different values of the threshold α. For these data most effects appears to be zero or small as assumed in our result. The requirement of a low number of expected false positives leave us with no differential expressed variables. The data still points to a number of differential expressed variables, but the expression is small and

Table 2.1: Threshold and differential expression needed 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 and 5 respectively. The scaled differential expression δ has been chosen such that an expressed variable is included in the classifier with probability 0.9.

		$\omega_n = 0.1$		$\omega_n = 5$	
n	p	α	δ	α	δ
40	1000	1.37	1.82	0.94	1.37
80	1000	0.92	1.22	0.65	0.95
160	1000	0.64	0.85	0.46	0.67
160	4000	0.69	0.90	0.52	0.73
160	20000	0.75	0.96	0.60	0.81

Table 2.2: Number of observed positives and expected false positives for various values of the threshold α in bladder cancer data from Dyrskjøt et al. (2003).

$\sqrt{m}\alpha$	Observed positives	Expected false positives
1	1195	987
1.5	615	438
2	282	167
2.5	100	56
3	35	17
3.5	5	5
4	1	1

the false discovery rate seems to be above 50%. For these data it is far from possible to obtain perfect separation between true positives and false positives. In Dyrskjøt et al. (2003) a classifier with 26 variables was built based on these data, but it did not prove successful in a later follow up study (Dyrskjøt et al., 2003).

2.5 Appendix A: Lemma

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

Lemma 2.5. *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

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

(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 $E\big[\mathbf{1}\{|t| > \sqrt{m}\alpha\}U^2\big] \leq a_1 e^{-a_2\alpha^2 n}$.

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

Proof.

(i). This follows from

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

(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.

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

(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 $\delta < \alpha/2$ we use $\omega = 3/4$ 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 $\delta \leq \alpha/2$, 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.

(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

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

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

(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 (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. □

2.6 Appendix B: Selecting α to control the expected number of false discoveries

Theorem 2.1 does not report on how the threshold α should be determined for given values of p and n. Optimally, α should be as small as possible while the expected number of false positives is still converging to zero. In practice it is acceptable as long as the expected number of false positives is upper bounded by a constant as in Table 2.1, which allows α to converge slightly faster to zero.

We continuously examine the thresholded independence rule on Θ as in (2.1) with B either equal to B_1, or equal to B_2 with $K_n \to \infty$, which is rarely a restriction since when K_n is bounded, the situation is in most cases covered by B_1. The following theorem, similar to Theorem 2.1, holds:

Theorem 2.6. *Let V be a selected constant and p tend to infinity with n in such a way that $\log(p)/n \to 0$. Choose the sequence of α's such that*

$$p \cdot P_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha) = V. \tag{2.10}$$

Consider the parameter space given through Θ and B_1 or B_2 with $K_n \to \infty$. Then

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

Proof. The proof is based on Lemma 2.7 below. Due to Lemma 2.4 we are to prove

$$\frac{\sum_{k:\hat{\Delta}_k\neq 0}(\hat{\mu}_{0k}-\mu_{0k})^2/\sigma_k^2}{\sum_{k:\hat{\Delta}_k\neq 0}\hat{\Delta}_k^2/\sigma_k^2} \xrightarrow{P_\Theta} 0. \tag{2.11}$$

At first we note that due to (2.10), we still have $\sqrt{m}\alpha \geq t_{1-2V/p}(n) \to \infty$, where $t_{1-2V/p}(n)$ is the $(1-2V/p)$th fractile of a t-distribution with n degrees of freedom. Lemma 2.7(ii) gives

$$\begin{aligned}\sum_{k:|\delta_k|>2\alpha} P(|t_k| < \sqrt{m}\alpha) &\leq pP_{\delta=2\alpha}(|t| < \sqrt{m}\alpha) \\ &\leq pP_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha)a_1e^{-a_2\alpha^2 m} \\ &\leq Va_1e^{-a_2\alpha^2 m} \to 0,\end{aligned}$$

and furthermore from Lemma 2.7(iii) we have

$$\begin{aligned}\sum_{k:\delta_k\geq 2\alpha} P(|d_k|/\sigma_k < \alpha) &\leq pP_{\delta=2\alpha}\big(\frac{1}{\sqrt{m}}U + \delta < \alpha\big)\\ &\leq pP_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha)a_1e^{-a_2\alpha^2 m}\\ &\leq Va_1e^{-a_2\alpha^2 m} \to 0.\end{aligned}$$

Hereby the denominator in (2.11) is bounded below on B_1 as well as B_2 by the same constants as in (2.8).

Next we study the numerator in (2.11):

$$\begin{aligned}&E\Big[\sum_{k:\hat{\Delta}_k\neq 0}\frac{(\hat{\mu}_{0k}-\mu_{0k})^2}{\sigma_k^2}\Big]\\ &\quad\leq \sum_{k:|\delta_k|<\frac{\alpha}{2}}\Big\{P(|t_k|>\alpha\sqrt{m})\frac{1}{n_0+n_1}+\frac{n_1^2}{(n_0+n_1)^2}E\Big[\mathbf{1}\{|t_k|>\alpha\sqrt{m}\}\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big]\Big\}\\ &\quad\quad+\sum_{k:|\delta_k|>\frac{\alpha}{2}}\frac{1}{n_0+n_1}+\frac{n_1^2}{(n_0+n_1)^2}E\Big[\frac{(d_k-\Delta_k)^2}{\sigma_k^2}\Big].\end{aligned}$$

The last sum is upper bounded as in (2.9) for both B_1 and B_2. For B_1 we therefore obtain by Lemma 2.7(i):

$$\begin{aligned}E\Big[\sum_{k:\hat{\Delta}_k\neq 0}\frac{(\hat{\mu}_{0k}-\mu_{0k})^2}{\sigma_k^2}\Big] &\leq V\Big(\frac{1}{n_0+n_1}+a_1\alpha^2\Big)+e^{-n(a_2-\frac{\log(p)}{n})}+b_nn\Big(\frac{1}{n_0+n_1}+\frac{1}{m}\Big)\\ &\to 0.\end{aligned}$$

Combining the numerator and denominator provides the desired convergence to zero for B_1.

We now turn to B_2. With $K=(c_1+1)(1+1/\kappa_3)$ the quotient in (2.4) is bounded above:

$$\begin{aligned}&\frac{V(\frac{1}{n_0+n_1}+a_1\alpha^2)+\alpha^2e^{-n(a_2-\frac{\log(p)}{n})}+K\frac{K_n}{n}}{K_n\alpha^2}\\ &\quad\leq\frac{V}{K_n}\left(\frac{1}{(n_0+n_1)\alpha^2}+a_1+e^{-n(a_2-\frac{\log(p)}{n})}\right)+\frac{K}{n\alpha^2}\\ &\quad\to 0,\end{aligned}$$

so (2.11) holds by the extra assumption $K_n\to\infty$.

The inspection of $|D^{-1/2}\hat{\Delta}|$ is accomplished exactly as in Theorem 2.1. □

Lemma 2.7. *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 (i) to (iii) below hold:*

(i) If $|\delta|<\frac{\alpha}{2}$ then $E[\mathbf{1}\{|t|>\sqrt{m}\alpha\}U^2]\leq a_1m\alpha^2\big(e^{-na_2}+P(|t|>\sqrt{m}\alpha)\big)$.

(ii)

$$\frac{P_{\delta=2\alpha}(|t|<\sqrt{m}\alpha)}{P_{\delta=\alpha/2}(|t|>\sqrt{m}\alpha)}\leq a_1e^{-a_2\alpha^2 m}.$$

(iii)

$$\frac{P_{\delta=2\alpha}\left(\frac{1}{\sqrt{m}}U+\delta<\alpha\right)}{P_{\delta=\alpha/2}(|t|>\sqrt{m}\alpha)} \leq a_1 e^{-a_2\alpha^2 m}.$$

Proof.

(i). First note for $0 < s \leq t$ it holds

$$\begin{aligned}
E[U^2 \mid U > s] &= \frac{1}{P(U>s)}\int_s^\infty u^2\phi(u)du \\
&= \frac{1}{P(U>s)}\big(s\phi(s)+P(U>s)\big) \\
&= \frac{s\phi(s)}{P(U>s)}+1 \\
&\leq \frac{s\phi(s)}{s/(s^2+1)\phi(s)}+1 \\
&= s^2+2 \\
&\leq 2(t^2+1) \\
&= 2\left(\frac{t\phi(t)}{\phi(t)/t}+1\right) \\
&\leq 2\left(\frac{t\phi(t)}{P(U>t)}+1\right) \\
&\leq 2E[U^2 \mid U > t], \qquad (2.12)
\end{aligned}$$

where the first inequality is from Cook (2009). It is easily seen that (2.12) also holds for $-t \leq s < 0$.

Next we study the second moment of the normal distribution when conditioning on the upper tail of the t-distribution:

$$\begin{aligned}
E[\mathbf{1}\{t>\sqrt{m}\alpha\}U^2] &= \int_0^\infty E[U^2\mathbf{1}\{U>\sqrt{m}(\sqrt{v}\alpha-\delta)\}]f_V(v)dv \\
&= \int_0^\infty E[U^2 \mid U>\sqrt{m}(\sqrt{v}\alpha-\delta)]P\big(U>\sqrt{m}(\sqrt{v}\alpha-\delta)\big)f_V(v)dv \\
&\leq 2\int_0^{1/4} E[U^2 \mid U>\sqrt{m}\frac{\alpha}{2}]P\big(U>\sqrt{m}(\alpha\sqrt{v}-\delta)\big)f_V(v)dv \\
&\quad + 2\int_{1/4}^{2} m(\alpha\sqrt{v}-\delta)^2 P\big(U>\sqrt{m}(\alpha\sqrt{v}-\delta)\big)f_V(v)dv \\
&\quad + 2\int_{2}^{\infty} m(\alpha\sqrt{v}-\delta)^2 e^{-m(\alpha\sqrt{v}-\delta)^2} f_V(v)dv \\
&\leq 4m\frac{\alpha^2}{4}\int_0^{1/4} f_V(v)dv + 4m\alpha^2\int_{1/4}^{2} P\big(U>\sqrt{m}(\alpha\sqrt{v}-\delta)\big)f_V(v)dv \\
&\quad + 2\int_2^\infty v f_V(v)dv \\
&\leq m\alpha^2 e^{-na} + 4m\alpha^2 P(t>\sqrt{m}\alpha) + e^{-na},
\end{aligned}$$

where the first inequality follows from (2.12). The result follows.

(ii) : Define $A = \{|\sqrt{V} - 1| \le \frac{1}{8}\}$. Then

$$\begin{aligned} P_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha) &\ge P_{\delta=\alpha/2}(t > \sqrt{m}\alpha) \\ &\ge P\big(U > \sqrt{m}\Big(\alpha\sqrt{v} - \frac{\alpha}{2}\Big), V \in A\big) \\ &\ge P\big(U > \tfrac{5}{8}\sqrt{m}\alpha, V \in A\big) \\ &= P\big(U > \tfrac{5}{8}\sqrt{m}\alpha\big)P(V \in A), \end{aligned}$$

and

$$\begin{aligned} P_{\delta=2\alpha}(|t| < \sqrt{m}\alpha) &\le P(t < \sqrt{m}\alpha) \\ &\le P\big(U < \sqrt{m}(\alpha\sqrt{v} - 2\alpha), V \in A\big) + P(V \in A^C) \\ &\le P\big(U < -\tfrac{7}{8}\sqrt{m}\alpha\big)P(V \in A) + P(V \in A^C). \end{aligned}$$

Thereby:

$$\begin{aligned} \frac{P_{\delta=2\alpha}(|t| < \sqrt{m}\alpha)}{P_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha)} &\le \frac{P(U > \frac{7}{8}\sqrt{m}\alpha)}{P(U > \frac{5}{8}\sqrt{m}\alpha)} + \frac{P(V \in A^C)}{P(U > \frac{5}{8}\sqrt{m}\alpha)P(V \in A)} \\ &\le k_1 e^{-m\alpha^2(\frac{49-25}{2\cdot 64})} + k_2\sqrt{m}\alpha e^{-m(a-\frac{5}{8}\alpha^2)} \\ &\le a_1 e^{-a_2 m\alpha^2}. \end{aligned}$$

(iii) : Since $P_{\delta=2\alpha}\big(\frac{1}{\sqrt{m}}U + \delta < \alpha\big) = P(U > \sqrt{m}\alpha)$ we have

$$\begin{aligned} \frac{P_{\delta=2\alpha}\big(\frac{1}{\sqrt{m}}U + \delta < \alpha\big)}{P_{\delta=\alpha/2}(|t| > \sqrt{m}\alpha)} &\le \frac{P(U > \sqrt{m}\alpha)}{P(U > \frac{5}{8}\sqrt{m}\alpha)P(V \in A)} \\ &\le k e^{-m\alpha^2(1-25/64)/2} \\ &= a_1 e^{-a_2\alpha^2 m}. \end{aligned}$$ □

Bibliography

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.

Bühlmann, P. and S. van de Geer (2011). *Statistics for high-dimensional data: Methods, theory and applications*. Springer.

Cook, D. D. (2009). Upper and lower bounds for the normal distribution function. http://www.johndcook.com/normalbounds.pdf.

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 classification 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.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 58*(1), pp. 267–288.

3

High dimensional classifiers in the imbalanced case

Britta Anker Bak and Jens Ledet Jensen

Abstract

We consider the binary classification problem in the imbalanced case where the number of samples from the two groups differ. The classification problem is considered in the high dimensional case where the number of variables is much larger than the number of samples, and where the imbalance leads to a bias in the classification. A theoretical analysis of the independence classifier reveals the origin of the bias and based on this we suggest two new classifiers that can handle any imbalance ratio. The analytical results are supplemented by a simulation study, where the suggested classifiers in some aspects outperform multiple undersampling. For correlated data we consider the ROAD classifier and suggest a modification of this to handle the bias from imbalanced group sizes. In an appendix we see that oversampling increases the bias of the independence rule.

3.1 Introduction

During the last decade much research in the statistical community has been on classifiers for high dimensional data where the sample size is small, see e.g. Donoho and Jin (2009), Cai and Liu (2011) and Fan et al. (2012). Typically, this research has not focussed on the imbalance problem where the sample sizes of the groups differ. In real life experiments, on the other hand, imbalanced data sets are the norm rather than the exception. Even if scientists decide to collect a balanced data set, missing data due to for example patients dropping out of the experiment or invalid measurements commonly leads to imbalance.

Faced with imbalance most classifiers tend to classify observations from a binary classification problem to the majority group at the expense of the minority group. It appears to be overlooked or neglected that this imbalance problem becomes much more pronounced in high dimensional settings. To briefly illustrate this Table 3.1 gives the mean and standard deviation of the probability of correct classification for both groups in a few instances for the thresholded independence classifier. It is clearly seen that even rather small imbalances seriously harm classification, pointing to the need of correcting for all imbalances.

The imbalance problem has, however, been addressed recently in the computer science and engineering communities. Here the focus has been on reducing to the balanced case by either undersampling or oversampling. Lin et al. (2009), Yang et al. (2014) and Liu et al. (2009) introduced Meta Imbalanced Classification Ensemble (MICE), Sample Subset Optimization (SSO) and BalanceCascade, respectively. Those are all ensemble methods, where several classifiers are build on all observations in the minority group and wisely selected subsamples of the majority group. Chawla et al. (2002) propose a technique where the minority group is extended by adding observations on the line segments between an existing minority observation and its nearest neighbours. The above classifiers are studied empirically rather than theoretically, and are all shown to handle imbalanced classification problems well. Typically, the high dimensional situation is not addressed as a problem in itself.

The aim of the present paper is to analyse the imbalance problem in relation to high dimensional binary classification and, building on this analysis, to suggest classifiers that are not based on undersampling or oversampling. Ideally, we want our classifiers to involve a small number of variables only, while maintaining a high probability of correct classification. To this end we consider a simple classification problem between two groups with independent normally distributed variables. The assumption of independent variables is a simplification in relation to most data sets, but the setting is useful for studying the imbalance problem in high dimensional settings, and the classifiers are also of practical relevance for correlated variables.

After detecting the origin of the bias problem for imbalanced data in Section 3.2, we suggest in Section 3.3 two new classifiers with, practically, no bias. We discuss the properties of the suggested classifiers both theoretically and empirically. Turning to a situation with correlated variables in Section 3.6, we find that the corrections introduced for the case of independent variables can be combined with the ROAD classifier of Fan et al. (2012) for the imbalanced case. This suggests that the introduced correction methods can be helpful for a range of linear classifiers in more general situations. In the appendix in Section 3.8 we show that a general form of oversampling increases the bias instead of reducing it.

3.2 The bias problem for imbalanced data

The model we consider is as follows. Let $x_1, x_2, \dots, x_n$ and $y_1, y_2, \dots, y_m$ be p-dimensional observations from group 1 and group 2, respectively. Assume all observations and variables are independent with distributions $x_{ij} \sim N(\mu_j, \sigma_j^2)$ and $y_{ij} \sim N(\mu_j + \delta_j \sigma_j, \sigma_j^2)$.

Table 3.1: Average probability of correct classification of the thresholded independence classifier for a new observation from each of two groups. There are n samples from group 1 and m samples from group 2. Each observation has 1000 variables of which only 10 have a differential expression of size 1. Values are based on 1000 simulated data sets.

		Group 1		Group 2	
n	m	Mean	Std	Mean	Std
15	15	70.5	7.0	70.3	7.1
16	14	76.8	6.2	63.2	7.7
18	12	87.4	4.5	44.9	8.7
20	10	94.9	2.2	24.7	7.2

Let $\bar{x}_j$ and $\bar{y}_j$ denote the sample means of variable j for each of the two groups, and let s^2_{xj} and s^2_{yj} be the corresponding sample variances. Define the imbalance factor as $\rho = (n - m)/(nm)$, and let $f = n + m - 2$ be the degrees of freedom for the joint sample variance. We call δ_j the (scaled) differential expression.

To describe the independence classifier with thresholding we first define for $j = 1, \ldots, p$

$$s_j^2 = \frac{(n-1)s^2_{xj} + (m-1)s^2_{yj}}{n+m-2}, \qquad t_j = \frac{\bar{y}_j - \bar{x}_j}{\sqrt{s_j^2(1/n + 1/m)}},$$

and let $w(t)$ be a weight function. Hard thresholding, which we use throughout this paper, corresponds to $w(t) = \mathbf{1}\{|t| > \Delta\}$. The independence classifier with thresholding allocates a new observation z to group 1 if $D(z) < 0$ and to group 2 if $D(z) > 0$, where

$$D(z) = \sum_{j=1}^{p} \frac{\bar{y}_j - \bar{x}_j}{s_j^2} \left[z_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\right] w(t_j). \tag{3.1}$$

The probability of correct classification for a new observation from either group 1 or group 2 is

$$\Phi\Big(\frac{\xi_D}{\tau_D}\Big) \quad \text{and} \quad \Phi\Big(\frac{\tilde{\xi}_D}{\tau_D}\Big), \tag{3.2}$$

where $\xi_D = -D(\mu) = \sum_{j=1}^p \xi_{Dj}$, $\tilde{\xi}_D = D(\mu + \delta\sigma) = \sum_{j=1}^p \tilde{\xi}_{Dj}$, $\tau_D^2 = \sum_{j=1}^p w(t_j)(\bar{y}_j - \bar{x}_j)^2 \sigma_j^2 / s_j^4$ and

$$\begin{aligned} \xi_{Dj} &= -\frac{\bar{y}_j - \bar{x}_j}{s_j^2} \left[\mu_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\right] w(t_j), \\ \tilde{\xi}_{Dj} &= \frac{\bar{y}_j - \bar{x}_j}{s_j^2} \left[\mu_j + \delta_j\sigma_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\right] w(t_j). \end{aligned}$$

To describe the means of these terms define

$$T_{a,b}(\delta; n, m) = E\Big[\frac{(d+\delta)^a}{v^b} w(t)\Big],$$

where $d \sim N(0, 1/n + 1/m)$, $v \sim \chi^2(f)/f$ with $f = n + m - 2$ and $t = (d + \delta)/\sqrt{v(1/n + 1/m)}$.

Proposition 3.1. *Let ξ_D^0 and $\tilde{\xi}_D^0$ be generic terms in the sums ξ_D and $\tilde{\xi}_D$. Then*

$$\begin{aligned} E(\xi_D^0) &= \tfrac{1}{2}\left[(1-\rho)\delta T_{1,1}(\delta, n, m) + \rho T_{2,1}(\delta, n, m)\right], \\ E(\tilde{\xi}_D^0) &= \tfrac{1}{2}\left[(1+\rho)\delta T_{1,1}(\delta, n, m) - \rho T_{2,1}(\delta, n, m)\right]. \end{aligned}$$

When $\delta = 0$ we simply get $E(\xi_D^0) = -E(\tilde{\xi}_D^0) = \frac{1}{2}\rho T_{2,1}(0, n, m)$. For the case of no thresholding, $w(t) \equiv 1$, we get in the general case

$$E(\xi_D^0) = \frac{f}{2(f-2)}[\delta^2 + \rho(1/n + 1/m)], \quad E(\tilde{\xi}_D^0) = \frac{f}{2(f-2)}[\delta^2 - \rho(1/n + 1/m)].$$

Proof. Letting $u = (\bar{x} + \bar{y} - 2\mu - \delta)/\sigma \sim N(0, 1/n + 1/m)$, $d = (\bar{y} - \bar{x} - \delta)/\sigma \sim N(0, 1/n + 1/m)$ and $v = s^2/\sigma^2 \sim \chi^2(f)/f$ with $f = n + m - 2$, we can write

$$\xi_D^0 = \frac{d+\delta}{2v}(u+\delta)w(t) \quad \text{and} \quad \tilde{\xi}_D^0 = \frac{d+\delta}{2v}(\delta - u)w(t),$$

with $t = (d+\delta)/\sqrt{v(1/n+1/m)}$. Had u and d been independent, ξ_D^0 and $\tilde{\xi}_D^0$ would have the same mean and there would be no bias problem. However, in the imbalanced case we have

$$u \mid d \sim N\Big(\rho d, \frac{4}{n+m}\Big). \tag{3.3}$$

We then obtain

$$E(\xi_D^0) = E\Big[\frac{d+\delta}{2v}(\rho d+\delta)w(t)\Big] = \tfrac{1}{2}E\Big\{\Big[\rho\frac{(d+\delta)^2}{v} + \delta(1-\rho)\frac{d+\delta}{v}\Big]w(t)\Big\},$$

and $E(\tilde{\xi}_D^0)$ is calculated in the same way.

In the case of no thresholding, $w(t) \equiv 1$, we use that $E(1/v) = f/(f-2)$ so that

$$E\Big(\frac{(d+\delta)^2}{v}\Big) = \Big(\frac{1}{n}+\frac{1}{m}+\delta^2\Big)\frac{f}{f-2} \quad \text{and} \quad E\Big(\frac{d+\delta}{v}\Big) = \delta\frac{f}{f-2}. \qquad \square$$

The case of no differential expression ($\delta = 0$) in the proposition shows that if the expected number $pE(w(t))$ of variables with $\delta = 0$ included in the classifier is nonnegligible, then also the bias of the classifier is nonnegligible with the majority class being strongly favoured. In the general case, with $\delta \neq 0$, the formulae point to a bias in the same direction as in the $\delta = 0$ case. This is seen more directly for the case of no thresholding. Overall, the thresholding does not remove the bias problem for the imbalanced case. This can be seen more clearly from the left part of Figure 3.1. The two dotted curves illustrate the bias for the case of no differential expression. The figure shows the mean for a single term of ξ_D and $\tilde{\xi}_D$, conditional on this term being included in the classifier. The two dashed curves show the bias when the differential expression is one. The virtue of increasing the threshold is that we include much fewer of the $\delta = 0$ cases and keep most of the $\delta = 1$ cases. There are, however, a number of opposing effects. When the threshold is increased, the bias for each of those null cases included actually increases. Also, since the mean of τ_D^{02} is increasing with the threshold, the effect of each of the $\delta = 1$ cases in the probability (3.2) is diminished as the threshold is increased. The right part of Figure 3.1 relates to the classifiers proposed in the next section.

3.3 Bias adjusted classifiers

In this section we describe two ways of circumventing the bias problem in the imbalanced case. The origin of the bias problem is the lack of independence of $\bar{x}_j + \bar{y}_j$ and $\bar{y}_j - \bar{x}_j$ as stated in (3.3).

The first proposal is simply to subtract the conditional mean from (3.3). Thus we consider

$$B_0(z) = \sum_{j=1}^{p} \frac{\bar{y}_j - \bar{x}_j}{s_j^2}\Big[z_j - \frac{1}{2}(\bar{x}_j+\bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j)\Big]w(t_j).$$

Let $\xi_{B_0}^0$ be minus a generic term in the sum with z_j replaced by μ_j, and let $\tilde{\xi}_{B_0}^0$ be a generic term with z_j replaced by $\mu_j + \delta_j\sigma_j$. Then, with calculations as in Proposition 3.1, we find

$$E(\xi_{B_0}^0) = \frac{1-\rho}{2}\delta T_{1,1}(\delta,n,m), \quad \text{and} \quad E\big(\tilde{\xi}_{B_0}^0\big) = \frac{1+\rho}{2}\delta T_{1,1}(\delta,n,m).$$

Most importantly, we see here that the bias originating from those variables with $\delta = 0$ has been removed. However, there remains a bias for variables with $\delta \neq 0$, where now the minority group is favoured.

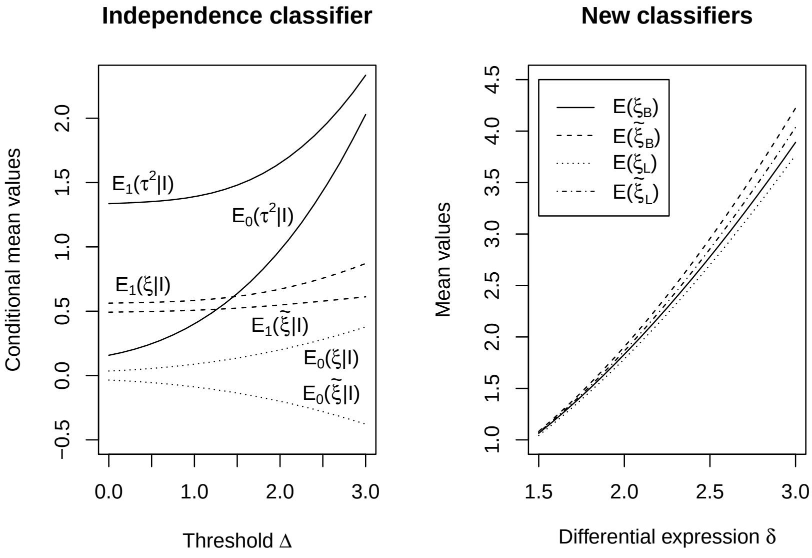

Figure 3.1: The left part shows the mean of a generic term ξ_D^0, $\tilde{\xi}_D^0$ and τ_D^{02} conditionally on the term being included, that is, given that $w(t) = 1$. Two cases of the differential expression are shown: $\delta = 0$ and $\delta = 1$ shown by the subscript on the mean value sign. The right part shows the mean value of ξ and $\tilde{\xi}$ for the two classifiers proposed in Section 3.3. The threshold here depends on the differential expression: $\Delta = \delta/\sqrt{1/n + 1/m} - 1$. In both figures $n = 30$ and $m = 10$.

We therefore consider a classifier on the form $B_0(z) - \epsilon$ for some constant ϵ. Optimally, we want $\xi_{B_0} + \epsilon = \tilde{\xi}_{B_0} - \epsilon$ or $\epsilon = (\tilde{\xi}_{B_0} - \xi_{B_0})/2$. We estimate ξ_{B_0} and $\tilde{\xi}_{B_0}$ by a leave-one-out cross-validation and use these to correct the classifier. To this end we define $B_0(z; x_i)$ to be the classifier based on the reduced sample with x_i excluded and, similarly, $B_0(z; y_i)$ is based on the reduced sample with y_i excluded. Define

$$\bar{\epsilon} = \frac{1}{2}\Big[\frac{1}{n}\sum_{i=1}^{n} B_0(x_i; x_i) + \frac{1}{m}\sum_{i=1}^{m} B_0(y_i; y_i)\Big].$$

Since B_0 is a sum over all p variables, we can also write $\bar{\epsilon}$ as a sum $\bar{\epsilon} = \sum_{j=1}^{p} \bar{\epsilon}_j$, where $\bar{\epsilon}_j$ depends on the j'th coordinate of the data only. The *bias adjusted independence* classifier (BAI classifier) is now defined as

$$B(z) = B_0(z) - \bar{\epsilon} = \sum_{j=1}^{p}\Big\{\frac{\bar{y}_j - \bar{x}_j}{s_j^2}\Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j)\Big]w(t_j) - \bar{\epsilon}_j\Big\}.$$

Defining ξ_B^0 and $\tilde{\xi}_B^0$ as the generic terms of $-B(\mu)$ and $B(\mu + \delta\sigma)$, we see that

$$\begin{aligned} E(\xi_B^0) &= \frac{\delta}{2}\Big\{(1-\rho)T_{1,1}(\delta, n, m) - \frac{1-\rho_1}{2}T_{1,1}(\delta, n-1, m) + \frac{1+\rho_2}{2}T_{1,1}(\delta, n, m-1)\Big\}, \\ E(\tilde{\xi}_B^0) &= \frac{\delta}{2}\Big\{(1+\rho)T_{1,1}(\delta, n, m) + \frac{1-\rho_1}{2}T_{1,1}(\delta, n-1, m) - \frac{1+\rho_2}{2}T_{1,1}(\delta, n, m-1)\Big\}, \end{aligned} \tag{3.4}$$

where $\rho = (n-m)/(n+1)$, $\rho_1 = (n-m-1)/(n+m-1)$ and $\rho_2 = (n-m+1)/(n+m-1)$. Since $\bar{\epsilon}$ is based on one less observation than B_0, the BAI classifier is not exactly unbiased, but the remaining bias is of no practical concern. The bias of the BAI classifier is illustrated in the right part of Figure 3.1 for the case $n = 30$ and $m = 10$. When the differential expression δ is less than 1.5, the bias is very small.

When calculating the probability of correct classification as in (3.2), the denominator is $\tau_B^2 = \sum_{j=1}^{p} w(t_j)(\bar{y}_j - \bar{x}_j)^2\sigma_j^2/s_j^4$, that is, the same expression as τ_D^2.

We next consider a different approach for removing the bias of the independence classifier in the imbalanced case. First, we rewrite the independence classifier as

$$D(z) = \frac{1}{2}\sum_{j=1}^{p}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{y}_j - \bar{x}_j}{s_j^2}(z_j - x_{ij})w(t_j) + \frac{1}{m}\sum_{i=1}^{m}\frac{\bar{y}_j - \bar{x}_j}{s_j^2}(z_j - y_{ij})w(t_j)\Big].$$

The origin of the bias problem, as given in (3.3), is here seen as the lack of independence of x_{ij} (or y_{ij}) and $\bar{y}_j - \bar{x}_j$. We suggest to solve this by removing x_{ij} (or y_{ij}) when calculating the difference $\bar{y}_j - \bar{x}_j$. Thus let $\bar{x}_j(i)$ and $\bar{y}_j(i)$ be the group averages when the i'th observation is left out, and let $s_j^2(x_i)$ and $s_j^2(y_i)$ be the within group variance when either x_i or y_i is left out. The corresponding t-value is denoted either $t_j(x_i)$ or $t_j(y_i)$. The *leave one out* independence classifier (LOUI classifier, originally suggested in Jensen (2006)) is defined as

$$\begin{aligned} L(z) = \frac{1}{2}\sum_{j=1}^{p}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)}(z_j - x_{ij})w(t_j(x_i)) \\ + \frac{1}{m}\sum_{i=1}^{m}\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)}(z_j - y_{ij})w(t_j(y_i))\Big]. \end{aligned}$$

Defining ξ_L^0 and $\tilde{\xi}_L^0$ as a generic term in $-L(\mu)$ and $L(\mu + \delta\sigma)$, we see that

$$E(\xi_L^0) = \tfrac{1}{2}\delta T_{1,1}(\delta, n, m-1) \quad \text{and} \quad E(\tilde{\xi}_L^0) = \tfrac{1}{2}\delta T_{1,1}(\delta, n-1, m).$$

The difference between these two terms is very small so that the LOUI classifier is almost unbiased. An example is shown in the right part of Figure 3.1 for the case $n = 30$ and $m = 10$.

When calculating the probability of correct classification, as in (3.2), the denominator is now

$$\tau_L^2 = \frac{1}{4}\sum_{j=1}^{p}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{\bar{y}_j - \bar{x}_j(i)}{s_j^2(x_i)/\sigma_j^2}w(t_j(x_i)) + \frac{1}{m}\sum_{i=1}^{m}\frac{\bar{y}_j(i) - \bar{x}_j}{s_j^2(y_i)/\sigma_j^2}w(t_j(y_i))\Big]^2,$$

which is somewhat more complicated than for the independence classifier and the BAI classifier.

A comparison of the two proposed classifiers BAI and LOUI is given in Section 3.5.

3.4 Distribution approximation of the error probability

We are mostly interested in situations where the number of variables with a nonzero differential expression is quite small, and the sample sizes n and m are not sufficiently large for a complete separation between the variables with a nonzero differential expression and those with no differential expression. The classifier therefore typically includes a limited number of variables and a part of these are null variables. The probability of correct classification given through ξ/τ and $\tilde{\xi}/\tau$ in (3.2) therefore has a fairly large variance, and part of this variance stems from the variance of the denominator τ. Actually, both ξ and τ turn out to have fairly large variances and a strong correlation.

We want to be able to look at the mean and variance of ξ/τ and $\tilde{\xi}/\tau$ for various combinations of the differential expressions δ_j in an easy computable way for the case of independent variables. This means that we want to use only moment values of generic terms ξ^0, $\tilde{\xi}^0$ and τ^{02}. For this purpose we use the following rough approximation

$$\xi \mid \tau^2 \approx N(\alpha + \beta\tau^2, \omega^2), \;\; \tilde{\xi} \mid \tau^2 \approx N(\tilde{\alpha} + \tilde{\beta}\tau^2, \tilde{\omega}^2), \;\; \tau^2 \approx \Gamma(\lambda, \kappa). \tag{3.5}$$

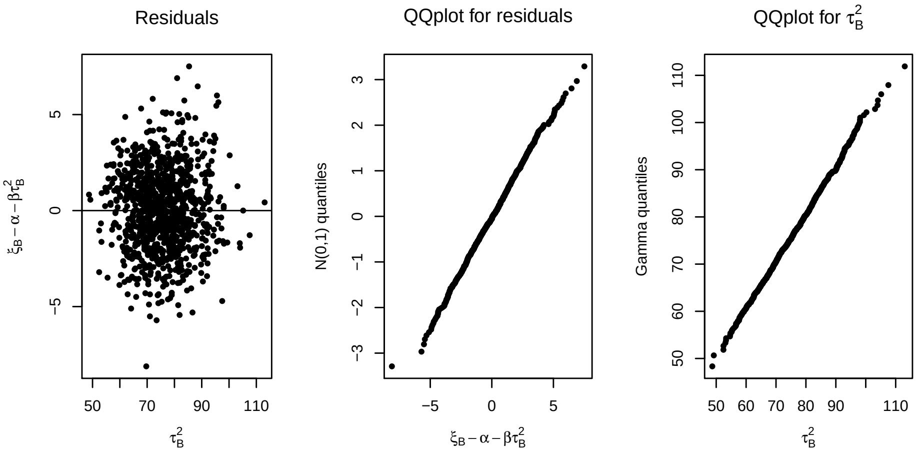

Figure 3.2: Illustration of the approximation (3.5) for the BAI classifier. The 1000 simulated values of ξ_B and τ_B^2 are for the case $n = 30$, $m = 10$, $\delta = 1$ and $\Delta = 2$. There are $p = 1000$ variables of which $k = 20$ are differentially expressed. The left subfigure shows the approximate linearity of the conditional mean of ξ_B given τ_B^2, the center figure shows the conditional normality and the right subfigure illustrates the Gamma approximation to the distribution of τ_B^2.

The approximation is illustrated in Figure 3.2. The left subfigure shows the approximate linear relationship $E(\xi_B \mid \tau_B^2) \approx \alpha + \beta\tau_B^2$, the center figure shows approximate normality of ξ_B given τ_B^2 and the right subfigure shows the Gamma approximation to the distribution of τ_B^2. Plots for the thresholded independence classifier and the LOUI classifier show that the approximation also works well in these cases.

Lemma 3.2. *Under the above approximation* (3.5) *we have*

$$E\Big(\frac{\xi}{\tau}\Big) \approx \alpha\sqrt{\kappa}\frac{\Gamma(\lambda - \frac{1}{2})}{\Gamma(\lambda)} + \beta\frac{\Gamma(\lambda + \frac{1}{2})}{\sqrt{\kappa}\Gamma(\lambda)},$$

$$\operatorname{Var}\Big(\frac{\xi}{\tau}\Big) \approx (\omega^2 + \alpha^2)\frac{\kappa}{\lambda - 1} + \beta^2\frac{\lambda}{\kappa} + 2\alpha\beta - \Big\{E\Big(\frac{\xi_N}{\tau_N}\Big)\Big\}^2,$$

with similar expressions for $\tilde{\xi}$ with (α, β) replaced by $(\tilde{\alpha}, \tilde{\beta})$.

Proof. We have $E(\xi/\tau) = \alpha E(1/\tau) + \beta E(\tau)$ and the first result follows from the properties of a gamma distribution. Next,

$$\begin{aligned}\operatorname{Var}(\xi/\tau) &= \operatorname{Var}(\alpha/\tau + \beta\tau) + E(\omega^2/\tau^2)\\ &= (\omega^2 + \alpha^2)E(1/\tau^2) + \beta^2 E(\tau^2) + 2\alpha\beta - [E(\xi/\tau)]^2.\end{aligned}$$

and the result for the variance again follows from properties of the gamma distribution. □

To use this in practice we choose the parameters in (3.5) from moment relations:

$$\frac{\lambda}{\kappa} = E(\tau^2), \quad \frac{\lambda}{\kappa^2} = \operatorname{Var}(\tau^2), \quad \operatorname{Cov}(\xi, \tau^2) = \beta\operatorname{Var}(\tau^2),$$
$$E(\xi) = \alpha + \beta E(\tau^2), \quad \operatorname{Var}(\xi) = \beta^2\operatorname{Var}(\tau^2) + \omega^2.$$

We write a generic term of the sums ξ_B and τ_B^2 as

$$\xi_B^0 = \xi_{B_0}^0 + \frac{1}{n}\sum_{i=1}^{n} B_0^0(x_i) + \frac{1}{m}\sum_{i=1}^{n} B_0^0(y_i), \quad \tau_B^{02} = \frac{(\bar{y} - \bar{x})^2}{s^4},$$

where $B_0^0(x_i)$ is a generic term in the sum $B_0(x_i; x_i)$ and $B_0^0(y_i)$ is a generic term in the sum $B_0(y_i; y_i)$. The first two moments can be simulated directly from standard normal variables $x_1, \ldots, x_n$ and $y_1, \ldots, y_m$, and calculating all the terms in ξ_B^0. However, the computational complexity can be reduced on writing variances and covariances as sums involving at most two terms from ξ_B^0. In that case we only need to simulate x_1, x_2, $\bar{x}(3) = \sum_{i=3}^n x_i/(n-3)$, $\sum_{i=3}^n (x_i - \bar{x}(3))^2$ (and similar y-terms), and calculate $\xi_{B_0}^0$, $B_0^0(x_1)$ and $B_0^0(x_2)$ from these. To this end, and supplementing the mean values in (3.4), we note the following simplifications.

Proposition 3.3. *For the case of hard thresholding we have the following moment relations:*

$$E(\tau_B^{02}) = T_{2,2}(\delta; n, m), \qquad E(\tau_B^{04}) = T_{4,4}(\delta; n, m),$$

$$E(\xi_{B_0}^{02}) = \Big[\frac{1}{n+m} + \frac{(1-\rho)^2}{4}\delta^2\Big] T_{2,2}(\delta; n, m), \quad E(\xi_{B_0}^0 \tau_B^{02}) = \frac{1-\rho}{2}\delta T_{3,3}(\delta; n, m),$$

$$E[B_0^0(x_1)] = -\frac{1-\rho_1}{2}\delta T_{1,1}(\delta; n-1, m),$$

$$E[B_0^0(x_1)^2] = \Big[1 + \frac{1}{n-1+m} + \frac{(1-\rho_1)^2}{4}\delta^2\Big] T_{2,2}(\delta; n-1, m).$$

Proof. The proof follows the same lines as the proof of Proposition 3.1. The only extra element used is that $E[(u-\rho d)^2 \mid d] = 4/(n+m)$ from (3.3). The requirement of hard thresholding is used for the simplification $w(t)^2 = w(t)$. □

3.4.1 Mean and variance investigations

In Figures 3.3 and 3.4 we compare the independence classifier D, the BAI-classifier B and the LOUI-classifier L. There are k differentiable expressed variables all with the same differential expression $\delta = 1$. We consider the two cases $k = 20$ and $k = 80$. In all cases we have $n = 30$ and $m = 10$. To calculate the mean and variance of ξ/τ we use the approximation in Lemma 3.2. To this end, we must calculate moments of generic terms ξ^0, $\tilde{\xi}^0$ and τ^{02} for the chosen value of δ for the expressed variables, as well as the case $\delta = 0$ for the nonexpressed variables. These moments cannot be calculated analytically, and we use 10^6 simulated values to estimate the moments. Note that the mean values μ_j and variances σ_j^2 do not enter the distribution of ξ^0, $\tilde{\xi}^0$ and τ^{02} so that we can fix these at zero and one, respectively.

In Figure 3.3 the threshold is fixed at $\Delta = 2$, and we consider the dependency on the differential expression δ for the k expressed variables in the range $0 < \delta < 1.5$. We consider the two cases $k = 20$ and $k = 80$, and either $p = 1000$ or $p = 10\,000$ variables. It is clearly seen that the independence classifier D performs much better on the majority group than on the minority group. For both BAI and LOUI there is practically no difference between the two groups, and also practically no difference between BAI and LOUI for the considered range of δ. For this reason only the BAI classifier is shown in Figure 3.3. Taking into account the random variation, and looking at the case $p = 10\,000$, we will indeed encounter simulations where the classifier is worse than a random guess unless the differential expression δ is large. For $k = 20$ and $\delta = 1$ this will happen in approximately 7% of the simulations. For $p = 1000$ variables the classifier is much more useful, although there is a considerable variation in ξ/τ giving a considerable variation in the probability of correct classification.

In Figure 3.4 the differential expression is fixed at $\delta = 1$, and we consider the dependency on the threshold Δ. As in Figure 3.3 the curves for the two classifiers BAI and LOUI as well as the curves for the two groups for each classifier are indistinguisable,

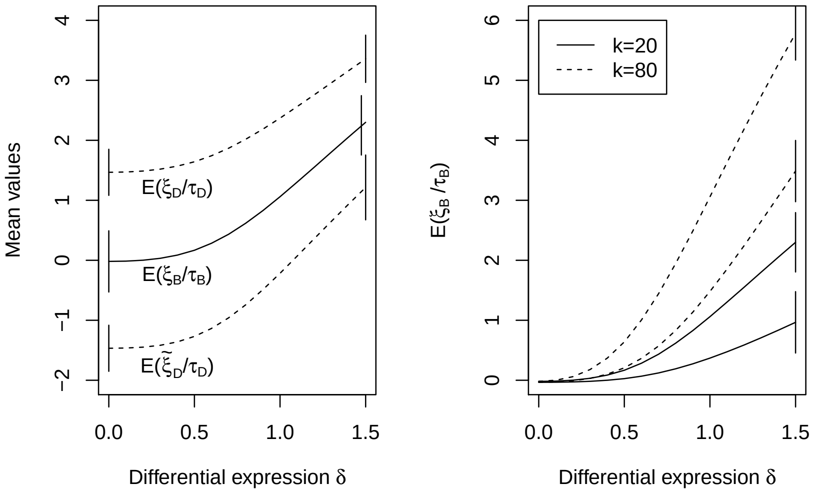

Figure 3.3: Performance of different classifiers for the case $n = 30$, $m = 10$ and with the threshold $\Delta = 2$. In the left part the means of the classification indices ξ_D/τ_D and $\tilde{\xi}_D/\tau_D$ are compared to the mean of ξ_B/τ_B for the case of $p = 1000$ variables with $k = 20$ having the differential expression δ, the remaining variables having no differential expression. In the right part the mean of the classification index ξ_B/τ_B is shown for different values of p and k. The value of k is shown in the legend, and the lower and upper curves of a specific line type correspond to $p = 10000$ and $p = 1000$, respectively. For the chosen settings of the parameters, the means of $\tilde{\xi}_B/\tau_B$, ξ_L/τ_L and $\tilde{\xi}_L/\tau_L$ are indistinguisable from the mean of ξ_B/τ_B. The vertical lines show plus and minus two times the standard deviation.

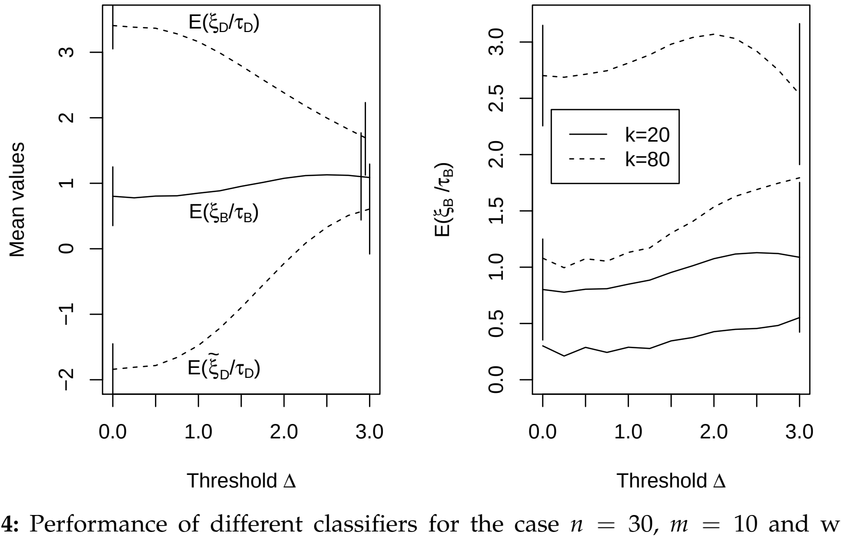

Figure 3.4: Performance of different classifiers for the case $n = 30$, $m = 10$ and with the differential expression $\delta = 1$. In the left part the means of the classification indices ξ_D/τ_D and $\tilde{\xi}_D/\tau_D$ are compared to the mean of ξ_B/τ_B for the case of $p = 1000$ variables with $k = 20$ having the differential expression $\delta = 1$, the remaining variables having no differential expression. In the right part the mean of the classification index ξ_B/τ_B is shown for different values of p and k. The value of k is shown in the legend, and the lower and upper curves of a specific line type correspond to $p = 10000$ and $p = 1000$, respectively. For the chosen settings of the parameters, the means of $\tilde{\xi}_B/\tau_B$, ξ_L/τ_L and $\tilde{\xi}_L/\tau_L$ are indistinguisable from the mean of ξ_B/τ_B. The vertical lines show plus and minus two times the standard deviation.

and only one curve is shown. Clearly, a high threshold reduces the strong bias of the independence classifier D. Still, in most cases the median probability of correct classification for the minority group is below 0.5. Looking for the value of the threshold Δ, where the mean value of ξ_B/τ_B is maximized, no clear optimal choice is seen for the case of $p = 10\,000$ variables. For $p = 1000$ the optimal value is between 2 and 2.5. However, the gain in mean value is partly reduced by having a large spread of ξ_B/τ_B when the threshold is increased.

3.5 Simulations

In this section we report on simulations to compare the suggested classifiers BAI and LOUI for the case of imbalanced data. We include also in the comparison a commonly used undersamling classifier, namely EasyEnsemble from Liu et al. (2009) built on top of the thresholded independence classifier. To write this explicitly, assume $n > m$ and let $D(z; A)$ be the independence classifier from (3.1) based on a subset A of the observations $x_1, \ldots, x_n$ from group 1 and all the observations from group 2, and with $|A| = m$. The undersampling classifier is based on

$$Q(z) = \frac{1}{q}\sum_{i=1}^{q} D(z; A_i), \tag{3.6}$$

where $A_1, \ldots, A_q$ are independent random subsets. In the results in Table 3.2 below we use a value of q such that the probability of using all the samples in the training of the classifier is at least 0.95.

We include the case of a fixed threshold in the comparisons, but we are mostly interested in the situation where the threshold Δ is chosen suitably for each simulated data set. In the simulations we have searched for a value of Δ in the range where a t-test will give between 1 and 30 false positives among p independent tests. For any classifier $H(z)$ we have used a leave-one-out cross-validation to choose Δ. Instead of using the number of correctly classified samples we use a measure that depends continuously on the threshold Δ. Define

$$\hat{\xi} = -\frac{1}{n}\sum_{i=1}^{n} H(x_i; x_i) \quad \text{and} \quad \hat{\tilde{\xi}} = \frac{1}{n}\sum_{i=1}^{m} H(y_i; y_i),$$

where $H(z; x_i)$ is the classifier constructed from the reduced sample with x_i left out and $H(z; y_i)$ defined similarly. Also let $\hat{\tau}^2$ be the empirical variance of the terms that enters $\hat{\xi}$ and $\hat{\tilde{\xi}}$. We then use $\Phi(\hat{\xi}/\hat{\tau})$ and $\Phi(\hat{\tilde{\xi}}/\hat{\tau})$ to choose Δ. Since we often see strong negative correlation between ξ and $\tilde{\xi}$, we have opted against using the average of the two terms for selecting Δ. Instead we use

$$\arg\max_{\Delta} \min\{\Phi(\hat{\xi}/\hat{\tau}), \Phi(\hat{\tilde{\xi}}/\hat{\tau})\}.$$

For the LOUI classifier it is easy to see that $\hat{\xi} = \hat{\tilde{\xi}}$ so that it is immaterial how the two terms are combined to choose Δ. We compare the above cross-validation choice with an optimal oracle selected threshold based on the true mean values, where we maximize $\min\{\Phi(\xi/\tau), \Phi(\tilde{\xi}/\tau)\}$ in the same range of Δ values as in the cross-validation approach.

The numbers in Table 3.2 are based on 1000 simulated data sets for each setting. It is clear from the table that the independence classifier D has an unacceptable large bias, even for the case of the optimal threshold. The bias for each of the LOUI, BAI and EasyEnsemble classifiers is very small, favouring the minority group in the fixed threshold and optimal threshold cases, and favouring the majority group in the cross-validation case. The EasyEnsemble classifier has the smallest bias, but at the same time also the smallest probability of correct classification for both groups, making it less optimal than the BAI and LOUI classifiers. The LOUI classifier typically has a slightly larger probability of correct classification as compared to the BAI classifier. However, this comes at the cost of including many more variables in the classifier. The EasyEnsemble classifier includes even more variables than the LOUI classifier.

Generally, the fixed threshold and the cross-validation threshold gives approximately the same probability of correct classification, but with the use of fewer variables for the cross-validation approach. Also, the cross-validation approach typically lowers the negative correlation between ξ/τ and $\tilde{\tilde{\xi}}/\tau$. For the BAI and LOUI classifiers the optimal threshold gives rise to a fairly large positive correlation. The reason for this is that in many instances the threshold will be chosen close to where the two curves for ξ/τ and $\tilde{\tilde{\xi}}/\tau$, as a function of Δ, intersects, so that the two values are almost identical.

In general, the BAI classifier is our preferred method since the bias is small, it has a comparable good probability of correct classification, and it uses only a small number of the variables for constructing the classifier.

Table 3.2: Comparison of the classifiers D, LOUI, BAI and EasyEnsemble for various values of p, n and m based on 1000 simulated data sets. There are $k = 20$ differential expressed variables with $\delta = 1$. The *Fixed* columns have the threshold fixed at $\Delta = 2.5$ when $p = 1000$ and $\Delta = 3$ when $p = 10000$. In the *CV* columns the threshold is selected by leave-one-out cross-validation for each data set, while *Opt* denotes the optimal threshold calculated from the true parameters.

	D			LOUI			BAI			EasyEnsemble		
p n m	fixed	CV	opt	Fixed	CV	Opt	Fixed	CV	Opt	Fixed	CV	Opt
10^3 30 10												
$E(\xi/\tau)$	2.00	1.55	1.56	1.19	1.20	1.21	1.14	1.14	1.19	1.13	1.10	1.12
$E(\tilde{\tilde{\xi}}/\tau)$	0.32	0.58	0.81	1.25	1.14	1.27	1.18	1.14	1.25	1.14	1.08	1.16
$\mathrm{Std}(\xi/\tau)$	0.23	0.38	0.32	0.29	0.28	0.21	0.29	0.28	0.22	0.26	0.27	0.23
$\mathrm{Cor}(\xi/\tau,\tilde{\tilde{\xi}}/\tau)$	−0.09	−0.10	0.08	−0.28	−0.15	0.42	0.06	−0.13	0.44	−0.17	−0.12	0.06
$E(N)$	28.5	12.3	10.6	68.0	51.7	55.1	28.5	22.6	22.2	165.8	121.7	144.6
$\mathrm{Std}(N)$	4.7	9.1	5.2	7.3	34.0	27.0	4.7	14.1	11.6	11.4	74.3	65.1
10^3 50 10												
$E(\xi/\tau)$	2.28	1.76	1.77	1.30	1.28	1.34	1.24	1.22	1.31	1.20	1.17	1.21
$E(\tilde{\tilde{\xi}}/\tau)$	0.31	0.65	0.87	1.41	1.28	1.41	1.35	1.30	1.39	1.24	1.20	1.26
$\mathrm{Std}(\xi/\tau)$	0.22	0.36	0.30	0.29	0.24	0.21	0.29	0.25	0.21	0.26	0.25	0.23
$\mathrm{Cor}(\xi/\tau,\tilde{\tilde{\xi}}/\tau)$	−0.12	−0.08	0.11	−0.33	−0.11	0.43	0.10	−0.13	0.48	−0.23	−0.12	0.03
$E(N)$	27.9	11.7	10.9	62.3	38.0	50.3	27.9	19.3	22.1	254.0	175.4	226.5
$\mathrm{Std}(N)$	4.4	7.1	4.3	7.0	28.8	24.1	4.4	12.4	10.6	13.6	110.0	93.1
10^4 30 10												
$E(\xi/\tau)$	2.35	1.03	0.99	0.60	0.68	0.64	0.55	0.58	0.63	0.53	0.50	0.52
$E(\tilde{\tilde{\xi}}/\tau)$	−1.22	−0.02	0.21	0.63	0.53	0.68	0.58	0.54	0.68	0.52	0.46	0.52
$\mathrm{Std}(\xi/\tau)$	0.23	0.41	0.35	0.32	0.34	0.24	0.31	0.31	0.24	0.26	0.29	0.23
$\mathrm{Cor}(\xi/\tau,\tilde{\tilde{\xi}}/\tau)$	−0.35	−0.21	0.15	−0.56	−0.21	0.58	−0.15	−0.09	0.58	−0.49	−0.14	0.17
$E(N)$	55.7	6.3	4.5	193.0	63.8	60.1	55.7	17.6	15.5	639.1	177.3	252.7
$\mathrm{Std}(N)$	7.2	7.1	3.2	13.5	50.3	40.8	7.2	14.3	11.5	25.1	146.1	148.6
10^4 50 10												
$E(\xi/\tau)$	2.77	1.29	1.27	0.70	0.84	0.80	0.64	0.72	0.77	0.59	0.59	0.61
$E(\tilde{\tilde{\xi}}/\tau)$	−1.41	0.05	0.26	0.79	0.70	0.86	0.72	0.73	0.85	0.60	0.59	0.64
$\mathrm{Std}(\xi/\tau)$	0.24	0.42	0.35	0.32	0.31	0.25	0.30	0.30	0.25	0.28	0.30	0.25
$\mathrm{Cor}(\xi/\tau,\tilde{\tilde{\xi}}/\tau)$	−0.32	−0.26	0.11	−0.44	−0.18	0.52	−0.02	−0.14	0.52	−0.44	−0.20	0.05
$E(N)$	48.8	6.1	5.0	157.9	54.2	54.4	48.8	18.2	16.0	1066.5	335.2	480.3
$\mathrm{Std}(N)$	6.7	5.5	3.0	12.0	44.3	36.4	6.7	14.3	11.2	30.7	250.7	277.4

3.5.1 Breast Cancer Data

We illustrate the imbalance bias problem with the breast cancer data from Sotiriou et al. (2003). There are 99 women in the study divided into two groups according to their estrogen receptor status. The ER+ group (65 women) are those women where the cancer has receptors for estrogen, and the ER- group (34 women) are those without receptors. In the original data there are 7650 variables, but we use here only the subset with $p = 4327$ variables measured in all 99 samples. One hundred times we split the data into a training set and a test set, the latter consisting of 20 randomly chosen observations from each group. The training set thus has 45 women in the ER+ group and 14 in the ER- goup, an imbalance ratio around 3. The threshold in the different classifiers is chosen through leave-one-out cross-validation, where the range considered corresponds to an expected number of false positives out of 4327 variables to be between 1 and 30.

In Table 3.3 we compare BAI, LOUI and EasyEnsemble to the thresholded independence classifier. The table gives the percentage of correctly classified samples, both when evaluated on the training set and on the test set. As expected, the independence classifier shows no bias on the training set, but has a considerable bias when evaluated on the test set. This bias is removed for all three alternatives BAI, LOUI and EasyEnsemble. The bias correction has the consequence that on the training set BAI, LOUI and EasyEnsemble perform best on the minority group. BAI obtains the same performance as LOUI and EasyEnsemble using much less variables, roughly one half of the variables used in LOUI and one third of the variables used in EasyEnsemble. It seems slightly astonishing for this data set, that although a large number of variables seem to be true positives, the classification error is still around 16%.

Table 3.3: Comparison of the thresholded independence classifier, BAI, LOUI and EasyEnsemble on the Breast Cancer data from Sotiriou et al. (2003). The data are randomly divided into a training set with $n = 45$ and $m = 14$ observations in the two groups ER+ and ER-, and a test set with 20 observations in each group. Numbers in the table are based on 100 random splits. The row N gives the number of variables included in the classifier and the remaining entries are percentage correctly classified samples.

	D		LOUI		BAI		EasyEnsemble	
Variable	Mean	Std	Mean	Std	Mean	Std	Mean	Std
Training ER+	94.8	2.2	89.7	3.7	89.3	3.9	89.6	3.3
Training ER-	92.1	6.2	94.4	5.1	94.4	5.0	94.7	4.6
Test ER+	90.5	5.5	84.1	7.2	83.9	7.1	84.6	6.9
Test ER-	75.5	8.2	83.3	6.2	83.8	5.9	83.0	6.0
N	229	127	409	198	232	121	602	270

3.6 Correlated data: BA-ROAD and LOU-ROAD

In many high dimensional settings the variables will be correlated, and classifiers build on the independence classifier will be suboptimal. The Fisher classifier based on an estimate of the inverse covariance matrix is not directly applicable when $p \gg n$. As an alternative Fan et al. (2012) suggested the *Regularized Optimal Affine Discriminant* (ROAD) classifier based on

$$R(z) = \sum_{j=1}^{p} r_j \big[z_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j) \big],$$

where

$$r = \underset{(\bar{y}-\bar{x})^{\mathsf{T}} r=1, |r|_1 \le c}{\arg\min} \; r^{\mathsf{T}} \hat{\Sigma} r, \tag{3.7}$$

with $\hat{\Sigma}$ the $p \times p$ estimated covariance matrix, and with the tuning parameter c chosen by cross-validation. Fan et al. (2012) introduced an efficient algorithm for calculating r, and simulations with $n = m$ show that ROAD performs better for correlated data as compared to a number of alternative classifiers including the independence classifier. However, as seen from the first two columns of Table 3.4, in the imbalanced case the ROAD classifier can have an appreciable bias. Inspired by the BAI and the LOUI corrections to the independence classifier, we propose the following adjustments to the ROAD classifier. First define

$$B_{0,R}(z) = \sum_{j=1}^{p} r_j \Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) + \frac{\rho}{2}(\bar{y}_j - \bar{x}_j) \Big],$$

$$\bar{\epsilon}_R = \frac{1}{2}\Big[\frac{1}{n} \sum_{i=1}^{n} B_{0,R}(x_i; x_i) + \frac{1}{m} \sum_{i=1}^{m} B_{0,R}(y_i; y_i) \Big],$$

where $B_{0,R}(x_i; x_i)$ and $B_{0,R}(y_i; y_i)$ are defined from $B_{0,R}$ in the same way as $B_0(x_i; x_i)$ and $B_0(y_i; y_i)$ are defined from B_0, that is, $B_{0,R}$ is constructed from a reduced sample with one observation left out and then evaluated on the excluded observation. The BA-ROAD classifier is next defined as

$$B_R(z) = B_{0R}(z) - \bar{\epsilon}_R.$$

In a similar spirit we define the LOU-ROAD classifier as

$$L_R(z) = \frac{1}{2} \sum_{j=1}^{p} \Big[\frac{1}{n} \sum_{i=1}^{n} r_j(x_i)(z_j - x_{ij}) + \frac{1}{m} \sum_{i=1}^{m} r_j(y_i)(z_j - y_{ij}) \Big],$$

where $r(x_i)$ and $r(y_i)$ are calculated as in (3.7) based on the reduced sample with either x_i or y_i left out.

For each of the above classifiers the probability of correct classification is evaluated through ξ, $\tilde{\xi}$ and τ^2 as in (3.2). Here ξ is minus the value of the classifier evaluated at μ, and $\tilde{\xi}$ is the value at $\mu + \delta\sigma$. For both of R and B_R we have $\tau^2 = \sum_{j=1}^{p} \sigma_j^2 r_j^2$, and for L_R the formula becomes

$$\tau^2 = \frac{1}{4} \sum_{j=1}^{p} \sigma_j^2 \Big[\frac{1}{n} \sum_{i=1}^{n} r_j(x_i) + \frac{1}{m} \sum_{i=1}^{m} r_j(y_i) \Big]^2.$$

We evaluate BA-ROAD and LOU-ROAD via a set of simulations. For comparison we include the EasyEnsemble undersampling classifier built on top of ROAD, that is, the classifier (3.6) with D replaced by R. In each simulation the value of c in (3.7) is determined by five-fold cross-validation for each of the classifiers. Also, we include the BAI independence classifier where the threshold Δ is chosen by five-fold cross-validation searching over a region with 5 to 30 expected false positives. We consider the setting with $n = 30$, $m = 10$ and $p = 1000$ variables of which the first 20 variables have differential expression 1, the remaining variables having no differential expression. The numbers in Table 3.4 are based on 100 simulated values. We consider three models for the covariance matrix Σ:

$$
\begin{aligned}
&\text{Model 1:} \quad \Sigma_{ii} = 1, \ \ \Sigma_{ij} = 0.2, \ i \neq j, \\
&\text{Model 2:} \quad \Sigma_{ij} = 0.8^{|i-j|}, \\
&\text{Model 3:} \quad \Sigma = \mathrm{Cor}\Big(\hat{\Sigma}_p + \sqrt{\frac{\log(p)}{n+m}} I_p\Big),
\end{aligned}
$$

where $\hat{\Sigma}_p$ is the empirical variance based on the data in Golub et al. (1999), I_p is the identity matrix and Cor is the function that transforms a variance matrix to a correlation matrix ($\hat{\Sigma}_p$ has been obtained by choosing p consequtive variables where the distribution of the correlations resembles the distribution for all variables).

First of all, Table 3.4 shows that ROAD itself has a considerable bias in the imbalanced case. The bias is almost eliminated with the use of BA-ROAD, LOU-ROAD or the EasyEnsemble-ROAD classifier. Generally, the performance of BA-ROAD is comparable to that of ROAD in terms of the number of variables included in the classifier. LOU-ROAD and EasyEnsemble-ROAD perform slightly better on average, but at the cost of including many more variables than BA-ROAD. In terms of mean values the BAI independence classifier performs as good as the ROAD based classifiers. However, it has a somewhat larger spread. A clear message from this small simulation study is that the bias of the ROAD classifier can be handled by using the classifiers we propose in this paper.

3.7 Conclusion

In this paper we have analyzed the independence classifier in order to study the bias originating from imbalanced data sets. It has been found that a correction for bias is needed also for minor imbalances when considering classification in the high dimensional case. The thresholded independence classifier favours the majority group, and in the high dimensional case this can lead to classifying practically all observations

Table 3.4: Comparison of ROAD, BA-ROAD, LOU-ROAD, EasyEnsemble-ROAD (EE-ROAD) and the BAI independence classifier for the case $n = 30$, $m = 10$ and $p = 1000$ variables of which the first $k = 20$ have differential expression $\delta = 1$. Values are based on 100 simulated data sets. The variable N is the number of variables included in the classifier and *Cor* is the correlation between ξ/τ and $\tilde{\xi}/\tau$.

		ROAD		LOU-ROAD		BA-ROAD		EE-ROAD		BAI	
Model	Variable	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
1	ξ/τ	1.28	0.33	1.06	0.26	0.92	0.27	1.16	0.28	1.06	0.52
	$\tilde{\xi}/\tau$	0.40	0.25	1.05	0.28	0.84	0.29	1.12	0.31	1.34	0.55
	N	24	21	138	65	35	20	147	63	15	12
	Cor	−0.22		−0.10		−0.22		−0.01		−0.00	
2	ξ/τ	0.97	0.27	0.63	0.32	0.57	0.28	0.77	0.37	1.04	0.73
	$\tilde{\xi}/\tau$	0.22	0.31	0.66	0.35	0.58	0.32	0.74	0.35	1.13	0.52
	N	10	11	49	57	16	18	69	71	15	9
	Cor	−0.26		−0.33		−0.22		−0.11		0.10	
3	ξ/τ	1.45	0.29	1.22	0.21	1.13	0.25	1.23	0.23	1.13	0.48
	$\tilde{\xi}/\tau$	0.74	0.25	1.28	0.29	1.12	0.28	1.21	0.33	1.21	0.61
	N	28	18	116	64	34	17	105	64	16	18
	Cor	0.19		0.04		0.06		−0.05		−0.49	

to the majority group. The two suggested classifiers virtually remove the bias and have almost the same error rate.

The BAI classifier performs better in the sense that it obtains the same error rate as the LOUI classifier using much fewer variables. This can be of some practical value when implementing a classifier as a diagnostic tool in a medical setting. Simulations reveal that both classifiers have a slightly lower error rate than a variant of multiple undersampling, which is currently considered among the best methods for correcting imbalance (Blagus and Lusa, 2013). Multiple undersampling uses a high number of variables which also makes it less attractive.

For the case of correlated variables the ROAD classifier turns out to have a bias in the imbalanced case. We have suggested a modification of the ROAD classifier that removes the bias, and simulations show a good performance of this classifier. Overall, our way of correcting for bias seems of value for a broad range of linear classifiers.

3.8 Appendix: Oversampling can increase bias

In Section 3.5 we saw an example of undersampling, meaning that the dataset is made balanced by removing observations from the majority group. The opposite strategy, where observations are added to the minority group, is called oversampling. The added observations can either be replications of original minority observations, or generated from the minority data in a more complicated way.

We consider applying oversampling when using the independence classifier for two normally distributed groups. We show that the bias in classification is increased for a range of oversampling procedures compared to when performing no correction for imbalance.

We consider the setup of Section 3.2, but for simplicity we further assume σ_j is known and equal to one for all $j = 1, \dots p$. The classifier in this case is

$$D(z) = \sum_{j=1}^{p} (\bar{y}_j - \bar{x}_j)\big(z_j - \tfrac{1}{2}(\bar{x}_j + \bar{y}_j)\big) w(u_j) \qquad \text{with } u_j = (\bar{y}_j - \bar{x}_j)/\sqrt{1/n + 1/m}.$$

Let $y_1, \dots, y_m$ denote the original minority observations, while $y_{m+1}, \dots, y_n$ are the added observations. Assume that the average of all minority observations has the form

$$\bar{y}_{OS} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\Big(\sum_{i=1}^{m} a_i y_i + V\Big), \qquad \sum_{i=1}^{m} a_i = n, \tag{3.8}$$

where $a_i \geq 1$ are random, V is random with $EV = 0$, and a, V and y are independent.

As in Section 3.2 we consider general terms ζ^0_{OS} and $\tilde{\zeta}^0_{OS}$, where $OS(z)$ is the classifier $D(z)$ with $\bar{y}$ replaced by $\bar{y}_{OS}$. For simplicity we consider a term corresponding to a variable with $\delta = 0$.

Given a, and defining $\mathrm{Var}(V) = \tau^2$, we have

$$\mathrm{Var}(\bar{y}_{OS}) = \frac{1}{n^2}\big(\sum_{i=1}^{m} a_i^2 + \tau^2\big) \geq \frac{1}{n^2}\Big(\frac{n^2}{m} + \tau^2\Big) \geq \frac{1}{m} = \mathrm{Var}(\bar{y}),$$

where the first inequality follows from $\sum_{i=1}^{m} a_i^2 \geq m\bar{a}^2 = n^2/m$. We write b_a for this conditional variance of $\bar{y}_{OS}$.

Given a, the conditional mean of

$$2(\bar{y}_{OS} - \bar{x})(\mu - \tfrac{1}{2}(\bar{x} + \bar{y}_{OS})\mathbf{1}\{|\bar{y}_{OS} - \bar{x}| > \Delta\}$$

is

$$\Big(b_a - \frac{1}{n}\Big)E\big(d_a^2\mathbf{1}\{|d_a| > \tilde{\Delta}\}\big),$$

where

$$d_a = \frac{\bar{y}_{OS} - \bar{x}}{\sqrt{b_a + 1/n}}, \quad \text{and} \quad \tilde{\Delta} = \frac{\Delta}{\sqrt{b_a + 1/n}}.$$

From Proposition 3.1 it is known that the independence classifier without oversampling gives a similar term with the factor $b_a - 1/n$ replaced by $1/m - 1/n$. Thus, for a given a, when we scale Δ such that the probability of including the variable is the same as for the independence classifier $D(z)$, the difference between $E(\bar{\xi}^0_{OS})$ and $E(\tilde{\xi}^0_{OS})$ is larger than the difference between E($\bar{\xi}^0_D$) and $E(\tilde{\xi}^0_D)$. That is, oversampling procedures fulfilling (3.8) increase the imbalance bias in classification.

We now give a few examples of such oversampling procedures.

Example 1: Random oversampling (ROS). Random oversampling, where $y_{m+1}, \ldots, y_n$ are randomly drawn from $y_1, \ldots, y_m$ with replacement. In this case $V = 0$, and a_i is the number of times, y_i is in the dataset.

Example 2: SMOTE-like oversampling. Each added observation is obtained by randomly choosing two (different) samples y_r and y_s from $y_1, \ldots, y_m$, and using $\alpha y_r + (1-\alpha)y_s$ as the new observation, where α is uniformly distributed on $(0, 1)$. In this case $V = 0$, and a_i is one plus the sum of the α and $1 - \alpha$ terms for those cases where y_i has been chosen.

Note that the original SMOTE of Chawla et al. (2002) does not fit into our framework since the a_i's and y_i's are not independent, when pairs of observations are selected through a nearest neighbour algorithm. Most likely, though, the original SMOTE will have the same bias problem as our SMOTE-like version. For further discussion of the SMOTE-method in a general setting, see Blagus and Lusa (2013).

Example 3: RWO-like oversampling. This procedure is inspired by the idea in Zhang and Li (2014) of generating artificial observations by adding normally distributed noise to existing observations. More specifically, the procedure consists of choosing randomly two (different) samples y_r and y_s from $y_1, \ldots, y_m$, and using $(y_r + y_s)/2 + U/\sqrt{2}$ as the new sample, where $U \sim N(0, 1)$. In this case V is proportional to a sum of $n - m$ standard normally distributed variables, and $a_i = 1 + q_i/2$, where q_i is the number of times y_i has been chosen.

3.8.1 Simulations

In Table 3.5 we compare the independence classifier with the oversampling procedures from Example 1-3 through simulations. For comparison we include also SMOTE of the form in Chawla et al. (2002). Contrary to the theoretical analysis above we do not consider the variance to be known, and all classifiers are build from D in (3.1). For the independence classifier we fix the threshold at 2.5 and 3.0 for $p = 1000$ and $p = 10000$, respectively. In the oversampling methods we select the threshold such that the number of included variables is on average close to the number of variables included in the independence classifier.

Our results clearly show that all classifiers constructed through oversampling increase the bias. One reason is that a slightly smaller fraction of the included variables are true detections.

There is no major difference in the behaviour of the classifiers build on SMOTE and SMOTE-like oversampling. Even though SMOTE does not fit into our framework, it still appears useless in correcting for imbalance.

Table 3.5: Comparison of D and the oversampling procedures ROS, SMOTE-like, SMOTE and RWO-like. All numbers are calculated over 1000 repetitions with $n = 30$, $m = 10$, and $k = 20$ differential expressed variables with $\delta = 1$. N is the number of variables included in the classifier, while N_{true} is the number of variables with $\delta = 1$ included in the classifier.

	D		ROS		SMOTE-like		SMOTE		RWO-like	
	Mean	Std	Mean	Std	Mean	Std	Mean	Std	Mean	Std
$p = 1000$										
$E[\xi/\tau]$	1.96	0.24	2.01	0.28	2.03	0.26	2.00	0.26	2.11	0.34
$E[\tilde{\xi}/\tau]$	0.40	0.32	0.11	0.37	0.18	0.34	0.23	0.34	0.06	0.34
N	25.50	4.47	25.02	5.27	26.58	5.00	25.93	4.80	26.53	4.77
N_{true}	11.54	2.17	10.28	2.25	10.95	2.17	10.98	2.17	10.62	2.20
Δ	2.5		3.9		4.0		4.0		3.8	
$p = 10000$										
$E[\xi/\tau]$	2.18	0.23	2.35	0.34	2.32	0.30	2.26	0.27	2.51	0.34
$E[\tilde{\xi}/\tau]$	−1.05	0.29	−1.38	0.41	−1.29	0.37	−1.21	0.34	−1.48	0.34
N	46.19	6.80	45.80	10.60	46.95	9.17	45.20	8.25	46.93	7.64
N_{true}	7.63	2.22	6.49	2.14	6.92	2.11	6.92	2.11	6.82	2.10
Δ	3.0		4.7		4.8		4.8		4.6	

Bibliography

Blagus, R. and L. Lusa (2013). Smote for high-dimensional class-imbalanced data. *BMC Bioinformatics 14*(106).

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Chawla, N. V., K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer (2002). Smote: Synthetic miniority over-sampling technique. *Journal of Artificial Intelligence Research 16*, 321–357.

Donoho, D. and J. Jin (2009). Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Philosophical transactions of the royal society A 367*, 4449–4470.

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Jensen, J. L. (2006). Maximum likelihood classifiers in microarray studies. Research Report 474, University of Aarhus.

Lin, S.-C., Y.-c. I. Chang, and W.-N. Yang (2009). Meta-learning for imbalanced data and classification ensemble in binary classification. *Neurocomputing 73*(1-3), 484–494.

Liu, X.-Y., J. Wu, and Z.-H. Zhou (2009). Exploratory undersampling for class-imbalance learning. *IEEE Transations on Sysetems, Man, and Cybernetics Part B: Cybernetics 39*(3), 539–550.

Sotiriou, C., S.-Y. Neo, L. M. McShane, E. L. Korn, P. M. Long, A. Jazaeri, P. Martiat, S. B. Fox, A. L. Harris, and E. T. Liu (2003). Breast cancer classification and prognosis based on gene expression profiles from a population-based study. *Proceedings of the National Academy of Sciences 100*(18), 10393–10398.

Yang, P., P. D. Yoo, J. Fernando, B. B. Zhou, Z. Zhang, and A. Y. Zomaya (2014). Sample subset optimization techniques for imbalanced and ensemble learning problems in bioinformatics applications. *IEEE Transations on Cybernetics 44*(3), 445–455.

Zhang, H. and M. Li (2014). Rwo-sampling: A random walk over-sampling approach to imbalanced data classification. *Information Fusion 20*(1), 99–116.

4

On oracle efficiency of the ROAD classification rule

Britta Anker Bak and Jens Ledet Jensen

Abstract

For high-dimensional classification Fishers rule performs poorly due to noise from estimation of the covariance matrix. Fan et al. (2012) introduced the ROAD classifier that puts an ℓ_1-constraint on the classification vector. In their Theorem 1 Fan et al. (2012) show that the ROAD classifier asymptotically has the same misclassification rate as the corresponding oracle based classifier. Unfortunately, the proof contains an error. Here we restate the theorem and provide a new proof.

4.1 Introduction

We consider classification among two groups based on a p-dimensional normally distributed variable. Let the means in the two groups be μ_1 and μ_2, and let the common covariance matrix be Σ. Also, let the probability of belonging to either of the two groups be $\frac{1}{2}$. Defining $\mu_a = (\mu_1 + \mu_2)/2$ and $\mu_d = (\mu_1 - \mu_2)/2$, the Bayes discriminant rule becomes

$$\delta_w(x) = 1 + \mathbf{1}\{w^T(x - \mu_a) < 0\}, \qquad \text{with } w = w_F = \Sigma^{-1}\mu_d,$$

where x is classified to group 1 or 2 according to the value of $\delta_w(x)$. The misclassification rate of the rule δ_w is

$$W(\delta_w) = \bar{\Phi}\big(\tfrac{1}{2}w^T\mu_d/(w^T\Sigma w)^{1/2}\big),$$

where $\bar{\Phi}(z) = 1 - \Phi(z)$ is the upper tail probability of a standard normal distribution. The interpretation of the Bayes rule is that w_F is the vector that minimizes the misclassification rate. Fan et al. (2012) suggest to use an ℓ_1 regularized version of w_F, that is,

$$w_c = \mathop{\arg\min}_{|w|_1 \le c,\ w^T\mu_d = 1} w^T\Sigma w.$$

Its sample version

$$\hat{w}_c = \mathop{\arg\min}_{|w|_1 \le c,\ w^T\hat{\mu}_d = 1} w^T\hat{\Sigma} w,$$

yields the ROAD classifier

$$\hat{\delta} = 1 + \mathbf{1}\{\hat{w}_c^T(x - \hat{\mu}_a) < 0\}.$$

Theorem 1 of Fan et al. (2012) states that the misclassification rate $W(\hat{\delta})$ of the ROAD classifier approaches the misclassification rate of the oracle classifier $W(\delta_{w_c})$. Unfortunately, an essential step in the proof uses an inequality which is not valid, see Section 4.3 for details. We reformulate the theorem and give a new proof.

For a matrix A, define $|A|_\infty = \max_{i,j} A_{i,j}$.

Theorem 4.1. *Let ϵ be a positive constant such that $\max_j\{|\mu_{dj}|\} > \epsilon$, and $c > \epsilon + 1/\max_j\{|\mu_{dj}|\}$. Let a_n be a sequence tending to zero such that $|\hat{\Sigma} - \Sigma|_\infty = O_p(a_n)$, and $|\hat{\mu}_i - \mu_i|_\infty = O_p(a_n)$, $i = 1, 2$. Then, as $n \to \infty$:*

$$W(\hat{\delta}) - W(\delta_{w_c}) = O_p(d_n),$$

with $d_n = c^2 a_n(1 + c^2|\Sigma|_\infty)$.

Prior to proving the theorem we comment on the differences compared to Theorem 1 of Fan et al. (2012). Contrary to us, Fan et al. (2012) require that the smallest eigenvalue of Σ is bounded from below. The upper bound on $W(\hat{\delta}) - W(\delta_{w_c})$ in Fan et al. (2012) depends on the sparsity of w_c and of $w_c^{(1)}$, where $w_c^{(1)}$ is given by

$$w_c^{(1)} = \underset{|w|_1 \le c,\ w^T\hat{\mu}_d = 1}{\arg\min} \quad w^T \Sigma w,$$

whereas our bound depends on the regularizing parameter c only. In the formulation of both theorems c is allowed to depend on n. We require a lower bound on $\max_j\{|\mu_{dj}|\}$, which is not part of the theorem in Fan et al. (2012). However, it enters indirectly in that we must have $c > 1/\max_j\{|\mu_{dj}|\}$ in order for w_c to exist. Thus, if $\max_j\{|\mu_{dj}|\} \to 0$, we have $c \to \infty$, and c enters the upper bound of Fan et al. (2012). The reason for our more restrictive condition $c > \epsilon + 1/\max_j\{|\mu_{dj}|\}$ is that the theorem only makes sense if $\hat{w}_c$ exists with probability tending to one. Similarly, whereas Fan et al. (2012) have the condition $|\hat{\mu}_d - \mu_d|_\infty = O_p(a_n)$, we have $|\hat{\mu}_i - \mu_i|_\infty = O_p(a_n)$, $i = 1, 2$, in order to handle a term in the misclassification rate that has been neglected in Fan et al. (2012). Finally, $|\Sigma|_\infty$ appears in our bound. However, requiring that the variances Σ_{ii}, $i = 1, \ldots, p$, are bounded is often encountered in high dimensional settings.

4.2 Proof of Theorem 1

In the proof we use the following inequalities:

$$|\bar{\Phi}(a(1+\epsilon)) - \bar{\Phi}(a)| \le 2\epsilon \text{ for } a > 0 \text{ and } |\epsilon| < 1, \tag{4.1}$$

$$|\bar{\Phi}((a+\epsilon)^{-1/2}) - \bar{\Phi}(a^{-1/2})| \le \epsilon \text{ for } a > 0 \text{ and } a + \epsilon > 0. \tag{4.2}$$

The misclassification rate consists of two terms corresponding to an observation from each of the two groups. The proofs for the two terms are identical, so to simplify we consider the misclassification rate of an observation from group 1 only. Using (4.1) the misclassification rate of $\hat{\delta}$ becomes

$$\begin{aligned} W(\hat{\delta}) &= \bar{\Phi}\Big(\frac{1}{2}\frac{\hat{w}_c^T\hat{\mu}_d + \hat{w}_c^T(\hat{\mu}_1 - \mu_1)}{\sqrt{\hat{w}_c^T\Sigma\hat{w}_c}}\Big) = \bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{\hat{w}_c^T\Sigma\hat{w}_c}}\Big) + O(|\hat{w}_c^T(\hat{\mu}_1 - \mu_1)|) \\ &\le \bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{\hat{w}_c^T\Sigma\hat{w}_c}}\Big) + O(c|\hat{\mu}_1 - \mu_1|_\infty). \end{aligned} \tag{4.3}$$

Next,

$$|\hat{w}_c^T \Sigma \hat{w}_c - \hat{w}_c^T \hat{\Sigma} \hat{w}_c| \le c^2 |\hat{\Sigma} - \Sigma|_\infty,$$

and from (4.2) we get

$$\bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{\hat{w}_c^T \Sigma \hat{w}_c}}\Big) = \bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{\hat{w}_c^T \hat{\Sigma} \hat{w}_c}}\Big) + O(c^2 |\hat{\Sigma} - \Sigma|_\infty). \tag{4.4}$$

From the proof in Fan et al. (2012) we see that

$$|\hat{w}_c^T \hat{\Sigma} \hat{w}_c - w_c^{(1)T} \Sigma w_c^{(1)}| \le c^2 |\hat{\Sigma} - \Sigma|_\infty,$$

and thus

$$\bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{\hat{w}_c^T \hat{\Sigma} \hat{w}_c}}\Big) = \bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{w_c^{(1)T} \Sigma \hat{w}_c^{(1)}}}\Big) + O(c^2 |\hat{\Sigma} - \Sigma|_\infty). \tag{4.5}$$

Combining (4.3–4.5) we have

$$W(\hat{\delta}) = \bar{\Phi}\Big(\frac{1}{2}\frac{1}{\sqrt{w_c^{(1)T} \Sigma \hat{w}_c^{(1)}}}\Big) + O(c^2 |\hat{\Sigma} - \Sigma|_\infty + c|\hat{\mu}_1 - \mu_1|_\infty). \tag{4.6}$$

Since the oracle misclassification rate is $W(\delta_{w_c}) = \bar{\Phi}\big(1/(2\sqrt{w_c^T \Sigma w_c})\big)$, we need to compare $w_c^T \Sigma w_c$ with $w_c^{(1)T} \Sigma \hat{w}_c^{(1)}$.

To this end let

$$A_1 = \{w : w^T \mu_d = 1,\ |w|_1 \le c\},$$
$$A_2 = \{w : w^T \hat{\mu}_d = 1,\ |w|_1 \le c\}.$$

We want to show that for any $w \in A_1$ there exists $\tilde{w} \in A_2$ such that $w^T \Sigma w$ is close to $\tilde{w}^T \Sigma \tilde{w}$ and vice versa. This means that the minimum of $w^T \Sigma w$ over the set A_1 is close to the minimum over the set A_2.

Let $w \in A_1$, and define $\tilde{w} = w/(w^T \hat{\mu}_d)$. If $|\tilde{w}|_1 \le c$, we have $\tilde{w} \in A_2$, and

$$w^T \Sigma w = (w^T \hat{\mu}_d)^2 \tilde{w}^T \Sigma \tilde{w} = \big(1 + O(c|\hat{\mu}_d - \mu_d|_\infty)\big)^2 \tilde{w}^T \Sigma \tilde{w}.$$

If instead $|\tilde{w}|_1 > c$, we first define $\bar{w} \in A_1$ and then $w^* = \bar{w}/(\bar{w}^T \hat{\mu}_d) \in A_2$. To define $\bar{w}$ assume without loss of generality that $\mu_{d1} = \max_j\{|\mu_{dj}|\}$. Write $w = (w_1, w_{(2)})$, where $w_{(2)}$ is $(p-1)$-dimensional, and define $\bar{w} = (\bar{w}_1, r w_{(2)})$ with $0 < r < 1$, and $\bar{w}_1$ chosen such that $\bar{w}^T \mu_d = 1$. The latter requirement implies

$$\bar{w}_1 \mu_{d1} = 1 - r w_{(2)}^T \mu_{d(2)} = 1 - r(1 - w_1 \mu_{d1}).$$

We will show that with $r = 1 - c^2|\hat{\mu}_d - \mu_d|_\infty/(c - 1/\mu_{d1}) = 1 - O(c^2|\hat{\mu}_d - \mu_d|_\infty)$, we have $|w^*|_1 \le c$. From the definition of $\bar{w}$ we have

$$|\bar{w}|_1 = |\bar{w}_1| + r|\bar{w}_{(2)}|_1 = \frac{|1 - r(1 - w_1 \mu_{d1})|}{\mu_{d1}} + r(|w|_1 - |w_1|).$$

If $1 - r(1 - w_1 \mu_{d1}) > 0$ we get

$$|\bar{w}|_1 = \frac{1}{\mu_{d1}} + r\big(|w|_1 - \frac{1}{\mu_{d1}} + w_1 - |w_1|\big) \le \frac{1}{\mu_{d1}} + r\Big(c - \frac{1}{\mu_{d1}}\Big).$$

This shows that $\bar{w} \in A_1$ and $w^* \in A_2$ since

$$|w^*|_1 = \frac{|\bar{w}|_1}{\bar{w}^T \hat{\mu}_d} \le \frac{\frac{1}{\mu_{d1}} + r\big(c - \frac{1}{\mu_{d1}}\big)}{1 - c|\hat{\mu}_d - \mu_d|_\infty} \le c,$$

when $r \leq 1 - c^2|\hat{\mu}_d - \mu_d|_\infty/(c - 1/\mu_{d1})$. If instead $1 - r(1 - w_1\mu_{d1}) < 0$ we find

$$|\bar{w}|_1 = \frac{r-1}{\mu_{d1}} + r|w|_1 \leq rc - \frac{1-r}{\mu_{d1}} \leq rc,$$

and $|w^*|_1 \leq rc/(1 - c|\hat{\mu}_d - \mu_d|_\infty) \leq c$ for $r \leq 1 - c|\hat{\mu}_d - \mu_d|_\infty$. The latter condition is satisfied with $r \leq 1 - c^2|\hat{\mu}_d - \mu_d|_\infty/(c - 1/\mu_{d1})$. Comparing $\bar{w}$ and w we get

$$\begin{aligned}|w^T\Sigma w - \bar{w}^T\Sigma\bar{w}| &\leq 2c|w - \bar{w}|_1|\Sigma|_\infty \leq 2c[(1-r)|w|_1 + (1-r)\frac{1}{\mu_{d1}}]|\Sigma|_\infty \\ &= O(c^4|\hat{\mu}_d - \mu_d|_\infty|\Sigma|_\infty),\end{aligned}$$

and also

$$|\bar{w}^T\Sigma\bar{w} - w^{*T}\Sigma w^*| \leq (w^{*T}\Sigma w^*)O(c|\hat{\mu}_d - \mu_d|_\infty).$$

We have now shown that any value of $w^T\Sigma w$ for $w \in A_1$ is close to the corresponding value for some $\tilde{w} \in A_2$. The other way around, starting with $w \in A_2$, is treated in the same way. The only difference is that instead of using $c - 1/\mu_{d1} > \epsilon$, we use that when $|\hat{\mu}_{d1} - \mu_{d1}| < \min\{\epsilon, \epsilon^3/(2+\epsilon^2)\}$, which happens with probability tending to 1 (exponentially fast), we have $\hat{\mu}_{d1} > 0$ and $c - 1/\hat{\mu}_{d1} > \epsilon/2$. Therefore, the minimum $w_c^T\Sigma w_c$ of $w^T\Sigma w$ over the set A_1 is close to the minimum $w_c^{(1)T}\Sigma w_c^{(1)}$ over the set A_2:

$$\begin{aligned}w_c^T\Sigma w_c &= w_c^{(1)T}\Sigma w_c^{(1)} + O(c^4|\hat{\mu}_d - \mu_d|_\infty|\Sigma|_\infty) + O(c|\hat{\mu}_d - \mu_d|_\infty w_c^T\Sigma w_c) \\ &= w_c^{(1)T}\Sigma w_c^{(1)} + O(c^4|\hat{\mu}_d - \mu_d|_\infty|\Sigma|_\infty).\end{aligned}$$

Combining the latter with (4.6) we conclude

$$|W(\hat{\delta}) - W(\delta_{w_c})| = O\big(c^2a_n(1 + c^2|\Sigma|_\infty)\big).$$

4.3 Appendix

An essential step in the proof in Fan et al. (2012) is the inequality (used in equation (21) of that paper)

$$\frac{w_c^T\hat{\mu}_d}{\sqrt{w_c^T\Sigma w_c}} \leq \frac{1}{\sqrt{w_c^{(1)T}\Sigma w_c^{(1)}}}.$$

Unfortunately, this inequality is not correct. We illustrate this by a concrete example. We consider the two-dimensional case with

$$\mu_d = (1,0)^T, \quad \Sigma = \begin{pmatrix}1 & 1\\ 1 & \sigma\end{pmatrix}, \quad c = 1+\epsilon \qquad \text{with } \epsilon < 1/\sigma.$$

In this case we have

$$w_c = (1, -\epsilon)^T \quad \text{and} \quad w_c^T\Sigma w_c = 1 - 2\epsilon + \sigma\epsilon^2.$$

Consider next $\hat{\mu}_d = (1+a, b)$ with a and b small. For a and b sufficiently small we obtain

$$w_c^{(1)} = \Big(\frac{1 + b[a + \epsilon(1+a)]/(1+a+b)}{1+a}, -\frac{a + \epsilon(1+a)}{1+a+b}\Big)^T,$$

and

$$w_c^{(1)T}\Sigma w_c^{(1)} = (w_{c1}^{(1)})^2 + 2w_{c1}^{(1)}w_{c2}^{(1)} + \sigma(w_{c2}^{(1)})^2.$$

For a and b small and including $O(a)$ and $O(b)$ terms only we get

$$\frac{1}{\sqrt{w_c^{(1)T}\Sigma w_c^{(1)}}} = \frac{1}{\sqrt{1-2\epsilon+\sigma\epsilon^2}}\{1+a-b\epsilon+(a-b\epsilon)\frac{1+\epsilon-\epsilon\sigma(1+\epsilon)}{1-2\epsilon+\sigma\epsilon^2}\}, \tag{4.7}$$

which must be compared to

$$\frac{w_c^T\hat{\mu}_d}{\sqrt{w_c^T\Sigma w_c}} = \frac{1+a-b\epsilon}{\sqrt{1-2\epsilon+\sigma\epsilon^2}}. \tag{4.8}$$

We thus see that (4.7) is less that (4.8) when $a-b\epsilon$ has the opposite sign of $1+\epsilon-\epsilon\sigma(1+\epsilon)$. Since $(a-b\epsilon)\sim N\big(0,(1-2\epsilon+\sigma\epsilon)c_0\big)$ for some constant c_0, the probability of a particular sign of $a-b\epsilon$ is one half.

Bibliography

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

5

A numeric comparison of sparse linear classifiers incorporating covariance

Britta Anker Bak and Jens Ledet Jensen

Abstract

Recent results have shown incorporating covariance into linear classifiers in highdimensional settings can improve their performance. In this report, we explore the properties of a new proposal, the linear lasso discriminant, along with three existing classifiers: Regularized optimal affine discriminant, linear programming discriminant, and sparse linear discriminant. The comparison is done through an extensive simulation study in a variety of settings, as well as on two datasets. The study shows that no classifier is generally the best with respect to classification error. In most situations the computational cost can be reduced by a preliminary selection procedure without losing much in performance. In our implementation the regularized optimal affine discriminant (ROAD) is much faster to compute compared to the other procedures, especially when the dimension is large.

5.1 Introduction

Scientific and technological developments during the last decades have caused the emergence of high dimensional data in a large variety of fields, including medical imaging (Garzon et al., 2011) and microarray analysis (Dyrskjøt et al., 2003). This is a challenge to the statistical community since it leads to situations where the number of variables p is much larger than the number of observations n, invalidating traditional assumptions based on n tending to infinity.

An important problem in high-dimensional data situations is classification where the data arise from multiple groups. Based on a training data set with known group labels, one wants to determine a function f, such that $f(x)$ can be used to predict the group label of an observation x. Due to simplicity a linear f is a popular choice, and we write a linear f as $f(x) = w^{\mathrm{T}}x + \text{constant}$, where w is called the classification vector.

The optimal, linear classification rule with respect to classification error is Bayes rule, see e.g. Mardia et al. (1979) for a proof. Bickel and Levina (2004) proved that Fishers rule, that is, the sample version of the optimal Bayes rule, is no better than a random guess when $p/n \to \infty$, indicating that estimation of the covariance matrix is infeasible in such high dimensional situations. Fan and Fan (2008) further showed that variable selection might be necessary to obtain useful classifiers, unless the distance between the group means is very large.

When estimation of the covariance matrix cannot be done with sufficient precision, there has been some focus on independence classifiers that simply ignore correlation, and perform variable selection based on marginal properties. However, as emphasized by Fan et al. (2012) and Cai and Liu (2011), correlation does matter in classification, and variables that marginally are irrelevant can be helpful in reducing the classification error. As a remedy, those two papers introduced the regularized optimal affine discriminant (ROAD) and linear programming discriminant (LPD), respectively. Both procedures search directly for an optimal linear p-dimensional classification vector, and thus avoid the estimation of the covariance matrix. ROAD and LPD have attractive properties theoretically as well as empirically.

Wu et al. (2011) introduced sparse linear discriminant analysis (sLDA), which has a theoretical formulation equivalent to ROAD, but an alternative algorithm is applied for its estimation, making it different from ROAD.

In this report, ROAD, LPD, sLDA, and our new suggestion, the linear lasso discriminant (LLD), are compared through an extensive simulation study. We analyze the results in search for the best classifier with respect to classification error, computational cost, and estimation of the oracle classification vector. The results designate no uniquely best classifier, but we do detect different behaviours among them.

In Section 5.2 we introduce the classifiers and summarize theoretical results from the literature. Section 5.3 describes the models for our simulations, while the results are given in Section 5.4 along with an analysis hereof. Section 5.5 states the main conclusions. The theory behind the algorithms for estimation is given in the appendix in Section 5.6.

5.2 Sparse linear classifiers incorporating covariance

We consider discrimination between two p-dimensional normally distributed groups with equal covariance matrix, and means μ_1 and μ_2. Thus, our training data are $X_{ij} \sim N_p(\mu_i, \Sigma)$ for $i = 1, 2$, and $j = 1, \dots, n_i$. We define $n = n_1 + n_2$, and $\Delta = \mu_2 - \mu_1$. Let $\hat{\mu}_1 = \bar{x}_1$ and $\hat{\mu}_2 = \bar{x}_2$ denote the sample group averages, $\hat{\Delta} = \bar{x}_2 - \bar{x}_1$, and $\hat{\Sigma}$ the within group sample variance.

We consider linear discriminants where we calculate

$$\delta_w(x) = \mathbf{1}\big\{w^{\mathrm{T}}\big(x - \tfrac{1}{2}(\mu_1 + \mu_2)\big) > 0\big\}, \tag{5.1}$$

for a new observation x, and classify x to group 1 when δ_w is zero, and to group 2 when δ_w is one. The probability of a wrong classification is

$$W(\delta_w) = \overline{\Phi}\bigg(\frac{w^{\mathrm{T}}\Delta}{2(w^{\mathrm{T}}\Sigma w)^{1/2}}\bigg), \tag{5.2}$$

both when x comes from group 1 and group 2. In this formula $\overline{\Phi}$ denotes the upper tail probability of a standard normal distribution. The smallest classification error is obtained with $w_{\text{Bayes}} = \Sigma^{-1}\Delta$ (Mardia et al. (1979)). This classifier is known as Bayes rule, and the classification error $W(\delta_{\text{Bayes}}) = \overline{\Phi}\big((\Delta\Sigma^{-1}\Delta)^{1/2}/2\big)$ is called Bayes risk. It is customary to evaluate the performance of a classifier by comparing with Bayes risk.

In practice, μ_1, μ_2 and w are replaced by estimates so that (5.1) becomes

$$\delta_{\hat{w}} = \mathbf{1}\big\{\hat{w}^{\mathrm{T}}\big(x - \tfrac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\big) > 0\big\}.$$

The classification error for a new observation from group 1 is then

$$W(\delta_{\hat{w}}) = \overline{\Phi}\left(\frac{\hat{w}^{\mathrm{T}}\hat{\Delta} + 2\hat{w}^{\mathrm{T}}(\hat{\mu}_1 - \mu_1)}{2(\hat{w}^{\mathrm{T}}\Sigma\hat{w})^{1/2}}\right). \tag{5.3}$$

Fishers rule corresponds to Bayes rule with estimated parameters, that is, $\hat{w} = \hat{\Sigma}^{-1}\hat{\Delta}$. For fixed p and large n Fishers rule is close to Bayes rule. When p is increasing with n, this is no longer the case. For $p/n \to \infty$ and Σ estimated by a generalized inverse, Bickel and Levina (2004) proved that the classification error of Fishers rule approaches one half, the classification error of a random guess. Intuitively, this occurs through error accumulation from estimating p^2 entries of Σ. One possible remedy is to use only the variances and disregard all the covariances, that is, using $\hat{w}_D = \hat{D}^{-1}\hat{\Delta}$, where $\hat{D}$ is the diagonal matrix of $\hat{\Sigma}$. This is known as the independence rule or naive Bayes corresponding to the naive assumption of independence between variables.

Often relatively few variables are relevant for classification, and to avoid noise accumulation variable selection is needed, as pointed out by Fan and Fan (2008). Among independence classifiers performing variable selection are nearest shrunken centroids (Tibshirani et al., 2003), features annealed independence rule (Fan and Fan, 2008), and the thresholded independence rule (Bak et al., 2015).

An independence classifier ignores the information in the correlation structure, and is therefore suboptimal. This is illustrated by the following example from Cai and Liu (2011). Write

$$\Delta = \begin{pmatrix}\Delta_1\\ \Delta_2\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix},$$

where Δ_1 is k-dimensional, Σ_{11} is $k \times k$-dimensional, and $\Delta_2 =$ is a $(p-k)$-dimensional zero vector. The classification error of the independence rule using the true parameter values is $\overline{\Phi}(\Gamma_p/2)$, where

$$\Gamma_p = \frac{\Delta_1^{\mathrm{T}} D_{11}^{-1}\Delta_1}{(\Delta_1^{\mathrm{T}} D_{11}^{-1}\Sigma_{11} D_{11}^{-1}\Delta_1)^{1/2}}.$$

From Lemma A.3 of Mardia et al. (1979), we have

$$\Delta_1^{\mathrm{T}}\Sigma_{11}^{-1}\Delta_1 = \max_{z\in\mathbf{R}^{\mathbf{k}}} \frac{z^{\mathrm{T}}\Delta_1^{\mathrm{T}}\Delta_1 z}{z^{\mathrm{T}}\Sigma_{11} z},$$

and thus $\Delta_1^{\mathrm{T}}\Sigma_{11}^{-1}\Delta_1 \geq \Gamma_p^2$. This shows that Bayes rule based on the first k variables is better than the independence rule. Furthemore,

$$\Delta^{\mathrm{T}}\Sigma^{-1}\Delta = \Delta_1^{\mathrm{T}}\Sigma_{11}^{-1}\Delta_1 + (\Sigma_{22}^{-1}\Sigma_{12}\Delta_1)^{\mathrm{T}}(\Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21})^{-1}(\Sigma_{22}^{-1}\Sigma_{12}\Delta_1),$$

and since $\Sigma_{22} - \Sigma_{12}\Sigma_{11}^{-1}\Sigma_{21}$ is positive definite, Bayes rule based on the first k variables is strictly worse than Bayes rule based on all coordinates, unless $\Sigma_{22}^{-1}\Sigma_{12}\Delta_1 = 0$. Thus, we see that the inclusion of variables with no differential expression between the two groups actually improves the classifier.

5.2.1 Regularized optimal affine discriminant (ROAD)

The classifier (5.1) is invariant to scaling of w. Choosing a scale by requiring $\Delta^{\mathrm{T}}w = 1$, minimizing (5.2) is equivalent to minimizing $w^{\mathrm{T}}\Sigma w$ subject to $\Delta^{\mathrm{T}}w = 1$. Fan et al. (2012) regularizes this minimization problem by putting an upper bound on the ℓ_1 norm of w:

$$w_c = \operatorname*{arg\,min}_{\Delta^{\mathrm{T}}w=1, |w|_1\leq c} w^{\mathrm{T}}\Sigma w. \tag{5.4}$$

The purpose of the ℓ_1 norm is to reduce the number of nonzero entries of w. The bound c can be chosen from scientific considerations or by cross-validation. There is no direct link between c and the sparseness of w, the latter depends on Δ and Σ as well as c. The existence of a non-empty search set for w requires $c \geq 1/|\Delta|_\infty$, and the solution differs from w_{Bayes} only when $c < |\Sigma^{-1}\Delta|_1/(\Delta^{\mathrm{T}}\Sigma^{-1}\Delta)$. The classifier in (5.4) is called the regularized optimal affine discriminant (ROAD), and its sample version is

$$\hat{w}_c = \underset{\hat{\Delta}^{\mathrm{T}}w=1, |w|_1\leq c}{\arg\min} \; w^{\mathrm{T}}\hat{\Sigma}w. \tag{5.5}$$

The minimization in ROAD can be reformulated as a Lagrangian problem, better suited for calculation:

$$w_\lambda = \underset{\Delta^{\mathrm{T}}w=1}{\arg\min}\{w^{\mathrm{T}}\Sigma w + \lambda|w|_1\}. \tag{5.6}$$

Large values of λ correspond to small values of c.

The following theorem summarizes Theorem 1 in Bak and Jensen (2014) (which is a corrected and improved version of Theorem 1 in Fan et al. (2012)), and Theorem 2 in Fan et al. (2012).

Theorem 5.1. *Let ϵ and K be positive constants such that $\max_j\{|\Delta_j|\} > \epsilon$, $c > \epsilon + 1/\max_j|\Delta_j|$, and $\lambda_{\min}(\Sigma) > K$ where $\lambda_{\min}(\Sigma)$ denotes the smallest eigenvalue of Σ. Assume that for some sequence $a_n \to 0$, we have $|\hat{\Sigma} - \Sigma|_\infty = O_p(a_n)$, and $|\hat{\Delta} - \Delta|_\infty = O_p(a_n)$. Then*

$$W(\delta_{\hat{w}_c}) - W(\delta_{w_c}) = O_p\big(c^2 a_n(1 + c^2|\Sigma|_\infty)\big). \tag{5.7}$$

Next, let w_B be w_{Bayes} scaled to fulfil $\Delta^T w_B = 1$, and consider the minimization problem in (5.6). We have

$$|w_\lambda - w_B|_2 \leq \frac{\lambda\sqrt{|w_B|_0}}{\lambda_{\min}(\Sigma)}, \tag{5.8}$$

where $|w|_0$ is the number of nonzero entries of w.

In practice, a_n can be chosen as $\sqrt{\log(p)/n}$ under mild conditions. For a fixed value of c, (5.7) shows that the difference between the classification error of the sample ROAD and the true optimal ROAD tends to zero, if $|\Sigma|_\infty$ is upper bounded. Equation (5.8), on the other hand, considers directly the difference between ROAD and Bayes rule. If w_{Bayes} is sparse ($|w_{\mathrm{B}}|_0$ is bounded), and we choose $\lambda = \sqrt{\log(p)/n} \to 0$, we see that ROAD approaches Bayes rule. Note, though, that c gets large when $\lambda \to 0$, so (5.7) gives no positive result for the sample ROAD in this situation.

In Fan et al. (2012), ROAD is implemented through a fast coordinate descent algorithm, described in Section 5.6.1, which minimizes

$$w_{\lambda,\gamma} = \underset{w^{\mathrm{T}}\Delta=1}{\arg\min}\{w^{\mathrm{T}}\Sigma w + \lambda|w|_1 + \tfrac{1}{2}\gamma(w^{\mathrm{T}}\Delta - 1)^2\}. \tag{5.9}$$

When referring to ROAD in the following sections, we mean the implementation through (5.9). When γ is fixed, and λ_γ is chosen by cross-validation, simulations (Fan et al. (2012)) show that the choice of γ is not crucial to obtain a low classification error. However, the choice of γ has an effect on the sparsity of $w_{\lambda,\gamma}$ and its ability to estimate w_{B}.

5.2.2 Sparse linear discriminant analysis (sLDA)

In Wu et al. (2011) the starting point is Fishers rule defined by $\arg\max_w w^{\mathrm{T}}\hat{\Sigma}_B w / w^{\mathrm{T}}\hat{\Sigma} w$, where $\hat{\Sigma}_B$ is the estimated between-group variance. This starting point relates to variances, and not directly to a distributional assumption. In the two group setting the criteria reduces to minimizing $w^{\mathrm{T}}\hat{\Sigma}w$ subject to $\hat{\Delta}^{\mathrm{T}}w = 1$. Using an ℓ_1 constraint the sparse linear discriminant (sLDA) is defined as in (5.5), the latter being the background for the ROAD algorithm which solves (5.9).

The implementation of sLDA in Wu et al. (2011) solves (5.5), and is therefore different from the ROAD algorithm. The authors first argue that the solution $\hat{w}_c$ is piecewise linear, as a function of c, and next describe an algorithm to produce the entire solution path. Their work is based on the results in Rosset and Zhu (2007). Details of the algorithm are given in Section 5.6.2.

Mai and Zou (2012) proved that sLDA is equivalent to sparse optimal scoring (Clemmensen et al., 2011) and direct sparse discriminant analysis (Mai et al., 2012).

5.2.3 Linear programming discriminant (LPD)

Bayes rule $w_{\text{Bayes}} = \Sigma^{-1}\Delta$ can be written as the solution to $\Sigma w - \Delta = 0$. Motivated by this Cai and Liu (2011) define the linear programming discriminant (LPD) as

$$\hat{w}_{\text{LPD}} = \mathop{\arg\min}_{|\hat{\Sigma}w - \hat{\Delta}|_\infty \leq \lambda_n} |w|_1. \tag{5.10}$$

To see the asymptotic properties of LPD we restate Theorem 2 and Theorem 3 of Cai and Liu (2011).

Theorem 5.2. *Assume $n_1 \asymp n_2$, $\log(p) \leq n$, $\max_i \Sigma_{ii} \leq K_1$, and $\Delta^T\Sigma^{-1}\Delta \geq K_2$ for positive constants K_1 and K_2, and let $\lambda = C\sqrt{\Delta^T\Sigma\Delta\log(p)/n}$ for a sufficiently large C.*

(i) If

$$\frac{|\Sigma^{-1}\Delta|_1}{(\Delta^T\Sigma^{-1}\Delta)^{1/2}} + \frac{|\Sigma^{-1}\Delta|_1^2}{(\Delta^T\Sigma^{-1}\Delta)^2} = o\Big(\sqrt{\frac{n}{\log(p)}}\Big),$$

then

$$W(\delta_{\text{LPD}}) - W(\delta_{\text{Bayes}}) \to 0,$$

in probability as $n \to \infty$.

(ii) If

$$|\Sigma^{-1}\Delta|_1(\Delta^T\Sigma^{-1}\Delta)^{1/2} + |\Sigma^{-1}\Delta|_1^2 = o\Big(\sqrt{\frac{n}{\log(p)}}\Big),$$

then

$$\frac{W(\delta_{\text{LPD}})}{W(\delta_{\text{Bayes}})} - 1 = O\Big((|\Sigma^{-1}\Delta|_1(\Delta^T\Sigma\Delta)^{1/2} + |\Sigma^{-1}\Delta|_1^2)\sqrt{\frac{\log(p)}{n}}\Big),$$

with probability greater than $1 - O(p^{-1})$.

(iii) If $|\Sigma^{-1}\Delta|_0\Delta^T\Sigma^{-1}\Delta = o\big(\sqrt{n/\log(p)}\big)$, and the eigenvalues of Σ are bounded above and below, it holds that:

$$\frac{W(\delta_{\text{LPD}})}{W(\delta_{\text{Bayes}})} - 1 = O_P\Big(|\Sigma^{-1}\Delta|_0\Delta^T\Sigma\Delta\sqrt{\frac{\log(p)}{n}}\Big).$$

Theorem 5.2 states that under various sparseness assumptions on $w_{\text{Bayes}} = \Sigma^{-1}\Delta$, the LPD classifier performs well as compared to the Bayes classifier. Part (iii) uses direct sparseness of $\Sigma^{-1}\Delta$, whereas parts (i) and (ii) use approximately sparseness through a restriction on the ℓ_1 norm of $\Sigma^{-1}\Delta$.

LPD is implemented using linear programming, see Section 5.6.3 for details.

5.2.4 Linear lasso discriminant (LLD)

The LPD in (5.10) is defined through the supremum norm of $\hat{\Sigma}w - \hat{\Delta}$. Using instead the ℓ_2 norm we get a classifier that we term *linear lasso discriminant* (LLD):

$$w_{\text{LLD}} = \operatorname*{arg\,min}_{w:|w|_1 \leq c} \frac{1}{p} |\hat{\Sigma}w - \hat{\Delta}|_2^2, \tag{5.11}$$

Numerically, this procedure is easily implemented through the LARS-algorithm (Efron et al., 2004) without intercept, see Section 5.6.4 for details. The setting here is quite different from the setting in Efron et al. (2004), and the properties shown in that paper cannot be directly transferred to our setting.

5.3 Methods and models

We compute ROAD, LPD, sLDA, and LLD on simulated datasets in 19 settings. Additionally, all classifiers are computed after preliminary variable selection: Selection version number 1 (S1) uses the 50 variables with the largest t-statistics, while selection version number 2 (S2) furthermore incorporates the variables which are most correlated with each of these 50 variables. This is inspired by Fan et al. (2012) where a closely related selection procedure turns out to produce a classifier almost as good as the original ROAD, while the computational burden is much smaller.

From experience it is known that LPD and LLD are better when $\hat{\Sigma}$ is replaced by $\tilde{\Sigma} = \hat{\Sigma} + \log(p)/nI_p$. For LPD this estimate furthermore enables us to find an initial solution.

ROAD, sLDA, and LLD result in sparse classification vectors whereas the algorithm applied to find w_{LPD} produces a solution with many coefficients very close to zero. For w_{LPD} we set coefficients smaller than e^{10}/τ equal to zero, where τ is described in Section 5.6.3.

For all classifiers the optimal value of the regularizing parameter is found by fivefold cross-validation with the same splitting across classifiers for each simulation, and searching over 100 values of the regularizing parameter.

All settings are performed for $p = 100$, 500, and 1000, while the number of observations from each group is constantly $n_1 = n_2 = 50$. The number of repeated simulations in each setting is $n_{\text{sim}} = 100$. When $p = 1000$, LPD is not calculated due to the very long computing time needed.

Let a_b denote a b-dimensional vector with a on all entrances. Our settings are as follows:

Setting 1 – Equal correlation: $\Delta = (1_{10}, 0_{p-10})$, $\Sigma_{ii} = 1$, and $\Sigma_{ij} = \rho \ i \neq j$ for $\rho = 0.1, 0.5, 0.9$.

Setting 2 – Blockwise correlation 1: $\Delta = (0.5_5, 0_5, 0.5_5, 0_{p-15})$, and $\Sigma = \text{diag}(\Sigma_0)$, where Σ_0 is a 10×10 equicorrelated matrix with correlation $\rho = -0.1, 0.1, 0.5, 0.9$.

Setting 3 – Blockwise correlation 2: $\Delta = \big((0.5_1, 0_9)_{10}, 0_{p-100}\big)$, and Σ as in blockwise correlation 1.

Setting 4 – Descending correlation 1: $\Delta = (1_{10}, 0_{p-10})$, and $\Sigma_{ij} = \rho^{|i-j|}$ for $\rho = 0.1, 0.5, 0.9$.

Setting 5 – Descending correlation 2: $\Delta = \big((1_1, 0_9)_{10}, 0_{p-100}\big)$, and Σ as in descending correlation 1.

Setting 6 – Leukaemia 1: $\Delta = \hat{\Delta}_p$ and $\Sigma = \mathrm{Cor}\Big(\hat{\Sigma}_p + \sqrt{\frac{\log(p)}{n_1+n_2}} I_p\Big)$, where $\hat{\Delta}_p$ is the empirical mean difference, and $\hat{\Sigma}_p$ is the empirical variance based on the data in Golub et al. (1999), I_p is the identity matrix, and $\mathrm{Cor}()$ is the function that transforms a variance matrix to a correlation matrix. The p variables have been selected by choosing p consecutive variables where the distribution of the correlations resembles the distribution for all variables.

Setting 7 – Leukaemia 2: $\Delta = (1_{10}, 0_{p-10})$, and Σ as in Leukaemia 1.

5.3.1 Previous simulations

We now give a brief overview over existing empirical studies of the relevant classifiers.

Fan et al. (2012) have performed a simulation study with $p = 1000$, $n_0 = n_1 = 300$, and 10 variables with differential expression. The differential expression is either 1 or 0.5 depending on the setting. With these parameter values an almost complete separation, based on the t-values, between the differentially expressed variables and the nondifferentially expressed variables, is obtained. Setting 1 above is considered with differential expression 1, and Setting 2 above with negative correlation, $\rho = -0.1$ is studied. Furthermore, two settings not related to Setting 1 to 6 are considered: First a two-block setting, with all the differentially expressed variables in a block of size 20, and next a setting with a 10-factor form of the covariance matrix, where the factors are generated from a uniform distribution, and the differential expressions is taken from a double exponential distribution.

Also three real datasets are considered, including the data of Golub et al. (1999) and Gordon et al. (2002), which both have many variables with a very high differential expression.

Cai and Liu (2011) have performed a simulation study with p in the range from 100 to 800, $n_1 = n_2 = 200$, and 10 variables with differential expression equal to 1. In such situations we have an almost complete separation, based on the t-values, between the differentially expressed variables and the nondifferentially expressed variables. Setting 1 above is considered with $\rho = 0.5$, and Setting 4 with $\rho = 0.8$. Also a setting with random entries of Σ^{-1} is considered. The datasets of Golub et al. (1999) and Gordon et al. (2002) are analyzed as well.

Note that due to our lower number of observations compared to the above studies, we cannot obtain the same quality of our resulting classifiers. In our opinion this lower sample size is more realistic in many microarray experiments.

Wu et al. (2011) only perform low-dimensional simulations with p equal to 10 and 40 respectively.

5.4 Results and discussion

In this section we compare the performance of the classifiers, with particular emphasis on classification error.

5.4.1 Classification error

Table 5.1 shows the average of the classification error from 100 simulations, each calculated from (5.3).

None of the procedures are uniformly best. Generally, ROAD and sLDA are very close, although mostly with sLDA having the larger classification error of the two. Often ROAD has the smallest classification error, or is close to the smallest values.

When considering the procedures starting from all variables, the errors of the classifiers are almost equal for $p = 100$. When $p = 500$ and $p = 1000$, ROAD and sLDA remain equally good, while LPD and LLD are worse in some situations, particularly when the correlation is large.

Regarding the variable selection method S1, the classifiers tend to be of the same quality, apart from LLD being slightly better than the rest in some of the settings with a small correlation.

The situation for the variable selection method S2 looks somewhat like the procedures with all variables. There is no significant difference for $p = 100$. For $p = 500$ and $p = 1000$, LPD and LLD tend to be worse than ROAD and sLDA when the correlation is large.

When considering the various versions of ROAD, we see that S1 is generally worse than the procedure including all variables, while S2 in some cases is better and in some cases worse. The latter happens mainly for large correlations. These conclusions are parallel to the ones from the simulations in Fan et al. (2012). When comparing S1, S2, and the procedure with all variables for the other classifiers, we arrive at similar conclusions as for ROAD.

5.4.2 Computational cost

In Table 5.2 the average running times over 100 simulations are seen. From this perspective ROAD is the most attractive classifier. Theoretical arguments suggest that the running time is $O(p^2)$ for ROAD. However, for the settings we have considered, the increase with p is much slower. LLD and sLDA seem more to obey a p^2 relationship, with sLDA using roughly twice the time of LLD, and LPD increases even faster. For one of the $p = 500$ cases, the simulation took around 5 minutes, and for that reason we did not calculate LPD for $p = 1000$. Note, however, that following James et al. (2009), it is possible to implement LPD with the same computational effort as sLDA and LLD (we have not pursued this in the present report).

5.4.3 Variable selection

Looking at the sparsity of the classifiers in Table 5.3, we see that ROAD and sLDA are almost equally sparse in most situations, and significantly more sparse than LPD and LLD, with LLD being the least sparse classifier. This is true both for the procedure with all variables and the variable selection method S2. For the S1 selection method the different classifiers are almost equally sparse. Slightly surprisingly we find that S1 and S2 in some cases are less sparse than the corresponding classifier build on all variables. From this we conclude that the variable selection methods should be seen as a help to

Table 5.1: Average number of nonzero coefficients over 100 simulations for all estimators. Standard deviations are given in parentheses.

Setting	ρ			$p = 100$				$p = 500$				$p = 1000$		
			Bayes	ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Equal	0.1	All	5.6	12.4 (2.5)	11.1 (2.4)	13.4 (3.3)	10.9 (2.7)	13.1 (2.7)	11.8 (2.6)	14.6 (4.0)	12.6 (2.3)	12.5 (2.8)	14.0 (3.8)	13.2 (3.3)
		S1		13.3 (3.0)	12.2 (2.5)	14.0 (3.2)	11.4 (2.6)	16.4 (4.0)	15.1 (3.3)	16.7 (4.3)	13.6 (2.9)	16.7 (4.0)	17.3 (4.0)	14.0 (3.6)
		S2						15.6 (4.2)	14.9 (3.9)	16.0 (4.4)	13.4 (3.2)	15.9 (4.8)	16.4 (4.3)	13.8 (3.4)
	0.5	All	1.7	6.1 (1.8)	4.2 (1.1)	7.1 (3.5)	4.1 (1.2)	5.3 (1.2)	4.0 (1.4)	7.1 (3.2)	4.3 (1.5)	4.9 (1.4)	7.6 (4.4)	4.5 (1.8)
		S1		6.5 (2.0)	4.8 (1.3)	7.2 (2.5)	4.9 (2.2)	7.2 (2.8)	5.7 (2.5)	8.1 (3.2)	5.7 (2.2)	8.1 (3.5)	8.8 (3.4)	6.6 (3.8)
		S2						7.2 (2.7)	5.7 (2.3)	8.0 (3.1)	5.1 (1.9)	8.3 (3.0)	9.5 (4.3)	5.8 (3.0)
	0.9	All	0.0	0.0 (0.0)	0.0 (0.0)	0.5 (0.5)	0.7 (1.0)	0.0 (1.0)	0.0 (0.0)	0.1 (0.3)	0.4 (0.5)	0.0 (0.0)	0.4 (0.8)	0.4 (0.5)
		S1		0.0 (0.0)	0.0 (0.0)	0.1 (0.1)	0.7 (0.9)	0.6 (5.4)	0.5 (5.4)	0.7 (5.4)	1.1 (5.4)	0.5 (5.2)	0.7 (5.5)	1.1 (5.3)
		S2						0.1 (1.1)	0.3 (2.5)	0.8 (1.9)	1.1 (2.3)	0.6 (5.7)	1.7 (6.6)	1.4 (5.7)
Blockwise 1	−0.1	All	3.2	20.6 (4.5)	24.6 (4.6)	20.2 (4.4)	26.3 (5.3)	32.1 (5.3)	33.5 (5.3)	33.0 (6.1)	36.4 (6.4)	36.3 (6.4)	35.7 (5.9)	39.3 (6.2)
		S1		23.7 (4.3)	24.2 (3.7)	24.2 (4.2)	25.4 (4.1)	35.2 (5.5)	34.8 (5.0)	35.4 (5.6)	34.4 (4.9)	38.4 (4.9)	38.5 (4.8)	37.5 (4.9)
		S2						35.3 (5.7)	35.0 (4.8)	35.2 (5.9)	35.0 (4.7)	39.3 (5.6)	38.9 (5.2)	38.4 (4.6)
	0.1	All	23.7	35.5 (5.1)	34.7 (4.6)	36.2 (5.0)	34.5 (5.0)	38.5 (5.0)	38.9 (5.2)	39.0 (5.3)	40.2 (4.9)	38.9 (5.4)	39.0 (5.5)	42.6 (5.5)
		S1		36.3 (5.0)	35.4 (4.3)	36.4 (4.7)	34.0 (4.7)	42.2 (4.6)	40.9 (4.5)	42.2 (4.6)	38.8 (4.2)	42.0 (4.5)	42.2 (4.6)	39.7 (4.5)
		S2						41.6 (4.5)	40.8 (4.4)	41.7 (4.9)	39.7 (4.9)	42.2 (4.5)	41.9 (4.7)	41.1 (4.7)
	0.5	All	20.4	35.5 (4.5)	34.8 (4.1)	35.1 (4.3)	34.6 (4.8)	41.0 (4.8)	41.7 (5.3)	41.0 (5.0)	43.3 (4.5)	42.3 (4.9)	42.4 (4.8)	44.8 (5.0)
		S1		36.2 (5.2)	35.4 (4.5)	35.9 (5.0)	35.8 (5.0)	43.3 (4.0)	42.5 (4.1)	43.4 (4.2)	42.3 (4.1)	44.8 (4.7)	44.6 (4.6)	43.1 (4.6)
		S2						40.7 (4.7)	39.8 (4.7)	40.6 (4.7)	40.0 (4.6)	42.0 (4.6)	42.4 (4.6)	41.8 (4.7)
	0.9	All	3.8	10.6 (3.1)	11.4 (3.3)	11.2 (3.5)	11.4 (3.3)	24.9 (3.3)	36.3 (6.1)	25.2 (6.8)	39.7 (6.9)	39.0 (6.0)	37.9 (7.5)	47.2 (4.7)
		S1		19.5 (10.2)	22.2 (10.0)	19.7 (10.0)	22.9 (10.6)	39.1 (9.3)	40.9 (6.6)	39.4 (9.3)	41.1 (6.3)	44.0 (6.7)	44.4 (6.6)	44.0 (5.5)
		S2						18.0 (6.6)	25.3 (7.8)	19.6 (7.3)	25.4 (7.9)	22.6 (8.7)	24.1 (8.7)	31.5 (8.5)
Blockwise 2	−0.1	All	14.3	32.0 (5.7)	30.9 (5.2)	31.9 (5.3)	32.1 (5.1)	36.5 (5.1)	36.5 (5.4)	36.8 (5.4)	39.6 (5.8)	37.3 (5.6)	38.1 (5.7)	41.0 (5.8)
		S1		32.3 (4.9)	31.7 (4.7)	32.2 (5.1)	30.8 (4.4)	39.5 (4.1)	38.4 (4.0)	39.4 (4.1)	37.6 (4.5)	40.9 (5.0)	40.6 (5.2)	38.3 (4.7)
		S2						39.5 (4.6)	39.0 (4.2)	39.2 (5.0)	38.4 (4.4)	41.2 (5.0)	41.1 (5.0)	39.8 (4.5)
	0.1	All	20.9	32.3 (5.8)	31.1 (5.4)	32.7 (5.4)	31.2 (5.2)	36.1 (5.2)	36.7 (4.9)	36.9 (5.3)	39.1 (6.0)	38.0 (5.3)	38.4 (5.7)	40.6 (5.1)
		S1		32.5 (5.1)	31.4 (4.7)	32.2 (4.7)	30.9 (4.6)	39.7 (4.5)	38.4 (4.3)	39.7 (4.8)	36.9 (4.5)	40.4 (4.8)	40.4 (4.9)	38.4 (4.8)
		S2						39.6 (4.7)	38.7 (4.5)	39.4 (4.9)	37.9 (4.5)	41.2 (5.0)	41.1 (5.1)	39.7 (4.4)
	0.5	All	14.3	26.6 (4.9)	25.6 (5.1)	27.4 (5.3)	25.3 (4.6)	33.8 (4.6)	34.8 (4.9)	34.2 (4.8)	37.4 (5.4)	36.4 (4.9)	36.7 (5.6)	42.8 (6.1)
		S1		26.1 (4.8)	25.8 (4.6)	26.7 (4.6)	25.7 (4.3)	35.3 (4.6)	35.0 (4.3)	35.6 (4.8)	35.1 (4.1)	39.0 (4.4)	39.0 (4.7)	38.3 (4.2)
		S2						34.4 (4.4)	34.2 (4.4)	34.4 (4.2)	34.9 (4.5)	38.3 (4.8)	38.0 (5.4)	38.6 (4.2)
	0.9	All	0.9	4.2 (1.4)	3.4 (1.4)	5.0 (2.2)	3.9 (2.3)	13.9 (2.3)	25.8 (5.9)	16.4 (5.5)	29.4 (6.7)	32.1 (6.4)	31.1 (6.6)	42.2 (5.6)
		S1		6.9 (2.9)	7.7 (4.4)	7.7 (3.2)	8.5 (4.4)	28.7 (6.7)	31.0 (5.5)	29.1 (7.3)	32.1 (5.8)	32.5 (6.3)	32.9 (6.6)	35.3 (4.9)
		S2						13.9 (4.8)	21.0 (5.9)	14.6 (5.2)	21.2 (6.2)	19.2 (5.3)	20.2 (5.3)	28.7 (6.1)
Descending 1	0.1	All	7.4	10.9 (2.7)	10.4 (2.8)	12.5 (3.8)	10.6 (2.7)	12.5 (2.7)	12.1 (3.1)	14.4 (4.5)	15.0 (3.1)	13.0 (3.0)	14.6 (4.8)	16.0 (3.0)
		S1		12.2 (3.5)	11.1 (2.8)	13.3 (3.6)	10.2 (2.2)	18.3 (5.4)	16.8 (4.2)	19.0 (5.6)	14.0 (3.7)	19.0 (4.6)	19.1 (4.8)	14.4 (2.8)
		S2						16.0 (5.3)	15.9 (4.7)	16.6 (5.5)	14.5 (3.9)	17.3 (5.0)	17.8 (4.9)	15.3 (3.6)
	0.5	All	14.9	20.7 (4.0)	20.0 (3.7)	22.7 (4.9)	20.3 (3.3)	21.4 (3.3)	21.0 (4.5)	22.2 (4.0)	23.2 (3.9)	21.2 (4.1)	21.8 (4.3)	25.3 (4.3)
		S1		22.3 (5.0)	21.3 (3.6)	23.5 (4.9)	20.7 (3.1)	27.9 (5.3)	26.1 (4.4)	28.2 (5.1)	24.5 (4.6)	30.2 (5.5)	29.9 (6.1)	26.0 (5.2)
		S2						25.4 (5.5)	25.4 (5.0)	26.3 (5.8)	24.7 (4.5)	28.0 (6.1)	28.4 (5.7)	26.3 (4.7)
	0.9	All	11.6	16.7 (3.4)	21.6 (3.4)	17.6 (3.6)	23.9 (4.2)	23.7 (4.2)	28.5 (4.3)	23.1 (4.5)	32.9 (4.2)	27.2 (5.6)	26.5 (4.9)	35.9 (5.4)
		S1		22.2 (7.2)	24.5 (5.1)	22.9 (7.5)	25.7 (4.9)	32.3 (6.6)	32.9 (5.1)	32.4 (7.0)	33.1 (5.0)	36.3 (6.1)	36.0 (6.6)	35.6 (4.9)
		S2						20.6 (3.8)	26.4 (3.5)	21.6 (4.9)	28.3 (3.9)	23.3 (4.2)	23.9 (4.2)	30.6 (4.8)
Descending 2	0.1	All	5.5	9.4 (2.9)	8.8 (2.4)	10.8 (3.1)	8.6 (2.3)	10.7 (2.3)	9.7 (2.4)	13.0 (4.4)	12.8 (2.9)	10.6 (2.9)	12.0 (4.3)	13.9 (3.0)
		S1		10.0 (3.0)	9.4 (2.6)	11.2 (3.2)	8.8 (2.3)	15.2 (4.9)	14.2 (4.2)	15.5 (5.0)	11.5 (3.4)	16.3 (5.2)	16.8 (5.3)	12.6 (3.4)
		S2						14.0 (5.2)	13.4 (4.3)	15.3 (5.0)	12.7 (4.0)	16.0 (6.2)	16.2 (5.5)	14.3 (4.0)
	0.5	All	2.2	7.6 (2.0)	7.1 (1.8)	8.2 (2.6)	7.3 (1.7)	9.9 (1.7)	9.8 (2.6)	11.1 (4.0)	13.0 (3.2)	10.1 (3.1)	12.1 (4.8)	13.9 (3.3)
		S1		8.2 (2.1)	7.7 (1.9)	8.5 (2.3)	7.7 (1.7)	13.3 (4.3)	12.6 (3.9)	13.7 (4.0)	11.7 (3.0)	15.6 (4.8)	16.0 (5.3)	12.3 (3.1)
		S2						11.8 (4.1)	11.6 (3.6)	12.5 (4.8)	12.4 (3.3)	13.5 (4.7)	14.5 (5.1)	13.1 (2.8)
	0.9	All	0.0	0.0 (0.0)	0.0 (0.0)	0.6 (1.4)	0.6 (0.7)	0.5 (0.7)	1.3 (0.9)	0.9 (0.9)	2.9 (1.6)	2.0 (2.0)	2.7 (2.3)	9.3 (3.8)
		S1		0.2 (0.1)	0.2 (0.2)	0.5 (0.7)	1.0 (0.9)	7.5 (5.3)	9.5 (4.7)	8.1 (5.6)	10.4 (4.9)	14.9 (6.6)	15.4 (6.9)	16.6 (5.1)
		S2						0.3 (0.2)	0.8 (0.6)	1.0 (1.0)	1.5 (1.5)	0.5 (0.4)	1.2 (1.2)	2.8 (2.2)
Leukaemia	1.0	All	0.2	2.5 (0.8)	1.9 (0.8)	2.7 (0.9)	2.4 (1.5)	0.6 (1.5)	0.1 (0.2)	1.4 (1.1)	0.9 (1.0)	0.1 (0.1)	0.1 (0.1)	0.9 (1.1)
		S1		3.0 (1.0)	2.5 (0.8)	3.3 (1.2)	2.9 (1.5)	3.6 (1.3)	3.2 (1.7)	4.4 (2.2)	3.5 (1.4)	0.4 (0.4)	0.9 (0.6)	1.3 (1.2)
		S2						2.3 (1.0)	1.5 (0.7)	3.0 (1.3)	2.2 (1.4)	0.3 (0.2)	0.9 (0.8)	1.2 (1.1)
	2.0	All	0.3	3.3 (1.4)	2.6 (1.0)	3.9 (1.6)	2.9 (1.4)	5.1 (1.4)	4.3 (1.8)	7.3 (3.8)	5.8 (2.1)	3.3 (1.2)	5.3 (4.1)	5.6 (2.1)
		S1		3.9 (1.2)	3.4 (1.0)	4.8 (1.9)	3.9 (1.4)	9.7 (3.0)	9.1 (2.8)	10.3 (3.2)	9.9 (3.3)	5.8 (1.9)	6.2 (2.3)	6.4 (2.1)
		S2						6.8 (1.8)	6.2 (1.8)	7.8 (2.7)	6.6 (2.2)	5.1 (1.8)	5.7 (2.9)	5.7 (2.2)

Table 5.2: Average computation time over 100 simulations for the classifiers calculated from all variables.

Setting	ρ	$p = 100$				$p = 500$				$p = 1000$		
		ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Equal	0.1	3.0	7.9	1.5	1.7	8.4	304.9	31.8	13.4	17.4	127.4	79.6
	0.5	3.1	9.3	1.4	1.7	8.8	374.9	30.7	11.6	17.5	119.0	72.3
	0.9	2.7	10.8	1.5	1.4	9.1	514.6	34.6	10.5	18.4	141.2	68.4
Blockwise 1	−0.1	4.0	7.6	1.6	1.7	8.9	274.5	33.1	13.0	17.4	136.4	73.9
	0.1	5.2	7.5	1.5	1.7	9.1	277.0	36.7	12.8	17.6	137.3	73.8
	0.5	5.9	7.8	1.5	1.7	10.1	288.6	35.7	13.1	18.8	136.8	74.4
	0.9	5.8	8.5	1.5	1.7	12.0	319.7	38.8	13.7	20.6	142.2	79.7
Blockwise 2	−0.1	5.0	7.6	1.6	1.7	9.0	290.3	36.8	13.1	17.6	136.0	74.3
	0.1	4.8	7.6	1.6	1.7	9.0	291.4	36.6	13.1	17.3	133.9	73.8
	0.5	5.2	7.9	1.5	1.7	9.5	289.0	34.6	13.7	17.9	132.8	74.4
	0.9	4.2	8.4	1.5	1.5	11.1	306.4	36.6	13.5	19.6	135.2	79.5
Descending 1	0.1	3.1	7.3	1.6	1.7	8.3	267.8	34.8	13.2	17.1	128.8	76.5
	0.5	4.0	7.3	1.6	1.7	8.9	260.0	33.5	13.1	17.8	128.0	76.1
	0.9	4.7	8.3	1.5	1.7	10.3	278.9	35.2	13.5	19.0	135.4	78.6
Descending 2	0.1	2.8	7.3	1.6	1.7	8.1	275.2	34.0	13.8	16.9	132.4	76.7
	0.5	2.7	7.5	1.6	1.7	8.1	266.7	34.6	13.4	17.0	136.0	77.9
	0.9	2.4	9.1	1.6	1.6	8.4	305.7	38.2	13.4	17.5	145.7	82.0
Leukaemia	1.0	2.3	8.2	1.5	1.7	8.1	297.0	35.1	11.7	14.9	118.9	75.0
	2.0	2.5	8.5	1.6	1.7	8.7	326.5	33.9	12.4	17.4	136.0	77.3

reduce computational cost rather than a way of reducing the dimension of the final classifier.

We now consider whether the classifiers find variables with large values of w_{Bayes} in those cases where w_{Bayes} has only few large values. More precisely, we look at those cases where the number of variables with w_{Bayes} larger than 10^{-4} is small. Setting the small entries of w_{Bayes} to zero we have the following cases with sparse w_{Bayes}:

Blockwise 1 with only the first 20 coordinates nonzero,

Descending 1 with only the first 11 coordinates nonzero,

Descending 2 with only the first 100 coordinates nonzero.

For the above settings we define a true positive rate (TPR) and a false positive rate (FPR) as follows:

$$\text{TPR} = \frac{\{i : w_i \neq 0, w_{\text{Bayes},i} \neq 0\}}{\#\{i : w_{\text{Bayes},i} \neq 0\}}, \quad \text{and} \quad \text{FPR} = \frac{\{i : w_i \neq 0, w_{\text{Bayes},i} = 0\}}{\#\{i : w_{\text{Bayes},i} = 0\}}.$$

These rates can be seen in Table 5.4 and Table 5.5. Both TPR and FPR are generally larger for LPD and LLD than for ROAD and sLDA in accordance with the previous observation that these methods are less sparse.

ROAD tends to have the lowest value of both TPR and FPR. TPR is always above 50% if $p = 100$, but gets as low as 17% for ROAD when $p = 1000$. For ROAD, LPD and sLDA the FPR are almost always below 50%, and when $p = 1000$, it is even below 10%. Summarizing, most procedures tend to include a large part of the relevant variables, but particularly LLD selects a lot of irrelevant ones simultaneously.

5.4.4 Estimation of w_{Bayes}

In this section we compare the estimates of w for the various classifiers. Figure 5.1 and Figure 5.2 plot pairs of estimates of w for four of the one hundred simulations. We see that the coefficients of sLDA and ROAD are strongly positively correlated. The

Table 5.3: Average number of nonzero coefficients over 100 simulations for all estimators. Standard deviations are given in parentheses.

Setting	ρ		$p = 100$				$p = 500$				$p = 1000$		
			ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Equal	0.1	All	34.5 (13.6)	42.1 (18.1)	38.0 (14.6)	42.6 (16.4)	48.6 (25.1)	81.5 (60.4)	43.3 (20.5)	105.0 (40.5)	51.7 (24.8)	48.6 (21.2)	148.1 (81.8)
		S1	30.6 (8.4)	35.9 (10.2)	32.5 (10.0)	32.6 (9.2)	33.4 (9.6)	39.1 (11.2)	33.7 (10.1)	38.0 (10.0)	35.0 (8.5)	36.1 (9.5)	39.6 (7.8)
		S2					47.5 (16.7)	68.7 (23.9)	47.2 (14.8)	64.5 (18.7)	50.4 (15.1)	51.6 (14.6)	64.3 (17.7)
	0.5	All	46.2 (11.8)	58.0 (17.7)	47.2 (12.7)	52.7 (14.2)	66.2 (19.3)	131.0 (85.8)	60.8 (18.1)	94.8 (38.8)	75.2 (22.1)	60.3 (18.6)	127.0 (65.3)
		S1	34.4 (6.0)	42.2 (6.3)	37.1 (7.3)	38.7 (7.8)	35.1 (7.4)	42.5 (7.9)	38.0 (8.6)	36.2 (9.7)	34.6 (8.3)	36.1 (9.3)	35.4 (10.0)
		S2					48.0 (11.7)	67.9 (21.5)	50.0 (13.1)	51.0 (14.5)	50.0 (13.3)	50.9 (14.8)	56.3 (17.8)
	0.9	All	48.1 (4.2)	95.9 (1.9)	78.6 (9.2)	25.4 (8.6)	71.9 (8.1)	452.6 (6.8)	85.1 (12.5)	38.6 (15.5)	81.4 (8.3)	82.3 (16.8)	42.6 (19.5)
		S1	35.5 (2.8)	49.6 (0.7)	48.6 (2.7)	24.7 (9.4)	35.3 (3.9)	49.2 (4.4)	48.2 (5.0)	28.5 (11.9)	35.2 (4.5)	47.8 (5.2)	27.3 (11.8)
		S2					46.3 (5.1)	95.6 (2.6)	78.7 (9.6)	25.9 (9.5)	46.1 (5.1)	77.1 (13.7)	24.9 (10.8)
Blockwise 1	−0.1	All	49.9 (24.1)	48.7 (29.3)	54.7 (21.4)	70.2 (23.0)	45.1 (34.9)	96.6 (114.3)	36.8 (26.9)	121.0 (55.1)	35.3 (32.5)	38.6 (27.2)	138.5 (92.4)
		S1	36.3 (9.2)	41.1 (7.8)	37.9 (9.9)	43.4 (5.5)	43.4 (3.0)	47.9 (2.4)	43.8 (5.1)	46.6 (3.7)	43.5 (2.8)	45.0 (3.7)	46.4 (4.7)
		S2					63.2 (13.0)	80.7 (15.6)	60.0 (12.3)	80.4 (11.7)	63.4 (11.0)	61.9 (10.8)	78.2 (12.9)
	0.1	All	29.4 (22.0)	33.7 (26.8)	36.5 (24.3)	41.9 (27.6)	34.2 (33.4)	98.8 (128.8)	35.3 (25.9)	105.8 (54.5)	33.6 (34.3)	29.9 (23.3)	130.7 (104.3)
		S1	32.6 (13.4)	39.6 (11.8)	34.9 (11.2)	37.7 (12.2)	43.5 (3.1)	47.9 (2.2)	43.3 (4.8)	46.1 (5.7)	43.5 (3.2)	44.8 (4.2)	47.1 (3.2)
		S2					62.6 (13.3)	82.0 (14.6)	60.9 (16.4)	75.6 (16.3)	65.4 (12.4)	63.4 (13.8)	80.3 (15.1)
	0.5	All	37.0 (25.8)	45.1 (31.4)	39.7 (22.0)	55.3 (29.5)	38.3 (36.3)	97.7 (137.9)	40.3 (28.6)	100.1 (71.2)	30.9 (35.5)	38.9 (28.2)	111.4 (78.4)
		S1	28.9 (11.6)	33.1 (12.9)	29.7 (10.4)	32.9 (12.6)	38.4 (6.2)	42.8 (6.0)	39.1 (7.4)	42.3 (7.4)	40.5 (4.1)	41.5 (5.8)	44.4 (4.9)
		S2					58.2 (14.7)	77.7 (17.8)	56.3 (17.0)	75.7 (17.3)	60.5 (11.3)	58.8 (14.7)	75.3 (16.2)
	0.9	All	41.7 (12.8)	72.2 (18.3)	43.1 (12.4)	71.5 (17.0)	76.1 (24.8)	98.4 (104.8)	70.1 (19.8)	141.6 (90.2)	53.1 (43.8)	58.2 (31.4)	132.1 (132.9)
		S1	25.1 (9.5)	34.3 (11.1)	28.4 (9.0)	34.6 (11.0)	24.5 (8.1)	34.0 (8.7)	29.2 (10.2)	30.6 (10.4)	23.6 (7.4)	28.5 (9.0)	30.9 (10.3)
		S2					47.9 (11.2)	78.6 (18.6)	52.1 (14.4)	82.6 (14.2)	51.0 (12.7)	54.0 (15.1)	80.1 (15.5)
Blockwise 2	−0.1	All	29.2 (23.9)	31.5 (25.7)	33.9 (23.7)	42.9 (26.1)	37.6 (34.7)	79.0 (107.7)	38.8 (25.9)	105.2 (54.2)	35.9 (36.8)	34.5 (25.4)	159.1 (124.1)
		S1	32.4 (13.5)	39.1 (11.7)	34.2 (11.3)	39.8 (10.2)	43.5 (3.8)	47.5 (2.4)	44.5 (4.9)	46.5 (3.9)	44.1 (2.9)	43.7 (4.3)	47.1 (3.4)
		S2					61.4 (15.4)	80.9 (15.9)	58.3 (16.5)	74.9 (16.7)	64.7 (12.7)	62.8 (12.9)	80.0 (12.3)
	0.1	All	31.8 (24.1)	38.6 (29.8)	37.9 (23.2)	51.1 (25.9)	36.5 (32.1)	114.3 (136.2)	37.7 (26.1)	111.2 (57.7)	36.9 (33.4)	37.4 (26.5)	158.4 (128.1)
		S1	34.3 (12.3)	38.8 (12.7)	35.8 (10.6)	40.7 (9.1)	44.2 (3.3)	48.2 (1.7)	44.3 (5.1)	47.4 (2.9)	44.1 (3.1)	44.3 (5.0)	47.1 (3.0)
		S2					65.3 (12.7)	84.5 (14.9)	62.0 (15.5)	77.5 (15.8)	64.3 (10.6)	63.0 (12.7)	79.3 (12.3)
	0.5	All	38.9 (20.9)	49.0 (27.9)	38.6 (21.2)	61.3 (21.9)	48.4 (34.7)	57.5 (73.1)	41.2 (25.9)	125.1 (51.1)	43.1 (37.8)	39.6 (28.3)	133.3 (101.5)
		S1	31.7 (10.2)	34.8 (11.5)	30.7 (10.7)	37.6 (7.2)	39.0 (5.6)	44.5 (4.3)	40.0 (6.6)	43.4 (4.9)	40.4 (4.5)	42.1 (5.7)	45.8 (2.8)
		S2					56.4 (15.4)	75.1 (19.8)	55.3 (15.3)	70.7 (18.8)	57.4 (15.0)	55.0 (15.7)	71.5 (16.1)
	0.9	All	51.9 (9.7)	85.4 (12.2)	52.7 (11.8)	73.8 (13.6)	84.6 (18.6)	103.9 (73.0)	72.2 (18.3)	188.9 (50.6)	51.1 (39.1)	48.1 (31.3)	176.7 (112.0)
		S1	34.3 (5.2)	45.5 (4.6)	36.9 (7.9)	41.8 (4.6)	26.3 (6.3)	34.1 (8.9)	28.5 (7.4)	33.0 (8.0)	28.2 (6.5)	30.7 (7.1)	34.6 (8.5)
		S2					50.0 (10.6)	75.7 (19.1)	52.1 (10.5)	79.7 (13.7)	50.8 (10.3)	52.8 (11.6)	78.7 (17.0)
Descending 1	0.1	All	19.9 (12.1)	23.7 (19.0)	25.8 (14.6)	27.2 (19.9)	24.8 (20.2)	47.1 (56.4)	29.6 (19.0)	104.4 (47.3)	29.3 (22.4)	32.5 (19.7)	133.2 (68.4)
		S1	23.4 (11.4)	26.0 (12.7)	27.3 (11.1)	26.1 (11.5)	37.0 (8.4)	43.2 (8.9)	37.1 (9.5)	40.1 (8.4)	39.4 (7.1)	39.1 (8.2)	41.7 (7.4)
		S2					40.9 (20.1)	61.5 (26.4)	41.5 (18.1)	59.0 (19.1)	47.4 (17.5)	48.1 (15.7)	60.9 (18.9)
	0.5	All	15.8 (13.7)	20.5 (18.8)	26.6 (18.3)	21.2 (19.1)	18.9 (16.8)	43.7 (93.1)	25.2 (19.4)	69.6 (55.6)	19.9 (17.7)	23.2 (15.8)	117.5 (89.3)
		S1	21.0 (11.9)	27.2 (12.7)	24.4 (11.8)	23.0 (12.6)	35.6 (9.5)	42.2 (7.4)	37.0 (8.0)	40.7 (8.9)	38.6 (7.1)	38.8 (7.7)	43.0 (7.4)
		S2					41.2 (19.5)	66.2 (22.9)	46.3 (18.3)	62.2 (22.4)	49.5 (17.7)	52.1 (14.8)	65.6 (17.4)
	0.9	All	22.1 (12.0)	31.0 (24.2)	26.4 (13.3)	45.7 (23.7)	37.6 (23.5)	37.3 (67.4)	34.5 (19.7)	69.7 (70.5)	35.2 (30.9)	36.4 (23.2)	111.2 (86.7)
		S1	16.7 (9.6)	22.6 (12.7)	21.7 (9.6)	23.4 (13.9)	19.3 (9.8)	26.7 (12.3)	22.6 (9.5)	24.8 (12.5)	25.8 (8.5)	27.8 (11.3)	32.9 (9.7)
		S2					32.4 (11.5)	52.3 (22.5)	36.1 (14.8)	57.4 (21.5)	37.6 (12.0)	41.0 (14.1)	64.6 (19.9)
Descending 2	0.1	All	23.9 (14.4)	29.8 (21.4)	30.8 (15.3)	32.9 (21.4)	27.1 (21.2)	42.0 (41.8)	32.5 (20.0)	101.6 (35.2)	31.3 (23.4)	29.6 (18.0)	133.7 (72.0)
		S1	25.3 (10.1)	28.5 (12.3)	28.1 (10.3)	26.3 (12.4)	36.1 (10.3)	43.4 (6.5)	36.6 (8.7)	39.2 (9.4)	37.7 (8.6)	38.1 (9.2)	40.5 (9.3)
		S2					42.5 (19.8)	60.8 (25.5)	44.9 (17.2)	58.2 (20.7)	47.3 (17.9)	47.9 (16.0)	64.9 (19.0)
	0.5	All	33.4 (16.0)	47.3 (24.8)	36.9 (16.0)	51.5 (19.9)	30.4 (22.2)	52.7 (53.3)	32.5 (19.6)	116.8 (38.8)	32.7 (22.8)	33.1 (20.2)	130.3 (56.5)
		S1	28.1 (9.6)	33.8 (12.0)	30.9 (8.8)	32.9 (10.2)	33.7 (8.9)	40.1 (9.3)	34.6 (8.7)	38.5 (8.5)	37.9 (7.7)	38.5 (7.9)	41.0 (7.7)
		S2					43.5 (17.0)	62.6 (25.3)	46.6 (17.0)	62.4 (18.9)	49.6 (15.8)	50.9 (15.4)	63.9 (17.7)
	0.9	All	48.2 (4.0)	95.5 (3.3)	72.7 (14.9)	57.6 (11.9)	80.9 (14.0)	120.7 (37.5)	76.6 (15.7)	152.7 (22.9)	87.4 (15.9)	66.8 (23.3)	179.8 (29.6)
		S1	34.9 (2.3)	48.5 (1.8)	46.1 (5.7)	40.0 (3.8)	30.0 (6.3)	37.0 (9.4)	31.1 (8.7)	37.8 (7.1)	30.2 (7.2)	32.3 (9.0)	35.1 (9.5)
		S2					50.2 (5.5)	84.5 (18.7)	64.3 (15.8)	67.3 (12.2)	50.6 (5.4)	64.5 (12.2)	70.3 (12.1)
Leukaemia	1.0	All	45.1 (13.3)	75.2 (25.6)	47.5 (14.1)	52.9 (22.9)	83.1 (15.3)	470.4 (72.9)	79.8 (16.0)	131.9 (39.0)	76.5 (7.1)	87.8 (9.9)	95.6 (25.9)
		S1	30.4 (8.1)	40.2 (10.7)	32.7 (10.3)	32.1 (8.8)	34.6 (7.4)	43.7 (8.6)	38.5 (7.5)	36.8 (5.2)	32.0 (4.9)	43.1 (9.2)	30.0 (5.6)
		S2					55.8 (9.8)	92.0 (14.5)	58.3 (14.3)	64.5 (19.0)	45.3 (8.1)	67.8 (16.7)	38.8 (14.0)
	2.0	All	45.0 (15.2)	67.6 (24.2)	48.3 (14.3)	52.3 (18.3)	70.9 (21.6)	119.5 (43.1)	62.2 (18.7)	133.8 (38.1)	69.9 (24.5)	62.2 (21.9)	145.8 (35.1)
		S1	32.2 (8.4)	37.8 (10.3)	34.3 (10.5)	31.9 (8.1)	34.3 (7.9)	41.7 (8.1)	35.5 (8.7)	38.5 (7.2)	31.9 (9.0)	34.4 (8.8)	35.1 (8.3)
		S2					49.8 (11.5)	74.5 (20.0)	50.0 (13.4)	66.0 (15.3)	46.9 (14.5)	48.0 (15.1)	60.7 (17.9)

Table 5.4: Average true positive rate in settings where w_{Bayes} is sparse.

Setting	ρ	$p = 100$				$p = 500$				$p = 1000$		
		ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Blockwise 1	−0.1	80.1	74.8	84.5	81.7	41.4	47.2	37.6	52.3	28.2	31.1	41.9
	0.1	48.5	52.0	54.0	59.5	28.3	40.7	28.1	45.2	21.7	21.7	33.4
	0.5	56.6	64.4	61.0	71.5	25.8	37.9	28.6	39.6	17.3	20.1	27.5
	0.9	88.5	99.8	87.4	99.9	68.1	64.7	62.1	66.5	28.2	30.5	28.8
Descending 1	0.1	87.1	89.0	84.3	88.2	79.8	84.2	74.6	89.1	81.2	75.5	88.5
	0.5	64.5	76.4	68.0	74.8	60.7	69.5	58.9	70.5	59.8	55.1	72.7
	0.9	47.1	69.6	51.9	73.5	40.9	55.4	45.6	50.9	36.7	43.9	55.0
Descending 2	0.1	44.4	49.1	48.3	49.9	34.3	37.7	32.9	45.9	33.5	31.1	41.8
	0.5	58.9	67.6	61.7	68.0	37.9	41.8	37.7	51.3	34.9	33.1	42.3
	0.9	86.7	99.6	90.1	72.9	75.4	85.3	69.4	84.3	66.0	52.6	72.1

Table 5.5: Average false positive rate in settings where w_{Bayes} is sparse.

Setting	ρ	$p = 100$				$p = 500$				$p = 1000$		
		ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Blockwise 1	−0.1	42.4	42.2	47.3	67.3	7.7	18.2	6.1	23.0	3.0	3.3	13.3
	0.1	24.7	29.1	32.1	37.5	5.9	18.9	6.2	20.2	3.0	2.6	12.7
	0.5	32.1	40.3	34.4	51.3	6.9	18.8	7.2	19.2	2.8	3.6	10.8
	0.9	30.0	65.3	32.1	64.5	13.0	17.8	12.0	26.7	4.8	5.3	12.9
Descending 1	0.1	11.6	15.6	18.6	19.7	3.3	7.7	4.4	19.4	2.1	2.4	12.5
	0.5	9.8	13.6	21.4	14.6	2.5	7.4	3.8	12.6	1.3	1.7	11.1
	0.9	19.0	26.3	23.2	42.3	6.8	6.4	6.0	13.1	3.1	3.2	10.6
Descending 1	0.1	15.5	21.9	23.7	26.0	3.6	6.6	4.9	18.7	2.2	2.1	12.5
	0.5	23.0	39.0	26.8	44.8	4.1	8.6	4.6	21.6	2.3	2.4	12.2
	0.9	32.5	93.9	65.5	51.4	12.5	20.4	12.0	27.2	7.0	5.3	16.4

correlation of the coefficients of LLD and LPD tend to be larger than the correlation of either LLD or LPD with ROAD or sLDA.

Though theoretical results do not guarantee that w_{Bayes} is estimated well by any of the classifiers, it is of interest to see how close the estimated classification vectors and w_{Bayes} are. In Table 5.6, $|w - w_{\text{Bayes}}|_2^2$ is given for all classifiers, while Figure 5.3 and Figure 5.4 compare the coefficients of w_{Bayes} to its estimates for all 100 simulations in two of the settings. We have scaled w_{ROAD} and w_{sLDA} by $\Delta^{\mathrm{T}} w_{\text{Bayes}}$ to make the results comparable.

We note that the estimation error of w_{Bayes} is larger when Bayes risk is low. ROAD and sLDA more often detect large coefficients of w_{Bayes} than LPD and LLD, particularly for settings with large correlations. Still, in many such situations, the estimation errors of ROAD and sLDA are larger than for LPD and LLD.

In Figure 5.5, the relationship between ROAD and sLDA is studied in more details. This is of interest because these classifiers solve closely related optimization problems. The number of nonzero coefficients is almost the same for the two methods, as seen in the left column. The fraction of common nonzero coefficients, defined as

$$\frac{2|w_{\text{ROAD,sLDA}}|_0}{|w_{\text{ROAD}}|_0 + |w_{\text{sLDA}}|_0},$$

where $w_{\text{ROAD,sLDA}}$ is the elementwise multiplication of w_{ROAD} and w_{sLDA}, is shown in the center column of Figure 5.5. This number is often, but far from always, above 50%. Right column shows the number of opposite signs when comparing ROAD and sLDA. This number is often zero, and across all settings never above 20.

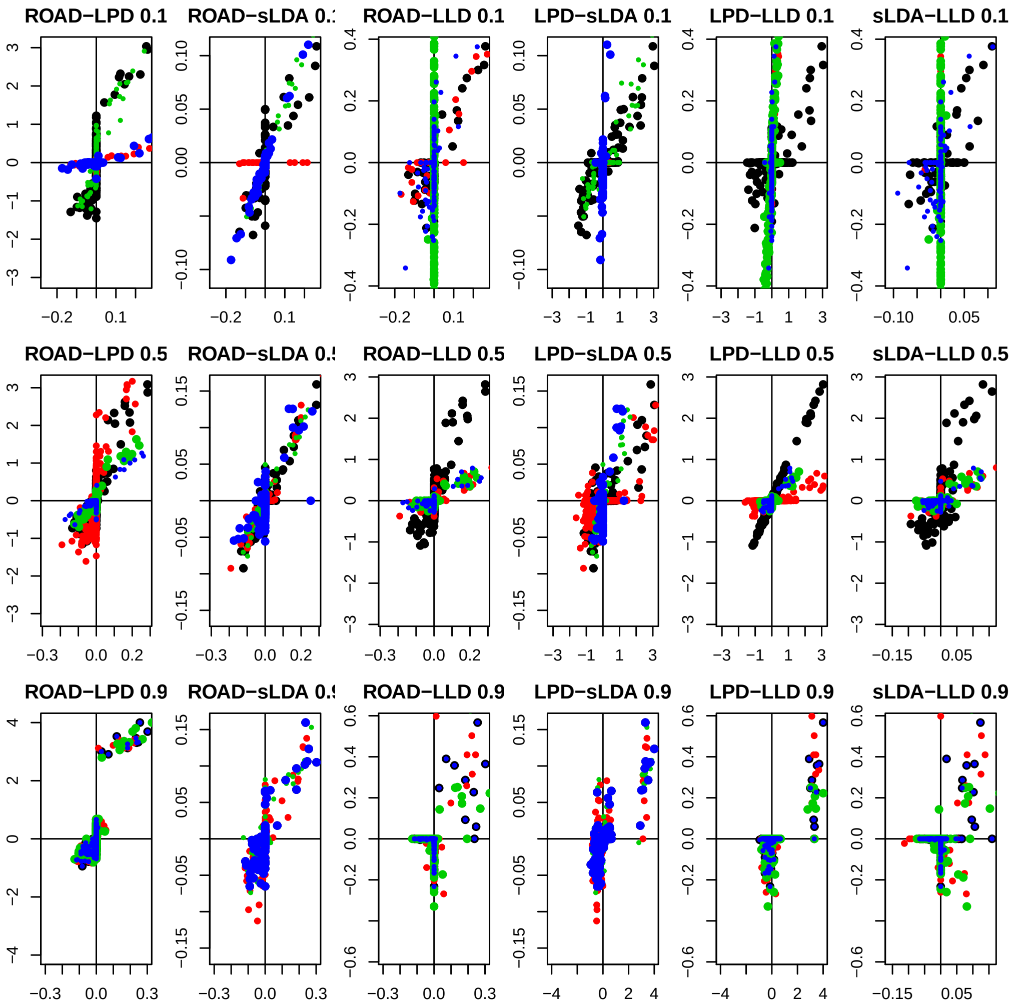

Figure 5.1: Pairs of estimates of the coordinates in w for four simulations in the setting of equal correlation with $p = 500$. In each row, points of the same color originate from the same simulation.

Table 5.6: Average estimation error, $|w - w_{\text{Bayes}}|_2^2$, across 100 simulations.

Setting	ρ	$p = 100$				$p = 500$				$p = 1000$		
		ROAD	LPD	sLDA	LLD	ROAD	LPD	sLDA	LLD	ROAD	sLDA	LLD
Equal	0.1	22.3	6.0	7.2	5.4	21.7	13.0	8.9	10.4	20.8	9.0	17.0
	0.5	97.0	14.6	27.0	15.7	83.7	22.6	29.6	22.1	77.7	35.7	24.3
	0.9	2425.0	402.6	1759.6	844.6	2558.3	453.7	1227.3	929.3	2478.8	1573.5	947.1
Blockwise 1	−0.1	104.7	95.4	55.5	89.7	134.4	127.0	118.4	115.5	143.6	121.0	123.0
	0.1	3.3	4.1	2.0	4.0	2.6	26.7	2.1	6.9	2.6	2.3	13.2
	0.5	9.2	6.7	4.7	6.4	7.1	24.3	5.6	10.0	6.7	6.0	9.5
	0.9	247.8	65.4	69.9	66.3	199.8	106.2	118.3	101.0	184.0	150.3	122.0
Blockwise 2	−0.1	29.7	24.4	23.8	24.9	29.6	39.9	26.9	29.5	29.5	27.8	41.4
	0.1	4.7	5.5	2.3	4.8	3.8	31.6	3.2	9.0	3.8	3.3	19.6
	0.5	15.0	7.2	7.1	6.9	11.6	11.6	8.2	10.7	11.6	9.7	15.1
	0.9	522.2	107.7	134.9	121.4	422.5	172.3	201.7	152.3	315.8	258.9	199.1
Descending 1	0.1	11.4	3.7	4.3	4.1	12.7	8.3	5.0	10.1	11.4	5.1	13.0
	0.5	6.1	2.4	3.1	2.7	5.8	14.4	3.0	5.4	5.4	2.8	11.4
	0.9	21.6	33.0	12.4	34.1	24.1	42.7	23.8	43.8	31.2	31.7	44.8
Descending 2	0.1	16.3	5.3	5.0	5.6	16.7	6.9	7.9	8.7	15.7	6.2	14.2
	0.5	52.7	17.7	17.8	17.5	47.2	25.0	21.9	25.3	46.1	26.4	27.8
	0.9	2183.2	718.5	1332.1	1043.1	2256.3	964.4	956.5	1008.0	2185.7	1291.5	1079.0
Leukaemia	1.0	213.0	48.5	66.3	57.8	3598.5	300.5	1362.0	514.3	13365.3	3801.9	1400.0
	2.0	212.0	45.3	64.9	60.8	341.1	77.9	138.1	91.2	295.5	129.6	97.7

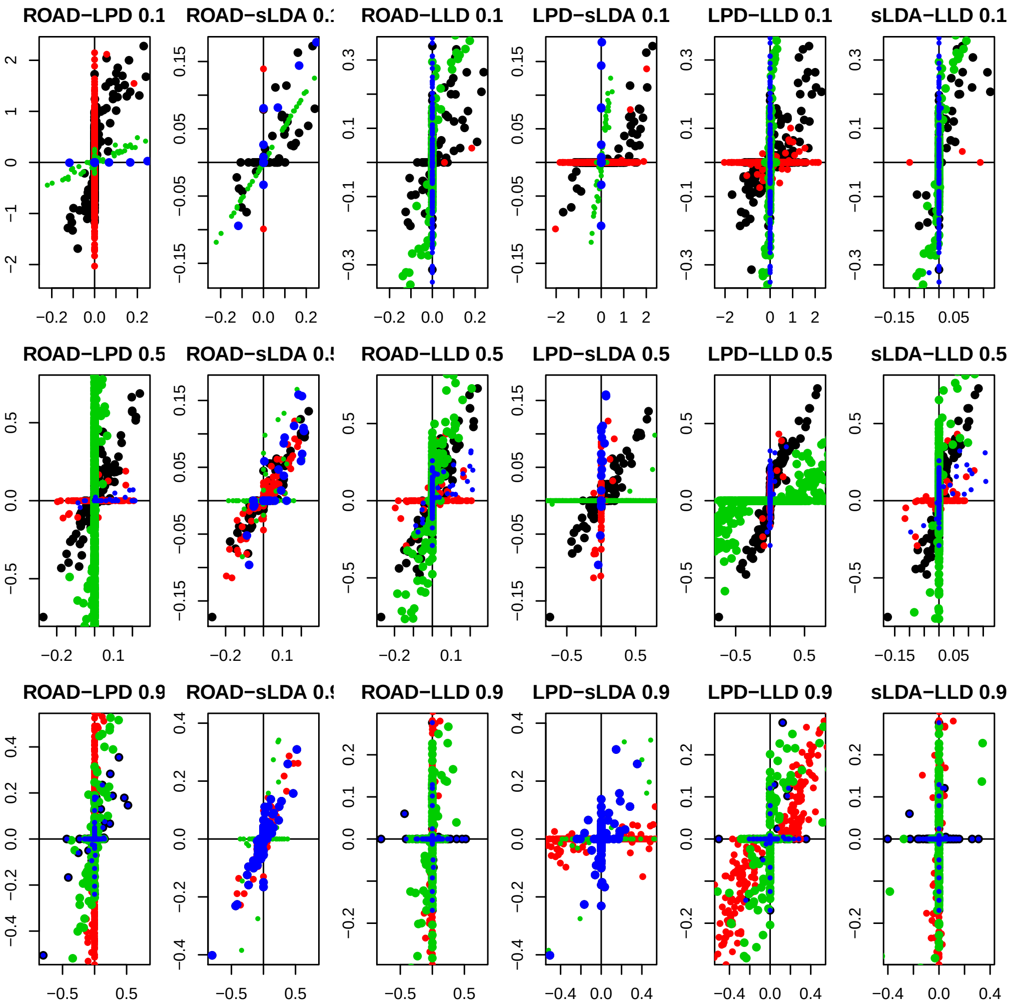

Figure 5.2: Pairs of estimates of the coordinates in w for four simulations in the setting of blockwise correlation 1 with p=500. In each row, points of the same color originate from the same simulation.

5.4.5 Analysis of real datasets

We compare the behaviour of the classification procedures on two real datasets: The leukaemia data (Golub et al., 1999), and the lung cancer data (Gordon et al., 2002). The leukaemia data consist of $p = 7129$ genes measured on a training set with $n_0 = 27$ samples with acute lymphoblastic leukaemia (ALL) and $n_1 = 11$ samples with acute myeloid leukaemia (AML), and a test dataset with 20 ALL and 14 AML samples. The lung cancer set contains $p = 12533$ genes measured on $n_0 = 16$ adenocarcinoma and $n_1 = 15$ mesothelioma samples in the training set, and 134 adenocarcinoma and 15 mesothelioma samples in the test set. For both datasets we only include the 3000 variables with largest t-statistics.

The results are shown in Table 5.7 and Table 5.8. Regarding the leukaemia dataset LPD, LLD and LLD-S2 perform best when considering test error, closely followed by ROAD, ROAD-S2 and LPD-S2. Those classifiers select between 24 and 60 variables. For the lung data ROAD, LPD and LLD have lowest test error, but LPD and LLD select extremely many variables, whereas ROAD selects 60 variables only and is therefore

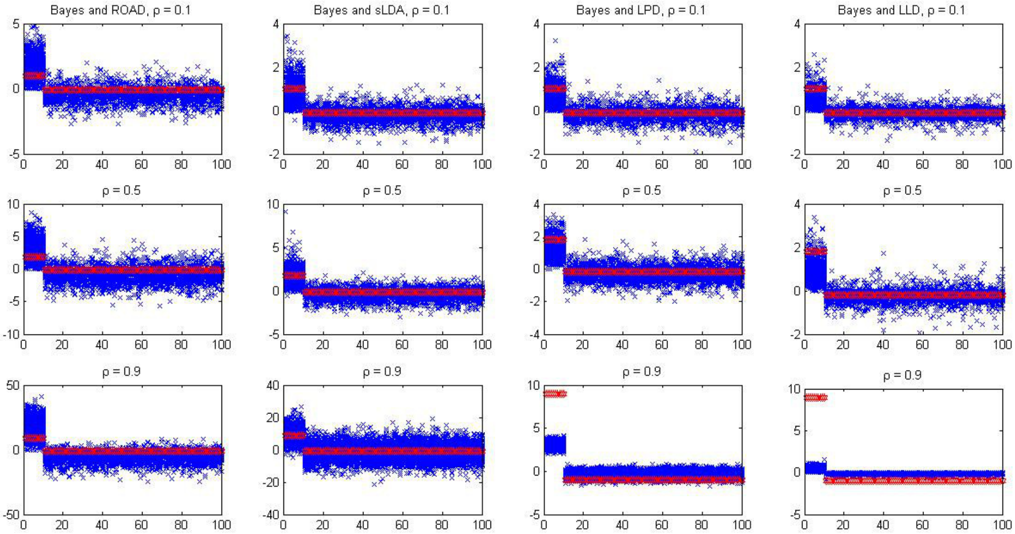

Figure 5.3: Comparison of the coefficients of w_{Bayes} (red) and the estimated values hereof (blue), for each of the p=100 coordinates in the equal correlation setting, and for all 100 simulations.

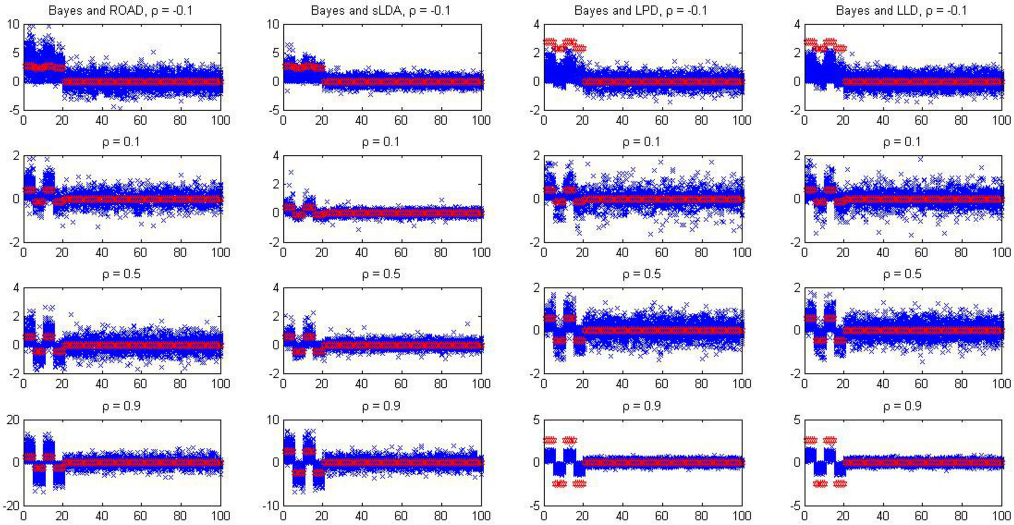

Figure 5.4: Comparison of the coefficients of w_{Bayes} (red) and the estimated values hereof (blue), for each of the p=100 coordinates in the blockwise correlation 1 setting, and for all 100 simulations.

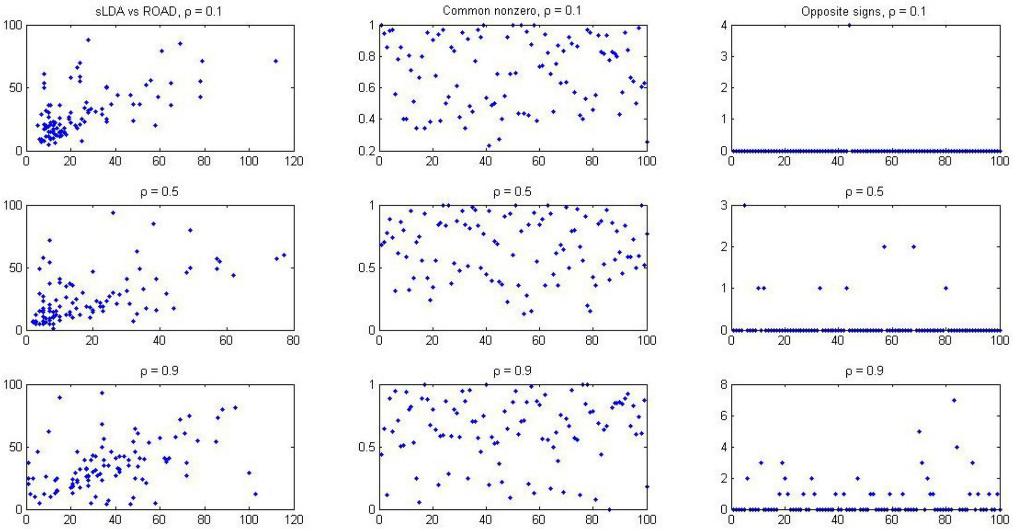

Figure 5.5: Comparison of the number of nonzero coefficients in ROAD and sLDA (left column), the fraction of common selections of ROAD and sLDA out of their total number of selections in each of the 100 simulations (center column), and the number of coefficients in w_{ROAD} and w_{sLDA} where the sign differ in each simulation (right column). All plots are in the descending correlation 1 setting with $p = 500$.

Table 5.7: Analysis of the leukaemia data. The S1 and S2 methods are based on an initial selection of 50 and 100 variables, respectively.

Procedure	Training Error	Testing Error	Number of selected genes
ROAD	0	2	60
ROAD-S1	0	5	13
ROAD-S2	0	2	26
LPD	0	1	37
LPD-S1	1	2	19
LPD-S2	0	2	38
sLDA	0	3	35
sLDA-S1	0	5	14
sLDA-S2	0	3	27
LLD	0	1	35
LLD-S1	3	6	2
LLD-S2	0	1	24

preferable. Since a large number of variables is needed to obtain a good classifier in this situation, the preselection methods do not work well.

5.5 Conclusions

Our analysis gives no unique recommendation of the most preferable classifier for high dimensional correlated data, but it does give new insight into the four classifiers considered. The origin of ROAD and sLDA is the same, and their performances are closely related, but far from equal. Though sLDA approaches the optimization problem in (5.5) more directly than ROAD, ROAD is preferable since it often gains slightly smaller classification error. LPD gives a less sparse classifier than ROAD and sLDA, but obtains comparable classification errors. LLD is primarily a good classifier when p

Table 5.8: Analysis of the lung data. The S1 and S2 methods are based on an initial selection of 50 and 100 variables, respectively.

Procedure	Training Error	Testing Error	Number of selected genes
ROAD	0	3	60
ROAD-S1	0	8	8
ROAD-S2	0	7	20
LPD	0	3	2719
LPD-S1	0	10	16
LPD-S2	0	9	32
sLDA	0	7	22
sLDA-S1	0	8	4
sLDA-S2	0	21	6
LLD	0	3	851
LLD-S1	1	25	3
LLD-S2	1	23	5

is small.

Considering the pre selection versions, S1 is not recommended in correlated situations, while S2 is often useful for reducing the computational cost without increasing the classification error. S2 cannot generally be expected to reduce the number of variables included in the resulting classifier though.

ROAD and sLDA identify large coefficients of w_{Bayes} better than LPD and LLD, but this do not result in a lower estimation error of w_{Bayes} in general.

When the computational time is a serious concern, the overall advice is to apply ROAD. In general, it is preferable to apply multiple classifiers, particularly ROAD and LPD, and either select the classifier that performs best on a test set, or use a combination of the suggested classifiers.

5.6 Appendix

This appendix considers the theory behind the numerical algorithms to compute the various classifiers.

5.6.1 ROAD: A coordinate descent algorithm

Fan et al. (2012) present a coordinate descent algorithm for minimizing (5.9). Before presenting the algorithm and its convergence properties, we describe the relation between the original minimization problem (5.4), and the two alternatives (5.6) and (5.9).

Transforming the original minimization problem

We outline here the reasoning behind rewriting (5.4) to (5.6), and refer to Chapter 5 of Boyd and Vandenberghe (2004) for a detailed description. Consider a general minimization problem

$$\min_{f_i(x)\leq 0, i\in I, h_j(x)=0, j\in J} f_0(x), \qquad x \in \mathbb{X}, \tag{5.12}$$

with $\mathbb{X}'$ the subset of $\mathbb{X}$ consisting of points satisfying the restrictions. Define the Lagrangian:

$$L(x,\lambda,\nu) = f_0(x) + \sum_{i\in I} \lambda_i f_i(x) + \sum_{j\in J} \nu_j h_j(x), \qquad \lambda_i \geq 0, \quad \nu_j \geq 0.$$

The Lagrange dual function is $g(\lambda, \nu) = \inf_{x \in \mathbb{X}} L(x, \lambda, \nu)$, which clearly satisfies

$$g(\lambda, \nu) \leq \min_{f_i(x) \leq 0, i \in I, h_j(x) = 0, j \in J} f_0(x) \qquad \forall \lambda, \nu \geq 0. \tag{5.13}$$

We seek (λ^*, ν^*) which maximizes g. If there is equality in (5.13) with $(\lambda, \nu) = (\lambda^*, \nu^*)$, we say that strong duality holds, and in this case minimizing $f_0(x)$ with respect to x is equivalent to maximizing $g(\lambda, \nu)$ with respect to (λ, ν).

For the general problem in (5.12) it is known that strong duality holds under the following two conditions:

(i) f_i, $i \in I$ are convex, and $h_j(x)$, $j \in J$, are linear,

(ii) there exists $x \in \text{relint}(\mathbb{X}')$ such that $f_i(x) < 0$, $i \in I$, and $h_j(x) = 0$, $j \in J$ (Slaters condition).

Transforming (5.4) to (5.6) corresponds to the general case with $f_0(w) = w^{\mathrm{T}}\hat{\Sigma}w$ and $f_1(w) = |w|_1 - c$. These functions are clearly convex, $|w|_1 < c$ is an open set, and the intersection with $\hat{\Delta}^{\mathrm{T}}w = 1$ is relative open.

Next consider replacing (5.6) by (5.9). Using a large value of γ, the two give almost the same result (this is true in the limit $\gamma \to \infty$ as shown in Theorem 6.7 of Ruszczynski (2006)), and in the simulations reported in this paper we use $\gamma = 10$. The latter value is based on the recommendation in Fan et al. (2012), where simulations show that the classification error is insensitive to the value of γ when λ is chosen by cross-validation.

The minimization step

For fixed λ and γ the minimization is performed iteratively by minimizing over each coordinate in turn. The minimization is done for a set of λ values, $\lambda_1 > \lambda_2 > \cdots > \lambda_{K-1}$, where the solution for $\lambda = \lambda_i$ is used as initial guess in the search when $\lambda = \lambda_{i+1}$, and the initial guess is $w = 0$ when $\lambda = \lambda_1$.

Minimizing (5.9) with respect to one coordinate of w, say w_1, while the $p - 1$ dimensional vector w_2 of the remaining coordinates is fixed, means minimizing the convex function:

$$\begin{aligned} g(w_1) = {} & \tfrac{1}{2} w_1^2 (\Sigma_{11} + \gamma \Delta_1^2) + + w_1 \{\Sigma_{12} w_2 + \gamma \Delta_1 (\Delta_2^{\mathrm{T}} w_2 - 1)\} + \lambda |w_1| + \lambda |w_2|_1 \\ & + \tfrac{1}{2} \{ w_2^{\mathrm{T}} \Sigma_{22} w_2 + \gamma (\Delta_2^{\mathrm{T}} w_2 - 1)^2 \}. \end{aligned} \tag{5.14}$$

The derivative $g'(w_1)$ has a jump at $w_1 = 0$, and from this one finds that $g(w_1)$ has minimum at

$$w_1 = \begin{cases} \frac{-(\Sigma_{12} + \gamma \Delta_1 \Delta_2^{\mathrm{T}}) w_2 + \gamma \Delta_1 + \lambda}{\Sigma_{11} + \gamma \Delta_1^2} & \text{if } (\Sigma_{12} + \gamma \Delta_1 \Delta_2^{\mathrm{T}}) w_2 - \gamma \Delta_1 - \lambda > 0, \\ \frac{-(\Sigma_{12} + \gamma \Delta_1 \Delta_2^{\mathrm{T}}) w_2 + \gamma \Delta_1 - \lambda}{\Sigma_{11} + \gamma \Delta_1^2} & \text{if } (\Sigma_{12} + \gamma \Delta_1 \Delta_2^{\mathrm{T}}) w_2 - \gamma \Delta_1 + \lambda < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.15}$$

From (5.15) it is seen that $w = 0$ minimizes (5.9) if and only if $\lambda \geq |\gamma \Delta_i|$ for all i. For this reason we start by calculating the solution with $\lambda = \lambda_1$, where λ_1 is slightly smaller than $\lambda_{\max} = \max_i |\gamma \Delta_i|$. For the calculations in this paper we have used $\lambda_1 = \lambda_{\max}(1 - \epsilon)$ and $\lambda_k = \epsilon \lambda_1$ with $\epsilon = 10^{-3}$.

Convergence properties

To study the convergence properties of the above algorithm we use a general result for the coordinate descent algorithm when minimizing

$$f(w) = f_0(w) + \sum_{k=1}^{p} f_k(w_k).$$

In our setup we take $f_0(w) = \frac{1}{2}w^{\mathrm{T}}\Sigma w + \frac{1}{2}\gamma(w^{\mathrm{T}}\mu_d - 1)^2$, and $f_k(w_k) = \lambda|w_k|$ for $k = 1, \dots, p$. Theorem 5.1 of Tseng and Mangasarian (2001) requires

(i) f_0 is continuous,

(ii) $w_k \to f(w_1, \dots, w_k, \dots, w_p)$ is quasiconvex and hemivariate,

(iii) for $i = 0, 1, \dots, p$ it holds that $\liminf_{w \to w_0} f_i(w) \geq f_i(w_0)\ \forall w_0$,

(iv) the domain of f_0 is open,

(v) $f_0 \to \infty$ on every boundary point of its domain,

(vi) $\{w : f(w) \leq f(w_0)\}$ is bounded, where w_0 is the initial value in the search.

In our case (i), (iii), (iv) and (v) are clearly satisfied. Condition (ii) holds if f is strictly convex in each of its arguments. From (5.14) this is clearly seen to be the case since $\lambda|w_1|$ is convex, and $w_1^2(\Sigma_{11} + \gamma\Delta_1^2)$ is strictly convex. For proving (vi) we note that in our case $f(w_0) > 0$, and since $w^{\mathrm{T}}\Sigma w \geq \lambda_{\min}(\Sigma)|w|_2^2$, we have that $f(w) \leq f(w_0)$ implies $|w|_2^2 \leq 2f(w_0)/\lambda_{\min}(\Sigma)$.

Consider the coordinate descent algorithm, with coordinates being updated at proportional rates, leading to a sequence $w^1, w^2, \dots$ Under the assumptions (i)–(vi) it is shown in Tseng and Mangasarian (2001) that the sequence of updates is bounded, and any accumulation point is a coordinatewise minimum of f. Since in our case f_0 is convex and smooth, and f_k is convex, it follows that the coordinatewise minimum of f is a global minimum.

5.6.2 Sparse linear discriminant analysis

We describe here the algorithm of Wu et al. (2011) for solving the minimization problem (5.5). As already mentioned the solution $\hat{w}_c$ is piecewise linear in c. What we need then is to describe the points in c, where there is a change from one linear relation to another, and to describe the linear relation itself. For notational simplicity we use the population version (5.4) instead of the sample version (5.5).

Writing $w = w^+ - w^-$, with w^+ and w^- denoting the positive and negative part of w, we consider the minimization problem

$$\min_{w^+, w^-} (w^+ + w^-)^{\mathrm{T}}\Sigma(w^+ + w^-) \qquad \text{subject to:} \tag{5.16}$$
$$\sum_{i=1}^{p}(w_i^+ + w_i^-) - c \leq 0, \quad \sum_{i=1}^{p}\Delta^{\mathrm{T}}(w_i^+ - w_i^-) - 1 = 0, \quad -w_i^+ \leq 0, \quad -w_i^- \leq 0.$$

From the description below it follows that the solution to the latter problem actually has $w_i^- = 0$ if $w_i^+ > 0$ and vice versa, so that the solution gives the solution to the original problem (5.4).

Introducing Lagrangian multipliers ν, $\lambda \geq 0$, $\lambda_i^+ \geq 0$ and $\lambda_i^- \geq 0$, $i = 1, \ldots, p$, the KKT conditions (Boyd and Vandenberghe, 2004) for the convex minorization problem (5.16) are

$$2\Sigma_{i*}w + \lambda - \lambda_i^+ + \nu\Delta_i = 0, \quad -2\Sigma_{i*}w + \lambda - \lambda_i^- - \nu\Delta_i = 0, \quad i = 1, \ldots, p,$$
$$\lambda\Big(\sum_{i=1}^{p}(w_i^+ + w_i^-) - c\Big) = 0, \quad \lambda_i^+ w_i^+ = 0, \quad \lambda_i^- w_i^- = 0, \quad i = 1, \ldots, p, \tag{5.17}$$

where Σ_{i*} is the i'th row of Σ. We are looking at these equations as a function of c, and speak of the active set $\mathcal{A} = \{i : w_i = w_i(c) \neq 0\}$.

Generally, we have

$$|2\Sigma_{i*}w + \nu\Delta_i| \leq \lambda, \tag{5.18}$$

since $\lambda_i^+ \geq 0$ and $\lambda_i^- \geq 0$. For $i \in \mathcal{A}$ we find

$$|2\Sigma_{i*}w + \nu\Delta_i| = \lambda \;\text{ and }\; \operatorname{sgn}(2\Sigma_{i*}w + \nu\Delta_i) = -\operatorname{sgn}(w_i(\lambda)), \tag{5.19}$$

since $w_i^+ > 0$ implies $\lambda_i^+ = 0$ and $\lambda = -(2\Sigma_{i*}w + \nu\Delta_i)$, and $w_i^- > 0$ implies $\lambda_i^- = 0$ and $\lambda = (2\Sigma_{i*}w + \nu\Delta_i)$. From (5.17) it is also seen that $w_i^+ > 0$ implies $\lambda_i^- = 2\lambda$, so that $w_i^- = 0$ when $\lambda > 0$ and, similarly, $w_i^- > 0$ implies $w_i^+ = 0$.

Since we have a piecewise linear solution as a function of c, we see from (5.19) that a variable j is removed from the active set $\mathcal{A}$, when w_j hits zero. A variable $j \notin \mathcal{A}$ is entered to the active set when an infinitesimal move in the current direction leads to $|\Sigma_{i*}w + \nu\Delta_i| > \lambda$.

The linear part

Defining $\xi_i = \operatorname{sgn}(w_i)$, and combining (5.17) and (5.19) we obtain

$$2\Sigma_{i\mathcal{A}}w_{\mathcal{A}} + \nu\Delta_{\mathcal{A}} + \xi_i\lambda = 0, \qquad i \in \mathcal{A},$$
$$\Delta_{\mathcal{A}}^{\mathrm{T}}w_{\mathcal{A}} = 1, \qquad \xi_{\mathcal{A}}^{\mathrm{T}}w_{\mathcal{A}} = c.$$

Differentiating with respect to c gives

$$2\Sigma_{i\mathcal{A}}\frac{\partial w_{\mathcal{A}}}{\partial c} + \frac{\partial \nu}{\partial c}\Delta_{\mathcal{A}} + \xi_i\frac{\partial \lambda}{\partial c}, \qquad i \in \mathcal{A},$$
$$\Delta_{\mathcal{A}}^{\mathrm{T}}\frac{\partial w_{\mathcal{A}}}{\partial c} = 0, \qquad \xi_{\mathcal{A}}^{\mathrm{T}}\frac{\partial w_{\mathcal{A}}}{\partial c} = 1,$$

which shows that the direction for a linear part is the solution to a system of $|\mathcal{A}| + 2$ linear equations.

We now calculate the step size until a variable is either included in or excluded from $\mathcal{A}$. Let d_i be the solution to

$$\begin{cases} \left|2\Sigma_{i\mathcal{A}}(w_{\mathcal{A}} + d\frac{\partial w_{\mathcal{A}}}{\partial c}) + (\nu + d\frac{\partial \nu}{\partial c}\Delta_i)\right| = \lambda + d\frac{\partial \lambda}{\partial c}, & i \notin \mathcal{A}, \\ w_i + d\frac{\partial w_i}{\partial c} = 0, & i \in \mathcal{A}, \end{cases}$$

and let the stepsize be $s = \arg\min_i d_i$.

The algorithm terminates when (5.18) is no longer fulfilled.

To start the algorithm we note that $c_0 = \min\{1/|\Delta_i|\}$ is the smallest possible value of c. Let $I = \arg\max_i\{|\Delta_i|\}$, and let w^0 be zero except $w_I^0 = 1/\Delta_I$. It is not difficult to see, that when c is increased above c_0, a better solution can be found including one more variable than I. The active set at c_0 therefore contains two variables where the second variable is the one giving the largest value of λ, such that (5.18) is satisfied for the remaining $p-2$ variables. For the two variables I and j at c_0, λ and ν are determined by

$$2\Sigma_{II}\frac{1}{\Delta_I}+\lambda+\nu\Delta_I=0, \quad 2\Sigma_{jI}\frac{1}{\Delta_I}+\lambda+\nu\Delta_j=0.$$

This gives $\lambda_j = 2(\Sigma_{II}\Delta_j/\Delta_I - \Sigma_{jI})/(\Delta_I-\Delta_j)$, and the active set therefore becomes I and $\arg\max_{j\neq I}\lambda_j$.

5.6.3 Linear programming discriminant: A linear program

The minimization of

$$\min_{|\Sigma w-\Delta|_\infty\leq\gamma}|w|_1 \tag{5.20}$$

is done through a linear program similar to the one for the Dantzig selector (Candes and Tao, 2007).

First we rewrite (5.20) in order to avoid absolute values,

$$\begin{array}{lll} \min & \sum_{i=1}^{p} u_i & \\ \text{subject to:} & -w_i-u_i\leq 0, & w_i-u_i\leq 0, \\ & -\Sigma_{i*}w+\Delta_i-\gamma\leq 0, & \Sigma_{i*}w-\Delta_i-\gamma\leq 0, \end{array}$$

for $i=1,\dots,p$.

With $z=(w,u)$ this is an instance of a general linear program of the form:

$$\min_z c_0^{\mathrm{T}}z,$$
$$f_i(z)=c_i^{\mathrm{T}}z+d_i\leq 0, \qquad i=1,\dots 4p,$$

where in our case $c_0=(0_p,1_p)^{\mathrm{T}}$.

In the following we work along the lines of Candes and Romberg (2005). For further information see also Chapter 11 of Boyd and Vandenberghe (2004). Introducing Lagrangian multipliers $\lambda_1,\dots,\lambda_{4p}$ a solution (w,λ) must fulfil the KKT-conditions:

$$c_{0j}+\sum_{i=1}^{m}\lambda_i c_{ij}=0, \qquad j=1,\dots,2p,$$
$$\lambda_i f_i(z)=0, \quad f_i(z)\leq 0, \qquad i=1,\dots,4p.$$

Even though we expect the solution to be on the boundary of the feasible set, we want to stay in the interior during the steps of the search. To this end, we search for (z,λ) such that $r_\tau=(\mathrm{r}_{\mathrm{dual}},\mathrm{r}_{\mathrm{cent}})$ with

$$\mathrm{r}_{\mathrm{dual}}=c_0+\sum_{i=1}^{4p}\lambda_i c_i, \qquad \mathrm{r}_{\mathrm{cent}}=\left(-\lambda_1 f_1(z)-\frac{1}{\tau},\dots,-\lambda_{4p}f_{4p}(z)-\frac{1}{\tau}\right)^{\mathrm{T}}$$

is close to 0_{2p+m}. Therefore, we search for a direction $(\Delta z,\Delta\lambda)$ and a step size s such that r_τ at the new point is closer to zero than r_τ at the old point.

Finding the search step

Using a first order Taylor approximation we find the direction by solving

$$r_\tau(z,\lambda)+\frac{\partial r_\tau(z,\lambda)}{\partial(z,\lambda)}\begin{pmatrix}\Delta z\\ \Delta\lambda\end{pmatrix}=0_{6p}.$$

The derivatives of r_τ are

$$\frac{\partial \mathrm{r}_{\text{dual}}}{\partial z_i} = 0, \qquad \frac{\partial \mathrm{r}_{\text{dual}}}{\partial \lambda_j} = c_j,$$
$$\frac{\partial \mathrm{r}_{\text{cent,j}}}{\partial z} = \lambda_j c_j^{\mathrm{T}}, \qquad \frac{\partial \mathrm{r}_{\text{cent,j}}}{\partial \lambda} = f_j(z) e_j,$$

where e_j is the zero vector of length p, except $e_{jj} = 1$.

To find the step size s our first requirement is that $z + s\Delta z$ and $\lambda + s\Delta\lambda$ is in the interior of the search set, meaning that $f_i(z + s\Delta z) \leq 0$ and $\lambda + s\Delta\lambda > 0$. To this end, we only need to keep track of coordinates where $c_i^{\mathrm{T}}\Delta z > 0$ and $\Delta\lambda_i < 0$. Define $\mathbf{I}_f^+ = \{i : c_i^{\mathrm{T}}\Delta z > 0\}$ and $\mathbf{I}_\lambda^- = \{i : \Delta\lambda < 0\}$, and set

$$s = 0.99 \cdot \min\Big\{1, \Big\{\frac{-f_i(z)}{c_i^{\mathrm{T}}\Delta z}, \quad i \in \mathbf{I}_f^+\Big\}, \Big\{\frac{-\lambda_i}{\Delta\lambda_i}, \quad i \in \mathbf{I}_\lambda^-\Big\}\Big\}.$$

The second requirement is that the step should lead to a value of r_τ closer to zero. For this we iteratively set $s = \beta s$, for some β strictly less than one, until s fulfils

$$|r_\tau(z + s\Delta z, \lambda + s\Delta\lambda)|_2 \leq (1 - \alpha)|r_\tau(z, \lambda)|_2.$$

We have used the algorithm with $\alpha = 0.01$ and $\beta = 0.5$.

The minimization procedure described above is applied iteratively for increasing values of τ. The *surrogate duality gap* $\eta = -\lambda^{\mathrm{T}} f(z)$ approximates $c_0^{\mathrm{T}} z^* - c_0^{\mathrm{T}} z$, and thereby indicates the extent of deviation from the optimal solution. We stop the iterations when η gets below 10^{-3}.

5.6.4 Linear lasso discriminant

Since LLD in (5.11) is simply an application of the LASSO, its path can be computed by the LARS-algorithm. We give a brief summary based on Efron et al. (2004). See also Rosset and Zhu (2007).

LARS utilizes that the solution path of the LASSO is piecewise linear as a function of c with the gradient altering only when a variable is either introduced to or excluded from the set of active variables $\mathcal{A}$. Both the algorithm and derivation of LARS is related to sLDA, but LARS has a different criterion for selecting variables in each step.

LARS includes variables according to a *correlation* measure $\omega_i = \Sigma_{i*}(\Delta - \Sigma w)$. A variable is included when the correlation measure reaches the value for the current active set. The first variable to enter the active set is therefore $\arg\max_i \Sigma_{i*}\Delta$.

Let $\mathcal{A}$ denote the current active set, meaning that $w_i = 0$ for $i \notin \mathcal{A}$. The active set fulfills the property

$$\mathcal{A} = \{i : |\omega_i| = \Omega\}, \qquad \text{where } \Omega = \max_i |\omega_i|.$$

For the linear part of $w_{\mathcal{A}}$ the direction of change is given as the direction for which all the active variables continue to have the same correlation measure. To calculate the direction we define $\Sigma_{\mathcal{A}}$ as the $p \times |\mathcal{A}|$ matrix with columns from Σ having column number in $\mathcal{A}$. The direction is then calculated as

$$v_{\mathcal{A}} = -(\Sigma_{\mathcal{A}}^{\mathrm{T}} \Sigma_{\mathcal{A}})^{-1} \operatorname{sgn}(w_{\mathcal{A}}).$$

The direction needs to be recalculated only when a variable is either introduced to or excluded from $\mathcal{A}$. A variable in $\mathcal{A}^{\mathrm{C}}$ is introduced when the correlation measure of that variable equals the correlation measure of the variables of the active set. This occurs for a step size $d = d_1$ where

$$d_1 = \min_{i \in \mathcal{A}^C} \left\{ \frac{\Omega - |\omega_i|}{1 - \operatorname{sgn}(\omega_i) \Sigma_{i*}^{\mathrm{T}} \Sigma_{\mathcal{A}} v_{\mathcal{A}}} \right\}.$$

A variable is excluded from $\mathcal{A}$ when a coordinate in $\mathcal{A}$ hits zero. This occurs when $d = d_2$ where

$$d_2 = \min_{i \in \mathcal{A}:\, \operatorname{sgn}(w_i) \operatorname{sgn}(v_i) = -1} \left| \frac{w_i}{v_i} \right|.$$

The final step size is therefore $d = \min(d_1, d_2)$, and the appropriate variable is either added to or removed from $\mathcal{A}$, before a new step direction is calculated.

Due to Theorem 1 of Efron et al. (2004), LARS results in the full path of the LASSO solution under the one-at-a-time assumption, meaning that only one variable is allowed to be added to or removed from $\mathcal{A}$ at the same time. When the mean corrected MLEs are used for Σ and Δ, this holds with probability one.

Bibliography

Bak, B. A., M. F. Grøn, and J. L. Jensen (2015). Classification error of the thresholded independence rule. *Scandinavian Journal of Statistics 42*, 34–42.

Bak, B. A. and J. L. Jensen (2014). On oracle efficiency of the road classification rule. Technical report, Aarhus University.

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.

Boyd, S. and L. Vandenberghe (2004). *Convex Optimization*. Cambridge University Press.

Cai, T. and W. Liu (2011). A direct estimation approach to sparse linear discriminant analysis. *Journal of the American Statistical Association 106*(496), 1566–1577.

Candes, E. and J. Romberg (2005). l1-magic: Recovery of sparse signals via convex programming. Technical report, Caltech.

Candes, E. and T. Tao (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313–2351.

Clemmensen, L., T. Hastie, D. M. Witten, and B. Ersbøll (2011). Sparse discriminant analysis. *Technometrics 53*, 406–413.

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.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *The Annals of Statistics 32*, 407–499.

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

Fan, J., Y. Feng, and X. Tong (2012). A road to classification in high dimensional space: The regularized optimal affine discriminant. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 745–771.

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.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, and C. D. Bloomfield (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science 286*, 531–537.

Gordon, G. J., R. V. Jensen, L.-L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy, W. G. Richards, D. J. Sugarbaker, , and R. Bueno (2002). Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer Research 62*, 4963–4967.

James, G. M., P. Radchenko, and J. Lv (2009). Dasso: connections between the dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*, 127–142.

Mai, Q. and H. Zou (2012). A note on the connection and equivalence of three sparse linear discriminant analysis methods. *Technometrics 55*(2), 243–246.

Mai, Q., H. Zou, and M. Yuan (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika 99*(1), 29–42.

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

Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. *The Annals of Statistics 35*(3), 1030.

Ruszczynski, A. (2006). *Nonlinear Optimization*. Princeton University Press.

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

Tseng, P. and C. O. L. Mangasarian (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications 109*(3), 475–494.

Wu, M. C., L. Zhang, and X. Lin (2011). Two-group classfication using sparse linear discriminants analysis. Technical report, Department of Biostatistics, Harvard School of Public Health.