Statistical inference for high dimensional data

Jens Ledet Jensen

Version number: 21-02-2023

Contents

Preface

These notes are meant to give an introduction to the problems one is faced with when analyzing high dimensional data. In some of your bachelor courses you have typically looked at low dimensional data (sometimes even one dimensional!) and made tests for a simple hypothesis, or making a sequential series of tests within a more complicated model. To back up this, you use type I and type II errors and the classical Neyman-Pearson test theory, trying to find a most powerful test. In these notes each measurement is high dimensional, say, 1000 variables or higher, and the number of samples is small. The analysis often has a more exploratory nature, meaning that we are not testing a predefined theory, rather we try to find a set of variables that can be of interest to us.

The notes are of a theoretical nature. The exercises, however, are mostly practical, involving a large amount of R-programming. This will give you a good training in R, but the main purpose of the exercises is to familiarize you with the methods, make the theory less abstract and to make you think about the problems in high dimensional analysis. Also, the exercises allow you to play around with the procedures to learn more on your own.

The notes collect material from a number of articles and the notation typically varies from one article to the next. I have tried to keep a uniform notation within the notes, which means that the notation here often differs from the underlying article. As an example I use here m for the number of variables, whereas in many articles the notation is p instead. I have voted for m since p appears as probability and p-values.

The notes center around finding differences between two groups and are divided into two parts. In the first part we try to find variables that we believe behave differently in the two groups. Two important aspects here are *family wise error rate* (FWER) and *false discovery rate* (FDR). In the second part we look at the possibility of building a classifier to distinguish between the two groups. The problem here is that often only a small number of variables behave differently in the two groups and the remaining variables add noise to a potential classifier.

CHAPTER

1 Multiple testing

1.1 Motivating example

Around year 2000 microarrays were introduced in medical research. The microarray allows for simultaneous determination of the expression level of a large number of genes from a cell sample. This can then give information on a molecular level for a disease, supplementing more traditional directly observable features. In cancer research this is seen as an important tool that can be used to divide patients into subgroups that can be given different treatments.

In Dyrskjøt et al. (2003) (Identifying distinct classes of bladder carcinoma using microarray) data from 40 patients with bladder cancer is considered. The patients are divided, from traditional observations, into three groups. The group *Ta* are non-invasive tumours, *T1* are those where the cancer has grown into the cells lining the bladder and the group *T2* are cancers that have grown into the muscle layer. In the study there are 19 patients in the Ta group, 11 in the T1 group and 10 from the T2 group. The microarray measures the expression level of around 7000 genes. However, this is reduced to 1767 genes on requiring a sufficient quality of the measurement and a sufficient variation in the measurements across samples.

In the paper a molecular classifier is build to separate patient into the three groups. Before building the classifier 5 Ta's are removed that look different from the overall Ta group. The classifier uses 32 genes found through a cross validation study. The classifier is tested on a separate dataset with 68 samples. However, the latter dataset is measured on a different platform, which makes it more difficult to apply the classifier. Interestingly, those Ta's that are classified as T1 or T2 seem to have a significantly higher probability of tumour progression as compared to the correctly classified Ta's.

A classifier is also build to separate the Ta group into the two subgroups with and without recurrence of the cancer. This classifier uses 26 genes and a permutation analysis were performed to evaluate the classifier.

Exercise 1.1 (Observational studies)
High dimensional data can easily lead to non-reproducible results. As a warning read the article Young and Karr (2011) (Deming, data and observational studies). ■

1.1.1 Typical setup

We have two groups of observations. In group 1 there is n_1 observations, denoted $x_i, \ldots, x_{n_1}$, and in group 2 there is n_2 observations, denoted $y_1, \ldots, y_{n_2}$. Each obser-

vation is m-dimensional so that as an example $x_i = (x_{i1}, \ldots, x_{im})$. The means in the two groups are $\mu_1 = (\mu_{11}, \ldots, \mu_{1m})$ and $\mu_2 = (\mu_{21}, \ldots, \mu_{2m})$. We speak of $\Delta_j = \mu_{1j} - \mu_{2j}$ as the *differential expression* for variable j. The group averages and the common variance estimates are for $j = 1, \ldots, m$:

$$\hat{\mu}_{1j} = \bar{x}_j = \frac{1}{n_1}\sum_{i=1}^{n_1} x_{ij}, \qquad \hat{\mu}_{2j} = \bar{y}_j = \frac{1}{n_2}\sum_{i=1}^{n_2} y_{ij},$$
$$s_j^2 = \frac{1}{n_1+n_2-2}\Big\{\sum_{i=1}^{n_1}(x_{ij}-\hat{\mu}_{1j})^2 + \sum_{i=1}^{n_2}(y_{ij}-\hat{\mu}_{2j})^2\Big\}.$$

The t-statistic for no differential expression, $\mu_{1j} = \mu_{2j}$ or $\Delta_j = 0$, of the j'th variable is

$$t_j = \frac{\hat{\mu}_{1j} - \hat{\mu}_{2j}}{s_j\sqrt{1/n_1 + 1/n_2}}. \tag{1.1}$$

If we believe data are normally distributed we use a t-distribution to find the (two-sided) p-value:

$$p_j = P\big(|T| \geq |t_j|\big), \quad T \sim t(n_1+n_2-2) \tag{1.2}$$

If we do not trust the t-distribution we can compute a permutation p-value. Consider for a moment one variable only, say variable j. Under the hypothesis of no difference between group 1 and 2, we can consider group 1 as a random selection of n_1 elements from the combined group with $n = n_1 + n_2$ samples. For the permutation distribution we consider the values $(z_1, \ldots, z_n) = (x_{1j}, \ldots, x_{n_1 j}, y_{1j}, \ldots, y_{n_2 j})$ fixed, and make new groups by randomly selecting n_1 of the elements. Let S_n be the set of all permutations of $1, 2, \ldots, n$. We write a $b \in S_n$ as $(b(1), b(2), \ldots, b(n_1+n_2))$, and let $\{b(1), b(2), \ldots, b(n_1)\}$ be the new group 1. To simulate $b \in S_n$ we simply make randow draws from $\{1, 2, \ldots, n_1+n_2\}$ without replacement. For $b \in S_n$ let t_j^b be the t-statistic based on the two new groups. The permutation p-value is then

$$p_j = \frac{1}{|S_n|}\sum_{b\in S_n} 1\big(|t_j^b| \geq |t_j|\big). \tag{1.3}$$

When n is too large it becomes computationally infeasible to sum over all of S_n, and a random subset of size B is used. This gives the random permutation p-value

$$p_j = \frac{1}{B}\sum_{b=1}^{B} 1\big(|t_j^b| \geq |t_j|\big). \tag{1.4}$$

For the permutation p-value we have from Theorem 1.3 below that $P(P_j \leq \alpha) \leq \alpha$, but for the random permutation p-value this is only true in an asymptotic setting as $B \to \infty$.

In the multiple testing situation we calculate for each permutation the t-statistics for all variables $j = 1, \ldots, m$, that is we calculate $t_1^b, \ldots, t_m^b$. In this way we imitate the correlation structure of the original t-values $t_1, \ldots, t_m$.

Exercise 1.2 (Permuation p-value)
Consider the one dimensional case, $m = 1$, with sample sizes $n_1 = 4$ and $n_2 = 5$. Simulate 1000 datasets and for each of these calculate the p-value (1.2) from the t-distribution and the permutation p-value (1.3). Make a 2×2 table with the number of cases where the two p-values are below or above 0.05.

Consider the following settings: normally distributed data with the same mean value, $\mu_1 = \mu_2$; normally distributed data with $\mu_1 - \mu_2 = \sigma$, where σ is the standard deviation in the normal distribution; t-distributed data with the same mean, $x_i \sim t(1)$, $y_i \sim t(1)$; and t-distributed data with $x_i \sim 1 + t(1)$ and $y_i \sim t(1)$. ■

1.1.2 P-value

Consider a test statistic W for a hypothesis H where large values are critical. Assume, that the distribution of W is known under H, that is, the distribution does not depend on unknown parameters. Then the p-value is

$$pval(w) = P(W \geq w), \tag{1.5}$$

where w is the observed value, and $P(\cdot)$ refers to the distribution under H.

Theorem 1.3
Let P_W be the stochastic variable corresponding to (1.5), $P_W = pval(W)$. Then we have, generally, that $P(P_W \leq \alpha) \leq \alpha$, and if the distribution of W is continuous we have $P(P_W \leq \alpha) = \alpha$.

Proof. For a given α, if there exists w_α with $P(W \geq w_\alpha) = \alpha$ we have

$$P(P_W \leq \alpha) = P(W \geq w_\alpha) = \alpha.$$

If instead $P(W \geq w_\alpha) > \alpha$ and $P(W > w_\alpha) < \alpha$, then

$$P(P_W \leq \alpha) = P(W > w_\alpha) < \alpha.$$

When W has a continuous distribution the first display shows that the p-value is uniformly distributed. □

The uniformly distributed p-value is used in Young et al. (2009) (Cereal-induced gender selection? Most likely a multiple testing false positive) as an argument for rejecting the findings in a previous article. However, the authors of that previous article have a strong counter argument you can see by following a link at the end of the paper being linked to above.

Exercise 1.4 (Binomial p-value)
Let x be an observation form a binomial distribution with success probablity p and number of trials n. For testing the hypothesis $p = \frac{1}{2}$ we use the test statistic $w = |x - \frac{n}{2}|$.

Make a plot showing the distribution function of the p-value for this test. Include the identity line in the plot. ■

Exercise 1.5 (Power of t-test)
If $U \sim N(0,1)$ and independently $V \sim \chi^2(f)/f$ the distribution of $U/\sqrt{(V)}$ is by definition a t-distribution with f degress of freedom. When δ is a constant the distibution of $(U+\delta)/\sqrt{(V)}$ is a noncentral t-distribution with noncentrality parameter δ. In **R** the cumulative distribution function of the latter is calculated as `pt(x,f,ncp=`δ`)`.

Consider the t-test for equal means of two normally distributed samples with sample sizes n_1 and n_2. The test rejects when $|t| > t_0$ with $t_0 = t_{\text{inv}}(1-\alpha/2, n_1+n_2-2)$, where α is a prespecified level. Make a plot of the power as a function of the mean difference $(\mu_1-\mu_2)/\sigma$, standardized by the standard deviation σ for the individual obserations, for the cases $\alpha = 0.05, 0.05/10, 0.05/100, 0.05/1000$, and with your own choices of n_1 and n_2. *Help:* first figure out that the noncentrality parameter is $(\mu_1 - \mu_2)/(\sigma\sqrt{1/n_1+1/n_2})$. ■

1.2 Type I error rates for multiple testing

Consider testing the m null hypotheses $H_1, \ldots, H_m$, where we use the notation that $H_i = 0$ when H_i is true and $H_i = 1$ when H_i is false. Define the following quantities

$$M_0 = \{i : H_i = 0\}, \qquad m_0 = |M_0|,$$
$$M_1 = \{i : H_i = 1\}, \qquad m_1 = |M_1|.$$

Generally, for a subset $A \subseteq \{1, \ldots, m\}$, we let H_A be the hypothesis that $H_j = 0$ for $j \in A$, irrespectively of the status of H_j, $j \notin A$.

Notation 1.6 (Ordered p-values)
To each hypothesis we have a p-value p_i (the corresponding random variable being P_i). The ordered p-values are $p_{(1)} \le p_{(2)} \le \cdots \le p_{(m)}$. In order to refer to the corresponding hypotheses we let r_i be the hypothesis corresponding to the i'th order variable, $p_{(i)} = p_{r_i}$. The p-values for the true hypotheses, H_i, $i \in M_0$, are denoted $q_1, \ldots, q_{m_0}$, and those for the false hypotheses, H_i, $i \in M_1$, are denoted $q_1^*, \ldots, q_{m_1}^*$. ■

We describe the outcome of the multiple testing problem as in Table 1.1. In the table m is known, R and N_0 are observed, and the rest are unobservable. The number of *false positives* is N_{01} and the number of *false negatives* is N_{10}. False positives are type I errors and false negatives are type II errors.

When testing a single hypothesis we typically perform a test at level α, meaning that the probability of an error of type I is less than or equal to α. We also formulate this by saying that the error of type I is *controlled* at α. Among tests at level α we aim at maximizing power or, equivalently, minimizing the probability of an error of type II.

For the multiple testing problem there is a multitude of type I error rates. The ones we will cover in these notes are listed in Table 1.2.

	Names		Numbers		
	Not rejected	Rejected	NR	R	Sum
Null cases(M_0)	True negative	False positive	N_{00}	N_{01}	m_0
Nonnull cases (M_1)	False negative	True positive	N_{10}	N_{11}	m_1
			N_0	R	m

Table 1.1: Outcome of a multiple testing problem. The headings are NR for *Not rejected* and R for *Rejected*.

Family-wise error rate	FWER = $P(N_{01} > 0)$
False discovery rate	FDR = $E(\frac{N_{01}}{R} 1(R > 0))$
k-Familywise error rate	kFWER = $P(N_{01} \geq k)$

Table 1.2: Type I error rates for a multiple testing problem

We clearly have the inequalities

$$\frac{N_{01}}{R} 1(R > 0) \leq 1(N_{01} > 0) \quad \text{and} \quad 1(N_{01} > k) \leq 1(N_{01} > 0).$$

Taking expectation we get

$$\text{FDR} \leq \text{FWER} \quad \text{and} \quad \text{kFWER} \leq \text{FWER}$$

A multiple testing procedure is aimed at controlling a type I error rate at a chosen level. However, one has to specify which distribution is used to calculate the error rate. For the classical situation of a single test we simply use the distribution under the null hypothesis. For the multiple testing case one speaks of *weak* and *strong* control.

Notation 1.7 (Weak and strong control)
A weak control considers probabilities under the complete null hypothesis H^c where H_i is true ($H_i = 0$) for all i. A strong control considers probabilities under the hypothesis H_{M_0}, where $H_i = 0$ for $i \in M_0$, and the control has to hold for every possible choice of M_0. ■

The classical error control in a multiple testing situation is the socalled Bonferroni correction (see https://en.wikipedia.org/wiki/Bonferroni_inequalities for the origin of the name).

Proposition 1.8 (Bonferroni correction)
In a multiple testing problem the Bonferroni correction consists of rejecting hypothesis H_i if $p_i \leq \alpha/m$. This procedure provides strong control of the FWER at level α.

Proof. Generalizing $P(A \cup B) \leq P(A) + P(B)$ to a finite collection of sets (Boole's inequality) we find

$$\begin{aligned} \text{FWER} = P\big(N_{01} > 0\big) = P\Big(\exists i \in M_0 : P_i \leq \frac{\alpha}{m}\Big) = P\Big(\cup_{i \in M_0} \big\{P_i \leq \frac{\alpha}{m}\big\}\Big) \\ \leq \sum_{i \in M_0} P\Big(P_i \leq \frac{\alpha}{m}\Big) \leq \sum_{i \in M_0} \frac{\alpha}{m} = m_0 \frac{\alpha}{m} \leq \alpha. \end{aligned}$$

Here we have used that under the hypothesis H_{M_0}, where $M_0 = \{j : H_j = 0\}$, the p-values for $j \in M_0$ are sub-uniformly distributed, $P(P_i \leq p) \leq p$. □

If one rejects that all means are equal in a one way anova analysis, one often makes pairwise comparisons of means to see which means are different. This is then a multiple testing situation, and two methods for doing this is known under the names *Scheffe* and *Tukey*.

Exercise 1.9 (Bonferroni correction for independent tests)
Explain, that when the Bonferroni procedure is used we have under the complete null hypothesis H^c that FWER $= P(P_{(1)} \leq \alpha/m)$, where $P_{(1)}$ is the smallest p-value among the m p-values. Show, that when the p-values are independent we have FWER $\leq 1 - (1 - \alpha/m)^m$ and

$$1 - \exp\{-\alpha\} \leq 1 - (1 - \alpha/m)^m \leq 1 - \exp\{-\alpha(1 + \alpha/m)\},$$

for $m \geq 2$. Make a drawing of $1 - (1 - \alpha/m)^m$ and the two bounding curves as function of α with your own choice of m. ■

Exercise 1.10 (Setup for dependent variables)
We introduce here a setting that will be used many times in these notes. We consider m-dimensional normally distributed observations from two groups with common $m \times m$ variance matrix Σ and m-dimensional means μ_1 and μ_2. The sample sizes are n_1 and n_2. We are interested in simulating the t-statistics for testing equal means for each of the m variables. To this end we need to simulate values of the averages $\bar{x}_1$ and $\bar{x}_2$ from the two groups, and independently of these the empirical variances $(s_1^2, \ldots, s_m^2)$ of the m coordinates. The independence comes from the wellknown independence of $\bar{x}$ and $\hat{\Sigma}$ for a sample from a multivariate normal distribution.

Consider the special case with Σ having block structure, with each diagonal block being the 10×10 matrix Σ_0 and all off diagonal blocks being zero. Also, let $\Delta = \mu_1 - \mu_2$ have block structure, where the first five blocks are $(\psi, \psi, 0, \ldots, 0)$ and the remaining blocks are all zero. To simulate $\bar{x} \sim N_m(0, \Sigma/n)$ we can simulate $m/10$ observations from $N_{10}(0, \Sigma_0/n)$ and stack these on top of one another. To simulate $s_1^2, \ldots, s_m^2$ we can simulate $m/10$ observation from a Wishart$(\Sigma_0, n_1+n_2-2)/(n_1+n_2-2)$ distribution and extract all the diagonal elements of these. In R the multivariate normal distribution can be simulated with `mvrnorm` from the `MASS` package, and the Wishart distribution with `rWishart`.

Consider the case with Σ_0 being 1 in the diagonal and ρ in all the off diagonal elements, and with the mean difference Δ being given through $\psi = 1$. Calculate the t-statistics and, using the Bonferroni correction, reject the hypotheses of equal means when $|t| > t_0(\alpha, m)$, where t_0 is such that the probability of rejecting is α/m when the two means are equal. Count the number of rejected variables with no difference in the mean (the number of false positives) and the number of rejected variables with difference $\psi = 1$ in the means (number of true positives).

Repeat the simulations many times in order to estimate the distribution of the number of true positives and the probability FWER of having at least one false positive.

Make the simulation for a range of values of ρ and plot the result as a function of ρ. Do this for $m = 50, 100, 1000$ and $\alpha = 0.05, 0.1$. ■

1.2.1 Adjusted p-values

First a warning: adjusted p-values to be described here are not p-values. They give a way of describing the result of a particular multiple testing procedure that makes it easy to compare different procedures.

The traditional p-value from a test of a single hypothesis (univariate test) is the smallest level of the test at which the hypothesis is rejected. More formally

$$p\text{-value} = \inf\{\alpha : H \text{ is rejected at level } \alpha\}. \tag{1.6}$$

For the univariate test the p-value is uniformly distributed (when the underlying distribution is continuous). The adjusted p-values are defined in a way similar to (1.6), but they do not have the property of being uniformly distributed. This is the reason that I started this subsection with a warning! When needed, the p-values from a univariate test will be called *raw* p-values.

Definition 1.11 (Adjusted p-values)
Consider a particular type I error rate, say multERR*, for a multiple testing situation with hypotheses $H_1, \ldots, H_m$, and a procedure for controlling this error rate. We say that the test is at nominal level α when the procedure controls the error rate at level α. The adjusted p-value related to multERR for hypothesis H_j, $j = 1, \ldots, m$, is*

$$\tilde{p}_j = \inf\{\alpha : H_j \textit{ is rejected at the multERR nominal level } \alpha\}. \tag{1.7}$$

The above definition is given in Wright (1992) (Adjusted p-values for simultaneous inference).

Example 1.12 (Bonferroni adjusted FWER p-values)
Using the Bonferroni correction, hypothesis H_j is rejected when $p_j \le \alpha/m$, and according to Proposition 1.8 this controls the FWER at level α. Thus,

$$\tilde{p}_j = \inf\{\alpha : p_j \le \alpha/m\} = mp_j.$$

Although slightly in conflict with the definition we will in most situations use $\min\{mp_j, 1\}$ instead of mp_j. ■

The adjusted p-values depend on the type I error rate considered, as well as the multiple testing procedure used. In the Bonferroni example above the adjusted value $\tilde{p}_j$ depends on p_j only, but in some of the procedures to be discussed below it depends also on the rank of p_j among all the raw p-values.

1.3 Single-step procedures

Multiple testing procedures that control a particular type I error rate can be divided into two categories. *Single-step* procedures use the same form of adjustment for all hypotheses, not using the ordering of the raw p-values. *Stepwise* procedures, on the other hand, use the ordering of the raw p-values with *stepdown* starting from the smallest p-value and *stepup* starting from the largest p-value. The origin of these names are difficult to trace, but they appear in the book Hochberg and Tamhane (1987).

The Bonferroni procedure from Proposition 1.8 is a single-step procedure.

1.3.1 Šidák's procedure

In a single step procedure we reject a hypothesis H_j when $p_j \le \alpha(m)$, where the level $\alpha(m)$ depends on m and the procedure used. It is clear in this setting that

$$\text{FWER} = P(N_{01} > 0) = P(P^0_{(1)} \le \alpha(m)),$$

where $P^0_{(1)}$ is the smallest of the p-values among the true hypothesis, $P^0_{(1)} = \min_{j\in M_0}\{P_j\}$. If the hypotheses (in M_0) are independent we have

$$\begin{aligned}\text{FWER} &= 1 - P(P_j > \alpha(m), j \in M_0) = 1 - \prod_{j\in M_0}\big(1 - P(P_j \le \alpha(m))\big)\\ &\le 1 - \prod_{j\in M_0}\big(1-\alpha(m)\big) = 1 - \big(1-\alpha(m)\big)^{m_0}\\ &\le 1 - \big(1-\alpha(m)\big)^{m}.\end{aligned}$$

If therefore we take $1 - \big(1-\alpha(m)\big)^m = \alpha$, or $\alpha(m) = 1-(1-\alpha)^{1/m}$ we have strong control at level α. The difference between $\alpha(m) = 1-(1-\alpha)^{1/m}$ and α/m is usually small:

	$\alpha = 0.05$			$\alpha = 0.01$		
m	2	10	100	2	10	100
$1-(1-\alpha)^{1/m}$	0.0253	0.00512	0.000513	0.00501	0.00100	0.000100
α/m	0.0250	0.00500	0.000500	0.00500	0.00100	0.000100

We always have $\alpha(m) \ge \alpha/m$ and the procedure, known as Šidák's procedure, using the limit $\alpha(m) = 1-(1-\alpha)^{1/m}$ therefore rejects more hypotheses than the Bonferroni procedure.

Proposition 1.13
If the p-values for the true hypotheses are independent, Šidák's procedure provides strong control of FWER.

If the p-values for the true hypotheses come from the multivariate normal with an arbitrary covariance matrix ($p_j = P(|U_j| \geq |u_j|)$, where $U \sim N_m(\mu, \Sigma)$, $\mu_j = 0, j \in M_0$), or from the multivariate t-distribution as in (1.1) with underlying correlations of the form $Cov(X_{ir}, X_{is}) = \rho_r \rho_s$, then the Šidák's procedure provides strong control of FWER.

Proof. The stong control under the assumption of independence has been proved above.

For the case of multivariate normal test statistics, $U \sim N_m(\mu, \Sigma)$ with $\mu_j = 0$ for $j \in M_0$, it is shown in Sidak (1967) (Rectangular confidence regions for the means of multivariate normal distributions) that

$$P\big(|U_j| \leq c_j, j \in M_0\big) \geq \prod_{j \in M_0} P\big(|U_j| \leq c_j\big). \tag{1.8}$$

This gives

$$\begin{aligned} P(N_{01} = 0) = P\big(|U_j| \leq c_j(\alpha, m), j \in M_0\big) &\geq \prod_{j \in M_0} P\big(|U_j| \leq c_j(\alpha, m)\big) \\ &= \prod_{j \in M_0} \Big\{1 - \big[1 - \big(1 - \alpha\big)^{1/m}\big)\big]\Big\} = (1 - \alpha)^{m_0/m}, \end{aligned}$$

where $c_j(\alpha, m)$ is determined so that $P\big(|U_j| \leq c_j(\alpha, m)\big) = 1 - \big[1 - \big(1 - \alpha\big)^{1/m}\big)\big]$.

Sidak (1971) considers the multivariate t-distribution, where $t_j = u_j / s_j$ with u and s^2 independent, $u \sim N_m(0, \Sigma)$ and $(s_1^2, \ldots, s_m^2)$ being the diagonal of a Wishart$(\Sigma, n)/n$ distributed matrix. For a correlation structure as stated in the proposition, it is shown that (1.8) holds. A slight extension of the result in Sidak (1971) is given in Jogdeo (1977). □

The multivariate t-distribution in Proposition 1.13 is the one that corresponds to the t-statistics we generate from data. It is, however, not the one you find from wikipedias page on the Multivariate t-distribution.

Exercise 1.14 (Šidák's inequality)
Let $(X, Y) \sim N_2(0, \Sigma_0)$, where Σ_0 is 1 at the diagonal and ρ off the diagonal. Show that

$$P(|X| \leq c_1,\ |Y| \leq c_2) \geq P(|X| \leq c_1)\, P(|Y| \leq c_2).$$

■

1.3.2 Singlestep minP procedure

In a single step procedure, where we reject a hypothesis H_j when $p_j \leq \alpha(m)$, we have FWER $= P(P_{(1)}^0 \leq \alpha(m))$, and the best we can do is to use the distribution of

$P^0_{(1)}$ to determine $\alpha(m)$. Here $P^0_{(1)}$ is the minimum over the p-values from the true hypotheses. The above approach is not possible though, as this involves the unknown set M_0.

For weak control, where $M_0 = M$, we can let F_1 be the distribution of the minimum of all the p-values,

$$F_1(z) = P(P_{(1)} \leq z | H^c),$$

and reject a hypothese H_j when $F_1(p_j) \leq \alpha$. Then under H^c

$$\text{FWER} = P\big(F_1(P_{(1)}) \leq \alpha\big) \leq \alpha.$$

Westfall and Young (1993) show that the procedure which rejects when $F_1(p_j) \leq \alpha$, known as the *minP* procedure, provides strong control under an additional assumption.

Proposition 1.15
The p-values satisfy subset pivotality *if the distribution of $\{P_i, i \in M_0\}$ is the same under the hypothesis H_{M_0} as under the complete hypothesis H^c. Thus, the distribution of $\{P_i, i \in M_0\}$ is the same irrespective of the truth or falsity of the hypotheses in $M_1 = M \setminus M_0$.*

Under the assumption of subset pivotality the minP procedure provides strong control at level α

Proof. The result follows from

$$\begin{aligned}
\text{FWER}(M_0) &= P\Big(F_1\big(P^0_{(1)}\big) \leq \alpha | H_{M_0}\Big) \\
&= P\Big(F_1\big(\min_{j \in M_0}\{P_j\}\big) \leq \alpha | H^c\Big) \quad \text{(by assumption)} \\
&\leq P\Big(F_1\big(\min_{j \in M}\{P_i\}\big) \leq \alpha | H^c\Big) = P\Big(F_1\big(P_{(1)}\big) \leq \alpha\Big) \leq \alpha. \qquad \square
\end{aligned}$$

The *maxT* procedure is based on the test statistics directly instead of the p-values. Define

$$G_1(z) = P\Big(\max_{j \in M}\{|T_j|\} \geq z | H^c\Big), \tag{1.9}$$

which is the probability of the maximum of all the test statistics being greater than or equal to z. The maxT procedure rejects all hypotheses H_j with $G_1(|t_j|) \leq \alpha$. If all the marginal distributions of $|T_1|, \ldots, |T_m|$ are identical under H^c the maxT and minP procedures are identical.

Remark 1.16
For the two sample setup of Section 1.1.1 the p-value p_j depends on the data only through x_{ij} and y_{ij}. The joint distribution of p_j, $j \in M_0$, therefore depends on the joint distribution of x_{i,M_0} and y_{i,M_0} only, and the latter is not dependent on H_j, $j \notin M_0$. This means that we have subset pivotality in this case. ■

1.4 Stepwise procedures

Imagine that we have only two hypotheses $m = 2$, and let us start off using the Bonferroni bound $\alpha/2$. If both hypotheses are true we have control of FWER from $P(\min\{P_1, P_2\} \leq \alpha/2) \leq \alpha$. If only one hypothesis is true it suffices with the limit α instead of $\alpha/2$ to get control of FWER. We can achieve this partly by using a stepwise procedure where we first consider if $P_{(1)} \leq \alpha/2$. If this is the case we reject the hypothesis and consider if $P_{(2)} \leq \alpha$. If $P_{(1)} > \alpha/2$ we do not reject any hypotheses. If $|M_0| = 2$ we have

$$\text{FWER} = P(P_{(1)} \leq \alpha/2) \leq \alpha,$$

and if $|M_0| = 1$, say $M_0 = \{1\}$, we get

$$\begin{aligned}\text{FWER} &= P(P_1 \leq P_2, P_1 \leq \alpha/2) + P(P_1 > P_2, P_2 \leq \alpha/2, P_1 \leq \alpha) \\ &\leq P(P_1 \leq P_2, P_1 \leq \alpha) + P(P_1 > P_2, P_1 \leq \alpha) = P(P_1 \leq \alpha) \leq \alpha.\end{aligned}$$

This way of thinking leads to the Holm's procedure given below. Before stating the procedure we introduce general noation for stepwise procedures.

Definition 1.17 (Stepdown procedure)
Let α_i, $i = 1, \ldots, m$ be a nondecreasing sequence of numbers. The stepdown procedure starts by rejecting H_{r_1} if $p_{(1)} \leq \alpha_1$. When H_{r_1} is rejected the next hypothesis H_{r_2} is rejected if $p_{(2)} \leq \alpha_2$. The procedure continues until the first hypothesis is accepted, and all the remaining hypotheses are then also accepted. Here are two equivalent formulations of the stepdown procedure. The first formulation is

$$\begin{aligned}&\textit{Let } I \textit{ be the first occurence of } p_{(i)} > \alpha_i\textit{:} \\ &\quad \textit{reject } H_{r_1}, \ldots, H_{r_{I-1}} \textit{ and accept } H_{r_I}, \ldots, H_{r_m}.\end{aligned}$$

and the second is

$$\begin{aligned}&\textit{Let } L \textit{ be the largest value such that } p_{(1)} \leq \alpha_1, \ldots, p_{(L)} \leq \alpha_L\textit{:} \\ &\quad \textit{reject } H_{r_1}, \ldots, H_{r_L} \textit{ and accept } H_{r_{L+1}}, \ldots, H_{r_m}.\end{aligned}$$

When the algorithm stops at the I'th comparison we have that all hypotheses with $p_i \leq \alpha_{I-1}$ are rejected (and there are $I-1$ of these), and all hypotheses with $p_i > \alpha_I$ are accepted.

Definition 1.18 (Stepup procedure)
Let α_i, $i = 1, \ldots, m$ be a nondecreasing sequence of numbers. The stepup procedure starts by accepting H_{r_m} if $p_{(m)} > \alpha_m$. When H_{r_m} is accepted the next hypothesis $H_{r_{m-1}}$ is accepted if $p_{(m-1)} > \alpha_{m-1}$. The procedure continues until the first hypothesis is rejected, and all the remaining hypotheses are then also rejected. Here are two equivalent formulations of the stepup procedure. The first formulation is

$$\begin{aligned}&\textit{Let } I \textit{ be the last occurence of } p_{(i)} \leq \alpha_i\textit{:} \\ &\quad \textit{reject } H_{r_1}, \ldots, H_{r_I} \textit{ and accept } H_{r_{I+1}}, \ldots, H_{r_m}.\end{aligned}$$

and the second is

Let L be the smallest value such that $p_{(L)} > \alpha_L, \dots, p_{(m)} > \alpha_m$:
reject $H_{r_1}, \dots, H_{r_{L-1}}$ and accept $H_{r_L}, \dots, H_{r_m}$.

When the algorithm stops at I with $p_{(I)} \le \alpha_I$ we have that all hypotheses with $p_i \le \alpha_I$ are rejected (and there are I of these), and all hypotheses with $p_i > \alpha_{I+1}$ are accepted.

Proposition 1.19
Let $\alpha_i = a_i \alpha$ for a known increasing sequence $a_1, \dots, a_m$. For a stepdown procedure the adjusted p-value is

$$\tilde{p}_{r_j} = \max_{k=1,\dots,j} \{\min\big(\frac{1}{a_k} p_{(k)}, 1\big)\},$$

and for a stepup procedure the adjusted p-value is

$$\tilde{p}_{r_j} = \min_{k=j,\dots,m} \{\min\big(\frac{1}{a_k} p_{(k)}, 1\big)\}.$$

Proof. Consider the stepdown procedure. If $\tilde{p}_{r_j} \le \alpha$ then $p_{(k)} \le a_k \alpha = \alpha_k$ for $k = 1, \dots, j$, which shows that $I > j$ and the stepdown procedure rejects at least $H_{r_1}, \dots, H_{r_j}$. In particular, H_{r_j} is rejected. If instead, $\tilde{p}_{r_j} > \alpha$ there exists $k \le j$ with $p_{(k)} > a_k \alpha = \alpha_k$, which gives that $H_{r_k}, \dots, H_{r_m}$ are not rejected by the stepdown procedure. In particular, H_{r_j} is not rejected. Thus H_{r_j} is rejected if and only if $\tilde{p}_{r_j} \le \alpha$.

For the stepup procedure the argument is similar. □

1.4.1 Holm's stepdown procedure

Holm's procedure (Holm (1979)) is the stepdown procedure with constants

$$\alpha_j = \frac{\alpha}{m - j + 1}.$$

It is clear that Holm's procedure rejects more hypotheses that the Bonferroni procedure since $\alpha/(m-j+1) > \alpha/m$ for $j = 2, \dots, m$.

Proposition 1.20
Holm's procedure provide strong control of FWER at level α.

Proof. We split the event $N_{01} \ge 1$ according to the value of v in $q_{(1)} = p_{(v)}$ (see Notation 1.6). Using that $\alpha_v = \alpha/(m - v + 1) \le \alpha/m_0$ we find

$$\begin{aligned}
P(N_{01} \ge 1) &= \sum_{v=1}^{m_1+1} P(P_{(1)} \le \alpha_1, \dots, P_{(v-1)} \le \alpha_{v-1}, P_{(v)} = Q_{(1)} \le \alpha_v) \\
&\le \sum_{v=1}^{m_1+1} P(P_{(1)} \le \alpha_1, \dots, P_{(v-1)} \le \alpha_{v-1}, P_{(v)} = Q_{(1)}, Q_{(1)} \le \alpha/m_0) \\
&\le \sum_{v=1}^{m_1+1} P(P_{(v)} = Q_{(1)}, Q_{(1)} \le \alpha/m_0) \\
&= P(Q_{(1)} \le \alpha/m_0) \le \alpha,
\end{aligned}$$

where the last inequality follows from the proof of Proposition 1.8. □

Exercise 1.21 (Holm's procedure for independent tests)
Consider the two sample setup in section 1.1.1 with independent variables. Simulate independent t-statistics, $t_1,\ldots,t_m$, where the scaled mean difference $(\mu_{1j}-\mu_{2j})/\sigma_j$ is 1 for k variables and zero for the remaining $m-k$ variables. Use Holm's stepdown procedure to reject hypotheses (if you calculate adjusted p-values the function `cummax` in R is convenient). Count the number of rejected variables with no differential expression (the number of false positives) and the number of rejected variables with nonzero differential expression (number of true positives).

Repeat the simulations many times in order to estimate the distribution of the number of true positives and the probability FWER of having at least one false positive.

Make your own choices of n_1, n_2, m and k. ■

Exercise 1.22 (Holm's procedure for dependent tests)
Simulate dependent t-statistics as in Exercise 1.10 and use Holm's stepdown procedure to reject hypotheses. Find the mean of the number of true positives and the probability FWER that there is at least one false positive. Make a plot similar to the plot in Exercise 1.10.

Make a plot that illustrate the gain from using Holm's stepdown procedure as compared to using Bonferroni correction. ■

1.4.2 Stepdown minP procedure

The stepdown procedure of Holm builds on the Bonferroni procedure. Can one make a stepdown procedure that build on the minP procedure instead? The answer is yes. When $F_1(p_{(1)}) \le \alpha$, so that the hypothesis H_{r_1} is rejected, we must consider the distribution of the minimum of the p-values for the remaining hypotheses. Consider a fixed value r_1, let $r(1)$ be the set $\{r_1\}$, and define

$$F_2(z;r(1)) = P\Big(\min_{j\in M\setminus r(1)}\{P_j\} \le z | H^c\Big).$$

We then reject H_{r_2} if $F_2(p_{(2)};r(1)) \le \alpha$. We continue this way untill we first encounter a hypothesis not being rejected. For fixed values $r_1,\ldots,r_{d-1}$, and with $r(d-1) = \{r_1,\ldots,r_{d-1}\}$, we use

$$F_d\big(z;r(d-1)\big) = P\Big(\min_{j\in M\setminus r(d-1)}\{P_j\} \le z | H^c\Big).$$

The stepdown minP procedure of Westfall and Young (1993) rejects $H_{r_1},\ldots,H_{r_L}$ where L is the largest index with

$$F_1\big(p_{(1)}\big) \le \alpha, F_2\big(p_{(2)};r(1)\big) \le \alpha,\ldots,F_L\big(p_{(L)};r(L-1)\big) \le \alpha,$$

Proposition 1.23
The stepdown minP procedure provides strong control of FWER at level α under the assumption of subset pivotality.

Proof. As before we let $P^0_{(1)}$ be the smallest p-value among the true hypotheses. Also we let $F^0(z) = P(P^0_{(1)} \le z|H^c)$. We split the event $N_{01} \ge 1$ according to the value of v in $P_{(v)} = P^0_{(1)}$. For a given value of v there are $v-1$ hypotheses in M_1 with p-values smaller than $P^0_{(1)}$. For notational reasons let B_j be the event $F_j(P_{(j)}; r(j-1)) \le \alpha$. For a given value of v the event $N_{01} \ge 1$ is

$$B_1, \ldots, B_{v-1}, F_v\big(P_{(v)}; r(v-1)\big) \le \alpha, \{r_1, \ldots, r_{v-1}\} \subseteq M_1, r_v \in M_0.$$

Also,

$$F_v\big(z; r(v-1)\big) = P\Big(\min_{j \in M_0 \cup M_1 \setminus r(v-1)} \{P_j\} \le z|H^c\Big) \ge P\Big(\min_{j \in M_0} \{P_j\} \le z|H^c\Big) = F^0(z),$$

so that $F_v\big(P_{(v)}; r(v-1)\big) \ge F^0\big(P^0_{(1)}\big)$. This last inequality gives us the first inequality below:

$$\begin{aligned} P(N_{01} \ge 1|H_{M_0}) &\le \sum_v P\Big(B_1, \ldots, B_{v-1}, F^0\big(P^0_{(1)}\big) \le \alpha, \{r_1, \ldots, r_{v-1}\} \in M_1, r_v \in M_0|H_{M_0}\Big) \\ &\le \sum_v P\Big(F^0\big(P^0_{(1)}\big) \le \alpha, \{r_1, \ldots, r_{v-1}\} \in M_1, r_v \in M_0|H_{M_0}\Big) \\ &= P\big(F^0\big(P^0_{(1)}\big) \le \alpha|H_{M_0}\big) = P\big(F^0\big(P^0_{(1)}\big) \le \alpha|H^c\big) \le \alpha, \end{aligned}$$

proving the strong control of FWER at level α. □

To describe the maxT stepdown procedure, order the test statistics from largest absolute value to smallest absolute value: $|t_{s_1}| \ge |t_{s_2}| \ge \cdots \ge |t_{s_m}|$. Note that, if the marginal distribution of the t-statistics are identical, then $(s_1, \ldots, s_m) = (r_1, \ldots, r_m)$, where $p_{r_1} \le p_{r_2} \le \cdots \le p_{r_m}$. Let now, with $s(j-1) = \{s_1, \ldots, s_{j-1}\}$,

$$G_j\big(z; s(j-1)\big) = P\Big(\max_{j \in M \setminus s(j-1)} |T_j| \ge z\Big|H^c\Big),\ j = 1 \ldots, m.$$

The maxT stepdown procedure rejects $H_{s_1}, \ldots, H_{s_L}$ where L is the largest index with

$$G_1\big(|t_{s_1}|\big) \le \alpha, \ldots, G_L\big(|t_{s_L}|; s(L-1)\big) \le \alpha.$$

Using permutations to evaluate the distribution functions the maxT procedure is much easier to implement as compared to the minP procedure. The reason is that the minP procedure requires a double simulation, first to evaluate the p-values and next to evaluate the distribution of the minimum of a set of p-values. For the maxT procedure we only need to evaluate the distribution of the maximum of a set of test statistics.

The following simulation procedure for the maxT procedure is a slight reformulation of the algorithm in Ge et al. (2003) (Resampling-based multiple testing for microar- ray data analysis), which in turn is based on algorithms in Westfall and Young (1993). The algorithm provides strong control of FWER with respect to probabilities from the permutation distribution. See section 1.6.2 for more details.

Algorithm 1.24 (Permutation algorithm for stepdown maxT procedure)

1. *Order the observed test statistics so that* $|t_{s_1}| \geq |t_{s_2}| \geq \cdots \geq |t_{s_m}|$, *consider* $s_1 \ldots, s_m$ *fixed in steps 2 and 3 below and let* $s(j) = \{s_1, \ldots, s_j\}$.

2. *Make B permutations of the group labels. For each permuation* $b = 1, \ldots, B$ *compute the test statistics* $t_{b,1}, \ldots, t_{b,m}$ *for the m hypotheses. Compute cumulative maxima*
$$u_{b,j} = \max_{w \in M \setminus s(j-1)} |t_{b,w}|, \quad j = m, m-1, \ldots, 1.$$

3. *Calculate tail probabilities*
$$\hat{G}_j\big(|t_{s_j}|; s(j-1)\big) = \frac{1}{B} \sum_{b=1}^{B} 1\big(u_{b,j} \geq |t_{s_j}|\big), \quad j = 1, \ldots, m,$$
and reject $H_{s_1}, \ldots, H_{s_L}$ *where L is the largest index with*
$$\hat{G}_1\big(|t_{s_1}|\big) \leq \alpha, \ldots, \hat{G}_L\big(|t_{s_L}|; s(L-1)\big) \leq \alpha.$$

In terms of adjusted p-values step 3 can be formulated as rejecting those hypotheses H_{s_j} with $\tilde{p}_{s_j} \leq \alpha$, where $\tilde{p}_{s_1} = \hat{G}_1\big(|t_{s_1}|\big)$ and

$$\tilde{p}_{s_j} = \max\big\{\tilde{p}_{s_{j-1}}, \hat{G}_j\big(|t_{s_j}|; s(j-1)\big)\big\}, \quad j = 2, \ldots, m.$$

Note that we do not need to store $(u_{b,1}, \ldots, u_{b,m})$ for $b = 1, \ldots, B$ since the sums in point 3 of the algorithm can be calculated recursively.

Exercise 1.25
Make an R programme that implements Algorithm 1.24. In R a random subset of size n_1 from $1, \ldots, n$ can be obtained by `sample(n,`n_1`)`.

Example 1.26
To illustrate some of the issues involved in multiple testing let us consider a simple situation with independent p-values. Also, let us say that all p-values originate from testing $\mu = 0$ in the model $X_i \sim N(\mu, \sigma^2)$, $i = 1 \ldots, n$, using a t-test. Under the null hypothesis the p-value is of course uniformly distributed, and under the alternative the distribution of the p-value is found through a noncentral t-distribution with noncentrality parameter $\sqrt{n}\mu$. When the test rejects for $|t| \geq t_0$, the probability of rejecting under the alternative is, with **R**-notation,

$$\beta(t_0) = \texttt{1-pt(t0,n-1,ncp=np)+pt(-t0,n-1,ncp=np)}, \tag{1.10}$$

where np is the value of the noncentrality parameter.

Consider the situation of m hypotheses and the minP procedure. To control the FWER at level α we reject the smallest p-value when this is below $\alpha_1 = 1-(1-\alpha)^{1/m}$, and the second smallest when this is below $\alpha_2 = 1-(1-\alpha)^{1/(m-1)}$. Imagine now that hypothesis H_1 is false with $\mu = 1$ and this is the only false hypothesis, $m_1 = 1$.

Let us consider the probability of rejecting hypothesis H_1 and accepting all other hypotheses (P(only H_1)). This is the event that $|t_1| \geq t_0(1)$, $t_0(1) = t_{\text{inv}}(1-\alpha_1/2, n-1)$, and that $\min\{P_2, \ldots, P_m\} > \alpha_2$. The probability is therefore given by $\beta(t_0(1))$ from (1.10), times $(1-\alpha_2)^{m-1}$. Examples of these probabilitites, for the case $\alpha = 0.05$ and $n = 10$, can be seen in Table 1.3 in the first row.

m	1	2	10	100	1000
$P(\text{TP} = 1, \text{FP} = 0)$, $m_1 = 1$	0.803	0.650	0.362	0.100	0.019
$P(\text{TP} = 1, \text{FP} = 0 \mid R = 1)$, $m_1 = 1$	1.000	0.992	0.932	0.698	0.281
$P(\text{TP} = 2, \text{FP} = 0)$, $m_1 = 2$		0.631	0.151	0.011	0.000
$P(\text{TP} \geq 1, \text{FP} = 0)$, $m_1 = 2$		0.900	0.589	0.189	0.037

Table 1.3: Various probabilities for detecting one or two false hypotheses. All p-values are independent. See text for detailed explanation. In all cases the FWER is controlled at level $\alpha = 0.05$ and $n = 10$.

Above we calculated the probability of rejecting H_1 and accepting all other hypotheses. Consider instead the probability of rejecting H_2 and accepting all other hypotheses. This probability is the product of three terms. The first term is the probability that $P_2 \leq \alpha_1$, the second is the probability that none of the remaining true hypotheses are rejected, $\min\{P_3, \ldots, P_m\} > \alpha_2$ and the third is the probability that H_1 is not rejected, $|t_1| < t_0(2)$ with $t_0(2) = t_{\text{inv}}(1-\alpha_2/2, n-1)$. We can therefore write the probability as

$$P(\text{only } H_2) = \alpha_1(1-\alpha_2)^{m-2}\big(1-\beta(t_0(2))\big).$$

We can now calculate the conditional probabillity of rejecting H_1 given that we reject exactly one hypothesis as

$$\frac{P(\text{only } H_1)}{P(\text{only } H_1) + (m-1)P(\text{only } H_2)}.$$

These conditional probabilities are included in Table 1.3.

Let us now consider the situation with two false hypotheses, $m_1 = 2$, both of which have $\mu = 1$. The probability of finding both of these and accepting all other hypotheses is

$$\big(\beta(t_0(2))^2 - \big(\beta(t_0(2)) - \beta(t_0(1))\big)^2\big)(1-\alpha_3)^{m-2}.$$

The probability of finding at least one of the two false hypotheses and accepting all the true hypotheses is

$$2\beta(t_0(1))\big(1-\beta(t_0(2))\big)(1-\alpha_2)^{m-2} + \big(\beta(t_0(2))^2 - \big(\beta(t_0(2)) - \beta(t_0(1))\big)^2\big)(1-\alpha_3)^{m-2}.$$

These probabilities are in the two last rows of Table 1.3. ∎

Exercise 1.27 (Breast cancer data)
Sotiriou et al. (2003) (Breast cancer classification and prognosis based on gene expression profiles from a population-based study) report on a study of breast cancer where 99 women are divided into two groups. The first group with 65 women consists of those women where the cancer has receptors for estrogen (ER+), and the second group with 34 women are those where the cancer is without receptors (ER-). The data are gene expressions from cDNA microarrays with 7650 probes (roughly corresponding to looking at 7650 different genes). Data are $\log_2$ values, and can be found in the file "2912Table3.csv". The first six columns are information on the gene considered, and the remaining 99 columns are the values for the patients in the study. The group labels for the 99 women can be found in column 14 of the file "2912Table2.csv", with the ER+ group coded as 1 and the ER- as 0. Use `read.csv(,header=T)` to read the files.

There are missing values in the dataset (in R you can use `is.na` to help you locate these). In the analysis you can exclude all variables with missing values. There are 4404 variables with no missing values.

Find for the breast cancer data the rejected hypotheses (variables declared differential expressed) on using Bonferroni correction and the theoretical t-distribution to calculate p-values.

Find the rejected hypotheses on using Holm's stepdown procedure and the theoretical t-distribution to calculate p-values.

Use Algorithm 1.24 to find the rejected hypotheses.

Illustrate the result through adjusted p-values and suitable plots. ■

Exercise 1.28 (Colorectal cancer data)
Kruhøffer et al. (2005) (Gene expression signatures for colorectal cancer microsatellite status and hnpcc) study the gene expression for 101 stage II and III colorectal cancers. Of these 34 are microsatellite instable (MSI) and 67 are microsatellite stable (MSS). Log transformed values are in the file "102-5082-logdata.txt" with $5082 \cdot 102$ values (data appears in the order row 1, row 2 and so on, where rows correspond to genes and columns to sample). The microsatellite status can be found in the file "102-5082-samples-tal.txt", where numbers below 50 are MSI and the rest are MSS. Sample 83 out of the 102 in the files should be deleted (error in data).

Find for the MSI/MSS data the rejected hypotheses (variables declared differential expressed) on using Bonferroni correction and the theoretical t-distribution to calculate p-values.

Find the rejected hypotheses on using Holm's stepdown procedure and the theoretical t-distribution to calculate p-values.

Use Algorithm 1.24 to find the rejected hypotheses.

Illustrate the result through adjusted p-values and suitable plots. ■

$m = 1000$ hypotheses of which $m_1 = 100$ are false

Significance point t_0	3	4	5	6	7	8
Expected FP	13	2.8	0.67	0.18	0.06	0.02
Expected TP	58	30	13	5.5	2.3	1.00
E(FP)/(E(FP)+E(TP))	0.188	0.085	0.048	0.032	0.024	0.020
FDR	0.187	0.085	0.047	0.031	0.022	0.012

Table 1.4: FDR for a setting with independent p-values. See text for details.

1.5 False discovery rate

In this section we study the Benjamini-Hochberg procedure for controlling the false discovery rate FDR as described in Benjamini and Yekutieli (2001) (The control of the false discovery rate in multiple testing under dependency). For cases with very many variables, and only a few of these have nonzero differential expression, a control of the FWER can lead to no variables being detected. The FDR instead allow a fraction of the detected variables to be false detections. In this section we consider a control of a mean value, whereas in the next section we consider a control of a probability.

Example 1.29
Let us start with a simple example to illustrate the concept of a false discovery rate. Let the m p-values be independent and derived from testing $\mu = 0$ in the model $X_i \sim N(\mu, \sigma^2)$, $i = 1, \ldots, n$, using the ordinary t-test. Consider the single step procedure that rejects when $|t| \geq t_0$. Using **R**-commands the probability of rejecting a hypothesis is

p_0 = 2*pt(-t0,n-1) when hypothesis is true,
p_1 = 1-pt(t0,n-1,ncp=np)+pt(-t0,n-1,ncp=np) when hypothesis is false.

Table 1.4 considers the situation with $m_0 = 900$ true hypotheses and $m_1 = 100$ false, all with $\mu = 1$, and gives the expected number of FP and TP.

In this simple setting the random number of false positives and true positives are both binomial distributed, FP $\sim$ binom$(900, p_0)$ and TP $\sim$ binom$(100, p_1)$. From this we can calculate the false discovery rate FDR. The result is shown in Table 1.4.

Figure 1.1 shows the distribution of FP conditioned on the total number of positives for two instances of the single step limit t_0. As an example if we use the limit $t_0 = 3$, where FDR = 0.19, and imagine that the number of positives is 85 then the number of false positives will be around 20 and almost surely below 30. In this case FDR times 85 is roughly 16 and the reason that this number is at the left side of the distribution of FP is that the total number of positives, 85, is above the expected number of 71. ■

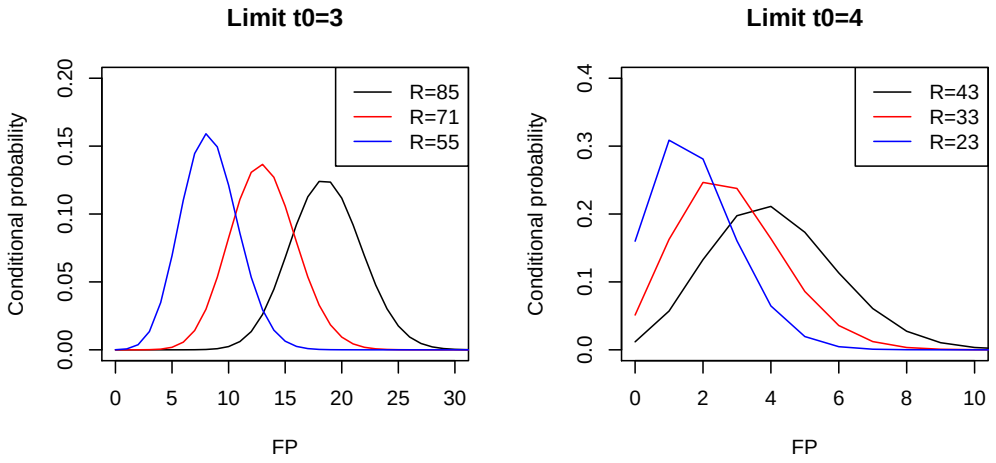

Figure 1.1: The distribution of FP when conditioned on the total number of positives. The two figures differ in the limit t_0 for rejecting a hypothesis. See text for the setting of the calculation.

The Benjamin-Hochberg procedure is a stepup procedure with the constants

$$\alpha_i = \frac{i\alpha}{m}.$$

Let R be the number of rejected hypotheses. Then R is determined as the largest index j with $p_{(j)} \le \alpha_j$, that is, $p_{(R)} \le \alpha_R$ and $p_{(l)} > \alpha_l$, $l = R+1, \ldots, m$.

The Benjamin-Yekutelli procedure is the stepup procedure with constants

$$\alpha_i = \frac{i\alpha}{m\sum_{i=1}^{m}\frac{1}{i}},$$

which is the Benjamin-Hochberg procedure with α replaced by $\alpha / \sum_{i=1}^{m}\frac{1}{i}$. Note, that the factor $\sum_{i=1}^{m}\frac{1}{i}$ gives a considerable decrease of the constants α_i as compared to the Benjamin-Hochberg procedure. If $m = 100$ we get $\sum_{i=1}^{m}\frac{1}{i} = 5.2$, $m = 1000$ gives 7.5 and $m = 10\,000$ gives 9.8.

When controlling FDR an important aspects is correlation among the m variables. In the original paper, Benjamini and Hochberg (1995), it is assumed that variables are independent. This is in many appplications (microarray experiments, say) not a realistic assumption. Benjamini and Yekutieli (2001) introduce *positive regression dependency on a subset* (PRDS) as follows.

Definition 1.30 (PRDS)
An m-dimensional set D is called increasing if $x \in D$ implies that $y \in D$ for all y with $y_i \ge x_i$, $i = 1, \ldots, m$. Try to make a drawing of an increasing set in $\mathbf{R}^2$.

An m-dimensional stochastic vector X is called PRDS on a subset $I_0 \subseteq \{1, \ldots, m\}$ if for each $i \in I_0$ and for any increasing set D the conditional probability $P(X \in D | X_i = x)$ is nondecreasing in x. When $I_0 = \{1, \ldots, m\}$ we use the notation PRD instead of PRDS.

Lemma 1.31

i) If X has independent coordinates then X is PRD.

ii) Let $Y \in \mathbf{R}$ be a random variable. If $P(X \in D|Y = y)$ is nondecreasing in y, so is $P(X \in D|Y \le y)$.

iii) Let $T = (T_1, \ldots, T_m)$ be test statistics and assume that T is PRDS on the subset I_0. Let the p-values $P = (P_1, \ldots, P_m)$ be the tail probabilitites of T. Then P is also PRDS on the subset I_0 (Benjamini and Yekutieli (2001) page 1171).

Proof. For the first result define

$$D_i(a) = \{(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m) : ((x_1, \ldots, x_{i-1}, a, x_{i+1}, \ldots, x_m) \in D\}.$$

When D is increasing and $b > a$ we have $D_i(a) \subseteq D_i(b)$ and so

$$\begin{aligned} P(X \in D|X_i = b) &= P((X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m) \in D_i(b)) \\ &\ge P((X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m) \in D_i(a)) = P(X \in D|X_i = a). \end{aligned}$$

ii) Write $h(y) = P(X \in D|Y = y)$ and $F(y) = P(Y \le y)$. Then with $y < z$,

$$\begin{aligned} P(X \in D|Y \le z) &= \frac{P(X \in D, Y \le z)}{P(Y \le z)} = \int_{-\infty}^{\infty} P(X \in D, Y \le z|Y = u)\frac{P_Y(du)}{F(z)} \\ &= \int_{-\infty}^{z} P(X \in D|Y = u)\frac{P_Y(du)}{F(z)} = \int_{-\infty}^{z} h(u)\frac{P_Y(du)}{F(z)} \\ &= \frac{F(y)}{F(z)}\int_{-\infty}^{y} h(u)\frac{P_Y(du)}{F(y)} + \frac{F(z)-F(y)}{F(z)}\int_{y}^{z} h(u)\frac{P_Y(du)}{F(z)-F(y)} \\ &= P(X \in D|Y \le y) + \frac{F(z)-F(y)}{F(z)}\left\{\int_{y}^{z} h(u)\frac{P_Y(du)}{F(z)-F(y)} - \int_{-\infty}^{y} h(u)\frac{P_Y(du)}{F(y)}\right\} \\ &\ge P(X \in D|Y \le y), \end{aligned}$$

where the last inequality uses that $h(u)$ is nondecreasing and the two terms within the curly parentheses are both mean values over $h(u)$.

iii) For the third property let $y > x$ and let D be an increasing set. We must prove that

$$P(P \in D|P_i = y) \ge P(P \in D|P_i = x).$$

To this end write $\bar{F}_i$ for the tail probability of T_i, $\bar{F}_i(z) = P(T_i > z)$, so that $P_i = \bar{F}_i(T_i)$. Let $\tilde{x} = \bar{F}_i^{-1}(x)$ and $\tilde{y} = \bar{F}_i^{-1}(y)$ with the property $\tilde{x} \ge \tilde{y}$ derived from $y > x$. Define

$$\tilde{D} = \{t : (\bar{F}_1(t_1), \ldots, \bar{F}_m(t_m)) \in D\}.$$

We must then prove

$$P(T \in \tilde{D}|T_i = \tilde{y}) \ge P(T \in \tilde{D}|T_i = \tilde{x}) \quad \text{or} \quad P(T \in \tilde{D}^c|T_i = \tilde{x}) \ge P(T \in \tilde{D}^c|T_i = \tilde{y}).$$

If $\tilde{D}^c$ is an increasing set the latter inequality follows from the PRDS assumption and $\tilde{x} \geq \tilde{y}$. To show that $\tilde{D}^c$ is an increasing set let $p^1 = (\bar{F}_1(t_1^1), \ldots, \bar{F}_m(t_m^1))$ for $t^1 \in \tilde{D}^c$ and observe that $p^1 \notin D$. Also $t^2 \geq t^1$ (coordinate wise) implies that $\bar{F}_j(t_j^2) \leq \bar{F}_j(t_1^1)$, $j = 1, \ldots, m$, so that $p^2 = (\bar{F}_1(t_1^2), \ldots, \bar{F}_m(t_m^2)) \leq p^1$. If $p^2 \in D$ then since D is an increasing set it follows that $p^1 \in D$, and since this is not the case we have $p^2 \notin D$ and $t^2 \in \tilde{D}^c$.□

Theorem 1.32 (Benjamini and Yekutieli)
If the joint distribution for the test statistics is PRDS on the set M_0 of true hypotheses, the Benjamini-Hochberg procedure controls the FDR at a level less than or equal to $\frac{m_0}{m}\alpha$.

The Benjamini-Yekutelli procedure always controls the FDR at level less than or equal to $\frac{m_0}{m}\tilde{\alpha}$.

Proof. The proof here is basically the same as in Benjamini and Yekutieli (2001), but I have tried to make some of the steps more transparent. Let us first note that when $R = j$ hypotheses are rejected we have $p_{(j)} \leq \alpha_j$ and $p_{(l)} > \alpha_l$, $l = j+1, \ldots, m$, which implies that we reject the hypotheses with $P_i \leq \alpha_j$ and accept those with $P_i > \alpha_{j+1}$. The false discovery rate FDR can therefore be written as

$$\begin{aligned} \text{FDR} &= E\Big\{\frac{N_{01}}{R}1(R>0)\Big\} = \sum_{j=1}^{m}\frac{1}{j}E\big\{N_{01}1(R=j)\big\} \\ &= \sum_{j=1}^{m}\frac{1}{j}E\Big\{\sum_{i\in M_0}1\big(P_i \leq \alpha_j\big)1(R=j)\Big\} \\ &= \sum_{i\in M_0}\sum_{j=1}^{m}\frac{1}{j}E\big\{1\big(P_i \leq \alpha_j\big)1(R=j)\big\}. \end{aligned} \tag{1.11}$$

I) When removing a value, P_i, from the set of p-values, we denote the remaining values as $p_w^{-i} = p_w$, $w \in M \setminus \{i\}$ and let $p_{(1)}^{-i}, \ldots, p_{(m-1)}^{-i}$ be the ordered values. The situation when $P_i \leq \alpha_j$ and $R = j$ is shown in figure 1.2.

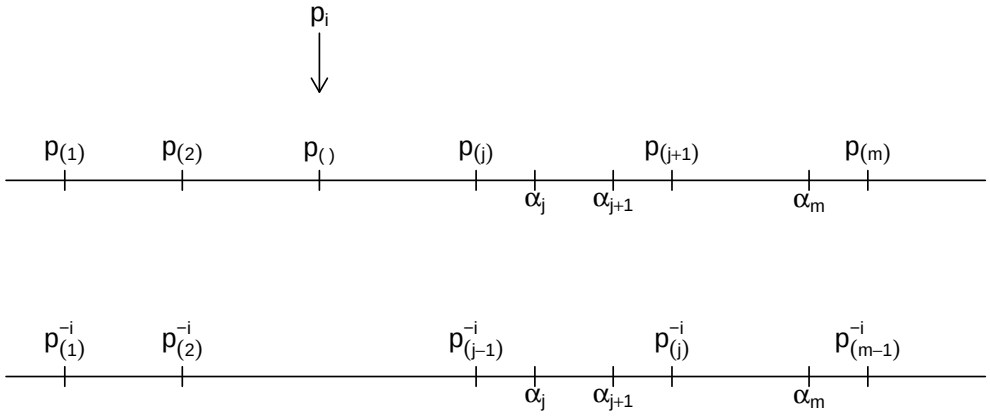

Figure 1.2: The event $R = j$ and $P_i \leq \alpha_j$.

Rewriting the event $R = j$ when $P_i \le \alpha_j$ we find

$$\begin{aligned} 1\big(P_i \le \alpha_j\big)\, 1(R = j) &= 1\big(P_i \le \alpha_j\big)\, 1\Big(P_{(j)} \le \alpha_j, P_{(l)} > \alpha_l, l = j+1, \ldots, m\Big) \\ &= 1\big(P_i \le \alpha_j\big)\, 1\Big(P^{-i}_{(j-1)} \le \alpha_j, P^{-i}_{(l)} > \alpha_{l+1}, l = j, \ldots, m-1\Big) \\ &= 1\big(P_i \le \alpha_j\big)\, 1\big(B^{-i}_{j-1}\big), \end{aligned}$$

where B^{-i}_{j-1} is simply a name for the event in the line above. Inserting this in (1.11), and using $P(P_i \le \alpha_j) \le j\alpha/m$ or $1/j \le (\alpha/m)/P(P_i \le \alpha_j)$, we get

$$\begin{aligned} \mathrm{FDR} &= \sum_{i \in M_0} \sum_{j=1}^{m} \frac{1}{j} E\big\{1\big(P_i \le \alpha_j\big) 1\big(B^{-i}_{j-1}\big)\big\} \qquad (1.12) \\ &\le \sum_{i \in M_0} \sum_{j=1}^{m} \frac{\alpha}{m} \frac{E\big\{1\big(P_i \le \alpha_j\big) 1\big(B^{-i}_{j-1}\big)\big\}}{P\big(P_i \le \alpha_j\big)} \\ &= \sum_{i \in M_0} \sum_{j=1}^{m} \frac{\alpha}{m} P\big(B^{-i}_{j-1} | P_i \le \alpha_j\big). \qquad (1.13) \end{aligned}$$

II) At this point in the proof let us derive the original result of Benjamini and Hochberg (1995) for the case of independent test statistics. In this case $\{P_i \le \alpha_j\}$ and B^{-i}_{j-1} are independent events (B^{-i}_{j-1} depends only on $\{P_l, l \ne i\}$) and this gives from (1.12)

$$\begin{aligned} \mathrm{FDR} &= \sum_{i \in M_0} \sum_{j=1}^{m} \frac{1}{j} P\big(P_i \le \alpha_j\big) P\big(B^{-i}_{j-1}\big) \\ &\le \sum_{i \in M_0} \frac{\alpha}{m} \sum_{j=1}^{m} P\big(B^{-i}_{j-1}\big) = \sum_{i \in M_0} \frac{\alpha}{m} P\big(\cup_{j=1}^{m} B^{-i}_{j-1}\big) \\ &= \sum_{i \in M_0} \frac{\alpha}{m} = \frac{m_0}{m}\alpha, \end{aligned}$$

where we have used that the events B^{-i}_{j-1}, $j = 1, \ldots, m$ are disjoint by construction:

$$\begin{aligned} B^{-i}_0 &= \{P^{-i}_{(1)} > \alpha_2, \ldots, P^{-i}_{(m-1)} > \alpha_m\} \\ B^{-i}_1 &= \{P^{-i}_{(1)} \le \alpha_2, P^{-i}_{(2)} > \alpha_3, \ldots, P^{-i}_{(m-1)} > \alpha_m\} \\ B^{-i}_2 &= \{P^{-i}_{(2)} \le \alpha_3, P^{-i}_{(3)} > \alpha_4, \ldots, P^{-i}_{(m-1)} > \alpha_m\} \\ &\vdots \\ B^{-i}_{m-1} &= \{P^{-i}_{(m-1)} \le \alpha_m\}. \end{aligned}$$

Thus, in the case of independent test statistics the Benjamini-Hochberg procedure has FDR at most $\frac{m_0}{m}\alpha$, so that there is strong control at level α.

III) We have now come to the point where the PRDS property is used. Let us first argue that $\bar{B}^{-i}_j = \cup_{s=0}^{j} B^{-i}_s = \{p^{-i}_{(l)} > \alpha_{l+1}$ for $l = j+1, \ldots, m-1\}$ is an increasing set.

If $(p_1,\dots,P_m) \le (q_1,\dots,q_m)$ (coordinate wise) then clearly $\{p_w^{-i}\}_{w\in M\setminus i} \le \{q_w^{-i}\}_{w\in M\setminus i}$, and this in turn implies that $\big(p_{(1)}^{-i},\dots,p_{(m-1)}^{-i}\big) \le \big(q_{(1)}^{-i},\dots,q_{(m-1)}^{-i}\big)$ (by definition of the ordered values there is at least $m-l$ values among p_w^{-i} greater than or equal to $p_{(l)}^{-i}$, then there is at least $m-l$ values among q_w^{-i} greater than or equal to $p_{(l)}^{-i}$, and therefore $q_{(l)}^{-i} \ge p_{(l)}^{-i}$). This shows that if $(p_1,\dots,p_m)$ is in $\bar{B}_j^{-i}$ then so is $(q_1,\dots,q_m)$. Using that $\bar{B}_j^{-i}$ is an increasing set and Lemma 1.31 we make the following calculation

$$
\begin{aligned}
\sum_{j=1}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) &= P\big(B_0^{-i}|P_i \le \alpha_1\big) + \sum_{j=2}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) \\
&\le P\big(B_0^{-i}|P_i \le \alpha_2\big) + P\big(B_1^{-i}|P_i \le \alpha_2\big) + \sum_{j=3}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) \\
&= P\big(\bar{B}_1^{-i}|P_i \le \alpha_2\big) + \sum_{j=3}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) \\
&\le P\big(\bar{B}_1^{-i}|P_i \le \alpha_3\big) + P\big(B_2^{-i}|P_i \le \alpha_3\big) + \sum_{j=4}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) \\
&= P\big(\bar{B}_2^{-i}|P_i \le \alpha_3\big) + \sum_{j=4}^{m} P\big(B_{j-1}^{-i}|P_i \le \alpha_j\big) \\
&\le \cdots \le P\big(\bar{B}_{m-1}^{-i}|P_i \le \alpha_m\big) = 1.
\end{aligned}
$$

From (1.13) we therefore see that

$$
\mathrm{FDR} \le \sum_{i\in M_0} \frac{\alpha}{m} = \frac{m_0}{m}\alpha. \tag{1.14}
$$

We have now proved the first part of the theorem

IV) We next prove the statement for the Benjamini-Yekutelli procedure. Define $p_{ijl} = P(\alpha_{l-1} < P_i \le \alpha_l, B_{j-1}^{-i})$. Then from (1.12) we obtain

$$
\mathrm{FDR} = \sum_{i\in M_0}\sum_{j=1}^{m}\frac{1}{j}\sum_{l=1}^{j} p_{ijl} = \sum_{i\in M_0}\sum_{l=1}^{m}\sum_{j=l}^{m}\frac{1}{j}p_{ijl} \le \sum_{i\in M_0}\sum_{l=1}^{m}\sum_{j=l}^{m}\frac{1}{l}p_{ijl} \le \sum_{i\in M_0}\sum_{l=1}^{m}\frac{1}{l}\sum_{j=1}^{m} p_{ijl}.
$$

Finally, we have for $i \in M_0$:

$$
\sum_{j=1}^{m} p_{ijl} = P\Big(\alpha_{l-1} < P_i \le \alpha_l, \cup_{j=1}^{m} B_{j-1}^{-i}\Big) = P\big(\alpha_{l-1} < P_i \le \alpha_l\big),
$$

where we have used that the sets B_j^{-i} are disjoint and together constitute the com-

plete space (see the detailed description under point II). This gives

$$\begin{aligned}
\text{FDR} &\le \sum_{i\in M_0}\sum_{l=1}^{m}\frac{1}{l}P\big(\alpha_{l-1}<P_i\le\alpha_l\big) = \sum_{i\in M_0}\sum_{l=1}^{m}\frac{1}{l}\big(P\big(P_i\le\alpha_l\big)-P\big(P_i\le\alpha_{l-1}\big)\big)\\
&= \sum_{i\in M_0}\big\{\frac{1}{m}P\big(P_i\le\alpha_m\big)+\big(\frac{1}{m-1}-\frac{1}{m}\big)P\big(P_i\le\alpha_{m-1}\big)+\cdots+\big(\frac{1}{1}-\frac{1}{2}\big)P\big(P_i\le\alpha_1\big)\big\}\\
&\le \sum_{i\in M_0}\big\{\frac{1}{m}\alpha_m+\big(\frac{1}{m-1}-\frac{1}{m}\big)\alpha_{m-1}+\cdots+\big(\frac{1}{1}-\frac{1}{2}\big)\alpha_1\big\}\\
&\le \sum_{i\in M_0}\big\{\frac{1}{m}\frac{m\alpha}{m}+\big(\frac{1}{m-1}-\frac{1}{m}\big)\frac{(m-1)\alpha}{m}+\cdots+\big(\frac{1}{1}-\frac{1}{2}\big)\frac{1\cdot\alpha}{m}\big\}\\
&= \sum_{i\in M_0}\frac{\alpha}{m}\big(1+\frac{1}{m}+\frac{1}{m-1}+\cdots+\frac{1}{2}\big)\\
&== \frac{m_0\alpha}{m}\sum_{l=1}^{m}\frac{1}{l} = \frac{m_0\tilde{\alpha}}{m},
\end{aligned}$$

We have then established the result of the theorem. □

Remark 1.33
Consider the case of a multivariate normal distribution $N_m(\mu,\Sigma)$ for the test statistics, where $M_0=\{i|\mu_i=0\}$ and where $\mu_i>0$ for $i\in M_1$. We thus consider one sided tests and reject when X_i is large. If all correlations with variables from M_0 are non-negative, $\Sigma_{ij}\ge 0$, $i\in M_0$, $j\in M$, the test statistics are PRDS on the subset M_0 (Benjamini and Yekutieli (2001), page 1172).

As far as I can see the multivariate t, as we use in our basic setup, is not treated in Benjamini and Yekutieli (2001), and it is not known (to mee) if PRDS on M_0 holds. From other examples it seems that PRDS does not hold. ■

Exercise 1.34 (Increasing set)
Show that a closed increasing set $D\subset\mathbb{R}^2$ can be written as

$$D=\{(x_1,x_2)|x_2\ge f(x_1)\}$$

for some nonincreasing function f. ■

Exercise 1.35 (PRD for two dimensional normal distribution)
Show that $X\sim N_2(0,\Sigma_0)$, with Σ_0 being 1 at the diagonal and ρ off the diagonal, is PRD when $\rho\ge 0$.

The two dimensional t distribution is defined through $t=(u/s_1,v/s_2)$, where $(u,v)\sim N_2(0,\Sigma_0)$ and independently hereof (s_1^2,s_2^2) is the diagonal of a Wishart$(\Sigma_0,n)/n$. Can you prove PRD for the two dimensional t distribution using ideas from the first question in this exercise? ■

Exercise 1.36 (Breast cancer data)
Consider the Sotiriou data in Problem 1.27. Use Benjamini-Hochberg procedure to find the rejected hypotheses (variables declared differential expressed). ■

Exercise 1.37 (Benjamini-Hochberg for dependent data)
Simulate dependent t-statistics as in Exercise 1.10 and use the Benjamini-Hochberg procedure to reject hypotheses. Find the mean of the number of true positives and the number of false positives. Find the false dicovery rate FDR and the positive false dicovery rate pFDR. ■

1.6 Generalized family wise error rate

This section is based on Lehmann and Romano (2005) (Generalizations of the familywise error rate). The probability of rejecting at least k true hypotheses is denoted kFWER:

$$\text{kFWER} = P(N_{01} \geq k).$$

We consider a stepdown analogue of Holm's procedure. As before let the ordered p-values be $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, and let r_i point to the hypothesis number corresponding to $p_{(i)}$, $p_{r_i} = p_{(i)}$. The procedure is defined as the stepdown algorithm with constants

$$\alpha_i = \begin{cases} \frac{k\alpha}{m} & i \leq k, \\ \frac{k\alpha}{m+k-i} & i > k, \end{cases} \tag{1.15}$$

for $i = 1, \ldots, m$.

Theorem 1.38
The stepdown procedure (1.15) *provides strong control of kFWER at level* α*:*

$$P(N_{01} \geq k) \leq \alpha.$$

Proof. If the number of true hypotheses m_0 is less than k the result holds trivially. Let therefore $m_0 \geq k$. Similarly to the proof of strong control for the Holm procedure we split the event $N_{01} \geq k$ according to the value of v in $q_{(k)} = p_{(k-1+v)}$ (see Notation 1.6), that is, there are $k-1$ true hypotheses with p-values less tham $q_{(k)}$ and $v-1$ false hypotheses with p-values less tham $q_{(k)}$. Using that

$$\alpha_{k-1+v} = \frac{k\alpha}{m+k-(k-1+v)} = \frac{k\alpha}{m_0+m_1-v+1} \leq \frac{k\alpha}{m_0},$$

for $v \leq m_1 + 1$, we find

$$\begin{aligned} P(N_{01} \geq k) &= \sum_{v=1}^{m_1+1} P(P_{(1)} \leq \alpha_1, \ldots, P_{(k-1+v-1)} \leq \alpha_{k-1+v-1}, P_{(k-1+v)} = Q_{(k)} \leq \alpha_{k-1+v}) \\ &\leq \sum_{v=1}^{m_1+1} P(P_{(1)} \leq \alpha_1, \ldots, P_{(k-1+v-1)} \leq \alpha_{k-1+v-1}, P_{(k-1+v)} = Q_{(k)}, Q_{(k)} \leq k\alpha/m_0) \\ &\leq \sum_{v=1}^{m_1+1} P(P_{(k-1+v)} = Q_{(k)}, Q_{(k)} \leq k\alpha/m_0) \\ &= P(Q_{(k)} \leq k\alpha/m_0). \end{aligned}$$

Let K be the number of p-values, q_i, among the true hypotheses with $q_i \le k\alpha/m_0$. Then $K \ge k$ is the same as $Q_{(k)} \le k\alpha/m_0$. Using Lemma 1.39 below we find

$$P(K \ge k) \le \frac{1}{k} \sum_{i \in M_0} P\Big(P_i \le \frac{k\alpha}{m_0}\Big) \le \frac{1}{k} \sum_{i \in M_0} \frac{k\alpha}{m_0} = \alpha.$$

We therefore have the bound kFWER $= P(N_{01} \ge k) \le \alpha$. □

Lemma 1.39
Let N be the sum of m Bernoulli variables, $N = \sum_{i=1}^m 1(X_i \in A_i)$. Then

$$P(N \ge k) \le \frac{\sum_{i=1}^m P(X_i \in A_i)}{k}.$$

Proof. By Markovs inequality

$$P(N \ge k) = E\big(1(N \ge k)\big) = \frac{1}{k} E\big(k1(N \ge k)\big) \le \frac{E(N)}{k} = \frac{\sum_{i=1}^m P(X_i \in A_i)}{k},$$

which is the stated result. □

1.6.1 False discovery proportion

We define the *false discovery proportion*, FDP, as

$$\text{FDP} = \frac{N_{01}}{R} 1(R > 0),$$

so that FDR $= E(\text{FDP})$. For a given γ in $(0,1)$ the next theorem provides a control at level α for the probability $P(\text{FDP} > \gamma)$, using a stepdown procedure with

$$\alpha_i = \frac{(\lfloor i\gamma \rfloor + 1)\alpha}{m + \lfloor i\gamma \rfloor + 1 - i}. \tag{1.16}$$

To state the theorem choose an ordering $1, \ldots, m_1$ of the set M_1 of false hypotheses, and let $u = (u_1, \ldots, u_{m_1})$ be the corresponding p-values. Similarly, choose an ordering of the set M_0 of true hypotheses, and let $q_1, \ldots, q_{m_0}$ be the corresponding p-values.

Theorem 1.40
Assume that for $i = 1, \ldots, m_0$,

$$P\big(q_i \le z | u_1, \ldots, u_{m_1}\big) \le z. \tag{1.17}$$

Then the stepdown procedure with limits (1.16) controls FDP at level α:

$$P(FDP > \gamma) \le \alpha.$$

Proof. Let R_i be the number of u_j's in the interval $(\alpha_{i-1}, \alpha_i]$, where we use $\alpha_0 = 0$, and define for $i = 1, \ldots, m$,

$$\mathrm{Fdp}_i = \frac{i - \sum_{l=1}^{i} R_l}{i}.$$

The stepdown procedure stops at step I (I is the first occurrence of $p_{(I-1)} \le \alpha_{I-1}$ and $p_{(I)} > \alpha_I$), where we then have $\mathrm{FDP} = \mathrm{Fdp}_{I-1}$.

If $\mathrm{Fdp}_v \le \gamma$ for all v we also have $\mathrm{FDP} \le \gamma$. We therefore consider the situation, defined entirely from $u_1, \ldots, u_{m_1}$, where Fdp_v is not less than or equal to γ for all v, and let j be the first instance where Fdp_v is above γ,

$$j = \min\{v : \mathrm{Fdp}_v > \gamma\}.$$

If now $\mathrm{FDP} > \gamma$ we know that $j \le I - 1$. In this case we know that at step j there are at least j p-values less than or equal to α_j and at most $\sum_{l=1}^{j} R_l$ of these can be from false hypotheses so that at least $j - \sum_{l=1}^{j} R_l$ are from true hypotheses. We have therefore obtained

$$P(\mathrm{FDP} > \gamma | u) \le P\Big(K_j \ge j - \sum_{l=1}^{j} R_l \Big| u\Big), \quad K_j = \big|\{i \in M_0 : p_i \le \alpha_j\}\big|. \tag{1.18}$$

We now show that $j - \sum_{l=1}^{j} R_l = 1 + \lfloor j\gamma \rfloor$. We first show by contradiction that $R_j = 0$. If $R_j > 0$ we find

$$\mathrm{Fdp}_{j-1} = \frac{j - 1 - \sum_{l=1}^{j-1} R_l}{j-1} = \frac{j - \sum_{l=1}^{j} R_l + R_j - 1}{j-1} \ge \frac{j - \sum_{l=1}^{j} R_l}{j-1} \ge \frac{j - \sum_{l=1}^{j} R_l}{j} > \gamma,$$

which contradicts the definition of j. Since now $\sum_{l=1}^{j-1} R_l = \sum_{l=1}^{j} R_l$, $\mathrm{Fdp}_{j-1} \le \gamma$ and $\mathrm{Fdp}_j > \gamma$ we have

$$j\gamma < j - \sum_{l=1}^{j} R_l \le 1 + (j-1)\gamma,$$

which implies $j - \sum_{l=1}^{j} R_l = 1 + \lfloor j\gamma \rfloor$. Returning to (1.18) we find from Lemma 1.39 and the assumption of the theorem

$$\begin{aligned} P(\mathrm{FDP} > \gamma | u) &\le \frac{\sum_{i \in M_0} P\big(P_i \le \alpha_j | u\big)}{1 + \lfloor j\gamma \rfloor} \\ &\le \frac{m_0}{1 + \lfloor j\gamma \rfloor} \cdot \frac{(1 + \lfloor j\gamma \rfloor)\alpha}{m + 1 + \lfloor j\gamma \rfloor - j} = \frac{m_0 \alpha}{m + 1 + \lfloor j\gamma \rfloor - j} \\ &= \frac{m_0 \alpha}{m - \sum_{l=1}^{j} R_j} \le \frac{m_0 \alpha}{m - m_1} = \alpha. \end{aligned}$$

Since the bound do not depend on u we have also $P(\mathrm{FDP} > \gamma) \le \alpha$. □

Remark 1.41
Clearly, condition (1.17) is fulfilled if the p-values for the true hypotheses and the p-values for the false hypotheses are independent. ■

The condition of the above theorem involves both the p-values of the true hypotheses and the p-values of the false hypotheses. The next theorem is formulated through a condition on the p-values of the true hypotheses only.

Theorem 1.42
Let $K = \min\{m_0, \lfloor m\gamma \rfloor + 1\}$, and let $q_{(1)}, \ldots, q_{(m_0)}$ be the ordered p-values for the true hypotheses. We consider the stepdown procedure with limits (1.16).

(i) We have the following bound:

$$P(FDP > \gamma) \leq P\Big(\cup_{i=1}^{K} \Big\{ q_{(i)} \leq \frac{i\alpha}{m_0} \Big\} \Big).$$

(ii) Simes inequality for the distribution of the p-values for the true hypotheses states that the right hand side of the above display is bounded by α. Therefore, if Simes inequality holds we have

$$P(FDP > \gamma) \leq \alpha.$$

Proof. At step i in the stepdown procedure, let v_i be the number of true hypotheses being rejected, and define $\mathrm{FDP}_i = \frac{v_i}{i}$. When the algorithm stops at step I we then have $\mathrm{FDP} = \mathrm{FDP}_{I-1}$. If $\mathrm{FDP} > \gamma$ there is a first instance j with $\mathrm{FDP}_j > \gamma$.

Since j is the first instance we have

$$\frac{v_{j-1}}{j-1} \leq \gamma, \quad \frac{v_j}{j} > \gamma,$$

and since we in each step up to $I-1$ add one more hypothesis being rejected, we see that $v_j = v_{j-1} + 1$. Combining the display with the latter statement we have

$$j\gamma < v_j \leq 1 + (j-1)\gamma = j\gamma + 1 - \gamma,$$

from which we find

$$v_j = 1 + \lfloor j\gamma \rfloor. \tag{1.19}$$

Using this we obtain

$$\alpha_j = \frac{(1 + \lfloor j\gamma \rfloor)\alpha}{m + 1 + \lfloor j\gamma \rfloor - j} = \frac{v_j \alpha}{m - (j - v_j)} \leq \frac{v_j \alpha}{m - m_1} = \frac{v_j \alpha}{m_0}.$$

Since $v_j = v_{j-1} + 1$ implies that the hypothesis rejected at step j is a true hypothesis we have $p_{(j)} = q_{(v_j)}$, and therefore

$$q_{(v_j)} \leq \alpha_j \leq \frac{v_j \alpha}{m_0}.$$

For any j we have a k with $k - 1 \leq j\gamma < k$, with the largest possible value of k being $\lfloor m\gamma \rfloor + 1$. With this k we find from (1.19) that $v_j = k$, and the above statement becomes

$$q_{(k)} \leq \frac{k\alpha}{m_0}.$$

This shows that we must also have $k \le m_0$. We now end up with

$$\begin{aligned} P(\text{FDP} > \gamma) &\le P\big(\cup_{j=1}^{m} \{\text{FDP}_l \le \gamma, l = 1,\dots,j-1, \text{FDP}_j > \gamma\}\big) \\ &= P\big(\cup_{k=1}^{K} \cup_{j:k-1\le j\gamma<k} \{\text{FDP}_l \le \gamma, l = 1,\dots,j-1, \text{FDP}_j > \gamma\}\big) \\ &\le P\big(\cup_{k=1}^{K} \{q_{(k)} \le \frac{k\alpha}{m_0}\}\big). \end{aligned}$$ □

The following theorem is a simplified version of a result in Sarkar (2008). We consider m p-values, $p_1,\dots,p_m$, with ordered values $p_{(1)},\dots,p_{(m)}$.

Theorem 1.43 (Simes inequality)
Let the multivariate distribution of the p-values for the true hypotheses be PRD. Then

$$P\Big(Q_{(i)} > \frac{i\alpha}{m_0}, i = 1,\dots,m_0\Big) \ge 1-\alpha,$$

or, equivalently,

$$P\Big(\cup_{i=1}^{m_0} \Big\{Q_{(i)} \le \frac{i\alpha}{m_0}\Big\}\Big) \le \alpha.$$

Proof. Let $\alpha_j = j\alpha/m_0$ and define $R^0 = \max\{j : q_{(j)} \le \alpha_j\}$ and $R^0 = 0$ when $q_{(j)} > \alpha_j$ for all j. When $R^0 = j$ there are exactly j p-values (among the true hypothese) less than or equal to α_j. Then

$$P\Big(Q_{(i)} > \frac{i\alpha}{m_0}, i = 1,\dots,m_0\Big) = P(R^0 = 0) = 1 - P(R^0 \ge 1).$$

Furthermore,

$$\begin{aligned} P(R^0 \ge 1) &= \sum_{j=1}^{m_0} P(R^0 = j) = \sum_{j=1}^{m_0} E\Big(\frac{R^0}{j} 1(R^0 = j)\Big) \\ &= \sum_{j=1}^{m_0} \frac{1}{j} E\Big(\sum_{i=1}^{m_0} 1(Q_i \le \alpha_j) 1(R^0 = j)\Big) \\ &= \sum_{i=1}^{m_0} \sum_{j=1}^{m_0} \frac{1}{j} E\big(1(Q_i \le \alpha_j) 1(R^0 = j)\big). \end{aligned}$$

This is exactly the expression (1.11) (with m replaced by m_0) in the proof of Theorem 1.32, and the result follows from (1.14) with $m = m_0$ in that expression. □

Exercise 1.44 (Increasing sequence)
Show that the sequence in (1.16) is increasing. ■

Exercise 1.45 (Lehmann-Romano kFWER for dependent tests)
Simulate dependent t-statistics as in Exercise 1.10. Use the Lehmann-Romano stepdown procedure with constant (1.15) and show that the procedure controls kFWER at level α. ■

Exercise 1.46 (Lehmann-Romano FDP for dependent tests)
Simulate t-statistics as in Exercise 1.10. Use the Lehmann-Romano stepdown procedure with constants (1.16) and find the probability $P(\text{FDP} > \gamma)$. ■

1.6.2 Alternative permutation based kFWER control

The Lehmann-Romano stepdown method (1.15) to control kFWER, although very simple, is not good in practice. This can be seen as follows. Assume that all hypotheses are true ($m = m_0$) and that variables are independent. The Bonferroni correction rejects when $p_i \le \alpha/m$. The actual FWER=1FWER is $P(N_{01} \ge 1) = 1 - (1-\alpha/m)^m \approx 1 - \exp(-\alpha) = \alpha + O(\alpha^2)$. The *single step version* of the Lehmann-Romano method rejects when $p_i \le k\alpha/m$. For $k = 2$ and independent variables we find $P(N_{01} \ge 2) = 1 - (1-2\alpha/m)^m - m(2\alpha/m)(1-2\alpha/m)^{m-1} \approx 1 - \exp(-2\alpha)(1+2\alpha) = 2\alpha^2 + O(\alpha^3)$. Therefore, the actual 2FWER for the case of independent variables is of order $2\alpha^2$, which is much smaller than the nominal level α. For $k = 3$ the actual 3FWER is of order $18\alpha^3$.

Korn et al. (2004) (Controlling the number of false discoveries: application to high-dimensional genomic data) have provided a simple alternative to the Lehmann-Romano method to control kFWER. It is based on the permutation distribution and can be seen as a generalization of the maxT permutation algorithm. What is needed is to get closer to the actual distribution of $q_{(k)}$. Formally, consider the distribution P where the distribution of $q_1, \ldots, q_{m_0}$ is evaluated under the permutation distribution P_0 (the distribution obtained by randomly permuting the group labels for the subvector with variables from M_0) and the conditional distribution of $q_1^*, \ldots, q_{m_1}^*$ given $q_1, \ldots, q_{m_0}$ is arbitrary. Let P_{00} be the permutation distribution for all variables. We use P_{00} in the construction, but obtain a control of kFWER under P.

For an index set $I \subseteq M$, let $p_{(k)}^I$ be the k'th order statistic of the p-values p_i, $i \in I$, and let

$$u(z; I) = P_{00}(P_{(k)}^I \le z)$$

be the P_{00} permutation distribution of this statistic. Define for $s \ge k$ and for $1 \le i_1 < \cdots < i_{k-1} < s$,

$$w(z; s, i_1, \ldots, i_{k-1}) = u(z; \{r_{i_1}, \ldots, r_{i_{k-1}}, r_s, \ldots, r_m\})$$

and

$$w_{\max}(z; s) = \max_{1 \le i_1 < \cdots < i_{k-1} < s} w(z; s, i_1, \ldots, i_{k-1}).$$

The procedure in Korn et al. (2004) is a stepdown procedure that rejects $H_{r_1}, \ldots, H_{r_{k-1}}$ and from k and onwards rejects as long as $w_{\max}(p_{(s)}; s) \le \alpha$. For $k = 1$ this corresponds to the maxT permutation algorithm.

It is clear that $w_{\max}(z; s)$ is decreasing in s since the k'th order statistic increases when less variables are used. A conservative stategy is therefore to replace $w_{\max}(z; s)$ by $w_{\max}(z; s_0)$ for $s \ge s_0$. The most extreme version of this is to replace $w_{\max}(z; s)$ by $w_{\max}(z; k) = u(z; M)$.

Theorem 1.47
The stepdown permutation procedure of Korn et al. (2004) provides strong control of kFWER under the permutation distribution P described above.

Proof. As in the proof of strong control for the Lehmann-Romano stepdown method to control kFWER we split the event $N_{01} \ge k$ according to the value of v in $q_{(k)} =$

$p_{(k-1+v)}$ (see Notation 1.6). Since the k'th order statistic over a set decreases when the set is expanded we have

$$w_{\max}(z;s) \geq w(z;s,i_1,\ldots,i_{k-1}) \geq u(z;M_0) \text{ whenever } M_0 \subseteq \{r_{i_1},\ldots,r_{i_{k-1}},r_s,\ldots,r_m\}.$$

We then find

$$\begin{aligned}
P(N_{01} \geq k) &= \sum_{v=1}^{m_1+1} P\big(w_{\max}(P_{(k)};k) \leq \alpha,\ldots,w_{\max}(P_{(k-1+v)};k-1+v) \leq \alpha, P_{(k-1+v)} = Q_{(k)}\big) \\
&\leq \sum_{v=1}^{m_1+1} P\big(w_{\max}(P_{(k)};k) \leq \alpha,\ldots,w_{\max}(P_{(k-1+v-1)};k-1+v-1) \leq \alpha, P_{(k-1+v)} = Q_{(k)}, \\
&\qquad u(Q_{(k)};M_0) \leq \alpha\big) \\
&= P\big(\{H_i \text{ is rejected if } P_i < Q_{(k)}\}, u(Q_{(k)};M_0) \leq \alpha\big) \\
&\leq P_0\big(u(Q_{(k)};M_0) \leq \alpha\big) \leq \alpha,
\end{aligned}$$

which shows that FWER $\leq \alpha$. □

The following algorithm is based on the algorithm described in Korn et al. (2004).

Algorithm 1.48 (Permutation algorithm kFWER control)

1. *For the observed p-values we use Notation 1.6. Let the first $k-1$ adjusted p-values be $\tilde{p}_{r_1} = \cdots = \tilde{p}_{r_{k-1}} = 0$. When $\tilde{p}_{r_{v-1}}$ has been found use the steps below to find $\tilde{p}_{r_v}$.*

2. *Make B permutations of the group labels. For each permuation $b = 1,\ldots,B$ compute the p-values $p_{b,1},\ldots,p_{b,m}$ for the m hypotheses. Let C_v be the collection of all subsets of the form*

$$\{r_{i_1},\ldots,r_{i_{k-1}},r_v,\ldots,r_m\}, \quad 0 < i_1 < \cdots < i_{k-1} < v,$$

implying that C_v has cardinality $\binom{v-1}{k-1}$. For $I \in C_v$ let $p^I_{b,(k)}$ the k'th order value for $p_{b,j}$, $j \in I$.

3. *Calculate probabilities*

$$\hat{u}(p_{(v)};I) = \frac{1}{B}\sum_{b=1}^{B} 1\big(p^I_{b,(k)} \leq p_{(v)}\big), \quad I \in C_v$$

and the maximum

$$\hat{u}(p_{(v)}) = \max_{I \in C_v} \hat{u}(p_{(v)};I)$$

4. *Calculate adjusted probability*

$$\tilde{p}_{r_v} = \max\{\tilde{p}_{r_1},\ldots,\tilde{p}_{r_{v-1}},\hat{u}(p_{(v)})\}.$$

Because steps 2 to 4 must be repeated for each new adjusted p-value, the algorithm is computationally complicated. The conservative strategy, where $u(z;s)$ is replaced by $u(z;M)$, looks as follows.

Algorithm 1.49 (Permutation algorithm kFWER control (conservative version))

1. *For the observed p-values we use Notation 1.6. Let the first $k-1$ adjusted p-values be $\tilde{p}_{r_1} = \cdots = \tilde{p}_{r_{k-1}} = 0$.*
2. *Make B permutations of the group labels. For each permuation $b = 1,\ldots,B$ compute the p-values $p_{b,1},\ldots,p_{b,m}$ for the m hypotheses and calculate the k'th order value $p_{b,(k)}$.*
3. *Calculate probabilities*
$$\hat{u}(p_{(j)}) = \frac{1}{B}\sum_{b=1}^{B} 1\big(p_{b,(k)} \le p_{(j)}\big), \quad j = k,\ldots,m.$$
4. *Calculate adjusted probabilities*
$$\tilde{p}_{r_j} = \max\{\tilde{p}_{r_{j-1}}, \hat{u}(p_{(j)})\}, \quad j = k,\ldots,m.$$

1.6.3 Exceedance control

The probalistic control of the false discovery proportion *FDP* as in section 1.6.1 is also known under the name *exceedance control*. General methods for obtaining exceedance control are discussed in Genovese and Wasserman (2006) (Exceedance Control of the False Discovery Proportion).

To be investigated further.

1.7 SAM procedure

One of the first approaches to using permutation analysis in connection with microarray data (the typical setting in this chapter) is the Significance Analysis of Microarrays (section 12). The original reference is Tusher et al. (2001).

If $t_{b,1},\ldots,t_{b,m}$ are the t-values from a permuation b, the SAM procedure calculates expected values of order statistics,

$$\bar{d}_{(j)} = \frac{1}{B}\sum_{b=1}^{B} t_{b,(j)}, \quad j = 1,\ldots,m,$$

where $t_{b,(1)},\ldots,t_{b,(m)}$ are the ordered values. Then a plot of $t_{(j)}$ versus $\bar{d}_{(j)}$ is made and a sort of double sided stepup procedure is used.

The sam method does not use the ordinary t-statistic, but replaces s in the denominator by $s + s_0$ for a chosen constant s_0.

Try to form your own opinion on the method.

CHAPTER

2 Efron's two group model

Efron in a series of papers and in his book, Efron (2010b) (you can download chapters from https://doi.org/10.1017/CBO9780511761362), looks directly at the empirical distribution of the p-values (or the test statistic) to disentangle the part from the true hypotheses and the part from the false hypotheses. Efron uses a Bayesian setup where a hypothesis is either true or false with probabilities π_0 or $1-\pi_0$. For a true hypothesis the density of the test statistic is $f_0(z)$ and for a false hypothesis the density is $f_1(z)$. This setup is not of much help in practice if all the terms are unknown, but the idea is that f_0 is known (or partly known).

We can formalize the setup as the data being (I_i, Z_i), where I is the group number (0 or 1) and Z_i is the test statistic. In the Bayesian setting I is random with $P(I = 0) = \pi_0$, whereas in the frequentist setting of the previous sections $I_1, \ldots, I_m$ are fixed. Furthermore, the density of Z given $I = j$ is $f_j(z)$. The pairs $(I_1, Z_1), \ldots, (I_m, Z_m)$ need not be independent. In the Bayesian setting we can calculate the following Bayesian false discovery rates

$$\begin{aligned}
\text{BFdr}(z_0) &= P(I = 0 | Z \leq z_0) = \frac{\pi_0 F_0(z_0)}{\pi_0 F_0(z_0) + \pi_1 F_1(z_0)} \\
\text{Bfdr}(z_0) &= P(I = 0 | Z = z_0) = \frac{\pi_0 f_0(z_0)}{\pi_0 f_0(z_0) + \pi_1 f_1(z_0)}.
\end{aligned}$$

These probabilities do not exist in the frequentist setting. When looking at the event $Z \leq z_0$ we can instead in the frequentist setting calculate the expected number among the true hypotheses and compare with the expected total number. this gives BFdr with $\pi_0 = m_0/m$, and with a similar interpretation of Bfdr.

We cannot calculate the above rates as we do not know f_1 or F_1. However, we can use the empirical distribution function to estimate $F(z) = \pi_0 F_0(z) + \pi_1 F_1(z)$, that is, we estimate $F(z)$ by $\bar{F}(z) = \frac{1}{m}\sum_i 1(z_i \leq z)$. We thus have the empirical Bayes false discovery rate $\text{EBFdr}(z) = \pi_0 F_0(z)/\bar{F}(z)$. We do not know π_0 in this expression, but we expect this to be close to one so that we have the upper bound $\text{EBFdr}(z) \leq F_0(z)/\bar{F}(z)$.

The Benjamini-Hochberg procedure can be given a simple (Efron (2010b), section 4.3) interpretation in terms of the empirical Bayes false discovery rate. The procedure looks for the maximal index i with $F_0(z_{(i)}) \leq \frac{i}{m}\alpha$ and controls the FDR at level $\frac{m_0}{m}\alpha$. Letting $\pi_0 = \frac{m_0}{m}$ and noting that $\bar{F}(z_{(i)}) = \frac{i}{m}$ we have $\text{EBFdr}(z_{(i)}) = \big(\frac{m_0}{m}F_0(z_{(i)})\big)/\big(\frac{i}{m}\big)$, and $\text{EBFdr}(z_{(i)}) \leq \tilde{\alpha}$ is equivalent to $F_0(z_{(i)}) \leq \frac{i}{m}\big(\frac{m}{m_0}\tilde{\alpha}\big)$, which controls the FDR at level $\frac{m_0}{m}\big(\frac{m}{m_0}\tilde{\alpha}\big) = \tilde{\alpha}$. This shows that we have control at level $\tilde{\alpha}$ when we look for the maximal index i with $\text{EBFdr}(z_{(i)}) \leq \tilde{\alpha}$.

Example 2.1
As a first illustration of the Efron two group model let us consider the same dataset as in Efron (2010b). The dataset consists of the expression levels for 6033 genes for each of 102 persons. The 102 persons are divided into two groups with 50 healthy persons and 52 with prostate cancer. The original source of the data is Singh et al. (2002).

For each gene we calculate the t-statistic for no differential expression between the two groups. If there is no differential expression we take as a starting point the distribution of the test statistic to be $t(100)$. As in Efron (2010b) we transform the t-values to a standard normal scale using (here shown with notation from R)

$$z_i = \text{qnorm}(\text{pt}(t_i, 100))$$

■

Figure 2.1 shows on the left a histogram for the z-values together with a standard normal density (black) and a standard normal density multiplied by 0.926 (red).

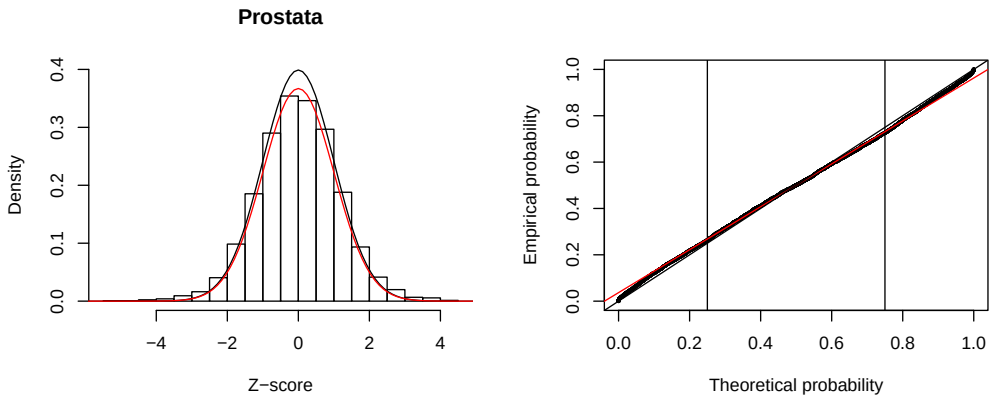

Figure 2.1: Histogram of the z-scores for the Prostata data together with the density of a standard normal distribution (black) and the latter multiplied by 0.926 (red). The right part of the figure is a pp-plot, where the theoretical distribution is a standard normal distribution.

Efron estimates the fraction π_0 of true hypotheses by assuming that the density $f_1(z)$ under the alternative does not contribute to the total density in a region around zero. Efron then uses the fraction of data points in the chosen region to estimate π_0. As an example, if we take the region to be the central 50% of a standard normal distribution the estimate of π is $\hat{\pi} = \frac{1}{m}\sum_i 1(u_{0.25} \leq z_i \leq u_{0.75})/(0.75 - 0.25)$ (u_p is the p-quantile of a standard normal distribution). This gives the value $\hat{\pi}_0 = 0.926$.

Another graphical illustration can be made with a pp-plot. The theoretical distribution function for the z-values in a region around zero is $\pi_{11} + \pi_0 F_0(z)$, where F_0 is the standard normal distribution and π_{11} is the fraction of non-null cases with a negative z-value (assuming that F_1 has no probability mass in a region around zero). Points on the form $\pi_{11} + \pi_0 F_0(z_{(i)}), (i - 0.5)/m)$ should therefore follow a

straight line with slope π_0 in the central part. The right part of figure 2.1 show this, where the black line is the identity line and the red line is a line with slope 0.926.

A very important observation made by Efron is that the theoretical null distribution of the test statitics often does not fit the central part of the data distribution, the actual distribution is wider than the theoretical. We start with an example from Efron (2010b) in order to compare his method of estimation with a proposal below. The data we consider consists of the expression levels of 7128 genes for 72 leukemia patients. The 72 persons are divided into two group: ALL (Acute lymphocytic leukemia) and AML (Acute myeloid leukemia). The original publication for the data is Golub et al. (1999). The data used here are taken from http://hastie.su.domains/CASI_files/DATA/leukemia_big.csv

For each gene we calculate the t-statistic for no difference between the two groups. The theoretical null distribution is a t-distribution with 70 degrees of freedom. Following Efron we consider instead the test statistic

$$z_i = \text{qnorm}(\text{pt}(t, 70)),$$

where the null distribution is a standard normal distribution. Figure 2.2 (which corresponds to Figure 6.1.a in Efron (2010b)) displays a density histogram for the z-values with the standard normal density multiplied by 0.655 as well as the density of a $N(0.0935, 1.637)$ distribution multiplied by 0.931 (these numbers are slightly different from Efrons numbers since also the histogram by itself deviates slightly from the histogram in Efron (2010b)). It is clear that in the central part of the distribution the latter distribution gives a better fit.

Efron discusses various explanations for this overdispersion in section 6.4 of his book. Small differences in the composition of the two groups can perhaps lead to small differences in the mean. Imagine as an example that the expression level of a gene has a slight correlation with age, then a difference in the age composition for the two groups can give a difference in the means. Similarly, the accumulated amount of a toxic substance in the body may differ between the two groups. Also technical aspects of the measurement technique can perhaps lead to a difference in the means for a gene for the two groups.

Efron (2010b) mentions two methods for estimating the null distribution so as to fit the central part of the distribution. In the first method the mixture density $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$ is estimated by fitting (typically) a sixth degree polynomial to the log of the density using a poisson model for histogram counts, that is, the histogram count is poisson distributed with mean $m\delta f(z)$, where δ is the bin width and z is the center of the bin. Next a second degree polynomial is fitted by least squares to the logarithm of the density estimate $\hat{f}(z)$ in a central region $z^1 \leq z \leq z^2$ using the same grid points as in the first part. Typically Efron uses the first and third quartiles for z^1 and z^2. The second degree polynomial is then equated to $\log(\pi_0/\sqrt{2\pi\sigma^2}) - (x-\mu)^2/(2\sigma^2)$. One may get the impression that only the central 50% central region matters, but this is not quite true. Using first a 6. degree polynomial means that the

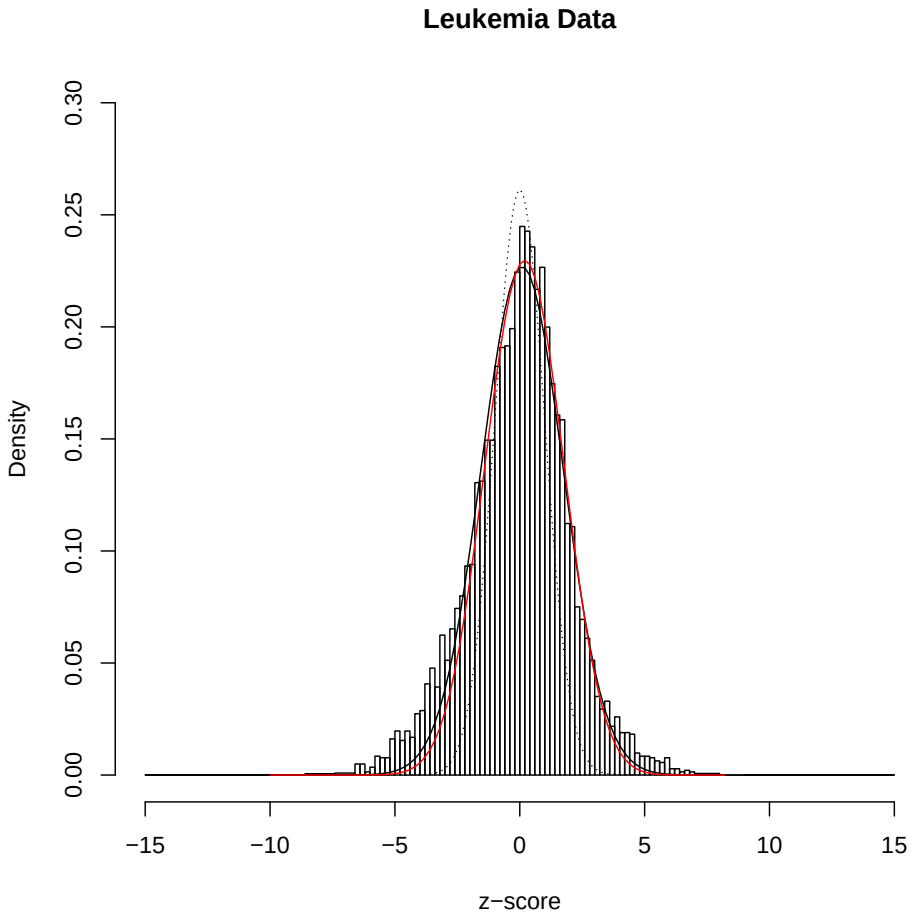

Figure 2.2: Histogram of z-scores for Leukemia data together with the density of a standard normal distribution multiplied by 0.655 (dotted) and the density of a $N($ 0.0935, 1.637) distribution multiplied by 0.931 (full drawn). The red curve displays the result of modelling the t-values directly.

estimate in the central 50% region is influenced by the neighgouring parts of the data.

In the second method a maximum likelihood approach, assuming independence, is used for data in the central region and using the normal distribution conditioned on falling in the central region. The central region here is different from the one mentioned above in the first method. It is taken as the median plus minus Ks_0, where s_0 is the standard deviation of a normal distribution having the same inter quartile range as the data, and K is 2 or smaller. For the Leukemia data it seems that Efron has used a value around 1.75, which implies that roughly 88% of the data are included. Finally, the fraction π_0 is estimated from the number of observations $N(z^1, z^2)$ in the central region and the probability $p(z^1, z^2)$ of the central region from

the estimated normal distribution,

$$\hat{\pi}_0 = \frac{N(z^1, z^2)}{m \cdot p(z^1, z^2)}.$$

For the Leukemia data the results of the two estimation methods are given in the following table

Method	π_0	μ	σ
Polynomial fitting	0.94	0.15	1.69
"Maximum likelihood"	0.93	0.09	1.64

2.1 A new estimation method

For a t-test situation Efron (2010b) uses the transformation

$$z_i = \Phi^{-1}\big(t_{\text{cdf}}(t, df)\big), \tag{2.1}$$

where Φ is the standard normal cumulative distribution function. However, when data t_i shows overdispersion as compared to the theoretical $t(df)$-distribution is is no longer obvious to use this transformation. Efron (2010a) shows in various ways that the distribution of the z-score can still be approximated by a normal distribution when t has a noncentral t-distribution. As I discuss next the situation we envisage is slightly different. Consider the two sample t-test based on the model $x_{1i} \sim N(\mu_1, \sigma^2)$, $i = 1, \ldots, n_1$, and $x_{2i} \sim N(\mu_2, \sigma^2)$, $i = 1, \ldots, n_2$, that is, $t = (\bar{x}_1 - \bar{x}_2)/\sqrt{s^2 \cdot (1/n_1 + 1/n_2)}$. If $v = \mu_1 - \mu_2$ is the difference of the two means we can write the distribution of t as the distribution of

$$t \sim \frac{\widetilde{u} + v/\sqrt{\sigma^2 \cdot (1/n_1 + 1/n_2)}}{\sqrt{w}}, \quad \widetilde{u} \sim N(0,1), \quad w \sim \chi^2(n_1 + n_2 - 2)/(n_1 + n_2 - 2).$$

Imagine now that the overdispersion originates from small random differences $v \sim N(\sigma\xi, \sigma^2\psi^2)$. This gives

$$t \sim \tau \frac{u + \delta}{\sqrt{w}}, \quad u \sim N(0,1), \quad \delta = \frac{\xi}{\sqrt{\psi^2 + 1/n_1 + 1/n_2}}, \quad \tau = \sqrt{1 + \psi^2/(1/n_1 + 1/n_2)}.$$

This shows that the distribution is a *scaled* noncentral t-distribution with noncentrality parameter δ and scaling τ. In Figure 2.3 the distribution of the z-score from (2.1) is shown for the case where t has a scaled noncentral t-distribution. The approximating normal distribution (red in the figure) is chosen to match the true distribution in terms of median and interquartile range. It appears from the figure that is the degrees of freedom is small the approximating normal distribution is not always good.

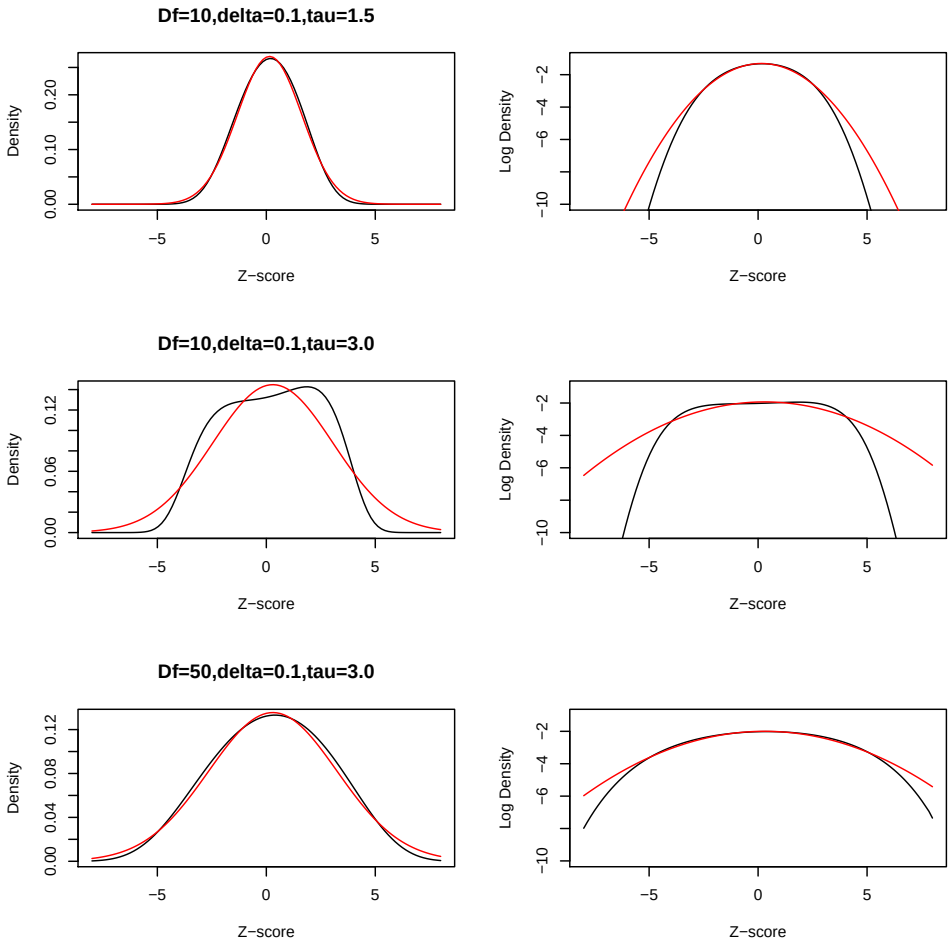

Figure 2.3: The black curve is the density of the z-score from (2.1) when t has a scaled noncentral t-distribution, and the red curve is an approximating normal density.

Like in Efron's model we have three parameters to describe the central part of the distribution, δ, τ and the null fraction π_0. In his own words Efron uses the central part of the distribution for estimation. However, this is perhaps an understatement. In his first method, where a quadratic polynomial is fitted to a 6. degree polynomial, the latter is fitted to all of the data and the central 50% part is therefore influenced by the data outside of this region. In Efrons 'maximum likelihood method' a fairly wide region is actually used for the estimation, in particular for the Leukemia data it seems that roughly 88% of the data are used. Can we perform the estimation differently so that we use the central 50% 'only', but at the same time ensure that the tails of $\pi_0 f_0(z)$ are lighter than the empirical tails?

We consider binned t-values and let a_i be the observed number in bin i. The expected number e_i from the null distribution is $m\pi_0 p(i,\delta,\tau)$ with parameters δ

and τ as described above. We choose the bins such that all the observed numbers are at least five. The three parameters are estimated by minimizing

$$C = \sum_{i \in I_0} \frac{(a_i - e_i)^2}{e_i} + \sum_{i \notin I_0} \psi\big((a_i - e_i)/e_i\big)^2,$$

where I_0 are those bins that constitute the central region of the distribution (roughly 50% of the distribution), and ψ is given by

$$\psi(x) = \begin{cases} x & x < 2, \\ 2 & x > 2. \end{cases}$$

The first part of C, corresponding to the central part of the distribution, ensures that the central part is described by $\pi_0 f_0$. The second part ensures that $\pi_0 f_0$ is consistent with the idea that outside the central region there is a contribution from the non-null cases. Thus, values of a_i much below the expected number e_i are not allowed and enters C in the usual way for the χ^2-statistic, whereas values of a_i much above e_i are consistent with the non-null cases entering and therefore should not be punished in the usual way in the χ^2-statistic. This is handled by replacing values of $(a_i - e_i)/\sqrt{e_i}$ above 2 by the bound 2.

For the Leukemia data the estimates becomes $\pi_0 = 0.898$, $\delta = 0.120$ and $\tau = 1.560$ based on bins that can be seen in figure 2.4. The upper part of the figure displays the bin counts and the expected numbers from the null cases. The points marked by red constitute the central part, corresponding to the first sum of C. The lower part of the figure shows the log counts. The estimated null density for the t-values has been transformed to the z-scale and shown also in figure 2.2.

2.2 Standard deviation of correlated tail counts

For a given cutoff, say $t_0 < 0$, we would like to know the distibution of the false positive number $N_0 = \sum_{i \in M_0} 1(t_i < t_0)$. As pointed out by Efron the correlation between the t-values is so strong that we cannot use the binomial distribution and the corresponding standard deviation.

In a general setting consider m test statistics z_i, $i = 1, \ldots, m$, and let $I_i = 1(z_i < z_0)$ together with $N = \sum_{i=1}^m I_i$. Then

$$\mathrm{Var}(N) = \sum_i \mathrm{Var}(I_i) + \sum_{i,j,j \neq i} \mathrm{Cov}(I_i, I_j).$$

Let all the test statistic have the same marginal distribution and let $p(z_0) = P(z_i < z_0)$ and $p_2(\rho_{i,j}, z_0) = P(z_i < z_0, z_j < z_0)$, where $\rho_{i,j}$ is the correlation describing cases i and j (not the correlation of I_i and I_j, but rather correlation of z_i and z_j). Also let $c(\rho, z_0) = p_2(\rho, z_0) - p_0(z_0)^2$, which equals $\mathrm{Cov}(I_i, I_j)$. Then

$$\mathrm{Var}(N) = mp(z_0)(1-p(z_0)) + \sum_{i,j,j \neq i} c(\rho_{i,j}, z_0) = mp(z_0)(1-p(z_0)) + m(m-1)E_\rho c_2(\rho, z_0),$$

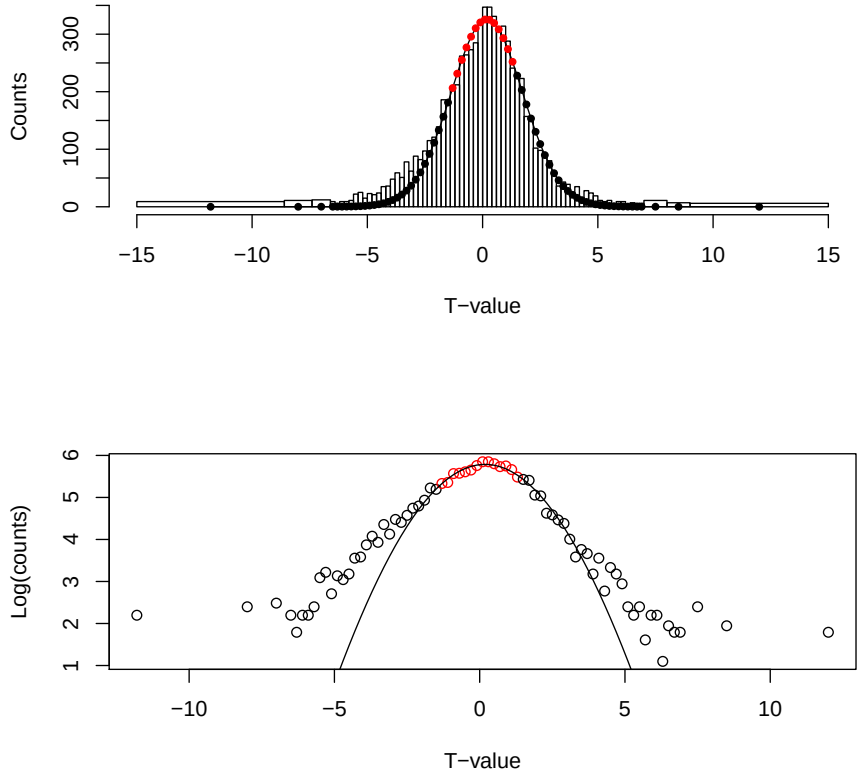

Figure 2.4: Histogram of t-values for Leukemia data. The fitted curve is the fitted values from a scaled noncentral t-distribution. The central part of the distribution is shown in red. The lower subfigure shows the log histogram values.

where E_ρ is simply a notation for the average over the $m(m-1)$ covariance terms.

To make the formula more concrete imagine that there are only three different correlations, zero and plus minus ρ_1, and that these three appear in the fractions α for $\rho = 0$ and $(1-\alpha)/2$ for both of $\rho = \rho_1$ and $\rho = -\rho_1$. Then we obtain

$$\mathrm{Var}(\frac{N}{m}) = \frac{p(z_0)(1-p(z_0))}{m} + \frac{m-1}{m}\frac{1-\alpha}{2}\big(c(\rho_1, z_0) + c(-\rho_1, z_0)\big), \tag{2.2}$$

where we have used the assumption that $c(0, z_0) = 0$. An important observation here is that if the correlation is strong, in the sense that α does not approach one as m increases, then the variance of the observed fraction N/m does not decrease to zero as m increases. More generally, the second term above often increases the standard deviation considerably as compared to the binomial term only.

To illustrate the effect consider the case of standard normally distributed test statistics and where the bivariate distribution of two test statistics is bivariate normal with unit variances and correlation ρ. Using the package *mvtnorm* in R we can

calculate $p_2(\rho, z_0)$ as

$$p_2^N(\rho, z_0) = \text{pmvnorm(corr=Sig,lower=c(-Inf,-Inf),upper=c}(z_0,z_0)), \ \text{Sig} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and calculate $p_0(z_0)$ as `pnorm(z_0)`. Figure 2.5 shows the standard deviation of N/m as function of z_0 when $m = 7000$ and ρ is concentrated on 0 and ± 0.2 with the fraction of zeros being $\alpha = 0.64$. Included in the figure is the binomial standard deviation and an approximation from Efron (2010a) where the second term in (2.2) in the setting here is approximated by

$$\frac{1}{2} E_\rho(\rho^2) z_0^2 \varphi(z_0)^2,$$

with $E_\rho(\rho^2) = 0.18 \cdot 0.2^2 + 0.18 \cdot (-0.2)^2 = 0.0144$. As appears from the figure Efrons approximation is almost indistinguisable from the exact term for the example here (Efron (2010a) shows an example with larger differences). For z_0 between -1 and -2 the actual standard deviation is of the order five times higher than the binomial term. From $z_0 = -3$ and further out the ratio decreases rapidly from a factor around three times bigger.

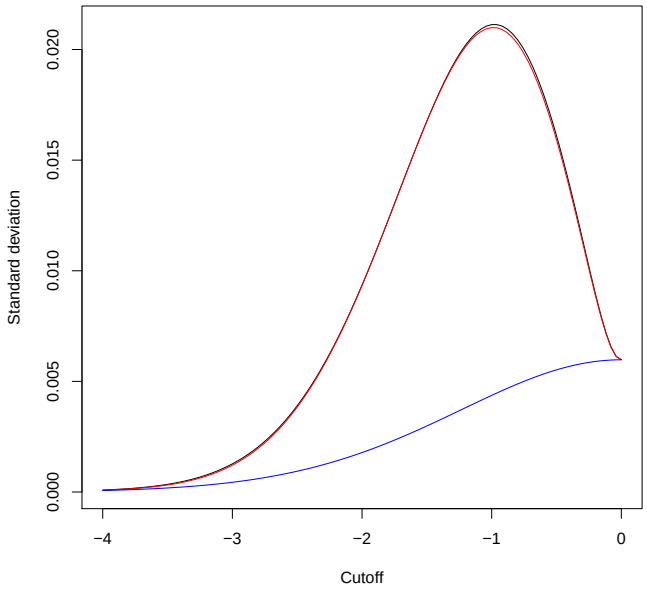

Figure 2.5: The black curve (almost hidden by the red curve) shows 2.2 for the situation described in the text. The red curve is the approximation from Efron (2010a) and the blue curve is the binomial part in 2.2.

We next turn to t-statistics based on correlated random variables. Consider variables $x_{11}, \ldots, x_{1n_1}$ and $x_{21}, \ldots, x_{2n_1}$ where within each sample variables are independent ('persons are independent'), but x_{1i} is correlated with x_{2i}, $\text{cor}(x_{1i}, x_{2i}) =$

ρ ('variables are correlated within the same person'). Similarly with the variables $y_{11},\ldots,y_{1n_2}$ and $y_{21},\ldots,y_{2n_2}$. From these variables we generate the two t-statistic

$$t_1=\frac{\bar{y}_1-\bar{x}_1}{\sqrt{s_1^2(1/n_1+1/n_2)}}\quad\text{and}\quad t_2=\frac{\bar{y}_2-\bar{x}_2}{\sqrt{s_2^2(1/n_1+1/n_2)}}.$$

The problem using (2.2) is that there does not seem to be a function in R that can calculate $p_2(\rho,t_0)=P(t_1\le t_0,t_2\le t_0)$. We therefore try to estimate this probability by simulations. What makes this difficult is that $p_2(\rho,t_0)$ is typically small and we subtract $p(t_0)^2$ making the result even smaller. A direct simulation approach is therefore not feasible. It is possible to make an importance sampling procedure, that increases the precision of the simulation, but a better approach is to simulate the variance terms s_1^2 and s_2^2 and then use the *mvtnorm* package to calculate the conditional probability $P(t_1\le t_0,t_2\le t_0|s_1^2,s_2^2)$. Due to the definition of the t-statistics we can take the variances of the observations to be one. We draw $(n_1+n2-2)(s_1^2,s_2^2)$ from a bivariate Wishart distribution with variance matrix Σ having one at the two diagonal elements and ρ as the off diagonal element. Since the averages are independent of the variance terms s_1^2 and s_2^2 we end up with

$$\begin{aligned}p_2(\rho,t_0)&=E\Big\{P\Big(\frac{\bar{y}_1-\bar{x}_1}{\sqrt{1/n_1+1/n_2}}\le s_1t_0,\frac{\bar{y}_2-\bar{x}_2}{\sqrt{1/n_1+1/n_2}}\le s_2t_0|s_1^2,s_2^2\Big)\Big\}\\&=E\Big\{\text{pmvnorm(corr=Sig,lower=c(-Inf,-Inf),upper=c}(s_1t_0,s_2t_0))\Big\},\qquad(2.3)\\\text{Sig}&=\begin{pmatrix}1&\rho\\\rho&1\end{pmatrix},\end{aligned}$$

(the alternative function *pbinom* from the package *VGAM* is not suffciently accurate for this calculation).

We compare below the results with those obtained by replacing the simulated value of $p_2(\rho,t_0)$ by the value from a two dimensional normal distribution

$$p_2^N(\rho,z_0),\quad z_0=\Phi^{-1}\big(t_{\text{cdf}}(t_0,n_1-n_2-2)\big),\qquad(2.4)$$

where z_0 is the t-value transformed to a standard normal scale. Figure 2.6 shows four situations all with $m=7000$ and ρ concentrated at zero and ±0.2, the fraction of the latter two being 0.18 each. The simulated values are based on 10^6 values. The upper row is for the situation above with t-distributions. The topleft is for the situation $n_1=47$ and $n_2=25$ whereas the topright has $n_1=n_2=10$. For the case $n_1=47$ and $n_2=25$ the normal approximation is at most of the order 6 percent larger than the simulated value, and for the case $n_1=10$ and $n_2=10$ the normal approximation is at most of the order 20 percent larger.

We finally turn to the situation of t-statistics showing overdispersion. Taking $\sigma^2=1$ the t-statistics take the form

$$T_1=\frac{\bar{Y}_1-\bar{X}_1+\psi U_1}{\sqrt{s_1^2(1/n_1+1/n_2)}}\ \text{ and }\ T_2=\frac{\bar{Y}_2-\bar{X}_2+\psi U_2}{\sqrt{s_2^2(1/n_1+1/n_2)}},$$

where $U_1 \sim N(\xi, 1)$ and the conditional distribution of U_2 given U_1 is $N\big(\xi + \rho_0(u_1 - \xi), 1 - \rho_0^2\big)$. The correlation between the two nominators are

$$\mathrm{Cor}\big(\overline{Y}_1 - \overline{X}_1 + \psi U_1, \overline{Y}_2 - \overline{X}_2 + \psi U_2\big) = \frac{\rho + \psi^2 \rho_0 n_1 n_2/n}{1 + \psi^2 n_1 n_2/n}.$$

The correlation is therefore ρ when $\psi = 0$, corresponding to no additional variation in the t-statistic, and when $\rho_0 = \rho$. When $\rho_0 = 0$, the correlation is reduced as compared to the case with $\psi = 0$.

The distribution of T_j is that of τT, where $T \sim t(n-2, \delta)$, $\tau^2 = 1 + \psi^2/(1/n_1 + 1/n_2)$, and the noncentrality parameter δ is $\delta = \psi\xi/(\tau\sqrt{(1/n_1 + 1/n_2)})$. Instead of (2.3) we get

$$p_2(\rho, t_0) = E\Big\{\text{pmvnorm(corr=Sig,lower=c(-Inf,-Inf),upper=c}(z_1, z_2))\Big\},$$
$$z_i = (s_i t_0 - \delta\tau)/\tau.$$

As above we compare the results with those obtained by replacing the simulated value of $p_2(\rho, t_0)$ by the value from a two dimensional normal distribution

$$p_2^N\big(\rho, z_0\big). \quad z_0 = \Phi^{-1}\big(t_{\mathrm{cdf}}(t_0/\tau, n_1 - n_2 - 2, \mathrm{ncp} = \delta)\big). \tag{2.5}$$

The lower part of figure 2.6 has $n_1 = 47$, $n_2 = 25$, $\psi = 0.3$ and $\delta = -0.16$ in the left part and $n_1 = 10$, $n_2 = 10$, $\psi = 0.3$ and $\delta = -0.5$ in the right part. The correlation structure is as in the upper part of the figure. For the case with $n_1 = 47$ and $n_2 = 25$ there is almost no difference between the simulated values and the normal approximation (the latter is approximately 6 percent larger in the left tail), whereas for the case with $n_1 = 10$, $n_2 = 10$ and $\delta = -0.5$ the normal approximation is of the order 30 percent larger in the left tail and roughly 10 percent smaller for z-scores close to zero.

The conclusion from the figures is that the approximation (2.5) based on a bivariate normal distribution can be used in most cases.

2.3 Permutation analysis

In the two group situation with $n = n_1 + n_2$ m-dimensional observations it is often suggested to take into account correlations between the m variables by using a permutation analysis. Thus we permute the n observation and pick n_1 of these as constituting group 1 and the remaining n_2 constitutes group 2. For each permutation the m t-statistics are evaluated and we can study the mean and variance of, say, the number of t-values below a given limit.

However, as pointed out in Efron (2010b) (section 6.5), the permutation distribution of the t-values resembles a t-distribution and so does not capture the overspread as seen often in the data. For the leukemia data the overspread is shown in figure 2.2 and the topleft part of figure 2.7 (based on 1000 permuted samples) shows

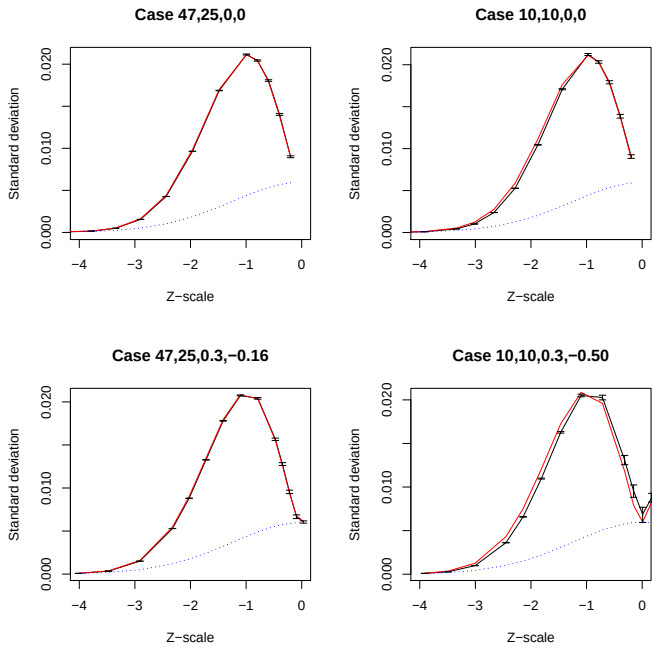

Figure 2.6: Simulated standard errors for the fraction of t-values below a given limit. The red curve is based on the approximation (2.5). In all cases the abscisse has been transformed to a standard normal scale. A 95% confidence interval for the simulated value is shown as well. The blue curve is the binomial part of the variation.

that the permutation distribution follows a $t(n_1+n_2-2)$ distribution quite closely, that is, the permutation distribution does not show the overspread.

The topright part of figure 2.7 shows the observed standard deviation of the tail fraction obtained from the permutation analysis. As can be seen the standard deviation is slightly larger that the approximation (2.4) to the theoretical expression, discussed in the previous section, using an average squared correlation of 0.12 in agreement with the numbers in Efron (2010a). The bottom part of figure 2.7 is included to discuss the use of the estimated standard deviations. Based on 1000 permuations, histograms are shown for the number of t-values being less that -3.0 and -3.4, respectively. Included in the figures is the density of a normal distribution with the same mean and standard deviation as the data in the histograms. As can be seen the normal distribution does not fit the data well. The heading shows this directly giving the number of instances with a value above a certain limit and the expected number from the approximating normal distribution. To make this more concrete when looking at the leukemia data and the number of t-values less than -3.4 a normal approximation would say that among 7128 null cases we would see 21 or more in less than one out of 1000 cases, whereas the permutation analysis says this happens in approximately 1 out of 100 cases.

The above discussion is not fully relevant for the leukemia data since the permutation distribution does not have the same overspread as the actual data. Can we improve on this? In section 2.1 we took the overspread into account ending up with a scaled noncentral t-distribution. In the permutation setting we could also add a small random component when constructing the t-values. The difficulties here is that the added component should have a variable specific variance and perhaps should have a correlation structure resembling the correlation structure in the data. However, we do not need to add a random component to obtain the larger variance, we can simply scale the t-values obtained from the permutation simulation. We still need to add a constant term scaled by the unknown standard deviation for each variable. Ideally this takes the form

$$T = \tau \frac{\overline{Y} - \overline{X} + \sigma\xi}{\sqrt{s^2(1/n_1 + 1/n2)}}$$

based on the permuted data. Since we do not know σ we will use an empirical Bayes approach. We model σ^2 as coming from a Gamma distribution (with mean λ/β and variance λ/β^2) and replace σ by $\sqrt{E(\sigma^2|s^2)}$. We estimate λ and β, for each permuted dataset, using that in the Bayesian setting

$$E(s^2) = \frac{\lambda}{\beta} \quad \text{and} \quad E(s^4) = \frac{\lambda(\lambda+1)}{\beta^2}\big(1 + \frac{2}{df}\big),$$

giving

$$\lambda = \frac{\hat{\mu}_1^2(1+2/df)}{\hat{\mu}_2 - \hat{\mu}_1^2(1+2/df)} \quad \text{and} \quad \beta = \frac{\lambda}{\hat{\mu}_1},$$

where $\hat{\mu}_1$ and $\hat{\mu}_2$ are averages of s^2 and s^4. The conditional distribution of σ^2 given s^2 is a generalized inverse gaussian distribution with mean

$$\sqrt{\frac{b}{a}} \frac{K_{p+1}(\sqrt{ab})}{K_p(\sqrt{ab})},$$

where

$$a = 2\beta, \quad b = df\, s^2 \quad \text{and} \quad p = \lambda - \frac{df}{2}.$$

Figure 2.8 corresponds to figure 2.7, but now with the permutation distribution being a scaled noncentral t-distribution. For the Leukemia data we use the scaling $\tau = 1.56$ and with the noncentrality parameter $\delta = 0.12$ we get $\xi = \delta\sqrt{1/n_1 + 1/n_2}$. The upper left part shows that the scaled noncentral t distribution fit the simulated permutation values very closely.

2.4 Conditional argument

Let $N_a = \sum_{i \in M_0} 1(t_i \le a)$ and $N_a^T = \sum_{i \in M_1} 1(t_i \le a)$ be the tail counts, using a threshold a, for the null cases and the nonnull cases respectively. Thus, if we reject a hypothesis when the t-value is below a, N_a is the number of false positives and N_a^T the

number of true positives. Also, let $\bar{N}_c$ be the central count, $\bar{N}_c = \sum_{i \in M} 1(c_1 \le t_i \le c_2)$. In the section above we studied the distibution of N_a through its mean and variance.

The motivation for studying the distribution of N_a is the following. If, in the original data, we have observed that $N_a + N_a^T = 100$, say, and if we know that it is very unlikely (the probability is small) that $N_a > 30$, say, we have some confidence in the statement that the number of true positives N_a^T is greater than $100 - 30 = 70$ (I will discuss this statement further below).

Figure 2.9 shows a plot of N_a^b against $\bar{N}_c^b$ for the Leukemia data when $a = -4$ and $\bar{N}_c^b = \sum_i 1(-1.2 \le t_i \le 1.2)$. Here $(N_a^b, \bar{N}_c^b)$ are values from a permutation analysis as descibed in section 2.3, $b = 1\ldots,1000$ (using $\tau = 1.55$, $\delta = 0.12$ and in each permutation a random fraction $\pi_0 = 0.895$ of the variables is selected). Figure 2.10 shows on the left a histogram of the N_a^b values. The range of these values are from 2 to 215 with 5 out of 1000 permutations having a value above 150. Thus values above 150 are very unlikely. For the Leukemia data there are 277 variables with a t-value below $a = -4$. Thus the distribution of N_a^b indicates that at least $277 - 150 = 127$ variables are true positives.

Can we improve the precision of this statement? The central t-values from the original data, that enter the central count $\bar{N}_c$, are used to estimate π_0, τ and δ. Afterwards we use the estimates to discuss the distribution of the tail count N_a. We could more directly consider a conditional argument conditioning on the value of $\bar{N}_c$ ($\bar{N}_c$ gives information on the actual spread for the data at hand that has a direct influence on tail counts). The conditional argument corresponds to not using all the N_a^b values seen in Figure 2.9, but concentrating on values with $\bar{N}_c^b$ in an interval around the observed value $\bar{N}_c$ as indicated by two vertical lines. The histogram of these N_a^b values are on the right of Figure 2.10. Now the range of values are from 8 to 41, based on 185 values, so that within this conditional distribution it is unlikely to see values much above 41, indicating that $277 - 41 = 236$ variables are true positives.

Unfortunately, the argument that a very low probability of the event N_a being greater than some number n_0 should imply that it is likely that N_a^T is greater than the combined tail count minus n_0 is not entirely correct. The problem is that when making such a statement we are actually working in the conditional distribution of N_a given the value of $N_a + N_a^T$. (The SAM procedure does not discuss this ambiguity defining a false discovery rate as a ratio between a quantity derived from the distribution of N_a and the observed number of positives, $N_a + N_a^T$. My point is that one should use the conditional distribution of N_a given $N_a + N_a^T$.) To get a feeling for this you can play around with the following toy example. Assume that N_a and N_a^T are independent and both having a negative binomial distribution,

$$N_a \sim \mathrm{NB}(\kappa, p), \quad N_a^T \sim \mathrm{NB}(\kappa_T, p)$$

then we have

$$N_a | N_a + N_a^T = n \sim \mathrm{BetaBinom}(n, \kappa, \kappa_T)$$

You can choose κ and p using the permutation experiment by equalling the mean and variance of N_a to the empirical counterparts from the permutations. Now

choose different values of κ_T and plot the conditional distribution of $N_a|N_a + N_a^T = n$. An initial value to look at can be the value for which $E(N_a) + E(N_a^T) = n$.

2.5 SAM procedure

The SAM-procedure (significance analysis of microarrays) was suggested in the paper Tusher et al. (2001). It is a method that compares the number of variables declared positive to the number obtained from a permutation analysis. As we have seen in section 2.3 the use of the permuation distribution may not always be the best approach (at least when using the ordinary t-statitic).

Using a threshold we can calculate the number N_{data} of test statistics above the threshold. Then we can for each permuted sample calculate the same number and find, for example, the median or the 90 percentile of these numbers among the permutations. A *false discovery rate* is then defined as the permutation percentile (50 or 90 percentile) divided by the number N_{data} declared positive in the data. The threshold is then chosen from these false discovery rates.

A special feature of the SAM-procedure is that the usual t-statistic is replaced by a statistic of the form

$$\frac{\bar{y}_i - \bar{x}_i}{s_i + s_0}.$$

For the data in Tusher et al. (2001) a value of $s_0 = 3.3$ was used, where the range of s_i values goes from 1 to 1000.

2.6 Project

2.6.1 Original article and data

Start by running through the article Spielman et al. (2007) to get a feeling for the data and the main conclusions, in particular the comparison of the two groups CEU and CHB+JPT.

Construct the dataset used in the article from information in the article and from the link to a data repository. This is not easy, but perhaps realistic as to the problems one encounters with real data. Your first result should be two 8793×208 data matrices, one with the expression levels and one where the entries are *A*, *M* or *P*, with *P* meaning that a signal is considered *present*. It is probably best to work in groups doing this work.

2.6.2 Reproducing results

Scale the expression levels for each sample as described in the article. Next use the information in the article to reduce the number of genes to about 4190 (you should also remove all genes starting with *AFFX*). Reduce the number of samples from 208 to 142 as used in the article.

Calculate t-statistic for difference between the two groups CEU and CHB+JPT for each gene. Find of the order 1097 significant genes as in Spielman et al. (2007).

2.6.3 Critique of the results

Read the article Akey et al. (2007). I do not want you to reproduce the results of this article, but you should understand the problem addressed and the result of the new analysis.

Read sections 1, 2 and 4.2 of Allen and Tibshirani (2012). The method in this article is general and as such an alternative to the use of Efron's model discussed above. Note that the conclusion for the data in Spielman et al. (2007) is slightly different from the conclusion in Akey et al. (2007).

2.6.4 Analysis via Efron's model

Make an analysis of the data along the lines in section 2.1. What is the conclusion from the estimation procedure?

Make a permutation analysis as described in section 2.3. Do you want to declare some of the genes differentially expressed between the two groups CEU and CHB+JPT?

2.7 Exercises

Exercise 2.2 (Z-score)
Let $T \sim t(df)$ and define $Z = \Phi^{-1}(t_{\text{cdf}}(T, df))$. Show that $Z \sim N(0,1)$.

Exercise 2.3 (Coefficients of second degree polynomial)
Let $f(x)$ be π_0 times the density of a normal distribution with mean μ and variance σ^2. Let $\log(f(x)) = a + bx + cx^2$ and express π_0, μ and σ^2 in terms of a, b and c.

Exercise 2.4 (Normal approximation on z-scale)
Consider the Kruhøffer data from exercise 1.28. Calculate t-statistics (group 1 with 67 samples and group 2 with 34 samples, nominator of t-statistic is average in group 1 minus average in group 2). Calculate corresponding z-scores (using qnorm(pt(t,df)) when $t \leq 0$ and -qnorm(pt(-t,df)) when $t > 0$). Remove values above ± 9. Make a histogram with narrow bins (bin width 0.2, first bin starting in -8.0 and last bin ending in 8.4) and use Efron's sixth degree polynomial to fit a normal distribution in the center of the data (using at least 50 percent of the data).

Make a figure with the empirical left tail counts in the range from -5 to -3, and insert a curve with the expected numbers from the normal distribution fitted to the center of the data.

Exercise 2.5 (Scaled noncentral *t*)
Consider the statistic

$$T = \frac{\bar{X}_1 - \bar{X}_2 + D}{\sqrt{s^2 \cdot (1/n_1 + 1/n_2)}},$$

where $\bar{X}_1 \sim N(\mu, \sigma^2/n_1)$, $\bar{X}_2 \sim N(\mu, \sigma^2/n_2)$, $D \sim N(\sigma\xi, \sigma^2\psi^2)$ and $s^2 \sim \sigma^2\chi^2(n_1+n_2-2)/(n_1+n_2-2)$, and all stochastic variables are independent. Show that

$$P(T \le t) = \mathrm{qt}(t/\tau, n_1+n_2-2, \mathrm{ncp} = \delta),$$
$$\delta = \frac{\xi}{\sqrt{\psi^2 + 1/n_1 + 1/n_2}}, \quad \tau = \sqrt{1 + \psi^2/(1/n_1 + 1/n_2)}.$$

■

Exercise 2.6 (Scaled noncentral *t* approximation)
Consider the Kruhøffer data from exercise 1.28. Calculate t-statistics (group 1 with 67 samples and group 2 with 34 samples, nominator of t-statistic is average in group 1 minus average in group 2). Remove values above ±10. Make a histogram with narrow bins (bin width 0.2, first bin starting in -9.6 and last bin ending in 10.0) and use Efron's sixth degree polynomial to fit a scaled noncentral t-distribution in the center of the data (using at least 50 percent of the data). You will need to define a function that calculates the sum of squared differences between the sixth degree polynomial and the logarithm of the scaled noncentral t density multiplied by π_0 for the central part of the data range. Then you can minimize this using *nlm*

Make a figure with the empirical left tail counts in the range from -5 to -3, and insert a curve with the expected numbers from the scaled noncentral t-distribution fitted to the center of the data.

Exercise 2.7 (Reproduce figure)
Reproduce the second row of figure 2.3. ■

Exercise 2.8 (Variance of indicator function)
Show that

$$\mathrm{Var}(1(T \le t)) = p(1-p), \quad p = P(T \le t),$$

and

$$\mathrm{Var}(1(T_1 \le t_1) \cdot 1(T_2 \le t_2)) = p_2 - p_{11}p_{12}, \quad p_{1j} = P(T_j \le t_j), \; p_2 = P(T_1 \le t_1, T_2 \le t_2).$$

Let I_i, $i = 1, \ldots, m$, be random indicator functions. Show that

$$\mathrm{Var}(\sum_i I_i) = \sum_i \mathrm{Var}(I_i) + 2 \sum_{i,j:j>i} \mathrm{Cov}(I_i, I_j)$$

■

Exercise 2.9 (Bivariate normal)
Let (X_1, X_2) be bivariate normal with mean zero, both having variance 1, and the correlation being $\rho = 0.2$. Calculate the probability

$$P(X_1 \le -2, X_2 \le -3).$$

Calculate the probability of the set

$$x_1 \le \begin{cases} -3 & -3 \le x_2 \le -2, \\ -2 & x_2 < -3, \\ -\infty & x_2 > -2. \end{cases}$$

■

Exercise 2.10 (Negative binomial)
The negative binomial distribution $\mathrm{NB}(\kappa, p)$, $\kappa > 0$, $0 \le p \le 1$, has point probabilities

$$P(X = k) = \frac{(k+\kappa-1)(k+\kappa-2)\cdots\kappa}{k!}(1-p)^k p^\kappa, \quad k = 0, 1, \ldots.$$

Assume that $N_1 \sim \mathrm{NB}(\kappa_1, p)$, $N_2 \sim \mathrm{NB}(\kappa_2, p)$, and that the two random variables are independent. Show that

$$N_1 + N_2 \sim \mathrm{NB}(\kappa_1 + \kappa_2, p).$$

(If you do not know about moment generating functions you should skip this question.)

Show that the conditional distribution of N_1, given that $N_1 + N_2 = n$, is a beta-binomial distribution,

$$N_1 | (N_1 + N_2 = n) \sim \mathrm{BeatBin}(n, \kappa_1, \kappa_2).$$

■

Exercise 2.11 (Rare event simulation)
Consider a direct simulation of $p_2 = P(X_1 \le x, X_2 \le x)$ in order to calculate $c = p_2 - p_1^2$, $p_1 = P(X_1 \le x) = P(X_2 \le x)$, in a situation where $p_1 = 0.001349898$ and $p_2 = 1.143459e-05$.

How many samples do you need to simulate in order that your estimate of c has relative standard deviation 0.1 ?

Exercise 2.12 (Correlation of empirical variances)
Let (X_i, Y_i) be two dimensional normally distributed with the variance matrix having one at the diagonal and ρ off the diagonal. Find the correlation of $\bar{X}$ and $\bar{Y}$.

Assume now that $E(X_i) = E(Y_i) = 0$ and let $\mathrm{SS}_x = \sum_i X_i^2$ and $\mathrm{SS}_y = \sum_i Y_i^2$. Can you find the correlation of SS_x and SS_y ?

Exercise 2.13 (Simulating correlated t-values)
Suggest alternative ways of simulating correlated t-values instead of the simple block structure of exercise 1.10.

Exercise 2.14 (Scaled noncentral t permutations)
This is a continuation of exercise 2.6. Use the parameters there to perform a permutation experiment as described in Sections 2.3 and 2.4.

Choose a threshold and look at the permutation distribution of the tail count. Compare with the conditional permutation distribution, where you condition on the central count (use an interval ± 50 around the observed value of the central count).

Define a "false discovery rate" as the 90% quantile of the distribution of the tail count from the permutations divided by the observed tail count. Do this for the two permutation distribution discussed above.

Repeat the calculations for a number of different threshold, and consider both lower tail counts and upper tail counts.

Exercise 2.15 (Beta-binomial)
Consider the Leukemia dataset as in Section 2.4 and the toy model with two negative binomial distributions. Consider the threshold -4.

Fit the negative binomial to the distribution of tail counts from the conditional permutation approach.

Find κ_T so that the mean of the sum of the two negative binomials equals the observed tail count, and calculate the conditional distribution of N_a given the value of the sum $N_a + N_a^T$.

Find κ_T so that the mean of the sum of the two negative binomials plus two times the standard deviation of the sum equals the observed tail count, and calculate the conditional distribution of N_a given the value of the sum $N_a + N_a^T$.

Exercise 2.16 (Open ended problem)
Consider combining the permutation method of Sections 2.3 and 2.4 with some of the methods of Chapter 1.

Exercise 2.17 (BYO)
Make your own exercise.

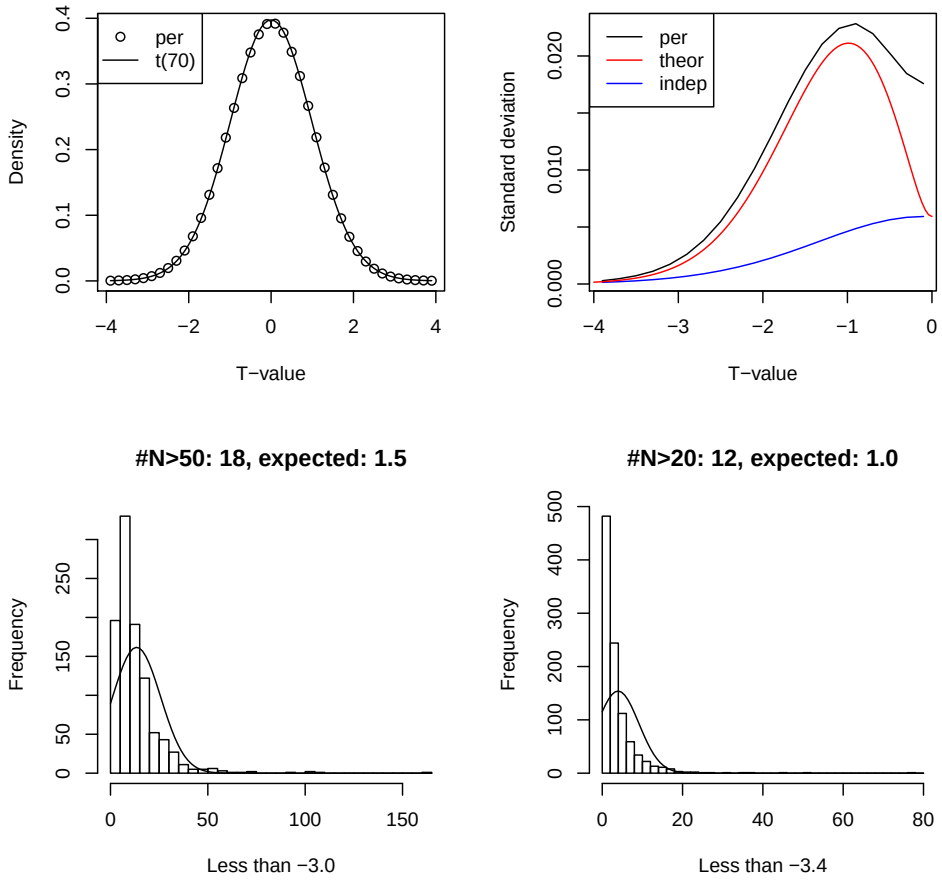

Figure 2.7: The figure is based on 1000 permutations of the leukemia dataset. The topleft shows the average histogram values, where each histogram is based on the 7128 t-values from a permuted sample, and the density of a t-distribution. The topright gives the standard deviation, from the 1000 permutations, of the fraction of t-values below a given value, together with a theoretical value based on the t-distribution and the observed correlation in the original dataset (red curve), and the theoretical value if the t-values were independent (blue curve). The bottom row gives the empirical distribution from the 1000 permutations for the number of t-values below a threshold, together with the density of a normal distribution with mean and variance as in the empirical distribution. The heading gives the number of instances in the 1000 permutations where the number of t-values below the threshold is above a certain value together with the expected number from a normal distribution.

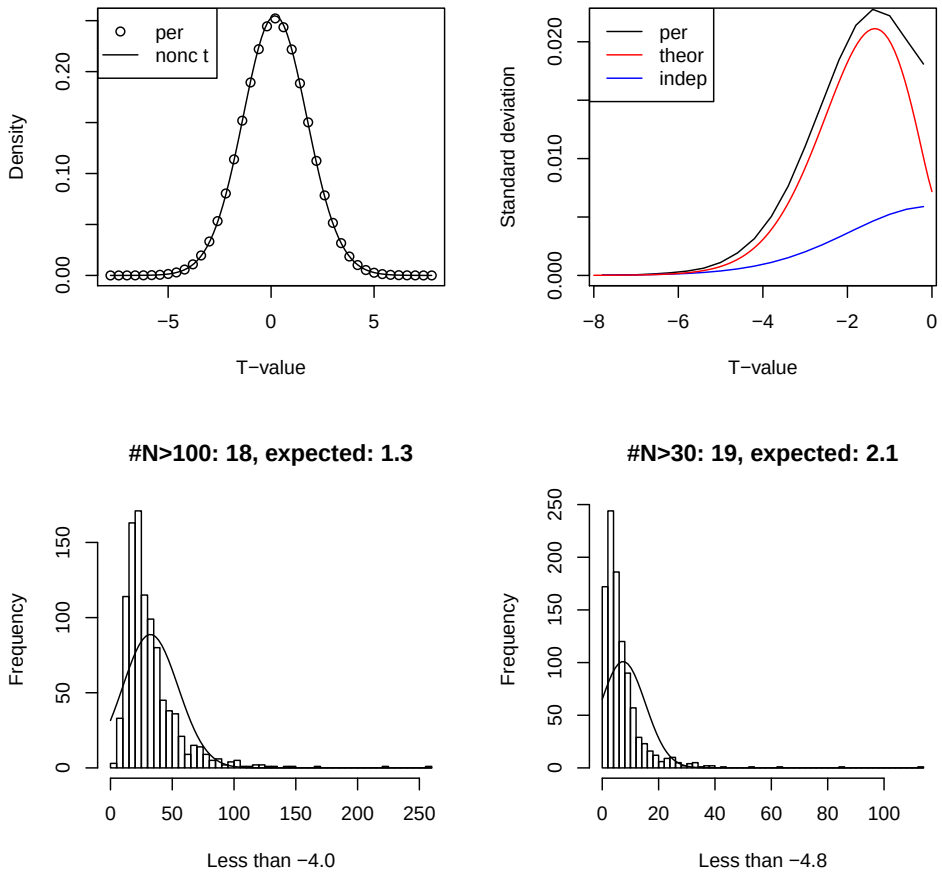

Figure 2.8: The figure is based on 1000 permutations of the leukemia dataset. The topleft shows the average histogram values, where each histogram is based on the 7128 scaled and displaced t-values from a permuted sample, and the density of a scaled noncentral t-distribution. The topright gives the standard deviation, from the 1000 permutations, of the fraction of t-values below a given value, together with a theoretical value based on the scaled noncentral t-distribution and the observed correlation in the original dataset (red curve), and the theoretical value if the t-values were independent (blue curve). The bottom row gives the empirical distribution from the 1000 permutations for the number of scaled and displaced t-values below a threshold, together with the density of a normal distribution with mean and variance as in the empirical distribution. The heading gives the number of instances in the 1000 permutations where the number of t-values below the threshold is above a certain value together with the expected number from a normal distribution.

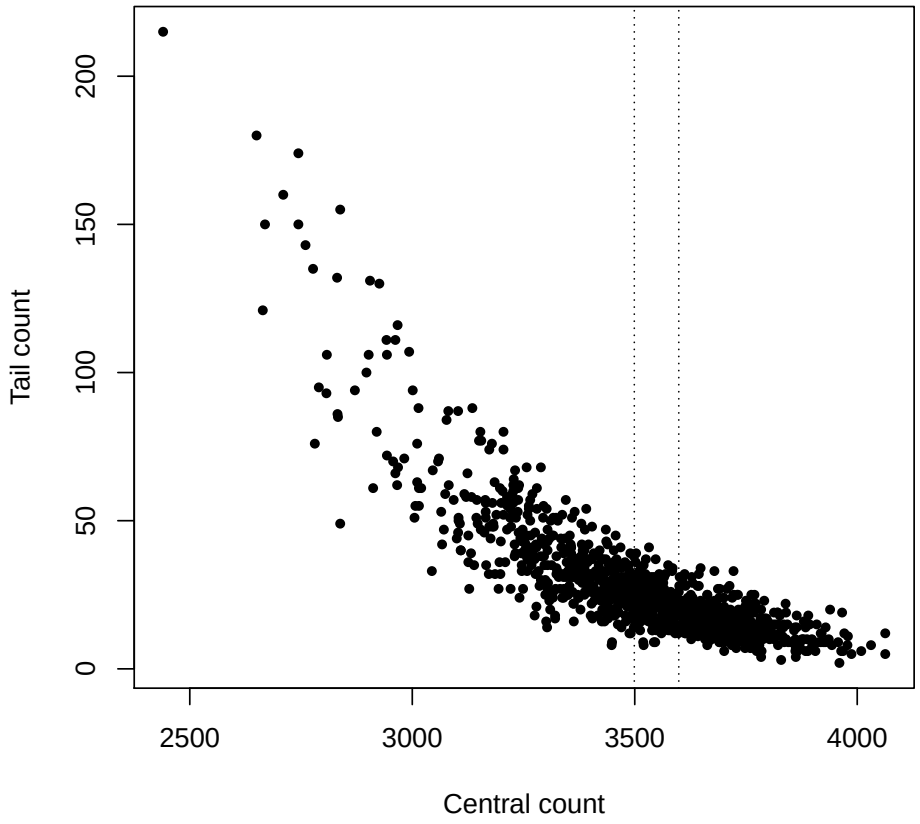

Figure 2.9: .

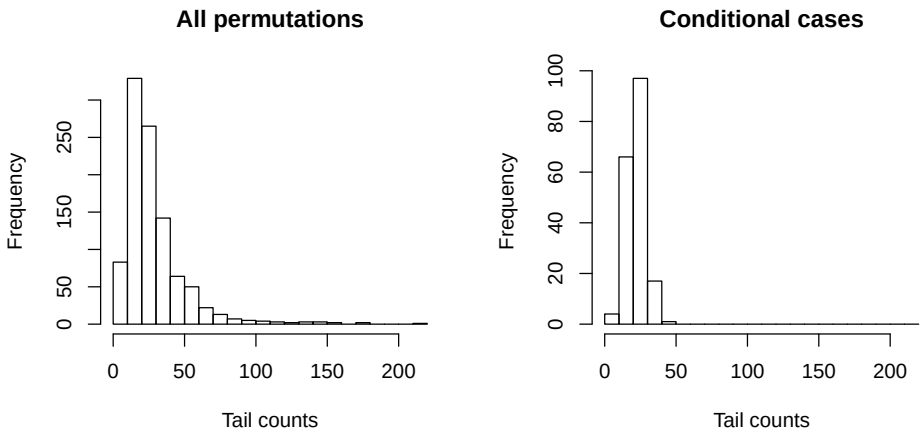

Figure 2.10: .

CHAPTER

3 Classification

We start this chapter with a discussion of classification into two (or more) groups when the distribution within each group is known. This situation is typically not relevant in practice, but it serves as motivation for the general situation where parameters of the distributions are unknown. Sometimes one uses the expression "oracle property" meaning that a procedure in some limit achieves the same properties as in the case where parameters are known.

A classifier is defined through a rule $\xi(x)$, stating that observation x is classified to the group $\xi(x) \in \{1,\dots,k\}$. The probability of classifying an individual from group i to group j is

$$p_{ij}(\xi) = \int 1(\xi(x) = j) f_i(x)\mu(dx),$$

where $f_i(x)$ is the density with respect to a measure μ for the distribution in group i, $P(X \in A) = \int_A f(x)\mu(dx)$. In particular $p_{ii}(\xi)$ is the probability of correct classification for observations from group i. For the two group case $p_{12}(\xi)$ and $p_{21}(\xi)$ are the probabilities of making a wrong classification. Note, that these probabilities refer to the distribution of a new observation x, and not the distribution of data that enters the construction of the classifier ξ.

3.1 Maksimum likelihood classifier

Consider k populations with densitites $f_i(x)$. Let a_i, $i = 1,\dots,k$ be positive constants. The (generalized) maksimum likelihood classifier is the rule

$$\xi_{\mathrm{ML}}(z) = \operatorname*{argmax}_i \{a_i f_i(z)\}.$$

When $a_i = 1$ for all i this is the usual maksimum likelihood classifier. The constants a_i are introduced to allow different emphasis on the probability of correct classification for the various groups. When there are two groups, $k = 2$, we have

$$\xi_{\mathrm{ML}}(z) = \begin{cases} 1 & \text{when } f_1(z)/f_2(z) \geq (a_2/a_1), \\ 2 & \text{when } f_1(z)/f_2(z) < (a_2/a_1). \end{cases}$$

We also write this as $\xi_{\mathrm{ML}}(z) = G\big(f_1(z)/f_2(z) - (a_2/a_1)\big)$ where G throughout this chapter is the function

$$G(u) = \begin{cases} 1 & u \geq 0, \\ 2 & u < 0. \end{cases}$$

The maksimum likelihood classifier is *admissible* as described in the following proposition. The result of the proposition is closely related to the Neymann-Pearson lemma that may be familiar to some of you. The two next propositions closely follow Theorems 11.2.2 and 11.2.3 in Mardia et al. (1979).

Proposition 3.1
There is no rule $\xi^*(x)$ such that the probabilities of correct classification $p_{ii}(\xi^*)$ satisfy

$$p_{jj}(\xi^*) > p_{jj}(\xi_{\mathrm{ML}}) \text{ for some } j, \text{ and } p_{ii}(\xi^*) \geq p_{ii}(\xi_{\mathrm{ML}}),\ i = 1,\ldots,k. \tag{3.1}$$

Proof. Assume that there exists ξ^* with the properties (3.1). Then $\sum_i a_i p_{ii}(\xi^*) > \sum_i a_i p_{ii}(\xi_{\mathrm{ML}})$. However,

$$\begin{aligned}
\sum_i a_i p_{ii}(\xi^*) &= \sum_i \int 1(\xi^*(x) = i) a_i f_i(x) \mu(dx) \leq \sum_i \int 1(\xi^*(x) = i) \max_j \{a_j f_j(x)\} \mu(dx) \\
&= \int [\sum_i 1(\xi^*(x) = i)] \max_j \{a_j f_j(x)\} \mu(dx) = \int \max_j \{a_j f_j(x)\} \mu(dx) \\
&= \sum_i \int 1\big(\xi_{\mathrm{ML}}(x) = i\big) \max_j \{a_j f_j(x)\} \mu(dx) = \sum_i \int 1\big(\xi_{\mathrm{ML}}(x) = i\big) a_i f_i(x) \mu(dx) \\
&= \sum_i a_i p_{ii}(\xi_{\mathrm{ML}}),
\end{aligned} \tag{3.2}$$

which is a contradiction. □

In a Bayesian setting let π_i, $i = 1,\ldots,k$, be prior probabilities for belonging to the k groups. Let I be a random group label with $P(I = i) = \pi_i$. The conditional distribution of X given $I = i$ has density $f_i(x)$. The *Bayes classifier*, denoted by ξ_B, selects the group i with the largest posterior probability $P(I = i|X)$. We can write this as

$$\xi_B(z) = \operatorname*{argmax}_i \{P(I = i|z)\} = \operatorname*{argmax}_i \{\frac{\pi_i f_i(z)}{\sum_j \pi_j f_j(z)}\} = \operatorname*{argmax}_i \{\pi_i f_i(z)\},$$

Which shows that the Bayes classifier is the maksimum likelihood classifier with $a_i = \pi_i$. Let us also note that for the case of two groups the posterior class probabilities are given as

$$P(I = 1|x) = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)} = \frac{1}{1 + \exp\big(\log(f_2(x)/f_1(x)) + \log(\pi_2/\pi_1)\big)},$$

and $P(I = 2|x) = 1 - P(I = 1|x)$. Below we will see that $\log(f_2(x)/f_1(x))$ takes a particular simple form for normal densitites with the same variance.

Let I is the random group label, with distribution given by the prior probabilities, and let X given $I = i$ have density $f_i(x)$. Then $p_{ii}(\xi_B) = P(\xi_B(X) = i|I = i)$ and the *Bayes risk* is defined as

$$R(\xi_B) = P(\xi_B(X) \neq I) = \sum_i \pi_i P(\xi_B(X) \neq i|I = i) = \sum_i (1 - p_{ii}(\xi_B)).$$

The Bayes risk is simply the probability of misclassification under the joint distribution of (I, X). We call $P(\xi_B(X) = I)$ the Bayes probability of correct classification.

Proposition 3.2
No other classification rule has larger posterior probability of correct classification than the Bayes rule.

Proof. For any other classification rule ξ^* it is shown in (3.2), with a_i replaced by π_i, that

$$\sum_i \pi_i p_{ii}(\xi^*) \leq \sum_i \pi_i p_{ii}(\xi_B),$$

which is the statement of the proposition. □

For the case of normally distributed data the classification probabilities for a set of socalled linear rules can be calculated explicitly. In the following proposition we look at the classification probablities for a given classifier, where the classifier itself can be either fixed or random based on a dataset. This means that the stated probabilities are with respect to the probabilities for a new observation that we want to classify.

Proposition 3.3 (Classification probabilities)
Let a and b be m-dimensional vectors and c a number. The classifier

$$\xi_L(z) = G(a^T(z-b)+c)$$

is called a linear rule. *Let $X \sim N_m(\mu_1, \Sigma)$ for observations from group 1 and $X \sim N_m(\mu_2, \Sigma)$ for observations from group 2. Then the probabilities of correct classification are*

$$p_{11} = \Phi\Big(\frac{\gamma^-}{2\tau}\Big), \quad p_{22} = \Phi\Big(\frac{\gamma^+}{2\tau}\Big), \tag{3.3}$$

with

$$\gamma^- = a^T(\Delta - d) + 2c, \quad \gamma^+ = a^T(\Delta + d) - 2c, \quad \tau^2 = a^T\Sigma a,$$

where $d = 2b - (\mu_1 + \mu_2)$ and $\Delta = \mu_1 - \mu_2$.

Proof. When $X \sim N_m(\mu_1, \Sigma)$ we have $a^{\mathrm{T}}(X-b) \sim N\big(a^{\mathrm{T}}(\mu_1 - b) + c, a^{\mathrm{T}}\Sigma a\big)$. We therefore have

$$P(a^{\mathrm{T}}(X-b)+c>0) = \Phi\Big(\frac{a^{\mathrm{T}}(\mu_1-b)+c}{\sqrt{a^{\mathrm{T}}\Sigma a}}\Big) = \Phi\Big(\frac{a^{\mathrm{T}}(\mu_1-\mu_2-(2b-\mu_1-\mu_2))+2c}{2\sqrt{a^{\mathrm{T}}\Sigma a}}\Big).$$

Similarly, with $X \sim N_m(\mu_2, \Sigma)$ we find

$$P(a^{\mathrm{T}}(X-b)+c<0) = \Phi\Big(-\frac{a^{\mathrm{T}}(\mu_2-b)-c}{\sqrt{a^{\mathrm{T}}\Sigma a}}\Big) = \Phi\Big(\frac{a^{\mathrm{T}}(\mu_1-\mu_2+(2b-\mu_1-\mu_2))-2c}{2\sqrt{a^{\mathrm{T}}\Sigma a}}\Big),$$

proving the proposition. □

Example 3.4 (One-dimensional normal distributions)
Suppose we have two groups with observations in $\mathbb{R}$ and with distributions $N(\mu_1,\sigma^2)$ and $N(\mu_2,\sigma^2)$. Then

$$\frac{f_1(z)}{f_2(z)} = \frac{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(z-\mu_1)^2\}}{\frac{1}{\sqrt{2\pi\sigma^2}}\exp\{-\frac{1}{2\sigma^2}(z-\mu_2)^2\}} = \exp\Big\{\frac{\mu_1-\mu_2}{\sigma^2}\big[z-\frac{1}{2}(\mu_1+\mu_2)\big]\Big\}. \tag{3.4}$$

Consider the maksimum likelihood classifier with equal weights $a_1 = a_2$:

$$\xi(z) = G\Big(\frac{\mu_1-\mu_2}{\sigma^2}\big[z-\frac{1}{2}(\mu_1+\mu_2)\big]\Big)$$

For this rule z is classified to group 1 when z is closer to μ_1 than to μ_2, and vice versa. Since

$$\frac{\mu_1-\mu_2}{\sigma^2}\big[X-\frac{1}{2}(\mu_1+\mu_2)\big] \sim \begin{cases} N\big(\frac{(\mu_1-\mu_2)^2}{2\sigma^2}, \frac{(\mu_1-\mu_2)^2}{\sigma^2}\big), & X \text{ from group 1,} \\ N\big(-\frac{(\mu_1-\mu_2)^2}{2\sigma^2}, \frac{(\mu_1-\mu_2)^2}{\sigma^2}\big), & X \text{ from group 2,} \end{cases}$$

we find the probabilities of correct classification to be

$$p_{11}(\xi) = p_{22}(\xi) = \Phi\Big(\frac{|\mu_1-\mu_2|}{2\sigma}\Big). \tag{3.5}$$

In a Bayesian setting the posterior probability of belonging to group 1 given an observation x becomes

$$P(I=1|x) = \frac{1}{1+\exp\big(\frac{\mu_2-\mu_1}{\sigma^2}\big[x-\frac{1}{2}(\mu_1+\mu_2)\big]+\log(\pi_2/\pi_1)\big)},$$

which is the logistic regression model. ∎

Example 3.5 (Maximum likelihood independence classifier)
Suppose we have two groups with observations in $\mathbb{R}^m$ and with distributions $N_m(\mu_1, D(\sigma^2))$ and $N_m(\mu_2, D(\sigma^2))$, where μ_1, μ_2 and σ^2 are m-dimensional vectors and $D(v)$ is a diagonal matrix with the diagonal given by the vector v (this says that the coordinates of the random vector are independent normally distributed, the j'th coordinate having variance σ_j^2 and mean μ_{1j} or μ_{2j}). Then from (3.4) we find

$$\frac{f_1(z)}{f_2(z)} = \exp\Big\{\sum_{j=1}^{m}\frac{\mu_{1j}-\mu_{2j}}{\sigma_j^2}\big[z_j-\frac{1}{2}(\mu_{1j}+\mu_{2j})\big]\Big\}.$$

For the maksimum likelihood classifier with equal weights $a_1 = a_2$ we find the probabilities of correct classification from (3.3),

$$p_{11} = p_{22} = \Phi\Big(\frac{1}{2}\sqrt{\sum_{j=1}^{m}\frac{(\mu_{1j}-\mu_{2j})^2}{\sigma_j^2}}\Big).$$

∎

Example 3.6 (Multidimensional normal distributions)
Suppose we have two groups with observations in $\mathbb{R}^m$ and distributions $N_m(\mu_1,\Sigma)$ and $N_m(\mu_2,\Sigma)$, where μ_1 and μ_2 are m-dimensional vectors and Σ is a positive definite $m\times m$ matrix. Define $\Delta=\mu_1-\mu_2$. Then instead of (3.4) we find

$$\frac{f_1(z)}{f_2(z)}=\exp\Big\{\Delta^{\mathrm{T}}\Sigma^{-1}\big[z-\frac{1}{2}(\mu_1+\mu_2)\big]\Big\}.$$

For the maksimum likelihood classifier with equal weights $a_1=a_2$ we find the probabilities of correct classification from (3.3) and get instead of (3.5)

$$p_{11}(\xi_{\mathrm{ML}})=p_{22}(\xi_{\mathrm{ML}})=\Phi\Big(\frac{1}{2}\sqrt{\Delta^{\mathrm{T}}\Sigma^{-1}\Delta}\Big), \tag{3.6}$$

since $\gamma^-=\gamma^+=\tau^2=\Delta^{\mathrm{T}}\Sigma^{-1}\Delta$. ■

Exercise 3.7 (One dimensional normal distribution)
Consider Example 3.4. Make a table with the probability of correct classification for the cases $|\mu_1-\mu_2|=1,2,4$ and $\sigma=0.1,1,2,100$. ■

Exercise 3.8 (One dimensional normal distribution)
Consider Example 3.4 with $\sigma=1$ and $\mu_2-\mu_1=1$. Make a figure with the Bayes probability of correct classification as a funtion of $0<\pi_1<1$.

Exercise 3.9 (Von Mises distribution)
The von Mises distribution is a distribution on the unit circle. The density is

$$f(x;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}\exp\big\{\kappa\cos(x-\mu)\big\},\quad 0\le x\le 2\pi,$$

where I_0 is the modfied Bessel funktion of order zero. The parameter $\mu\in[0,2\pi]$ is a location and κ is a scale parameter.

Find the maximum likelihood classifier when the two distributions are von Mises with parameters (μ_1,κ) and (μ_2,κ). Give a geometrical description of the classifier.

Make a table with the probability of correct classification for the cases $|\mu_1-\mu_2|=\frac{\pi}{4},\frac{\pi}{2},\pi$ and $\kappa=0,1,5,10$. ■

Exercise 3.10 (Independent normally distributed variables)
Consider Example 3.5. Let $\delta_j=(\mu_{1j}-\mu_{2j})/\sigma_j$ be the scaled differential expression. Assume that δ_j is 1 for k variables and zero for the remaining variables. Make a plot of the probability of correct classification as a function of k. ■

Exercise 3.11 (Dependent normally distributed variables)
Consider the probability of correct classification in (3.6). Let $\Delta=(0.5,1,0,\ldots,0)^{\mathrm{T}}$, $\Sigma_{ii}=1$ and for $i\neq j$ the correlation $\Sigma_{ij}=0$ except

$$\begin{aligned}\text{case I:}\quad &\Sigma_{12}=\Sigma_{21}=\rho,\ -1<\rho<1;\\ \text{case II:}\quad &\Sigma_{13}=\Sigma_{31}=\rho,\ -1<\rho<1.\end{aligned}$$

For both cases make a figure with the probability of correct classification as a function of ρ.

In the same setting as above consider also the linear rule $\xi(z) = G(\Delta^{\mathrm{T}}[z-(\mu_1+\mu_2)/2])$, $\Delta = \mu_1 - \mu_2$ (from the maksimum likihood independence classifier since $\sigma_j^2 = 1$), with probability of correct classification given as $\Phi\big(\frac{1}{2}\Delta^{\mathrm{T}}\Delta/\sqrt{\Delta^{\mathrm{T}}\Sigma\Delta}\big)$. Insert in the figure from before a line with this probability. ■

3.2 Fisher's rule

We now turn to the situation where parameters of the distributions are replaced by estimates. We shall often be using the folliwing setup. Let the m-dimensional independent observations from group 1 be $x_1,\ldots,x_{n_1}$ and those from group 2 be $y_1,\ldots,y_{n_2}$ with

$$\begin{aligned} &x_i \sim N_m(\mu_1,\Sigma) \text{ and } y_i \sim N_m(\mu_2,\Sigma) \\ &\mu_1 = (\mu_{11},\ldots,\mu_{1m}),\ \mu_2 = (\mu_{21},\ldots,\mu_{2m}),\ \Sigma_{jj} = \sigma_j^2 \\ &n = n_1+n_2,\ a_n = \frac{1}{n_1}+\frac{1}{n_2},\ \omega_n = \frac{n_1-n_2}{n}. \end{aligned} \tag{3.7}$$

Sometimes we will also use the assumption that n_1 and n_2 are of the same order: there exists a constant $c_s > 0$ such that

$$\frac{n_1}{n} \geq c_s \quad \text{and} \quad \frac{n_2}{n} \geq c_s. \tag{3.8}$$

Furthermore we define the *differential expression* for variable j as

$$\Delta_j = \mu_{1j} - \mu_{2j},\ \ \Delta = (\Delta_1,\ldots,\Delta_m) = \mu_1 - \mu_2, \tag{3.9}$$

and the scaled differential expression as

$$\delta_j = (\mu_{1j}-\mu_{2j})/\sigma_j,\ \ j = 1,\ldots,m.$$

The usual estimates in model (3.7) are

$$\hat{\mu}_1 = \bar{x},\ \ \hat{\mu}_2 = \bar{y},\ \ \hat{\Sigma} = \frac{SSD}{n-2} \text{ with } SSD = \sum_{i=1}^{n_1}(x_i-\bar{x})(x_i-\bar{x})^{\mathrm{T}} + \sum_{i=1}^{n_2}(y_i-\bar{y})(y_i-\bar{y})^{\mathrm{T}}.$$

The sample version of the rule from Example 3.6 then becomes

$$\xi_F(z) = G\Big((\bar{x}-\bar{y})^{\mathrm{T}}\hat{\Sigma}^{-1}\big[z-\frac{1}{2}(\bar{x}+\bar{y})\big]\Big). \tag{3.10}$$

This is known as Fisher's rule. Fisher (1936) did not derive this rule as a sample maximum likelihood classifier. Instead he started from the concept of a linear classifier: a *linear rule* is based on $a^{\mathrm{T}}z$ for some vector a, and assigns z to group 1 if $|a^{\mathrm{T}}z - a^{\mathrm{T}}\bar{x}| \leq |a^{\mathrm{T}}z - a^{\mathrm{T}}\bar{y}|$ and otherwise to group 2. This is the same as

$$\xi(z) = G\big(a^{\mathrm{T}}(z-\frac{1}{2}(\bar{x}+\bar{y})\big). \tag{3.11}$$

Fisher then took a to maximize the between group variation divided by the within group variation for $a^{\mathrm{T}}z$. This can be written as

$$a = \operatorname*{argmax}_{v} \frac{v^{\mathrm{T}}(\hat{\Delta}\hat{\Delta}^{\mathrm{T}})v}{v^{\mathrm{T}}SSD\,v}$$

with $\hat{\Delta} = \bar{x} - \bar{y}$. Now write v as $v = SSD^{-1/2}w$ (the square root of a positive semidefinite (and symmetric) matrix Σ is a positive semidefinite matrix B sucn that $BB = \Sigma$). We then want to maximize

$$\frac{w^{\mathrm{T}}SSD^{-1/2}\hat{\Delta}\hat{\Delta}^{\mathrm{T}}SSD^{-1/2}w}{w^{\mathrm{T}}SSD^{-1/2}SSDSSD^{-1/2}w} = \Big|\Big(\frac{w}{|w|}\Big)^{\mathrm{T}}SSD^{-1/2}\hat{\Delta}\Big|^2.$$

This shows that we can use $w = SSD^{-1/2}\hat{\Delta}$ and therefore $a = SSD^{-1}\hat{\Delta}$. The rule (3.11) therefore turns into (3.10). The above derivation indicate that the rule can be useful also when data are not normally distributed.

Define $\hat{a} = \hat{\Sigma}^{-1}\hat{\Delta}$. The probabilities of correct classification for Fisher's rule are from (3.3)

$$p_{11}(\xi_F) = \Phi\Big(\frac{\gamma_F^-}{2\tau_F}\Big),\quad p_{22}(\xi_F) = \Phi\Big(\frac{\gamma_F^+}{2\tau_F}\Big),\quad \gamma_F^{\pm} = \hat{a}^{\mathrm{T}}(\Delta \pm d),\quad \tau_F^2 = \hat{a}^{\mathrm{T}}\Sigma\hat{a}, \tag{3.12}$$

where $d = \bar{x} - \mu_1 + \bar{y} - \mu_2$. The misclassification probabilities $p_{12}(\xi_F) = 1 - p_{11}(\xi_F)$ and $p_{21}(\xi_F) = 1 - p_{22}(\xi_F)$ are obtained on replacing Φ by $\bar{\Phi}$ given by $\bar{\Phi}(z) = 1 - \Phi(z)$.

The customary asymptotic setting is to have the dimension m fixed and letting the sample size tend to infinity. If n_1 and n_2 tends to infinity we have $a_n = 1/n_1 + 1/n_2$ tends to zero, and

$$d \sim N_m\Big(0, a_n\Sigma\Big) \xrightarrow{P} 0,\quad \hat{\Delta} \sim N_m\Big(\Delta, a_n\Sigma\Big) \xrightarrow{P} \Delta,$$
$$\hat{\Sigma} \sim \text{Wishart}_m\big(\Sigma, n-2\big)/(n-2) \xrightarrow{P} \Sigma,$$

where the arrows indicate convergence in probability. From (3.12) and (3.6) we therefore find

$$p_{11}(\xi_F) \xrightarrow{P} p_{11}(\xi_{\mathrm{ML}}) \text{ and } p_{22}(\xi_F) \xrightarrow{P} p_{22}(\xi_{\mathrm{ML}}). \tag{3.13}$$

The above convergences in probability can be seen as follows. If $V \sim N_m(0, a_n\Sigma)$ then $V = \sqrt{a_n}V_0$ with $V_0 \sim N_m(0,\Sigma)$, and the convergence follows from $a_n \to 0$. A $\text{Wishart}_m(\Sigma, k)$ distribution is defined as the distribution of $U_1U_1^{\mathrm{T}} + \cdots + U_kU_k^{\mathrm{T}}$ with $U_1, \ldots, U_k$ independent and $U_i \sim N_m(0,\Sigma)$. When we divide a $\text{Wishart}_m(\Sigma, k)$ by k we therefore have an average and the law of large numbers give the convergence in probability (simply because the variance tends to zero).

Being an asymptotic result we cannot know how good the result in (3.13) is for a finite n and a particular parameter value, However, we can improve on the result by showing uniform convergence over a suitable parameter set. Let

$$\Theta(c) = \big\{(\mu_1, \mu_2, \Sigma) : \Delta^{\mathrm{T}}\Sigma^{-1}\Delta \geq c^2\big\},$$

which is the set where the length of the normalized difference $\Sigma^{-1/2}\Delta$ is bounded below by c.

Proposition 3.12
When the dimension m is fixed the convergence of the error probability (or, equivalently, the probability of correct classification) is uniform on Θ as n_1 and n_2 tends to infinity.

Proof. Let $a_n = 1/n_1 + 1/n_2$ and write

$$u = \Sigma^{-1/2}(\hat{\Delta} - \Delta) \sim N_m(0, a_n I), \quad v = \Sigma^{-1/2} d \sim N_m(0, a_n I),$$
$$w = \Sigma^{1/2}\hat{\Sigma}^{-1}\Sigma^{1/2} - I \sim \Big(\frac{\text{Wishart}_m(I, n-2)}{n-2}\Big)^{-1} - I.$$

Here u, v and w tend to zero in probability as n_1 and n_2 tends to infinity, and since the distributions do not depend on the parameters we have uniform convergence on Θ. In particular we can state the convergence as follows. For any $\epsilon > 0$ the probability that

$$|u_i| \le \epsilon \ \forall i, \ \ |v_i| \le \epsilon \ \forall i, \ \ |w_{ij}| \le \epsilon \ \forall i, j, \tag{3.14}$$

tends to one as n_1 and n_2 tends to infinity.

Consider the first term in (3.12) and rewrite the argument of Φ as

$$\begin{aligned}\frac{\hat{\Delta}^{\mathrm{T}}\hat{\Sigma}^{-1}(\Delta - d)}{2\sqrt{\hat{\Delta}^{\mathrm{T}}\hat{\Sigma}^{-1}\Sigma\hat{\Sigma}^{-1}\hat{\Delta}}} &= \frac{(\Sigma^{-1/2}\Delta + u)^{\mathrm{T}}(I+w)(\Sigma^{-1/2}\Delta - v)}{2\big[(\Sigma^{-1/2}\Delta + u)^{\mathrm{T}}(I+w)(I+w)(\Sigma^{-1/2}\Delta + u)\big]^{1/2}} \\ &= \frac{r}{2}\frac{(e + \frac{u}{r})^{\mathrm{T}}(I+w)(e - \frac{v}{r})}{\big[(e + \frac{u}{r})^{\mathrm{T}}(I+w)(I+w)(e + \frac{u}{r})\big]^{1/2}},\end{aligned}$$

where $r = |\Sigma^{-1/2}\Delta|$ and $e = \Sigma^{-1/2}\Delta/r$ is a unit vector. We first consider the nominator

$$a = \big(e + \frac{u}{r}\big)^{\mathrm{T}}(I+w)\big(e - \frac{v}{r}\big) = e^{\mathrm{T}}e - e^{\mathrm{T}}\frac{v}{r} + \big(\frac{u}{r}\big)^{\mathrm{T}}\big(e - \frac{v}{r}\big) + \big(e + \frac{u}{r}\big)^{\mathrm{T}} w \big(e - \frac{v}{r}\big).$$

Using the previous bounds (3.14) we have

$$\begin{aligned}&\big|e_i + \frac{u_i}{r}\big| \le 1 + \frac{\epsilon}{c}, \ \ \big|e_i - \frac{v_i}{r}\big| \le 1 + \frac{\epsilon}{c}, \ \ \big|e^{\mathrm{T}}\frac{v}{r}\big| \le \frac{m}{c}\epsilon, \\ &\big|\frac{u^{\mathrm{T}}}{r}\big(e - \frac{v}{r}\big)\big| \le \frac{m}{c}\big(1 + \frac{\epsilon}{c}\big)\epsilon, \ \ \big|\big(e + \frac{u}{r}\big)^{\mathrm{T}} w \big(e - \frac{v}{r}\big)\big| \le m^2\big(1 + \frac{\epsilon}{c}\big)^2\epsilon.\end{aligned}$$

Combining all these bounds we get that there exists a constant C_1 such that

$$|a - 1| \le C_1\epsilon.$$

Letting b be the term inside the square root in the denominator we find in a similar way that there exists a constant C_2 such that

$$|b - 1| \le C_2\epsilon.$$

If now ϵ is so small that $C_2\epsilon \le \frac{1}{2}$, say, there exists a constant C_3 such that

$$\Big|\frac{a}{\sqrt{b}} - 1\Big| \le C_3\epsilon.$$

Writing $a/\sqrt{b} = 1+\zeta$, with $|\zeta| \le C_3\epsilon$, we can now combine this result with one of the results in Appendix A.2 to get

$$p_{11}(\xi_F) - p_{11}(\xi_{\mathrm{ML}}) = \Phi\big(\frac{r}{2}(1+\zeta)\big) - \Phi\big(\frac{r}{2}\big) \le \frac{|\zeta|}{4} \le \frac{C_3}{4}\epsilon.$$

Since we have this bound with a probability tending to one, for any ϵ, we have convergence in probability uniformly on Θ. □

Exercise 3.13 (Fishers rule, $m < n-2$)
In this exercise you should construct an R programme for simulating the probability of correct classification of Fishers rule when $m < (n-2)$. If we transform by $\Sigma^{-1/2}$, as in the proof of Proposition 3.12, you will need to simulate

$$x = \Sigma^{-1/2}(\bar{x} - \mu_1) \sim N_m(0, I)/\sqrt{n_1}, \qquad y = \Sigma^{-1/2}(\bar{y} - \mu_2) \sim N_m(0, I)/\sqrt{n_2},$$
$$W = \Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} \sim \mathrm{Wishart}_m(I, n-2)/(n-2).$$

The Wishart distribution can be simulated by `rWishart` in R. Define the normalized difference $\delta_\Sigma = \Sigma^{-1/2}\Delta$. The probabilities of correct classification are $\Phi\big(\gamma^\pm/(2\tau)\big)$ with

$$\frac{\gamma^\pm}{2\tau} = \frac{(\delta_\Sigma + x - y)^{\mathrm{T}} W^{-1}\{\delta_\Sigma \pm (x+y)\}}{2\sqrt{(\delta_\Sigma + x - y)^{\mathrm{T}} W^{-1} W^{-1}(\delta_\Sigma + x - y)}}.$$

Keep $m = 10$ fixed and let $n_1 = n_2 = 10, 20, 40, 80$. Make your own favourite choices of Δ and Σ and perform the simulation. Report means and standard deviations (and histograms) of $\gamma^\pm/(2\tau)$ and correlation of $\gamma^-/(2\tau)$ and $\gamma^+/(2\tau)$.

Keep $m = 10$ fixed and let $n_1 = 3n_2$, $n_2 = 10, 20, 40$. Make your own favourite choices of Δ and Σ and perform the simulation again. ■

3.3 The curse of dimensionality

We now enter the discussion of the high dimensional case where the dimension m of the variables is bigger than the sample size $n = n_1 + n_2$. Formally, we consider m, μ_1, μ_2 and Σ in the model (3.7) to depend on n, with $m/n \to \infty$. In this section we assume throughout that n_1 and n_2 are of the same order as in (3.8). A measure of the total difference between the two groups is $\Sigma^{-1/2}\Delta$, $\Delta = \mu_1 - \mu_2$.

Let $|v|^2 = \sum_{j=1}^m v_j^2$ be the squared L_2-norm of a vector v, let $|v|_1 = \sum_j |v_j|$ be the L_1-norm, let $|v|_\infty = \max_j |v_j|$ be the supnorm, and for a set A let $|A|$ be the number of elements in the set.

When $m > (n-2)$ the estimate $\hat{\Sigma}$ is no longer positive definite and we cannot calculate $\hat{\Sigma}^{-1}$. To calculate Fisher's rule one uses instead what is known as a generalized inverse of a matrix. The Moore-Penrose inverse of the symmetric positive semi-definite matrix $\hat{\Sigma}$ is

$$\hat{\Sigma}^{-1} = \sum_{i=1}^{n-2} \frac{1}{\lambda_i}\zeta_i\zeta_i^{\mathrm{T}},$$

where λ_i, $i = 1,\dots,n-2$, are the positive eigenvalues with corresponding unit length eigenvectors ζ_i. This inverse has the property that $\hat{\Sigma}^{-1}b$ is the minimum length solution to $\operatorname{argmin}_w |\hat{\Sigma}w - b|^2$.

Exercise 3.14 (Generalized inverse)
Let λ_i, $i = 1,\dots,k$, be the positive eigenvalues of $\hat{\Sigma}$ with corresponding unit length eigenvectors ζ_i. Show that $\hat{\Sigma} = \sum_{i=1}^k \lambda_i \zeta_i \zeta_i^{\mathrm{T}}$.

Let V be the space spanned by $\zeta_1,\dots,\zeta_k$, and let $J = \sum_{i=1}^k (1/\lambda_i)\zeta_i\zeta_i^{\mathrm{T}}$. Show that $(J\hat{\Sigma})v = v$ for $v \in V$.

For a vector b write $b = b_0 + b_1$ where $b_0 \in V$ and $b_1 \in V^\perp$ ($V^\perp$ is the orthogonal complement to V). Similarly, write $w = w_0 + w_1$, $w_0 \in V$, $w_1 \in V^\perp$. Show that $\min_w |\hat{\Sigma}w - b|^2 = |b_1|^2$, and that the minimum is obtained with $w_0 = Jb_0$, and with no restriction on w_1. The minimum length minimizer w is therefore $w = w_0 = Jb_0 = Jb$. ∎

3.3.1 Estimation of the mean

The next proposition shows that that we cannot estimate Δ precisely when using the traditional mean square error. We have, however, convergence when using the supnorm. In Section 3.4 we will see that an estimator converging in mean square error can be constructed for a restricted parameter space. As above $\sigma_j^2 = \Sigma_{jj}$ is the variance of the j'th variable.

Proposition 3.15
Consider $\hat{\Delta} = \bar{x} - \bar{y}$ under the asymptotic regime (3.8) and $m/n \to \infty$. Then we have the following results.

(i) $E\left|\Sigma^{-1/2}(\hat{\Delta} - \Delta)\right|^2 \to \infty$.

(ii) *If $\sigma_j^2 \geq c_1$, $j = 1,\dots,m$, for some constant c_1, we have* $E|\hat{\Delta} - \Delta|^2 \to \infty$.

(iii) *If $\sigma_j^2 \leq c_2$, $j = 1,\dots,m$, for some constant c_2 and* $\log(m)/n \to 0$, *we have* $|\hat{\Delta} - \Delta|_\infty \xrightarrow{P} 0$.

Proof. For the first result we use $\hat{\Delta} \sim N_m(\Delta, a_n\Sigma)$, with $a_n = 1/n_1 + 1/n_2$. Therefore $\Sigma^{-1/2}(\hat{\Delta} - \Delta) \sim N_m(0, a_n I_m)$ and $|\Sigma^{-1/2}(\hat{\Delta} - \Delta)|^2 \sim a_n\chi^2(m)$. The latter gives

$$E|\Sigma^{-1/2}(\hat{\Delta} - \Delta)|^2 = m a_n \to \infty,$$

because n_1 and n_2 tends to infinity at the same rate as n.

For (ii) we have

$$E|\hat{\Delta} - \Delta|^2 = \sum_{j=1}^m a_n\sigma_j^2 \geq m\, c_1\, a_n \to \infty.$$

For (iii) we need the bound $\bar{\Phi}(x) \le \frac{1}{2}e^{-x^2/2}$, $x > 0$, from Appendix A.2. Then

$$P\big(|\hat{\Delta}-\Delta|_\infty > \epsilon\big) = P\big(\exists j : |\hat{\Delta}_j - \Delta_j| > \epsilon\big) \le \sum_{j=1}^{m} P\big(|\hat{\Delta}_j - \Delta_j| > \epsilon\big)$$
$$\le \sum_{j=1}^{m} \exp\Big\{-\frac{\epsilon^2}{2a_n\sigma_j^2}\Big\} \le m\cdot\exp\Big\{-\frac{\epsilon^2}{2a_nc_2}\Big\} \le m\cdot\exp\big\{-k_1 n\big\} \to 0,$$

where k_1 is a suitable constant. □

3.3.2 Failure of Fisher's rule

Bickel and Levina (2004) consider a restricted parameter space for the mean values μ_1 and μ_2 such that estimates can be constructed that converge in mean square error. Using these in the Fisher rule they show that when $m/n \to \infty$ the probability of correct classification tends to $\frac{1}{2}$. This can be formulated in the way that the rule is no better than a random guess. A proof that corrects some problems in the proof of Bickel and Levina (2004) can be found in Grøn (2007). In order not to go into the details of that proof, I now illustrate the large m problem for the case where it is known that $\Sigma = I$, that is, all variables are independent and have unit variance. In this case the maximum likelihood classifier is (*KV* for *known variance*)

$$\xi_{\mathrm{KV}}(z) = G\Big((\bar{x}-\bar{y})^{\mathrm{T}}\big[z-\frac{1}{2}(\bar{x}+\bar{y})\big]\Big),$$

with probabilities of correct classification from (3.3), $p_{11}(\xi_{\mathrm{KV}}) = \Phi\big(\gamma^-_{\mathrm{KV}}/(2\tau_{\mathrm{KV}})\big)$ and $p_{22}(\xi_{\mathrm{KV}}) = \Phi\big(\gamma^+_{\mathrm{KV}}/(2\tau_{\mathrm{KV}})\big)$, where

$$\gamma^{\pm}_{\mathrm{KV}} = \hat{\Delta}^{\mathrm{T}}(\Delta \pm d), \quad \tau^2_{\mathrm{KV}} = \hat{\Delta}^{\mathrm{T}}\hat{\Delta},$$

with $d = \bar{x}-\mu_1+\bar{y}-\mu_2$.

Below we use $\omega_n = (n_1-n_2)/(n_1+n_2)$ and note that under the asymptotic regime (3.8) we have $|\omega_n| \le 1-2c_s$.

Lemma 3.16
Let $|\Delta|^2 = \Delta^T\Delta = \sum_j \Delta_j^2$. *Then*

$$E\big(\gamma^-_{\mathrm{KV}}\big) = m\omega_n a_n + |\Delta|^2, \qquad \mathrm{Var}\big(\gamma^-_{\mathrm{KV}}\big) = ma_n^2(1+\omega_n^2) + 2a_n(1+\omega_n)|\Delta|^2,$$
$$E\big(\gamma^+_{\mathrm{KV}}\big) = -m\omega_n a_n + |\Delta|^2, \qquad \mathrm{Var}\big(\gamma^+_{\mathrm{KV}}\big) = ma_n^2(1+\omega_n^2) + 2a_n(1-\omega_n)|\Delta|^2,$$
$$E\big(\tau^2_{\mathrm{KV}}\big) = ma_n + |\Delta|^2, \qquad \mathrm{Var}\big(\tau^2_{\mathrm{KV}}\big) = 2ma_n^2 + 4a_n|\Delta|^2.$$

Proof. Let $u \sim N(0, 1/n_1)$ and $v \sim N(0, 1/n_2)$ with u and v independent. Then $\gamma^{\pm}_{\mathrm{KV},j} = (u-v+\Delta_j)(\Delta_j \pm (u+v))$ and $\tau^2_{\mathrm{KV},j} = (u-v+\Delta_j)^2$. A tedious calculation of moments gives the mean and variance on using that $E\big(N(0,1)^4\big) = 3$. For the formulas stated we have used

$$\frac{1}{n_2}-\frac{1}{n_1} = a_n\omega_n, \quad \frac{2}{n_1^2}+\frac{2}{n_2^2} = a_n^2(1+\omega_n^2), \quad \frac{2}{n_1} = a_n(1-\omega_n), \quad \frac{2}{n_2} = a_n(1+\omega_n). \quad \square$$

Proposition 3.17
Consider the asymptotic regime (3.8), $m/n \to \infty$ and $|\Delta|^2/\sqrt{ma_n} \to 0$. In the balanced case where $n_1 = n_2$ ($\omega_n = 0$) we have

$$p_{11}(\xi_{\mathrm{KV}}) \xrightarrow{P} \frac{1}{2}, \quad p_{22}(\xi_{\mathrm{KV}}) \xrightarrow{P} \frac{1}{2}.$$

In the imbalanced case with $n_1 > n_2$ and $\omega_n > 0$ fixed we have

$$p_{11}(\xi_{\mathrm{KV}}) \xrightarrow{P} 1, \quad p_{22}(\xi_{\mathrm{KV}}) \xrightarrow{P} 0.$$

Proof. We have that $\mathrm{Var}\big(\tau^2_{\mathrm{KV}}/E(\tau^2_{\mathrm{KV}})\big) \to 0$, which implies that $\tau^2_{\mathrm{KV}}/E(\tau^2_{\mathrm{KV}}) \to 1$ in probability. When $n_1 = n_2$ we have

$$\frac{E(\gamma^-_{\mathrm{KV}})}{\sqrt{E(\tau^2_{\mathrm{KV}})}} \to 0, \quad \frac{\mathrm{Var}(\gamma^-_{\mathrm{KV}})}{E(\tau^2_{\mathrm{KV}})} \to 0, \quad \frac{E(\gamma^+_{\mathrm{KV}})}{\sqrt{E(\tau^2_{\mathrm{KV}})}} \to 0, \quad \frac{\mathrm{Var}(\gamma^+_{\mathrm{KV}})}{E(\tau^2_{\mathrm{KV}})} \to 0,$$

so that $\gamma^-_{\mathrm{KV}}/\sqrt{E(\tau^2_{\mathrm{KV}})}$ and $\gamma^+_{\mathrm{KV}}/\sqrt{E(\tau^2_{\mathrm{KV}})}$ both tend to zero in probability.

When $\omega_n > 0$ is fixed we find instead

$$E(\gamma^-_{\mathrm{KV}})/\sqrt{E(\tau^2_{\mathrm{KV}})} \to \infty, \quad E(\gamma^+_{\mathrm{KV}})/\sqrt{E(\tau^2_{\mathrm{KV}})} \to -\infty,$$

so that $\gamma^-_{\mathrm{KV}}/\sqrt{E(\tau^2_{\mathrm{KV}})}$ tends to infinity and $\gamma^+_{\mathrm{KV}}/\sqrt{E(\tau^2_{\mathrm{KV}})}$ tends to minus infinity. □

Exercise 3.18 (Fishers rule, $m > n$)
Consider the two sample case with $n_1 = 10$, $n_2 = 10$, $m = 5, 10, 20, 40, 80, 160$ and where the mean difference is $\Delta = (1, 1, 1, 1, 1, 0, \ldots, 0)$.

For each value of m simulate 100 times the complete data set from the multivariate normal distribution (make your own choice for Σ). Calculate $\bar{x}$, $\bar{y}$ and $\hat{\Sigma}$. Next calculate $\hat{a} = \hat{\Sigma}^{-1}\hat{\Delta}$ using the generalized inverse (you can use `ginv` from the `MASS` package in R). Finally calculate $\gamma^{\pm}_F/(2\tau_F)$ from (3.12). Make a table with the mean and standard deviations of the 100 simulated data sets. ■

3.4 Thresholded independence classifier

We have seen in the previous section that Fisher's rule runs into problems when the dimension m is big. At the intuitive level the problem is that we must estimate m mean differences and $m(m+1)/2$ parameters in the variance Σ. This accumulates too much noise so that the true signal cannot be identified.

Consider the model (3.7) under the assumption of independent variables. The variance matrix Σ is then a diagonal matrix, which we denote D having entries σ^2_j. The maksimum likelihood classifier can be seen from Example 3.5, and replacing parameters by estimates we obtain the *independence classifier*:

$$\xi_I(z) = G\Big(\hat{\Delta}^{\mathrm{T}}\hat{D}^{-1}\big[z - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\big]\Big), \quad \hat{\Delta} = \hat{\mu}_1 - \hat{\mu}_2. \tag{3.15}$$

Here $\hat{D} = D(s^2)$ with

$$s_j^2 = \frac{1}{n-2}\Big\{\sum_{i=1}^{n_1}(x_{ij}-\bar{x}_j)^2 + \sum_{i=1}^{n_2}(y_{ij}-\bar{y}_j)^2\Big\}.$$

The idea is now that we use the independence classifier also in the dependent case where Σ is a general variance matrix. The hope is that this classifier has good properties even though it ignores dependency among the variables.

Bickel and Levina (2004) show that the indpendence classifier can indeed give useful results. The parameter space for the mean values μ_1 and μ_2 is restricted and special designed estimates for these are used. Their result gives an upper bound on the mean of the classification error. A result in the same spirit is given in Bak et al. (2015) which we will now study in detail. Instead of using a special designed estimate of the mean we instead exclude variables judged to have no differential expression. Variables are excluded based on the t-statistic for no group difference. Thus, let

$$t_j = \big(\bar{x}_j - \bar{y}_j\big)/\sqrt{s_j^2 a_n}, \quad w_j = 1(|t_j| > \kappa),$$

where κ is called the threshold for including the variable and $a_n = 1/n_1 + 1/n_2$. Before studying the classifier we mention a result that shows the possibility of a perfect separation between variables with differential expression and those with no differential expression. The result is a slightly changed version of Theorem 3 of Fan and Fan (2008). As before $\delta_j = (\mu_{1j} - \mu_{2j})/\sigma_j$ is the scaled differential expression. There is no assumption on the variance matrix Σ in the proposition.

Proposition 3.19
Let $0 < \lambda < 1$ be a constant and assume $\log(m)/n^\lambda \to 0$. Let $\beta < (1-\lambda)/2$ be a constant, define $A_0 = \{i : \delta_i = 0\}$ and $\delta_{\min} = \min_{i \notin A_0}\{|\delta_i|\}$, and assume that $\delta_{\min} \geq n^{-\beta}$. Then with $\kappa = n^{\lambda/2}$ we have

$$P\big(|t_j| \geq \kappa, j \notin A_0; |t_j| < \kappa, j \in A_0\big) \to 1.$$

Proof. We use Appendix A.2 (iii) with $b = 1/a_n$, where $c_s^2 n \leq b \leq (1-c_s)^2 n$. This gives the existence of constants a_1 and a_2 in the following bound:

$$\begin{aligned} P\big(\exists j \in A_0 : |t_j| \geq \kappa\big) &\leq \sum_{j \in A_0} P\big(|t_j| \geq \kappa\big) \leq m a_1 e^{-a_2 n\kappa^2/b} \leq m a_1 e^{-a_2\kappa^2/(1-c_s)^2} \\ &= a_1 \exp\big\{-n^\lambda\big(a_2/(1-c_s)^2 - \log(m)/n^\lambda\big)\big\} \to 0. \end{aligned} \tag{3.16}$$

For the variables with differential expression $|\delta_j| \geq \delta_{\min}$ we use Appendix A.2 (iv). To use the latter we note that

$$\delta_{\min} \geq n^{-\beta} \geq \frac{2\kappa}{\sqrt{b}} n^{-\beta+(1-\lambda)/2}\frac{c_s}{2},$$

which is greater than $2\kappa/\sqrt{b} = 2\alpha$ for large n on using $\beta < (1-\lambda)/2$. We then obtain

$$\begin{aligned} P\big(\exists j \notin A_0 : |t_j| < \kappa\big) &\leq \sum_{j \notin A_0} P\big(|t_j| < \kappa\big) \leq m a_1 e^{-a_2 n\kappa^2/b} \leq m a_1 e^{-a_2\kappa^2/(1-c_s)^2} \\ &= a_1 \exp\big\{-n^\lambda\big(a_2/(1-c_s)^2 - \log(m)/n^\lambda\big)\big\} \to 0. \end{aligned} \tag{3.17}$$

Combining the two displayed statements we have that one minus the probability in the proposition tends to zero. □

Proposition 3.19 seems to give a nice positive result, showing that even in high dimensional cases we can find the relevant information. However, a warning is needed here: the result is asymptotic. In Table 3.1 I have given some actual probabilities to show that the limit of Proposition 3.19 requires large sample sizes.

Setting	$\sum_{A_0}$	$\sum_{A_0^c}$	P_{A_0}	$P_{A_0^c}$
$n_1 = n_2 = 10$, $\kappa = 2.8089$	115.0	68.5	6.0e-51	6.1e-51
$n_1 = n_2 = 100$, $\kappa = 4.27$	0.30	0.30	0.74	0.74
$n_1 = n_2 = 200$, $\kappa = 5.53$	0.00058	0.00057	0.9994	0.9994

Table 3.1: In this table there are $m = 10000$ variables of which 100 has a nonzero differential expression of size $\delta = 1$. The t-statistics are thresholded at κ. The table gives the sum appearing after the first inequality in (3.16) ($\sum_{A_0} = \sum_{j\in A_0} P\big(|t_j| \geq \kappa\big)$) and (3.17) ($\sum_{A_0^c} = \sum_{j\notin A_0} P\big(|t_j| < \kappa\big)$). For the case of independent variables the table gives the probability that all absolute t-values for variables with differential expression are below κ ($P_{A_0} = P\big(|t_j| < \kappa, j \in A_0\big)$), and all absolute t-values for variables with no differential expression are above κ ($P_{A_0^c} = P\big(|t_j| \geq \kappa, j \notin A_0\big)$).

When considering the classification problem we do not need complete separation of null cases and nonnull cases as in Proposition 3.19. The next theorem is a variation on this theme. The setup is such that all the variables with no differential expression are eventually excluded, but we do not necessarily include all the variables with differential expression. To define the *thresholded independence classifier* let

$$\hat{\mu}_{1j} = \bar{x}_j, \qquad \hat{\mu}_{2j} = \bar{y}_j, \qquad s_j^2 = \frac{\sum_{i=1}^{n_1}(x_{ij}-\bar{x}_j)^2 + \sum_{i=1}^{n_2}(y_{ij}-\bar{y}_j)^2}{n_1+n_2-2},$$
$$t_j = \frac{\bar{x}_j - \bar{y}_j}{\sqrt{s_j^2 a_n}}, \qquad w_j = 1(|t_j| > \kappa) \qquad \hat{\Delta}_j = \bar{x}_j - \bar{y}_j, \quad \hat{\Delta}_{tj} = \hat{\Delta}_j w_j,$$

where $a_n = 1/n_1 + 1/n_2$. The classifier is

$$\xi_{\mathrm{TI}}(z) = G\Big(\hat{\Delta}_t^{\mathrm{T}} \hat{D}^{-1}\big[z - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\big]\Big), \tag{3.18}$$

where

$$\hat{\Delta}_t^{\mathrm{T}} \hat{D}^{-1}\big[z - \frac{1}{2}(\hat{\mu}_1 + \hat{\mu}_2)\big] = \sum_{j=1}^{m} \frac{\bar{x}_j - \bar{y}_j}{s_j^2}\big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j)\big] w_j.$$

The probabilities of correct classification are given in (3.3) with

$$\gamma_{\mathrm{TI}}^{\pm} = \sum_j a_{\mathrm{TI},j}\big(\Delta_j \pm (\bar{x}_j - \mu_1 + \bar{y}_j - \mu_2)\big), \quad \tau_{\mathrm{TI}}^2 = \hat{a}_{\mathrm{TI}}^{\mathrm{T}} \Sigma \hat{a}_{\mathrm{TI}} \quad \hat{a}_{\mathrm{TI},j} = w_j \frac{\bar{x}_j - \bar{y}_j}{s_j^2}. \tag{3.19}$$

We will compare the performance of the thresholded independence classifier with the corresponding oracle classfier where the parameters are known:

$$\xi_{\mathrm{OTI}}(z) = G\Big(\Delta_t^{\mathrm{T}} D^{-1}\big[z - \frac{1}{2}(\mu_1+\mu_2)\big]\Big), \quad \Delta_{tj} = \Delta_j 1(|\Delta_j|/\sigma_j > \alpha)$$

with the threshold related to α by $\kappa = \alpha/\sqrt{a_n}$. This classifier has probabilities of wrong classification $W(\xi_{\mathrm{OTI}}) = p_{12}(\xi_{\mathrm{OTI}}) = p_{21}(\xi_{\mathrm{OTI}})$ given by

$$W(\xi_{\mathrm{OTI}}) = \bar{\Phi}\Big(\frac{1}{2}\Delta_t^{\mathrm{T}} D^{-1}\Delta_t \Big/ \sqrt{\Delta_t^{\mathrm{T}} D^{-1}\Sigma D^{-1}\Delta_t}\Big)$$

If the eigenvalues of $D^{-1/2}\Sigma D^{-1/2}$ are bounded by a constant c_λ the largest classification error $W(\xi_{\mathrm{OTI}})$ is

$$\bar{\Phi}\Big(\frac{1}{2\sqrt{c_\lambda}}|D^{-1/2}\Delta_t|\Big) = \bar{\Phi}\Big(\frac{1}{2\sqrt{c_\lambda}}\sqrt{\sum_{j:|\delta_j|>\alpha}\delta_j^2}\Big). \tag{3.20}$$

Under a number of conditions we will show that the thresholded independence classifier has a limiting worst case classification error that is comparable to (3.20).

For the maximum likelihood classifier with classification error $\bar{\Phi}\big(\frac{1}{2}\sqrt{\Delta^{\mathrm{T}}\Sigma^{-1}\Delta}\big)$ from (3.6), we have a worst case error of

$$\bar{\Phi}\Big(\frac{1}{2\sqrt{c_\lambda}}\sqrt{\sum_j \delta_j^2}\Big),$$

which is slightly smaller than (3.20) because of the inclusion of all variables in the sum.

We now outline the parameter set for the mean values and the variance matrix to be considered. First we define a set of covariance matrices

$$B_0 = \big\{\Sigma : c_{\sigma 1} \le \sigma_j^2 \le c_{\sigma 2}, \lambda_{\max}(D^{-1/2}\Sigma D^{-1/2}) \le c_\lambda\big\},$$

where $c_{\sigma 1}$, $c_{\sigma 2}$ and c_λ are positive constants and $\lambda_{\max}$ is the maximal eigenvalue. Next, we define two sets of mean values. In these sets $c_{\delta 1}$ and $c_{\delta 2}$ are constants, b_n is a sequence tending to zero and α is a parameter that is allowed to depend on n. We express the sets through the scaled differential expression $\delta_j = \Delta_j/\sigma_j$. The sets are,

$$B_1(\alpha) = \big\{(\mu_1,\mu_2) : |\{j : |\delta_j| > \alpha/2\}| \le b_n n; |\{j : |\delta_j| > c_{\delta 1}\}| \ge 1\big\},$$
$$B_2(\alpha) = \big\{(\mu_1,\mu_2) : |\{j : \alpha/2 \le |\delta_j| \le 2\alpha\}| \le c_{\delta 2}K_n,\ K_n = |\{j : |\delta_j| > 2\alpha\}| \ge 1\big\}.$$

In the theorem below we consider the thresholded independence classifier with threshold $\kappa = \alpha/\sqrt{1/n_1 + 1/n_2}$. A variable with differentiable expression $\delta = \alpha$ will then be included in the classifier with probability around $\frac{1}{2}$. Both of B_1 and B_2 are introduced to extend a setting with a fixed number of differentiable expressed variables. The set B_1 says that the number of variables with a differential expression

bounded away from zero by $\alpha/2$ is of smaller order than n. The set B_2 does not restrict the number of differentiable expressed variables, instead the requirement is that the number of variables with a differentiable expression around α scales with the number of variables with a large differential expression. The theorem below gives a limiting upper bound for the error probabilities of the classifier based on the differential expression of those variables with a strong differential expression.

The set B_0 restricts the correlation structure. If the correlation between any two variables is the same, say $\rho > 0$, the maximal eigenvalue is $1 + \rho(m-1)$. In this case we do not have an upper bound on the eigenvalue. If, however, we restrict the row sums of the correlation matrix we have an upper bound on the maximal eigenvalue. A particular instance is when there is a constant K such that each variable is correlated with at most K other variables. The following result is attributed to Frobenius in two papers 1908 and 1909, or to Gershgorin circle theorem from 1931.

Lemma 3.20
Let V be a variance matrix. Then

$$\lambda_{\max}(V) \leq \max_i \sum_{j=1}^{m} |V_{ij}|.$$

Proof. Let λ be an eigenvalue with eigenvector v. Let $i = \text{argmax}_{j=1,\ldots,m} |v_j|$. Then

$$\lambda |v_i| = |(\lambda v)_i| = |\sum_j V_{ij} v_j| \leq \sum_j |V_{ij}||v_j| \leq \sum_j |V_{ij}||v_i|,$$

and dividing by $|v_i|$ we get the result of the lemma. □

Theorem 3.21
Let m tend to infinity with n such that $\log(m)/n = \phi_n \to 0$, and let $\alpha \geq c_\alpha \phi_n^{1/2-\gamma}$, where $c_\alpha > 0$ and $0 < \gamma < \frac{1}{2}$ are constants. From α we define the threshold as $\kappa = \alpha/\sqrt{a_n}$, where $a_n = 1/n_1 + 1/n_2$. Consider the parameter space Θ being either $\Theta = \{\Sigma \in B_0, (\mu_1, \mu_2) \in B_1(\alpha)\}$ or $\Theta = \{\Sigma \in B_0, (\mu_1, \mu_2) \in B_2(\alpha)\}$. Let $W(\xi_{\text{TI}}, \theta)$ be one of the probabilities of wrong classification $p_{12}(\xi_{\text{TI}})$ or $p_{21}(\xi_{\text{TI}})$. Then, for all $\epsilon > 0$

$$P\Big\{W(\xi_{\text{TI}}, \theta) - \bar{\Phi}\Big(\frac{1}{2\sqrt{c_\lambda}} \sqrt{\sum_{j:|\delta_j|>2\alpha} \delta_j^2}\Big) > \epsilon\Big\} \to 0,$$

uniformly for $\theta \in \Theta$.

Proof. See Bak et al. (2015). The uniformity in θ means that for any $\tilde{\epsilon} > 0$ there exists $n(\epsilon, \tilde{\epsilon})$ such that the probability is less than $\tilde{\epsilon}$ for $n > n(\epsilon, \tilde{\epsilon})$ for all $\theta \in \Theta$. □

Theorem 3.21 differs from the error probability in (3.20) for the oracle independence classifier in that not all variables above α are included. With estimated parameters we can only include variables with differential expression slightly above α, in the theorem formulated as being above 2α.

Exercise 3.22 (The set B_2)
Consider the situation where $m = n^2$. Assume that δ_j is zero except when j is a multiple of 10, and then $\delta_j = j^{-1/8}$. Show that the conditions of Theorem 3.21 are satisfied when using the set $B_2(\alpha)$. ∎

3.4.1 Efficiency

Consider the case of two variables only, $m = 2$. When parameters are known the maximum likelihood classifier has error probabilities

$$\bar{\Phi}\Big(\frac{\Delta^T\Sigma^{-1}\Delta}{2\sqrt{\Delta^T\Sigma^{-1}\Sigma\Sigma^{-1}\Delta}}\Big) = \bar{\Phi}\Big(\frac{1}{2}\sqrt{\Delta^T\Sigma^{-1}\Delta}\Big).$$

The independence classifier for the case where all variances are 1 has error probabilities

$$\bar{\Phi}\Big(\frac{\Delta^T D^{-1}\Delta}{2\sqrt{\Delta^T D^{-1}\Sigma D^{-1}\Delta}}\Big) = \bar{\Phi}\Big(\frac{\Delta^T\Delta}{2\sqrt{\Delta^T\Sigma\Delta}}\Big), \qquad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

To compare the two we let $\Delta^T\Sigma^{-1}\Delta = c^2$ be fixed. Let $\lambda_1 = 1+\rho$ and $\lambda_2 = 1-\rho$ be the eigenvalues of Σ with eigenvectors v_1 and v_2. Write $\Delta = \gamma_1 v_1 + \gamma_2 v_2$ so that $c^2 = \Delta^T\Sigma^{-1}\Delta = \gamma_1^2/\lambda_1 + \gamma_2^2/\lambda_2$, $\Delta^T\Delta = \gamma_1^2 + \gamma_2^2$ and $\Delta^T\Sigma\Delta = \lambda_1\gamma_1^2 + \lambda_2\gamma_2^2$. Expressing γ_2^2 in terms of γ_1^2 we end up with

$$f(\rho,\gamma_1^2) := \frac{\Delta^T\Delta}{\sqrt{\Delta^T\Sigma\Delta}} = \frac{c^2 + \frac{2\rho}{1-\rho^2}\gamma_1^2}{\sqrt{c^2 + \frac{4\rho}{(1-\rho^2)(1-\rho)}\gamma_1^2}}, \quad 0 \le \gamma_1^2 \le c^2(1+\rho).$$

Clearly, $f(\rho,\gamma_1^2) = c$ when $\gamma_1^2 = 0$ or $\gamma_1^2 = c^2(1+\rho)$ ($\gamma_2^2 = 0$). This means that when Δ is an eigenvector of Σ there is no increase in the classification error when using the indepencence classifier as compared to the maximum likelihood classifier. Furthermore,

$$f(\rho,\gamma_1^2) \ge f(\rho, c^2(1-\rho^2)/2) = c\sqrt{1-\rho^2}. \tag{3.21}$$

This shows that for small correlations we do not loose much from using the independence classifier (in the case $m = 2$), but for large correlations there exists Δ such that the classifier is almost useless.

Consider now a general m and let all the variances be 1. Let $\lambda_1,\dots,\lambda_m$ be the eigenvalues of Σ with eigenvectors $v_1,\dots,v_m$. Write $\Delta = \sum_j a_j v_j$. Then

$$c^2 = \Delta^T\Sigma^{-1}\Delta = \sum_j a_j^2/\lambda_j, \quad \Delta^T\Delta = \sum_j a_j^2, \quad \Delta^T\Sigma\Delta = \sum_j \lambda_j a_j^2.$$

We assume that we have the upper bound $\lambda_j \le c_\lambda$ and consider a situation where the load a_j of Δ on the eigenvector v_j is given by $a_j^2 = \lambda_j c^2/m$. In this case, using that $\sum_j \lambda_j = \mathrm{tr}(\Sigma) = m$, we get

$$\frac{\Delta^T\Delta}{\sqrt{\Delta^T\Sigma\Delta}} = \frac{c}{\sqrt{\frac{1}{m}\sum_j \lambda_j^2}} \ge \frac{c}{\sqrt{c_\lambda}}. \tag{3.22}$$

The argument is as follows. We write $\lambda = e + \alpha z$ with $e = (1,\ldots,1)$, $e \cdot z = 0$ and $z \cdot z = m$. We then have

$$\sum_j \lambda_j^2 = e^{\mathrm{T}} e + \alpha^2 z^{\mathrm{T}} z = m(1+\alpha^2), \quad 0 < 1 + \alpha z_i < c_\lambda.$$

Here we want to choose z to maximize α. The solution (most likely) is to take $z = (x,\ldots,x,y,\ldots,y)$ with $x < 0$, $y > 0$ and where x appears K times and y appears $m-K$ times. We then have

$$m = Kx^2 + (m-K)y^2, \quad Kx + (m-K)y = 0, \quad 1 + \alpha x > 0, \quad 1 + \alpha y < c_\lambda.$$

Taking x and y such that we get the same limit for α in the two last equations we find $y = (-x)(c_\lambda - 1)$. The second equation then gives $K = m(c_\lambda - 1)/c_\lambda$, and the first equations lead to $x^2 = 1/(c_\lambda - 1)$. Returning to the third equation we get the bound $\alpha^2 < c_\lambda - 1$ so that $1 + \alpha^2 < c_\lambda$. In this way we get the bound in (3.22).

Let us return to the case $m = 2$. The minimum in (3.21) is obtained when one of the coordinates of Δ is zero. Consider the case where one of the coordinates of Δ is zero and where $\Delta^{\mathrm{T}}\Sigma^{-1}\Delta = c^2$. The full maximum likelihood classifier has error probabilities $\bar{\Phi}(c/2)$, whereas using only the variables with nonzero entries of Δ leads to the error probabilities $\bar{\Phi}(c\sqrt{1-\rho^2}/2)$. The message here is that if we know the correlations, variables with no differential expression can actually help in improving the classifier. We know already that the Fisher rule runs into problems when trying to estimate all the correlations. The ROAD classifier in Section 3.5 below tries to incorporate correlation without using Fisher's rule.

3.5 The ROAD classifier

Suppose we have two groups with observations in $\mathbb{R}^m$ and with distributions $N_m(\mu_1, \Sigma)$ and $N_m(\mu_2, \Sigma)$, where μ_1 and μ_2 are m-dimensional vectors and Σ is a positive definite $m \times m$ matrix. Define $\Delta = \mu_1 - \mu_2$. For the linear classifier $\xi_L(x) = G\big(r^{\mathrm{T}}[x - (\mu_1 + \mu_2)/2]\big)$ we have from (3.3) that the probability of correct classification is $\Phi\big(\frac{1}{2} r^{\mathrm{T}}\Delta/\sqrt{r^{\mathrm{T}}\Sigma r}\big)$. To maximize this probability we must find the classification vector r that maximizes $r^{\mathrm{T}}\Delta/\sqrt{r^{\mathrm{T}}\Sigma r}$. Since the length of r does not enter the expression the maximization problem is equivalent to minimizing $r^{\mathrm{T}}\Sigma r$ subject to $r^{\mathrm{T}}\Delta = 1$.

We know from the maximum likelihood classifier that the solution is proportional to $\Sigma^{-1}\Delta$. We also know that the empirical version, namely Fisher's rule, does not work well in the high dimensional setting. The thresholded independence classifier can be seen as a regularized version of $\Sigma^{-1}\Delta$, where Σ is replaced by the diagoal D and Δ is replaced by a thresholded version Δ_t with $\Delta_{tj} = 0$ for many variables. In this section we consider a different regularization where the correlation structure is taken into account. The (theoretical/oracle) *regularized optimal affine discriminant* (ROAD) of Fan et al. (2012) is the linear classifier

$$\xi_R(z) = G\big(r_c^{\mathrm{T}}[z - (\mu_1 + \mu_2)/2]\big)$$

with

$$r_c = \underset{|r|_1 \le c,\ r^{\mathrm{T}}\Delta=1}{\operatorname{argmin}} \ r^{\mathrm{T}}\Sigma r. \tag{3.23}$$

Here c is a tuning parameter, just as the threshold of the thresholded independence classifier is a tuning parameter. We have the following limiting cases:

$$\begin{aligned} c < c_0 &= \frac{1}{\max_i |\Delta_i|} : \text{ there is no solution,} \\ c > c_1 &= \frac{|\Sigma^{-1}\Delta|_1}{\Delta\Sigma^{-1}\Delta} : \ r_c = \frac{\Sigma^{-1}\Delta}{\Delta\Sigma^{-1}\Delta}, \end{aligned} \tag{3.24}$$

where the first follows from

$$1 = r^{\mathrm{T}}\Delta \le \sum_{i=1}^{m} |r_{ci}||\Delta_i| \le \max_i |\Delta_i||r_c|_1 \le \max_i |\Delta_i| c,$$

and the second case says that we get the maximum likelihood classifier for sufficiently large c.

The empirial version of the ROAD classifier is obtained on replacing μ_1 by $\bar{x}$, μ_2 by $\bar{y}$ and Σ by $\hat{\Sigma}$:

$$\hat{\xi}_R(z) = G\big(\hat{r}_c^{\mathrm{T}}[z - (\bar{x}+\bar{y})/2]\big)$$

with

$$\hat{r}_c = \underset{|r|_1 \le c,\ r^{\mathrm{T}}\hat{\Delta}=1}{\operatorname{argmin}} \ a^{\mathrm{T}}\hat{\Sigma}a. \tag{3.25}$$

The probabilities of correct classifications are $\Phi\big(\gamma_R^{\pm}/(2\tau_R)\big)$ with

$$\gamma_R^{\pm} = \hat{r}_c^{\mathrm{T}}(\Delta \pm d), \qquad \tau_R^2 = \hat{r}_c^{\mathrm{T}}\Sigma\hat{r}_c,$$

with $d = \bar{x} - \mu_1 + \bar{y} - \mu_2$. Beware, that this differs from Fan et al. (2012), where $\gamma_R^{\pm}$ is replaced by $\hat{r}^{\mathrm{T}}\hat{\Delta}$ which is not correct:

$$\gamma_R^{-} = \hat{r}_c^{\mathrm{T}}(\hat{\Delta} - 2(\bar{x} - \mu_1)), \qquad \gamma_R^{+} = \hat{r}_c^{\mathrm{T}}(\hat{\Delta} + 2(\bar{y} - \mu_2)).$$

Fan et al. (2012) state a Theorem 1 on the closeness of the classification errors of the ROAD classisifer and the empirical ROAD classifier. However, the proof in Fan et al. (2012) uses a result that is not valid. Here follows an alternative version.

Theorem 3.23
Let ϵ be a positive constant such that $\max_j \{\frac{1}{2}|\Delta_j|\} > \epsilon$, and let c in the construction of $\hat{r}_c$ satisfy $c > \epsilon + 2/\max_j \{|\Delta_j|\}$. Let b_n be a sequence tending to zero such that $\|\hat{\Sigma} - \Sigma\|_\infty = O_p(b_n)$, and $\|\hat{\mu}_i - \mu_i\|_\infty = O_p(b_n)$, $i = 1,2$. Then, as $n \to \infty$:

$$p_{12}(\hat{\xi}_R) - p_{12}(\xi_R) = O_p(d_n)$$

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

Proof. See Bak and Jensen (2014). The notation $X_n = O_p(b_n)$ means that $|X_n|/b_n$ is stochastically bounded: for any $\epsilon > 0$ there exists K_ϵ and n_ϵ such that $P(|X_n|/b_n > K_\epsilon) \leq \epsilon$ for $n \geq n_\epsilon$. □

In section 3.7 below is a discussion of how to calculate the ROAD classifier.

Exercise 3.24
Simulate data where $\Delta_j \neq 0$ for a few variables only (need more specification?). Illustrate the full solution path (see section 3.7) for $\hat{r}_c$ as function of c and make a plot of the probabilities of correct classification as function of c. ■

3.6 The imbalance problem

In this section we consider the case where the number of samples in the two groups are different, say $n_1 > n_2$. We first show by a numerical example that ths can lead to a serious bias in the classification: the classification rule favours group 1. We next discuss the origin of the bias and suggest modifications of the classification rules to counterbalance the bias. This section is based on Bak and Jensen (2016).

Example 3.25
Let $x_i \sim N_m(\mu_1, I)$ and $y_i \sim N(\mu_2, I)$. The mean difference $\Delta = \mu_1 - \mu_2$ has $\Delta_j = 1$ for $j = 1, \ldots, 10$ and is zero otherwise. We consider the thresholded independence rule, with the threshold κ determined from the noncentral t-distribution such that the power of the t-test for detecting a mean difference of 1 is 0.8. The probabilities of correct classification are $p_{11}(\xi_{\mathrm{TI}}) = \Phi(\gamma_{\mathrm{TI}}^-/(2\tau_{\mathrm{TI}}))$ and $p_2(\xi_{\mathrm{TI}}) = \Phi(\gamma_{\mathrm{TI}}^+/(2\tau_{\mathrm{TI}}))$ with

$$\gamma_{\mathrm{TI}}^{\pm} = \sum_{j=1}^{m} \frac{\bar{x}_j - \bar{y}_j}{s_j^2}[\Delta_j \pm d_j]\, w_j, \quad \tau_{\mathrm{TI}} = \sum_{j=1}^{m} \frac{(\bar{x}_j - \bar{y}_j)^2}{s_j^4} w_j,$$

where $d = \bar{x} - \mu_1 + \bar{y} - \mu_2$. In Table 3.2 is a small simulation to illustrate the clasification bias when $n_1 \neq n_2$. It is clear that the bias is so large that a better classifier is needed. ■

Exercise 3.26 (Number of false positives)
Using the t-distribution and the noncentral t-distribution find for each entry in Table 3.2 the expected number of false positives in the variables that enter the classifier as well as the expected number of true positives.

Make a table similar to Table 3.2, but with the threshold determined such that the power of the t-test for detecting a mean difference of 1 is 0.5. ■

				Group 1		Group 2	
n_1	n_2	m	κ	Mean	Std	Mean	Std
15	15	1000	1.888	70.5	7.0	70.3	7.1
16	14	1000	1.882	76.8	6.2	63.2	7.7
18	12	1000	1.833	87.4	4.5	44.9	8.7
20	10	1000	1.734	94.9	2.2	24.7	7.2

Table 3.2: Average percentage of correct classification and the standard deviation based on 1000 simulated data sets for the thresholded independence classifier. The setting is described in the text. (Table 1 from Bak and Jensen (2016))

3.6.1 Origin of the bias problem

For the thresholded independence classifier the probabilities of correct classification in the case of independent variables are given via (3.19) with $\gamma_{\mathrm{TI}}^{\pm} = \sum_j \gamma_j^{\pm}$ and $\tau^2 = \sum_j \tau_j^2$ where

$$\gamma_j^{\pm} = \frac{\bar{x}_j - \bar{y}_j}{s_j^2}\{\Delta_j \pm d_j\}w_j, \quad d_j = \bar{x}_j - \mu_{1j} + \bar{y}_j - \mu_{2j} \qquad \tau_j^2 = \frac{(\bar{x}_j - \bar{y}_j)^2}{s_j^4}w_j. \tag{3.26}$$

For dependent variables $\gamma^{\pm}$ is as above, whereas τ^2 becomes more complicated, see (3.28) below. We have the following conditional distribution

$$(\bar{x}_j - \mu_{1j} + \bar{y}_j - \mu_{2j})|(\bar{x}_j - \bar{y}_j - \Delta_j) = z \sim N\Big(-\omega_n z, \frac{4\sigma_j^2}{n_1+n_2}\Big), \quad \omega_n = \frac{n_1 - n_2}{n_1 + n_2}. \tag{3.27}$$

In the imbalanced case with $n_1 > n_2$ we have $\omega_n > 0$ and the conditional mean in the above distribution is nonzero. We use this simple fact to understand the bias problem for the independence classifier.

The following proposition gives the mean of $\gamma_j^{\pm}$ and τ_j^2. Let $a_n = 1/n_1 + 1/n_2$ and define

$$T_{a,b}(\delta; n_1, n_2) = E\Big[\frac{(d+\delta)^a}{v^b}w(t)\Big] \quad \text{and} \quad S_{a,b}(\delta; n_1, n_2) = \omega_n T_{a,b}(\delta; n_1, n_2).$$

where $d \sim N(0, a_n)$, $v \sim \chi^2(f)/f$ with $f = n_1 + n_2 - 2$, d and v are independent, and $t = (d+\delta)/\sqrt{v a_n}$.

Proposition 3.27
We have with $\delta_j = \Delta_j/\sigma_j$:

$$\begin{aligned}
E(\gamma_j^-) &= \delta_j T_{1,1}(\delta_j, n_1, n_2) - \delta_j S_{1,1}(\delta_j, n_1, n_2) + S_{2,1}(\delta_j, n_1, n_2),\\
E(\gamma_j^+) &= \delta_j T_{1,1}(\delta_j, n_1, n_2) + \delta_j S_{1,1}(\delta_j, n_1, n_2) - S_{2,1}(\delta_j, n_1, n_2),\\
E\big(\tau_j^2\big) &= T_{2,2}(\delta_j; n_1, n_2), \quad E\big(\tau_j^4\big) = T_{4,4}(\delta_j; n_1, n_2).
\end{aligned}$$

Proof. The proof is based on first calculating the conditional mean of $\gamma_j^{\pm}$ given the difference $\bar{x}_j - \bar{y}_j$ using (3.27). See Bak and Jensen (2016) for details. □

When $\delta_j = 0$, we see that the mean of γ_j^- and γ_j^+ are nonzero with opposite sign. By themselves these means are small because of the threshold term w_j. However, if most variables have no differential expression, these means are multiplied by a number of order m to get the means of $\gamma^{\pm}$. For m large we can therefore have a large difference in γ^- and γ^+ leading to a bias in the classifier with the majority group being favoured. This is illustrated in Table 3.3

					$\delta = 0$			$\delta = 1$		
n_1	n_2	m	k	κ	$E\gamma_j^-$	$E\gamma_j^+$	$E\tau_j^2$	$E\gamma_j^-$	$E\gamma_j^+$	$E\tau_j^2$
20	10	1000	10	2.0	0.02	-0.02	0.07	1.00	0.86	1.36

$E\gamma^-$	$E\gamma^+$	$E\tau^2$	$\frac{E\gamma^-}{2\sqrt{E\tau^2}}$	$\frac{E\gamma^+}{2\sqrt{E\tau^2}}$	$E\frac{\gamma^-}{2\sqrt{\tau^2}}$	$E\frac{\gamma^+}{2\sqrt{\tau^2}}$
26.96	-8.35	79.42	1.51	-0.47	1.51	-0.47

Table 3.3: Calculations for the case of independent variables. Except for the two last entries, mean values are calculated from the formulas in Proposition 3.27 with the T and S functions found from simulations. The two last entries are simulated directly. There are k variables with a scaled differential expression 1, the remaining $m-k$ variables having zero differential expression.

Exercise 3.28 (Simulated mean value of $T_{a,b}$)
Repeat the calculations in Table 3.3, except for the two last entries, with $n_1 = 30$, $n_2 = 10$, $\kappa = 2.5$, $\delta = 0.5, 1, 2$ and $m = 1000, 10000$. Use simulations to find the mean values $T_{1,1}$, $T_{2,1}$ and $T_{2,2}$. ■

Exercise 3.29 (Imbalance with independent variables)
Consider the same settings as in Problem 3.28. Use simulations to find the the means $E\big(\gamma^{\pm}/(2\tau)\big)$. Compare these means to the values $E(\gamma^{\pm})/\big(2\sqrt{E(\tau^2)}\big)$ found in Problem 3.28. ■

Exercise 3.30 (Analytical mean value of $T_{a,b}$)
Consider the case of no thresholding, $\kappa = 0$. Calculate $T_{1,1}$, $T_{2,1}$ and $T_{2,2}$ analytically.

In the general case calculate $T_{1,1}$, $T_{2,1}$ and $T_{2,2}$ by numerical integration with respect to the density of v. ■

Exercise 3.31 (Imbalance with dependent variables)
In this problem you will evaluate the bias of the independence classifier for the case of dependent variables. In this case τ^2 from (3.26) must be replaced by

$$\tau^2 = \hat{\Delta}_t \hat{D}^{-1} \Sigma \hat{D}^{-1} \hat{\Delta}_t = \sum_{i,j} \frac{\bar{x}_i - \bar{y}_i}{s_i^2} \cdot \frac{\bar{x}_j - \bar{y}_j}{s_j^2} w_i w_j \Sigma_{ij}. \tag{3.28}$$

Consider the setup of Exercise 1.10 with a block structure of Σ, the diagonal blocks being Σ_0 and the off diagonal blocks being zero. Then (3.28) becomes a sum over blocks, and to evaluate the contribution from a block one can use

$$a^{\mathrm{T}}\Sigma_0 a = (1-\rho)\sum_{i=1}^{10} a_i^2 + \rho\Big(\sum_{i=1}^{10} a_i\Big)^2,$$

since Σ_0 is one at the diagonal and ρ on all off diagonal entries.

Use simulations to find the the means $E(\gamma^{\pm}/(2\tau))$ for the setting in Problem 3.28. Compare these means to the values $E(\gamma^{\pm})/\big(2\sqrt{E(\tau^2)}\big)$. ∎

3.6.2 Bias corrected classifiers

To remove the bias from the variables with no differential expression we subtract the conditional mean as given in (3.27) and consider

$$B_0(z) = \sum_{j=1}^{m} \frac{\bar{x}_j - \bar{y}_j}{s_j^2}\Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) - \frac{\omega_n}{2}(\bar{x}_j - \bar{y}_j)\Big]w_j. \tag{3.29}$$

Proposition 3.32
Define, with $d = \bar{x} - \mu_1 + \bar{y} - \mu_2$,

$$\gamma^{\pm}_{B_0,j} = \frac{\bar{x}_j - \bar{y}_j}{s_j^2}\{\Delta_j \pm [d + \omega_n(\bar{x}_j - \bar{y}_j)]\}w_j, \qquad \tau^2_{B_0,j} = \tau_j^2 = \frac{(\bar{x}_j - \bar{y}_j)^2}{s_j^4}w_j.$$

Then, with $\delta_j = \Delta_j/\sigma_j$,

$$E(\gamma^{\pm}_{B_0,j}) = \delta_j\big[T_{1,1}(\delta_j, n_1, n_2) \pm S_{1,1}(\delta_j, n_1, n_2)\big].$$

Proof. See Bak and Jensen (2016). □

It is clear that when $\delta_j = 0$ we have $E\big(\gamma^-_{B_0,j}\big) = E\big(\gamma^+_{B_0,j}\big) = 0$ meaning that there is no bias from these terms. However, for variables with $\delta_j \neq 0$ we have a bias since $E\big(\gamma^+_{B_0,j}\big) - E(\gamma^-_{B_0,j}) = 2\omega_n\delta_j T_{1,1}(\delta_j, n_1, n_2)$, showing that the minority group is now favoured. To remove most of this bias we use a cross validation idea. Let $B_0(z; x_i)$ be the classifier based on the reduced dataset with x_i excluded, and similarly with $B_0(z; y_i)$. Having excluded x_i from the dataset we can use x_i to test the classifier. Define

$$\bar{\epsilon} = \frac{1}{2}\Big[\frac{1}{n_1}\sum_{i=1}^{n_1} B_0(x_i; x_i) + \frac{1}{n_2}\sum_{i=1}^{n_2} B_0(y_i; y_i)\Big],$$

and write $\bar{\epsilon} = \sum_{j=1}^{m}\bar{\epsilon}_j$ where $\bar{\epsilon}_j$ corresponds to the j'th term of $B_0(\cdot;\cdot)$ in the sum (3.29). We define our final BAI classifier (*bias adjusted independence* classifier) as

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

For the next proposition we introduce the functions

$$T_d(\delta, n_1, n_2) = T_{1,1}(\delta, n_1 - 1, n_2) - T_{1,1}(\delta, n_1, n_2 - 1),$$
$$S_d(\delta, n_1, n_2) = S_{1,1}(\delta, n_1, n_2) - \frac{1}{2}S_{1,1}(\delta, n_1 - 1, n_2) - \frac{1}{2}S_{1,1}(\delta, n_1, n_2 - 1).$$

Proposition 3.33
Define

$$\gamma^{\pm}_{B,j} = \gamma^{\pm}_{B_0,j} - 2\bar{\epsilon}_j, \qquad \tau^2_{B,j} = \tau^2_j = \frac{(\bar{x}_j - \bar{y}_j)^2}{s^4_j} w_j.$$

The mean values of $\gamma^-_{B,j}$ and $\gamma^+_{B,j}$ are

$$E\big(\gamma^{\pm}_{B,j}\big) = \delta_j \Big\{ T_{1,1}(\delta_j, n_1, n_2) \pm \frac{1}{2} T_d(\delta_j, n_1, n_2) \pm S_d(\delta_j, n_1, n_2) \Big\}.$$

Proof. We simply use the mean values from Proposition 3.32 for B_0, and the same formula for $\bar{\epsilon}_j$ with n_1 or n_2 replaced by $n_1 - 1$ and $n_2 - 1$. □

It is clear that when $\delta_j = 0$ we have $E\big(\gamma^-_{B,j}\big) = E\big(\gamma^+_{B,j}\big) = 0$ meaning that there is no bias from these terms. For the variables with differential expression there is a small bias left caused by $\bar{\epsilon}$ being based on one less observation than B_0. Thus, $T_d(\delta, n_1, n_2)$ and $S_d(\delta, n_1, n_2)$ are small, but not zero.

Bak and Jensen (2016) give the following asymptotic result to show the benefits of the new classifier. As before $a_n = 1/n_1 + 1/n_2$.

Proposition 3.34
Consider a situation with K variables having scaled differential expression δ and the remaining $m - K$ variables having no differential expression. Let n_1 and n_2 tend to infinity with the imbalance factor $\omega_n = (n_1 - n_2)/(n_1 + n_2) > 0$ fixed. Also, let δ and the threshold κ be fixed. Finally, let the number of variables m and K tend to infinity in such a way that $K/\sqrt{m a_n} \to \alpha$ for some $\alpha > 0$.

In this setup we find for the independence classifier that $\gamma^-_D/\tau_D \to \infty$ and $\gamma^+_D/\tau_D \to -\infty$, implying that the probability of correct classification tends to either one or zero according to the new observation coming from either group 1 or group 2.

For the BAI classifier γ^-_B/τ_B and γ^+_B/τ_B have the same limiting positive value.

Proof. See Bak and Jensen (2016). □

From simulations it is shown in Bak and Jensen (2016) that the ROAD classifier, which takes the correlation among the variables into account, suffers from bias in the imbalanced case also. The authors therefore suggest an adjustment parallel to the BAI classifier. First introduce R_0 in the same way as B_0:

$$R_0(z) = \sum_{j=1}^{m} r_j \Big[z_j - \frac{1}{2}(\bar{x}_j + \bar{y}_j) - \frac{\omega_n}{2}(\bar{x}_j - \bar{y}_j) \Big],$$

where r is the ROAD classification vector from (3.25). Next let $R_0(z;x_i)$ and $R_0(z;y_i)$ be the corresponding classifiers based on reduced dataset with x_i or y_i excluded. Define

$$\bar{\epsilon}_R = \frac{1}{2}\Big[\frac{1}{n_1}\sum_{i=1}^{n_1} R_0(x_i;x_i) + \frac{1}{n_2}\sum_{i=1}^{n_2} R_0(y_i;y_i)\Big],$$

Finally, the BA-ROAD classifier is defined as

$$B_R(z) = R_0(z) - \bar{\epsilon}_R.$$

Proposition 3.35
Define

$$\gamma^{\pm}_{R_0,j} = r_j\{\Delta_j \pm [d + \omega_n(\bar{x}_j - \bar{y}_j)]\}w_j,$$

and

$$\gamma^{\pm}_{B_R,j} = \gamma^{\pm}_{R_0,j} - 2\bar{\epsilon}_{Rj}.$$

Then we have for $\Delta_j = 0$

$$E\gamma^{-}_{B_R,j} = E\gamma^{+}_{B_R,j} = 0.$$

The proposition shows that the BA-ROAD classifier has no bias for the variables with no differential expression.

Exercise 3.36 (BAI classifier, independent variables)
Repeat the calculations in Table 3.3, except for the two last columns, for the BAI classifier (underlying variables being independent) with $n_1 = 30$, $n_2 = 10$, $\kappa = 2.5$, $\delta = 0.5, 1, 2$ and $m = 1000, 10000$. Use simulations to find the mean values $T_{1,1}$, $T_{2,1}$ and $T_{2,2}$. ■

Exercise 3.37 (BAI classifier, independent variables)
Simulate the means of $\gamma^{\pm}_B/(2\tau_B)$ for the BAI classifier (underlying variables being independent) for the setting in Exercise 3.36.

You need a complicated programme for this!

Exercise 3.38 (BAI classifier, dependent variables)
Simulate the means of $\gamma^{\pm}_B/(2\tau_B)$ for the BAI classifier with block structure of variance matrix (remember that tau is more complicated). Use the setting from Exercise 3.31 and τ^2_B given in (3.28).

You need a complicated programme for this!

Now, redo columns *D,fixed* and *BAI,fixed* in table 2 of Bak and Jensen (2016).

Exercise 3.39 (Breast cancer data)
Consider breast cancer data introduced in Exercise 1.27. Divide the data into a training set and a test set. The training set should have 45 women from the ER+ group and 14 from the ER- group, whereas the test set then has 20 ER+ women and 20 ER- women.

Find the thresholded independence classifier and the BAI classifier from the training data with a threshold chosen by you. Evaluate the classifier on the test set.

Make a plot comparing the two classification vectors. ■

Exercise 3.40 (Breast cancer data)
Consider the same data as in Exercise 3.39 and the division into a training set and a test set. Find the classifier for the threshold equal to $\kappa = 2.0, 2.2, \ldots, 4.0$, and find for each of these the test error on the test data. Select the value of κ that gives the smallest test error (both for the independence classifier and the BAI classifier). Record κ, test error for both groups and the number of variables selected.

Repeat this 100 times with random selection of the training set and test set. Make suitable plots of the results. ■

Exercise 3.41 (Open questions)
Can we improve on the method used for the BAI classifier, that is, for the first part B_0 ? We are making a correction that is correct for the case $\delta_j = 0$, but at the same time we are hoping that we exclude most of the cases with $\delta_j = 0$ using thresholding. Can we subtract a total mean instead of a local condiional correction ? Can we use empirical Bayes ideas like in Efrons two group model ?

Exercise 3.42 (Critique of BAI)
The BAI classifier first corrects bias from nonrelevant variables by a theoretical argument and then corrects bias from relevant variables via a crossvalidation principle. Is this actually "walking across the creek to get water" (direct tranlation of danish expression) ? Why not forget the first part and make the crossvalidation from the start ?

You should investigate this by simulation. Make your own simulation setting so that we afterwards can pool your simulation results to get a general picture.

Recall that the probabilitites of correct classification for a linear rule $G(a^{\mathrm{T}}(z - b) + c)$ are given as $\Phi(\gamma^{\pm}/(2\tau))$, (3.3), with $\gamma^{\pm} = a^{\mathrm{T}}(\Delta \pm (d - c))$, $d = 2b - \mu_1 - \mu_2$ and $\tau^2 = a^{\mathrm{T}}\Sigma a$. For the BAI classifier we have

$$a_j = w_j \frac{\bar{x}_j - \bar{y}_j}{s_j^2}, \qquad b_j = \frac{1}{2}(\bar{x}_j + \bar{y}_j$$

$$c = -\frac{1}{2n_1}\sum_{i=1}^{n_1}\sum_j a_j(-x_i)(x_{ij} - b_j(-x_i)) - \frac{1}{2n_2}\sum_{i=1}^{n_2}\sum_j a_j(-y_i)(y_{ij} - b_j(-y_i)),$$

where $a(-x_i)$ and $b(-x_i)$ are as a and b calculated from the dataset with x_i excluded, and similarly with $a(-y_i)$ and $b(-y_i)$.

When using crossvalidation on the thresholded independence classifier directly we have a and c as above, but

$$b_j = \frac{1}{2}(\bar{x}_j + \bar{y}_j).$$ ■

3.6.3 Mojiri et al.

The bias correction aims at having $E(\gamma_j^- - \gamma_j^+) = 0$. From Proposition 3.27 we have

$$\frac{1}{2}E(\gamma_j^- - \gamma_j^+) = S_{2,1}(\delta_j, n_1, n_2) - \delta_j S_{1,1}(\delta_j, n_1, n_2).$$

Consider now the case of no thresholding, $\kappa = 0$, where we get

$$\frac{1}{2}E(\gamma_j^- - \gamma_j^+) = E\Big(\frac{d^2}{v}\Big).$$

In other words, the bias is not dependeny on the unknown δ_j and can therefore be subtracted from the classifier.

Mojiri et al. (2022) use this to construct a classifier where the bias has been subtracted. The idea is to split the data into two equal sized subsets and calculate the means and variances for the classifier from one subset and calculate the thresholding from the other subset. This is then averaged over several random choices of the division into two subsets.

3.7 Algorithm

The following points on constrained minimization problems can be found in Boyd and Vandenberghe (2004) sections 5.2.3 and 5.5.3.

We consider the constrained minimization problem

$$\begin{aligned}&\text{minimize } \ f(x) \ \text{ subject to:}\\ &g_i(x) \le 0, \ i = 1,\dots,I, \ \ h_j(x) = 0, \ j = 1,\dots,J.\end{aligned}$$

This is sometimes called the *primal problem*. Let x^* be the point minimizing f subject to the constraints. The *Lagrangian* for the problem is

$$L(x,\mu,\lambda) = f(x) + \sum_{i=1}^{I}\mu_i g_i(x) + \sum_{j=1}^{J}\lambda_j h_j(x), \ \ \mu_i \ge 0, \ \lambda \in \mathbb{R}^J,$$

with *Lagrange dual function*

$$d(\mu,\lambda) = \inf_x L(x,\mu,\lambda).$$

The *dual problem* is to maximize $g(\mu,\lambda)$. The point where the maximum is attained is denoted (μ^*,λ^*). Since

$$d(\mu,\lambda) \le \inf_{x:g_i(x)\le 0, h_j(x)=0} L(x,\mu,\lambda) \le \inf_{x:g_i(x)\le 0, h_j(x)=0} f(x),$$

we always have $d(\mu^*,\lambda^*) \le f(x^*)$. The latter inequality is known as *weak duality*, and if $d(\mu^*,\lambda^*) = f(x^*)$ this is called strong duality. If there is strong duality and all

functions f, g_i and h_j are differentiable the solution to the minimization problem has to satisfy the Karush-Kuhn-Tucker conditions (KKT conditions):

$$\begin{aligned}
&\frac{\partial f}{\partial x}(x) + \sum_{i=1}^{I} \mu_i \frac{\partial g_i}{\partial x}(x) + \sum_{j=1}^{J} \lambda_j \frac{\partial h_j}{\partial x}(x) = 0, \\
&\mu_i g_i(x) = 0 \ \text{(complementary slackness)}, \\
&g_i(x) \leq 0, \ \ h_j(x) = 0 \ \text{(primal feasibility)}, \quad \mu_i \geq 0 \ \text{(dual feasibility)}.
\end{aligned}$$

Slaters condition says that strong duality holds if f and g_i, $i = 1,\ldots,I$ are all convex functions, h_j, $j = 1,\ldots,J$ are affine functions, and there exists a point x in the relative interior of the region where f and g_i are defined with $g_i(x) < 0$, $i = 1,\ldots,I$, and $h_j(x) = 0$, $j = 1,\ldots,J$. In this case the KKT conditions are not only necessary, but also sufficient for optimality of the solution.

3.7.1 Solving for the ROAD classifier

This subsection is mostly based on Wu et al. (2008). In the paper the authors introduce the *sparse linear discriminant analysis* (sLDA), which is the same as the ROAD classifier given through (3.23), and give an efficient algorithm for calculating the solution to (3.23). This is different from the implementation of the ROAD classifier given in Fan et al. (2012): they solve another related minimization problem giving an approximation to the solution of (3.23). When we speak of the ROAD classifier or the sLDA classifier we mean (the empirical version) of the linear classifier given through (3.23). Wu et al. (2008) is an unpublished report, but the sLDA classifier is described in Wu et al. (2009). Furthermore, the algorithm given in Wu et al. (2008) is based on the work of Rosset and Zhu (2007). Today ROAD is implemented in R in the TULIP package using coordinate descent as in glmnet, that is, using an algorithm different from the one described here.

Consider the minimization problem

$$\begin{aligned}
&\text{minimize} \ \ (w^+ - w^-)\Sigma(w^+ - w^-) \ \ \text{subject to:} \qquad (3.30) \\
&-w_i^+ \leq 0, \ \ -w^- \leq 0, \ \ \sum_{i=1}^{m}(w_i^+ + w_i^-) - c \leq 0, \ \ \sum_{i=1}^{m}(w_i^+ - w_i^-)\Delta_i - 1 = 0.
\end{aligned}$$

This corresponds to finding the ROAD classifier (3.23) by decomposing a vector w into its positive and negative part: $w = w^+ - w^-$ with $w_i^+ = w_i 1(w_i >= 0)$ and $w_i^- = -w_i 1(w_i < 0)$. For w^+ and w^- derived from a vector w we of course have $w_i^+ > 0$ implies $w_i^- = 0$ and vice versa: $w_i^+ w_i^- = 0$. In the minimization problem above w^+ and w^- are free variable. However, the solution we find to the minimization problem satisfies $w_i^+ w_i^- = 0$, so that we have found the ROAD classifier (3.23).

The Lagrangian of the problem (3.30) is

$$(w^+ - w^-)^{\mathrm{T}}\Sigma(w^+ - w^-) + \lambda(\Delta^{\mathrm{T}}(w^+ - w^-) - 1) - \sum_{i=1}^{m} \mu_i^+ w_i^+ - \sum_{i=1}^{m} \mu_i^- w_i^-$$
$$+ \mu\big[\sum_{i=1}^{m}(w_i^+ + w_i^-) - c\big], \quad \lambda \in \mathbb{R},\ \mu_i^+ \geq 0,\ \mu_i^- \geq 0,\ \mu \geq 0.$$

Letting $\Sigma_{(i)}$ be the i'th row of Σ, the KKT conditions become

$$2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i - \mu_i^+ + \mu = 0, \quad -2\Sigma_{(i)}(w^+ - w^-) - \lambda\Delta_i - \mu_i^- + \mu = 0, \quad i = 1,\ldots,m,$$
$$\mu_i^+ w_i^+ = 0, \quad \mu_i^- w_i^- = 0, \quad \mu\big[\sum_{i=1}^{m}(w_i^+ + w_i^-) - c\big] = 0,$$

together with the primal and dual feasibility. The idea is to find the solution for all c by starting with the smallest possible c, and showing that the solution path is piecewise linear in c.

From the KKT conditions we note the following facts. Since $\mu_i^+ \geq 0$ and $\mu_i^- \geq 0$ we have for all i

$$2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i = \mu_i^+ - \mu = \mu - \mu_i^- \ \Rightarrow\ \big|2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i\big| \leq \mu. \tag{3.31}$$

The case $\mu = 0$ corresponds to $w = (w^+ - w^-) = \Sigma^{-1}\Delta/\Delta^{\mathrm{T}}\Sigma^{-1}\Delta$ or $c > c_1$ from (3.24). For $c_0 < c < c_1$ we have $\mu > 0$ which is the case we consider now. Also, since $\mu_i^+ w_i^+ = 0$ and $\mu_i^- w_i^- = 0$ we have

$$\begin{aligned} w_i^+ > 0 \ &\Rightarrow\ 2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i = -\mu,\ \ \mu_i^- = 2\mu \ \text{ and } \ w_i^- = 0, \\ w_i^- > 0 \ &\Rightarrow\ 2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i = \mu,\ \ \mu_i^+ = 2\mu \ \text{ and } \ w_i^+ = 0. \end{aligned} \tag{3.32}$$

Our first conclusion is now that the solution to the minimization problem (3.30) gives the solution to the ROAD classifier (3.23) with $w = w^+ - w^-$.

Let A (as a function of c) be the set of nonzero variables, $A = \{i : w_i^+ + w_i^- > 0\}$. (When we come to the actual implementation of the algorithm, the set A is at certain break points the limit from the right of the set of nonzero variables.) From (3.32) we have, since $\mu > 0$, that

$$w_i \neq 0 \ \Rightarrow\ \operatorname{sign}(w_i) = -\operatorname{sign}\big(2\Sigma_{(i)}w + \lambda\Delta_i\big). \tag{3.33}$$

To use this we define $\xi_i = -\operatorname{sign}\big(2\Sigma_{(i)}w + \lambda\Delta_i\big)$, and for $i \in A$ we have from (3.32) that $2\Sigma_{(i)}(w^+ - w^-) + \lambda\Delta_i + \mu\xi_i = 0$ Let now w_A be the subvector of $w = w^+ - w^-$ with indices in A, Δ_A and ξ_A being subvectors as well, and Σ_{AA} the submatrix with indices in A. We then have the following set of equations

$$2\Sigma_{AA}w_A + \lambda\Delta_A + \mu\xi_A = 0, \quad \Delta_A^{\mathrm{T}}w_A = 1, \quad \xi_A^{\mathrm{T}}w_A = c,$$

where the last equation comes from the KKT conditions using that $\mu > 0$. Differentiating with respect to c we get

$$2\Sigma_{AA}\frac{\partial w_A}{\partial c} + \frac{\partial \lambda}{\partial c}\Delta_A + \frac{\partial \mu}{\partial c}\xi_A = 0, \quad \Delta_A^{\mathrm{T}}\frac{\partial w_A}{\partial c} = 0, \quad \xi_A^{\mathrm{T}}\frac{\partial w_A}{\partial c} = 1,$$

Solving for $\partial w_A/\partial c$ in the first equation and inserting into the two remaining equaitons we find

$$\frac{\partial\lambda}{\partial c}(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)+\frac{\partial\mu}{\partial c}(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)=0 \text{ and } \frac{\partial\lambda}{\partial c}(\xi_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)+\frac{\partial\mu}{\partial c}(\xi_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)=-2. \quad (3.34)$$

Solving these equations we find

$$\begin{aligned}\frac{\partial\mu}{\partial c}&=-\frac{2\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A}{(\xi_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)-(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)^2},\\ \frac{\partial\lambda}{\partial c}&=\frac{2\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A}{(\xi_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)-(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)^2},\\ \frac{\partial w_A}{\partial c}&=-\frac{1}{2}\Big(\frac{\partial\lambda}{\partial c}\Sigma_{AA}^{-1}\Delta_A+\frac{\partial\mu}{\partial c}\Sigma_{AA}^{-1}\xi_A\Big).\end{aligned} \quad (3.35)$$

For this solution we need $(\xi_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)-(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\xi_A)^2>0$. From the Cauchy-Schwarz inequality this is true unless $\Delta_A=a\xi_A$ for some number a. However, if $\Delta_A=a\xi_A$ this leads from (3.34) to the contradiction

$$0=(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)\Big[\frac{\partial\lambda}{\partial c}+a\frac{\partial\mu}{\partial c}\Big] \text{ and } -2=a(\Delta_A^{\mathrm{T}}\Sigma_{AA}^{-1}\Delta_A)\Big[\frac{\partial\lambda}{\partial c}+a\frac{\partial\mu}{\partial c}\Big]=a\cdot 0=0.$$

Since the right hand side of (3.35) does not depend on c, we have our second conclusion that the solution to the minimization problem (3.31) is piecewise linear in c.

We can follow a linear path until the KKT conditions are violated. This can happen in two ways. According to (3.32) we must stop if for some $i\in A^c$, $|2\Sigma_{(i)}w_A+\lambda\Delta_i|$ is about to become greater than μ. We then add this variable to the active set. Also, we must stop if a variable w_i, $i\in A$, becomes zero, since a change of sign violates (3.33) where there is no change of sign of $2\Sigma_{(i)}w_A+\lambda\Delta_i$ since the absolute value of this quantity is $\mu>0$. In this case the variable is removed from the active set. Let (w_A,λ,μ) be the value at c and (w_A',λ',μ') be the derivatives as given in (3.35). The largest possible stepsize d is given by

$$d=\min_i d_i, \quad (3.36)$$

with $d_i>0$ being the solution to

$$\begin{cases}\big|2\Sigma_{iA}(w_A+d_iw_A')+(\lambda+d_i\lambda')\Delta_i\big|=\mu+d_i\mu', & i\notin A,\\ w_i+d_iw_i'=0, & i\in A,\ \operatorname{sign}(w_i)\operatorname{sign}(w_i')=-1,\\ \infty, & i\in A,\ \text{otherwise.}\end{cases}$$

The algorithm starts at the lowest possible value of c as given in (3.24): $c\geq c_0$ with $c_0=1/|\Delta_I|$, where $|\Delta_I|=\max_i|\Delta_i|$. Assume first that I is unique: $|\Delta_i|<|\Delta_I|$ for all $i\neq I$. Then the initial solution is w_0 with $w_{0i}=0$ except $w_{0I}=1/\Delta_I$. As soon as c becomes larger than c_0 a better solution can be found including another variable.

Also, when $c > c_0$ the KKT conditions have to be satisfied, and so they are also valid in the limit $c \to c_0$. We must therefore have

$$\left|2\Sigma_{II}/\Delta_I + \lambda\Delta_I\right| = \mu \text{ and } \left|2\Sigma_{iI}/\Delta_I + \lambda\Delta_i\right| = \mu \text{ for some } i \neq I.$$

In order to simplify the presentation assume that $\Delta_I > 0$. Since, generally, $-\mu \leq 2\Sigma_{iI}/\Delta_I + \lambda\Delta_i \leq \mu$ the first equation above together with these inequalities gives

$$\mu \geq \mu_i = \begin{cases} \frac{2v_i}{\Delta_I - \Delta_i}, & v_i > 0, \\ \frac{-2v_i}{\Delta_I - \Delta_i}, & v_i < 0. \end{cases} \quad \text{where } v_i = \Sigma_{II}\frac{\Delta_i}{\Delta_I} - \Sigma_{iI}.$$

From this it follows that we must have

$$\mu = \max_{i \neq I} \mu_i, \quad \lambda = -\frac{2\Sigma_{II}/\Delta_I + \mu}{\Delta_I}, \tag{3.37}$$

and the active set A_0 at $c = c_0$ becomes

$$A_0 = \{I\} \cup \left\{i \neq I : \mu_i = \mu\right\}. \tag{3.38}$$

Algorithm 3.43 (sLDA)
The following algorithm calculates the entire solution path of the problem (3.31).

Initial step: *For the case* $\Delta_I > 0$, $|\Delta_i < \Delta_I$ *for all* $i \neq I$, *the inital active set,* μ *and* λ *are given by* (3.37) *and* (3.38), *and the inital soultion is* $w_i = (1/\Delta_I)\,1(i = I)$.

Linear part: *For a given* c *and active set* A *calculate the direction from* (3.34). *Take a step of size* d *in this direction, where* d *is calculated from* (3.36).

Updating: *If the minimum, defining the distance* d *in* (3.36), *is attained for* $i \in A$ *remove the corresponding variable from the active set, and if* $i \notin A$ *add the corresponding variable to the active set. If* $\mu = 0$ *stop the algorithm, otherwise go back to the* Linear part.

Exercise 3.44 (sLDA algorithm, initial value)
Assume that there are $K > 1$ variables j with $\Delta_j = \max_i |\Delta_i|$. Assume also that all rowsums of Σ_{KK}^{-1} are positive. Find the initial value of (w, λ, μ, A) in the sLDA algorithm. ■

Exercise 3.45 (sLDA algorithm, initial value)
Asume that $m = 3$ and

$$\Delta = (1,1,1)^{\mathrm{T}}, \quad \Sigma = \begin{pmatrix} 1 & 0.5 & 0.5 \\ 0.5 & 1 & -0.2 \\ 0.5 & -0.2 & 1 \end{pmatrix}.$$

Find the initial value of (w, λ, μ, A) in the sLDA algorihm. ■

When using the algorithm in practice we replace Σ by $\hat{\Sigma}$ and Δ by $\hat{\Delta}$, as in the definition (3.25). If we run the sLDA algorithm all the way to $\mu = 0$ we have the problem that $\hat{\Sigma}_{AA}$ in (3.35) is not invertible when the active set becomes too large, $|A| > n-2$. This problem is often solved by replacing $\hat{\Sigma}$ by $\hat{\Sigma} + \nu I$, where I is the identity matrix. In regression analysis this is known as *ridge regression*. We can also think of this as a penalty on large values of the quadratic norm $|w|_2 = \sum_i w_i^2$. In Wu et al. (2009) the authors use the value $\nu = 2\log(m)/n$.

Exercise 3.46 (sLDA algorithm, solution path)
Asume that $m = 4$ and

$$\Delta = (4,3,2,1)^{\mathrm{T}}, \quad \Sigma = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Find the complete solution path for (w, λ, μ, A) using the sLDA algorithm as much as possible. ■

Exercise 3.47 (Implementation of sLDA)
Implement the sLDA algorithm in R. Run your algorithm on a suitable dataset (either real dataset or simulated data) and compare your solution to the one given in the R-package TULIP.

See if you can make a computer experiment to measure the time used with your own algorithm and compare to the time used by the implementation in TULIP. ■

Exercise 3.48 (Colorectal cancer data)
Consider the data introduced in Exercise 1.28. Divide the data into a training set and a test set. The training set should have 14 MSI samples and 47 MSS samples, whereas the test set then has 20 MSI and 20 MSS samples.

Find the BAI classifier and the ROAD classifier from the training data with a threshold κ and L_1-bound c chosen by you. Evaluate the classifier on the test set.

Make a plot comparing the two classification vectors. ■

Exercise 3.49 (Colorectal cancer data)
Consider the same data as in Exercise 3.48. Find the BAI classifier for the threshold equal to $\kappa = 2.0, 2.2, \ldots, 4.0$, and the ROAD classifier for the L_1-bound $c = ...$, and find for each of these the test error on the test data. Select the value of κ and c that gives the smallest test error. Record κ, c, test error for both groups and the number of variables selected.

Repeat this 100 times with random selection of the training set and test set. Make suitable plots of the results. ■

3.7.2 Adaptive sLDA

The sLDA penalized the classification vector by the constraint $\sum_i |w_i| \leq c$, with the idea that many variables are set to zero. When comparing with the maximum likelihood classifier where $w_{\text{MLE}} = \Sigma^{-1}\Delta$ we are hoping that sLDA retains the large entries of w_{MLE}. The *adaptive sLDA* uses weights in the penalty in order to make it easier to keep the large entries of w_{MLE}. First an initial estimate of w_{MLE} is made. This could be $\hat{w}_0 = \hat{\Sigma}^{-1}\hat{\Delta}$, where a generalized inverse is used, or a ridge estimate

$$\hat{w}_0 = \left(\hat{\Sigma} + \nu I\right)^{-1}\hat{\Delta}.$$

We then use the new constraint

$$\sum_i \frac{|w_i|}{|\hat{w}_{0i}} \leq c.$$

We can stil use Algorithm 3.43 by replacing $\hat{\Sigma}$ by $\tilde{\Sigma}$ and $\hat{\Delta}$ by $\tilde{\Delta}$ where

$$\tilde{\Sigma}_{ij} = \hat{\Sigma}|w_{0i}||w_{0j}|, \qquad \tilde{\Delta}_i = \hat{\Delta}_i|w_{0i}|.$$

One can now make theoretical asymptotic results of the form: as n tends to infinity the adaptive procedure achieves oracle property. This means that the solution converge to the vector obtained when parameters are known.

3.8 List of classifiers

For an easy reference to the names of the different classifiers I give her a list of the classifiers. In this list

$$\Delta = \mu_1 - \mu_2, \quad D = \text{diag}(\sigma_1^2, \ldots, \sigma_m^2), \quad \hat{D} = \text{diag}(s_1^2, \ldots, s_m^2).$$

Name	Formula
ML, $m = 1$	$\xi(z) = G\Big(\frac{\mu_1-\mu_2}{\sigma^2}\big[z - \frac{1}{2}(\mu_1+\mu_2)\big]\Big)$
ML	$\xi(z) = G\Big(\Delta^{\mathrm{T}}\Sigma^{-1}\big[z - \frac{1}{2}(\mu_1+\mu_2)\big]\Big)$
ML, independence	$\xi(z) = G\Big(\Delta^{\mathrm{T}}D^{-1}\big[z - \frac{1}{2}(\mu_1+\mu_2)\big]\Big)$
Fisher's rule	$\xi(z) = G\Big((\bar{x}-\bar{y})^{\mathrm{T}}\hat{\Sigma}^{-1}\big[z - \frac{1}{2}(\bar{x}+\bar{y})\big]\Big)$
Independence classifier	$\xi(z) = G\Big((\bar{x}-\bar{y})^{\mathrm{T}}\hat{D}^{-1}\big[z - \frac{1}{2}(\bar{x}+\bar{y})\big]\Big)$
Thresholded IC	$\xi(z) = G\Big(\sum_{j=1}^{m}\frac{\bar{x}_j-\bar{y}_j}{s_j^2}\big[z_j - \frac{1}{2}(\bar{x}_j+\bar{y}_j)\big]w_j\Big)$
ROAD	$\xi(z) = G\Big(\hat{r}_c^{\mathrm{T}}\big[z - \frac{1}{2}(\bar{x}+\bar{y})\big]\Big)$
BAI	$\xi(z) = G\Big(\sum_{j=1}^{m}\Big\{\frac{\bar{x}_j-\bar{y}_j}{s_j^2}\big[z_j - \frac{1}{2}(\bar{x}_j+\bar{y}_j) - \frac{\omega_n}{2}(\bar{x}_j-\bar{y}_j)\big]w_j - \bar{\epsilon}_j\Big\}\Big)$

3.9 More exercises

Exercise 3.50 (Thresholded independence classifier)
Simulate values of $\gamma_{\mathrm{TI}}^{\pm}/(2\tau_{\mathrm{TI}}$ for the tresholded independence classifier with formulas given in (3.19). Consider a setting similar to the one in Exercise 3.18 with $\Delta = \mu_1 - \mu_2$ equal to 1 at the first k entries and zero otherwise, and with $m = 20, 40, 80, 160, 320$. Choose the threshold κ such that on average k out of m variables with no differential expression are included in the classifier.

Do a similar calculation where on average $2k$ variables are included in the classifier.

Try also a run where instead of selecting variables based on whether the t-statistic is above a threshold κ you select the k (or $2k$) variables having the highest absolute values of the t-statistics.

Also compare with the independence classifier where all variables are included

Exercise 3.51 (Leukemia data)
Construct the thresholded independence classfier for Efrons Leukemia data for a value of the threshold κ chosen by you.

Next, consider a set of κ values and perform leave one out cross validation (LOOCV) to evaluate each value. The output from the cross validation is the number of correctly classified samples. See exercise 3.55 below for description of LOOCV.

Which value of κ do you prefer.

Exercise 3.52 (Kruhøffer data: thresholded)
Use LOOCV to choose the threshold κ in the thresholded independence classifier for the Kruhøffer data.

Compare the number of variables used in the classifier with the number of variables found in the multiple testing problem considered in many of the previous exercises for the Kruhøffer data.

Exercise 3.53 (TULIP)
Download the R-package TULIP.

The ROAD function in the TULIP package finds

$$\hat{\beta} = \underset{\beta^T\hat{\Delta}/2=1}{\operatorname{argmin}}\{\beta^T\hat{\Sigma}\beta + \lambda|\beta|_1\}.$$

■

This is equivalent to the definition of the ROAD classifier in these notes in the sence that for each λ there is a corresponding c such that the solution $\hat{\beta}(\lambda)$ is identical to $2\hat{r}_c$ from (3.25) for a suitable value of c

Run the ROAD function on Efrons Leukemia data for a suitable set of λ-values.

Exercise 3.54 (Your own ROAD)
This exercise are for those that likes playing around with R!

Implement algorithm 3.43. Use your algorithm on Efrons Leukemia data for for a value of c chosen by you. Try to see if you can find the corresponding λ using the TULIP function ROAD.

Exercise 3.55 (Prediction using TULIP)
Having run the ROAD function from TULIP you have an $m \times K$ matrix B of β values, where K is the number of λ values in the call to the function.

Assume now that you have a test observation (a single observation) consisting of the vector X and corresponding class label in Z. Then you can calculate your classifier for each of the K values of λ by looking at the signs of

```
rbind(X-(barx+bary)/2)%*%B
```

where *barx* and *bary* are the averages in the two groups for the data in the call to the ROAD function.

Thus, if $Z = 1$, all the λ values for which the above vector entry is positive provides a correct classification.

You can use this to construct a function that finds the *Leave One Out Cross Validation error* (LOOCV). Thus, you must make a for-loop over the observations in the dataset, exclude one observation, run the ROAD function on the remaining $n-1$ observations, use the constructed β matrix to calculate the classifier for the observation left out, and register whether the classification is correct.

Perform LOOCV on Efrons Leukemia data.

Exercise 3.56 (Kruhøffer data: ROAD)
Use LOOCV to choose the regularizing parameter λ in the ROAD classifier as implemented in TULIP for the Kruhøffer data.

Compare the variables used in the classifier with the variables used in the thresholded independence classifier from exercise 3.52. In particular, you can look at the rank of the t-values for the variables chosen by the ROAD classifier.

CHAPTER

4 LASSO

Most of you have probably encountered backward selection and forward selection in a multiple regressions model. For a model with a small number of regression variables a backward selection method is often used. When the number of regressors becomes large the forward selection method is considered. One may think of the forward method as a very restrictive way of searching for relevant subsets of regressors, not being able to try out all possible subsets. When the number of regression variables becomes larger than the number of observations backward selection is not possible and so, previously, forward selection was the method used. In implementations one typically uses a more advanced stepwise procedure than the strict forward selection, where theforward and backward methods are blanded.

In modern methods, like LASSO regression, the variable selection has become part of the estimation process. Thus, during the estimation some of the regression parameters are set to zero corresponding to excluding these variables from the model. Typically these methods introduce bias in the estimation of the remaining parameters and in this chapter we will consider a small part of the research into the properties of these methods.

Consider the regression model

$$y_i \sim N\big(x_i^{\mathrm{T}}\beta, \sigma^2\big), \quad \beta, x_i \in \mathbb{R}^m, \quad i = 1, \ldots, n,$$

where the number of covariates m is large as compared to the number of samples n. Tibshirani (1996) introduced the *LASSO* as a regularized estimation of β, having the property that many coefficients of the estimate are set to zero. The LASSO estimate $\hat{\beta}$ is the solution to

$$\text{minimize} \sum_{i=1}^{n} \big(y_i - \beta^{\mathrm{T}} x_i\big)^2 \ \text{ subject to } \ \sum_{j=1}^{m} |\beta_j| \le c. \tag{4.1}$$

The acronym LASSO comes from *least absolute shrinkage and selection operator*. The Lagrangian formulation considers

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{\frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \beta^{\mathrm{T}} x_i\big)^2 + \lambda \sum_{i=1}^{n} |\beta_i|\}. \tag{4.2}$$

Any solution to (4.1) for a given c corresponds to a solution to (4.2) for a suitabla value of λ, and vice versa (it is obvious that the solution $\hat{\beta}(\lambda)$ to (4.2) also gives the solution to (4.1) with $c = \sum_i |\hat{\beta}_i(\lambda)|$).

We can copy the sLDA algorithm 3.43 to find the complete solution path of the LASSO minimization problem.

Algorithm 4.1 (LARS)
The following algorithm calculates the entire solution path of the problem (4.1).

Initial step: *Define* $\psi = \max_j \psi_j$ *with* $\psi_j = |2\sum_i y_i x_{ij}|$, *and* $A_0 = \{j : \psi_j = \psi\}$. *The algorithm starts at* $c = 0$ *with* $\beta = 0$, $\mu = \psi$ *and active set* A_0.

Linear part: *For a given* c, *active set* A *and solution* β_A *define* $S_A = \sum_i x_{iA} x_{iA}^T$ *and* $\xi_j = \operatorname{sign}\left(2\sum_i (y_i - x_{iA}^T \beta_A) x_{ij}\right)$. *Calculate the direction of* μ *and* β *as*

$$\frac{\partial \mu}{\partial c} = -\frac{2}{\xi_A^T S_A^{-1} \xi_A}, \qquad \frac{\partial \beta_A}{\partial c} = \frac{S_A^{-1} \xi_A}{\xi_A^T S_A^{-1} \xi_A},$$

and stepsize $d = \min_j d_j$ *with* $d_j > 0$ *being the solution to*

$$\begin{cases} \left|2\sum_i \left(y_i - x_{iA}^T(\beta_A + d_j \frac{\partial \beta_A}{\partial c})\right) x_{ij}\right| = \mu + d_j \frac{\partial \mu}{\partial c}, & j \notin A, \\ \beta_j + d_j \frac{\partial \beta_j}{\partial c} = 0, & j \in A, \ \operatorname{sign}(\beta_j)\operatorname{sign}(\frac{\partial \beta_j}{\partial c}) = -1, \\ \infty, & j \in A, \ \textit{otherwise.} \end{cases}$$

Take a step of size d.

Updating: *If the minimum, defining the distance* d *is attained for* $j \in A$ *remove the corresponding variable from the active set, and if* $j \notin A$ *add the corresponding variable to the active set. If* $\mu = 0$ *stop the algorithm, otherwise go back to the* Linear part.

The LARS algorithm (*Least Angle Regression*) was introduced in Efron et al. (2004). There are more than one algorithm in the paper with Algorithm 4.1 being the *LASSO modification of the LARS algorithm*. In the original LASSO paper Tibshirani (1996) a different algorithm was used for finding the regularized estimate.

Exercise 4.2
Understand the algorithm used in the R-package *glmnet*. From https://glmnet.stanford.edu/articles/glmnet.html I have taken the following quote: "The glmnet algorithms use cyclical coordinate descent, which successively optimizes the objective function over each parameter with others fixed, and cycles repeatedly until convergence. The package also makes use of the strong rules for efficient restriction of the active set. Due to highly efficient updates and techniques such as warm starts and active-set convergence, our algorithms can compute the solution path very quickly." ■

4.0.1 LASSO in high dimensional setting

In chapter 11 of Hastie et al. (2015) (book can be downloaded at https://hastie.su.domains/StatLearnSparsity/) you find a very readable account of results for the LASSO in the highdimensional sparse setting. The chapter is based on "modern"

techniques using non-asymptotic bounds and a minimum of probabilistic bounds. Convexity is a key property in the approach. To get a feeling for the beauty in the approach I am here listing the (few and simple) inequalities used in the proofs. I use the following notation for various norms of a k-dimensional vector $u = (u_1 \dots u_k)$,

$$\|u\|_\infty = \max_j |u_j|, \quad \|u\|_1 = \sum_j |u_j|, \quad \|u\|_2 = \sqrt{\sum_j |u_j|^2}.$$

The inequalities are

$$(i) : |a+b| \geq |a| - |b|, \text{ and therefore } \|u+v\|_1 \geq \|u\|_1 - \|v\|_1,$$
$$(ii) : |\sum_j u_j v_j| \leq \max_j |u_j| \sum_j |v_j| = \|u\|_\infty \cdot \|v\|_1,$$
$$(iii) : \sum_j |v_j| \leq \sqrt{k} \sqrt{\sum_j v_j^2} \text{ or } \|v\|_1 \leq \sqrt{k}\|v\|_2 \text{ (Cauchy-Swartz)},$$
$$(iv) : v^{\mathrm{T}} A v \begin{cases} \leq \lambda_{\max}(A)\|v\|_2^2, \\ \geq \lambda_{\min}(A)\|v\|_2^2, \end{cases} \quad A \text{ positive definite},$$
$$(v) : v^{\mathrm{T}} A^{-1} v \leq \lambda_{\min}(A)\|v\|_2^2,$$
$$(vi) : P(X \geq t) \leq e^{-t^2/(2\sigma^2)}, \ X \sim N(0,\sigma^2) \text{ (Gaussian tail)},$$
$$(vii) : P(\max_j |X_j| \geq t) \leq k \cdot \max_j P(|X_j| \geq t) \text{ (union bound or Boole's inequality)}.$$

Now, go and read the chapter!

Let us take here a look at the assumptions. Above x_i is the vector with covariate values for the i'th observation. Let X be the $n \times m$ matrix with rows x_i^{T}, $i+1, \dots, n$. The j'th column of X, that is, the vector with all the values of the j'th covariate is denoted x_{*j}. The *restricted eigenvalue* bound says that there exists $\gamma > 0$ such that

$$v^{\mathrm{T}} X^{\mathrm{T}} X v \geq \gamma \|v\|_2^2, \quad v \neq 0, \ v \in \mathscr{C}(S,3),$$

where

$$\mathscr{C}(S,3) = \{v \in \mathbf{R}^m | \|v_S^c\|_1 \leq 3\|v_S\|_1\},$$

and S is the set of indices with nonzero coordinates of the true parameter vector β^*. Thus we only consider vectors that do not vary wildly in the coordinates off the set S.

With this assumption one obtains, with a proper choise of λ_n, an upper bound on $\|\hat{\beta} - \beta^*\|_2$ that tends to zero with n and that holds with a probability tending to one.

One also obtains an upper bound on the prediction error $(1/n)\|X(\hat{\beta} - \beta^*)\|_2^2$.

When it comes to finding the set S of nonzero regression coefficients the situation is more complicated. Here one needs the *irrepresentability condition* (or mutual incoherence), which says that there exists $\gamma > 0$ such that

$$\max_{j \in S^c} \|(X_S^{\mathrm{T}} X_S)^{-1} X_S^{\mathrm{T}} x_{*j}\|_1 \leq 1 - \gamma,$$

where x_{*j} is a column of X. Intuitively, this is a condition on "near" orthogonality of columns x_j, $j \notin S$, with columns $j \in S$. This condition is hard to fulfill, see Figure 11.5 in Hastie et al. (2015). Apart from this condition it is required that columns are standardized ($\|x_j\|_2/\sqrt{n}$ is bounded) and that the smallest eigenvalue of $X_S^{\mathrm{T}} X_S/n$ is bounded away from zero. Then with a proper choise of λ_n one has on a set, with a probability tending to one, that $\hat{\beta}$ is unique, the support of $\hat{\beta}$ is included in S, and $\|\hat{\beta}_S - \beta_S^*\|_\infty$ is bounded by a term tending to zero.

The latter property implies that $\hat{\beta}_j$ is nonzero for those $j \in S$ for which $|\beta_j^*|$ is not too small.

4.1 Adaptive LASSO: theory

In this section I present Theorem 2 of Zou (2006)concerning the oracle properties of the adaptive LASSO. The setup is that we have independent variables $Y_i \sim N(x_i^{\mathrm{T}}\beta^*, \sigma^2)$, and $\hat{\beta}$ is the solution to

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \{ \sum_i (y_i - x_i^{\mathrm{T}}\beta)^2 + \lambda_n \sum_j \hat{w}_j |\beta_j| \}.$$

where $w_j = 1/|\hat{\beta}_j(\mathrm{ols})|^\gamma$. If Y_i is not normally distributed we extend the assumptions below with the condition that

$$\max_i \|x_i\|_2^2/n \to 0,$$

in order to use the Lindeberg-Feller central limit theorem. We use the notation

$$S = \{j : \beta_j^* \neq 0\} \text{ and } \hat{S} = \{j : \hat{\beta}_j \neq 0\},$$

and will for convenience take S to be $S = \{1 \ldots, m_0\}$.

Theorem 4.3 (Oracle property)
Assume that

i) $\frac{1}{n}\sum x_i x_i^T \to C$, *$C$ is positive definite;*

ii) $\lambda_n/\sqrt{n} \to 0$ *and* $\lambda_n n^{(\gamma-1)/2} \to \infty$.

Then we have variable selection consistency, $P(\hat{S} = S) \to 1$, and the usual asymptotic normality of $\hat{\beta}_S$,

$$\sqrt{n}\big(\hat{\beta}_S - \beta_S^*\big) \xrightarrow{\sim} N_{m_0}\big(0, \sigma^2 C_{11}^{-1}\big).$$

Proof. Let us introduce the notation $M_n = \sum_i x_i x_i^{\mathrm{T}}/n$, $U_n = \sum_i \epsilon_i x_i^{\mathrm{T}}/\sqrt{n}$, and M_n^0 and U_n^0 for the parts relates to coordinates in S.

We scale β by writing $\beta = \beta^* + u/\sqrt{n}$ and introduce $\Psi_n(u)$ by

$$\Psi_n(u) = \sum_i (y_i - x_i^{\mathrm{T}}\beta)^2 + \lambda_n \sum_j \hat{w}_j |\beta_j| = \sum_i (y_i - x_i^{\mathrm{T}}(\beta^* + \frac{u}{\sqrt{n}}))^2 + \lambda_n \sum_j \hat{w}_j |\beta_j^* + \frac{u_j}{\sqrt{n}}|,$$

and $\hat{\beta} = \beta^* + \hat{u}/\sqrt{n}$, or $\hat{u} = \sqrt{n}(\hat{\beta} - \beta^*)$, where $\hat{u} = \operatorname{argmin}_u \Psi_n(u)$. Writing $V_n(u) = \Psi_n(u) - \Psi_n(0)$ we find

$$V_n(u) = u^{\mathrm{T}} M_n u - 2U_n^{\mathrm{T}} u + \frac{\lambda_n}{\sqrt{n}} \sum_j \hat{w}_j \sqrt{n}\big(\big|\beta_j^* + \frac{u_j}{\sqrt{n}}\big| - |\beta_j^*|\big).$$

This is a convex function since each term in the sum is a convex function.

By assumption the first term converge to $u^{\mathrm{T}} C u$. For the second term we have

$$U_n \sim N_m\big(0, \sigma^2 M_n\big) \xrightarrow{d} N_m(0, \sigma^2 C),$$

if the error terms ϵ_i's are normally distributed, and more generally due to the Lindeberg-Feller central limit theorem when none of the terms x_i dominate.

For the third sum we first argue that

$$\frac{\lambda_n}{\sqrt{n}} \hat{w}_j \xrightarrow{P} \begin{cases} 0 & \text{for } j \in S, \\ \infty & \text{for } j \notin S. \end{cases} \tag{4.3}$$

For $j \in S$ we have $\beta_j^* \neq 0$ and by consistency of the OLS estimate this gives $\hat{w}_j \xrightarrow{P} |\beta_j^*|^{-\gamma}$, and since $\frac{\lambda_n}{\sqrt{n}} \to 0$ by assumption this gives the first part. When $\beta_j^* = 0$ we have that $\sqrt{n}\hat{\beta}_j(\text{ols})$ is stochastically bounded, and so $1/|\sqrt{n}\hat{\beta}_j(\text{ols})|$ is stochatically bounded away from zero. Since $(\lambda_n/\sqrt{n})\hat{w}_j = \lambda_n n^{(\gamma-1)/2} |\sqrt{n}\hat{\beta}_j(\text{ols})|^{-\gamma}$, and the first part tends to infinity by assumption, the whole terms tends to infinity in probability.

Furthermore for the third sum we also have

$$\sqrt{n}\big(\big|\beta_j^* + \frac{u_j}{\sqrt{n}}\big| - |\beta_j^*|\big) \to \begin{cases} u_j \operatorname{sign}(\beta_j^*) & \text{for } j \in S, \\ |u_j| & \text{for } j \notin S. \end{cases}$$

We are now ready to argue that $P(\hat{\beta}_j = 0, j \notin S) \to 1$. For u in a neighbourhood of zero the derivative of $u^{\mathrm{T}} M_n u - 2U_n/\sqrt{n} u$ is stochastically bounded since U_n is stochastically bounded and $M_n \to C$. For $j \notin S$ the increase in $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j |u_j|$ when moving away from $u_j = 0$ is proportional to $\frac{\lambda_n}{\sqrt{n}} \hat{w}_j$ which tends to infinity in probability ((4.3)). Thus the increase from the latter term cannot be balanced by a possible decrease in $u^{\mathrm{T}} M_n u - 2U_n/\sqrt{n} u$ and the minimum is attained at $u_j = 0$.

We can now consider the behaviour of u_S on the random set with $\hat{\beta}_j = 0$, $j \notin S$, this set having a probability tending to one. On this set u_S is the minimizer of

$$V_n^0(u) = u_S^{\mathrm{T}} M_n^0 u_S - 2U_n^{0\mathrm{T}} u_S + \frac{\lambda_n}{\sqrt{n}} \sum_{j \in S} \hat{w}_j \sqrt{n}\big(\big|\beta_j^* + \frac{u_j}{\sqrt{n}}\big| - |\beta_j^*|\big),$$

and this function converge in probability to

$$V(u) = u_S^{\mathrm{T}} C_{11} u_S - 2u_S^{\mathrm{T}} W_S$$

where $W_S \sim N_{m_0}(0, \sigma^2 C_{11})$. From Corollary A.4 we get that

$$\hat{u}_S \xrightarrow{d} N_{m_0}(0, \sigma^2 C_{11}^{-1}).$$

In particular this shows that for $j \in S$ we have that $P(\hat{\beta}_j = 0) \to 0$. □

When it comes to results for the adaptive LASSO in the high dimensional sparse setting results become rather difficult to read. I haven't found a "easy to read" chapter like for the LASSO mentioned above. A comprehensive treatment is given in Bühlmann and van de Geer (2011), but this book is not easy to read. A couple of articles of interest are Huang et al. (2008) and Zhou (2010).

The theoretical results mentioned in this chapter are indeed *theoretical*. In practice the penalizing parameter λ_n is chosen by crossvalidation and the predicition property of the model is also evaluated via crossvalidation.

Exercise 4.4 (Gazoline data)
Compare stepwise regression, LASSO and adapted LASSO for the gazoline data in terms of prediction accuracy (use the *glmnet* package in R and ??)

Appendices

APPENDIX

A Asymptotics

A.1 Small o and big O

If $f_n = o(g_n)$ this means that $f_n/g_n \to 0$ as $n \to \infty$. The expression $o(g_n)$ does not refer to a particular value, but to a property of the term as $n \to \infty$. If, as an example, one writes

$$f_n = o(g_n) = 2 \cdot o(g_n)$$

this does not mean that $x = 2x$ for some value x (which of course is only true when $x = 0$), but that $f_n/g_n \to 0$ implies $f_n/(2g_n) \to 0$.

If $f_n = O(g_n)$ this means that there exist a constant c and number n_0 such that $f_n/g_n \leq c$ for $n \geq n_0$.

A.2 Normal distribution function and t-distribution

Let Φ be the standard normal distribution function. Let $U \sim N(0,1)$, $V \sim \chi^2(n)/n$ and $t = \sqrt{b}(\delta + U/\sqrt{b})/\sqrt{V}$ with $b \geq c_1 n$ for some constant $c_1 > 0$. Then

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

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

(iii) There exists constants $a_1, a_2 > 0$ such that for $\delta = 0$ we have $P(|t| \geq \alpha\sqrt{b}) \leq a_1 \exp(-a_2\alpha^2 n)$.

(iv) There exists constants $a_1, a_2 > 0$ such that for $|\delta| > 2\alpha$ we have $P(|t| \leq \alpha\sqrt{b}) \leq a_1 \exp(-a_2\alpha^2 n)$.

The proof can be found in Bak et al. (2015). The result here relates to general tail and concentration bounds, see for example chapter 2 in Wainwright (2019).

A.3 Multivariate normal distribution

A stochastic vector $X = (X_1, \ldots, X_m)$ where the coordinates are indpendent normally distributed, $X_j \sim N(\mu_j, \sigma_j^2)$, is said to have a $N_m(\mu, D(\sigma))$ distribution. Here $\mu = (\mu_1, \ldots, \mu_m)$ is the mean vector, and $D(\sigma)$ is the variance matrix which is diagonal with σ_j^2 the j'th diagonal element.

Let $U \sim N_m(0, I)$, with I the identity matrix, and let B be a $k \times m$ matrix and μ a k-dimensional vector. Define $X = \mu + BU$. Then we have $E(X) = \mu$ and $\mathrm{Var}(X) = BB^{\mathrm{T}} = \Sigma$ and we write

$$X \sim N_k(\mu, \Sigma).$$

If $X_1, \ldots, X_n$ are independent and identically distributed, $X_i \sim N_m(\mu, \Sigma)$, the distribution of $\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^{\mathrm{T}}$ is a Wishart distribution $W_m(\Sigma, n-1)$.

A.4 Integration

For a discrete stochastic variable X with values in a set S you know that the mean value of $h(X)$ is $E(h(X)) = \sum_{x \in S} h(x)P(X = x)$. For a continuous random variable with density $f(x)$ the mean value is $E(h(X)) = \int h(x)f(x)dx$.

Now, if X both has a part that is discrete and a part that is continuous the mean value is a sum of two terms as above, where $\sum_{x \in S} P(X = x) + \int f(x)dx = 1$.

In measure theory one defines an integral with respect to the probability measure P such that it covers the three cases above. The general notation is

$$E(h(X)) = \int h(x)P(dx) = \sum_{x \in S} h(x)P(X = x) + \int h(x)f(x)dx.$$

A.5 Change of measure

In many cases you will see expression of the form

$$\int h(x)P(dx) = \int h(x)\frac{dP}{dP_0}(x)P_0(dx) \quad \text{or} \quad E(h(X)) = E_0\Big(h(X)\frac{dP}{dP_0}(X)\Big).$$

To understand these expressions let us start with the case of a discrete stochastic variable with values in a set S. We then write

$$E(h(X)) = \sum_{x \in S} h(x)P(X = x) = \sum_{x \in S} h(x)\frac{P(X = x)}{P_0(X = x)}P_0(X = x) = E_0\Big(h(X)\frac{dP}{dP_0}(X)\Big)$$

when

$$\frac{dP}{dP_0}(x) = \frac{P(X = x)}{P_0(X = x)}.$$

For this to be valid we must require that $P_0(X = x) > 0$ whenever $P(X = x) > 0$.

For a continuous random variable we write

$$E(h(X)) = \int h(x)f(x)dx = \int h(x)\frac{f(x)}{f_0(x)}f_0(x)dx = E_0\Big(h(X)\frac{dP}{dP_0}(X)\Big)$$

when

$$\frac{dP}{dP_0}(x) = \frac{f(x)}{f_0(x)}.$$

For this to be valid we must require that $f_0(x) > 0$ on the set $\{x|f(x) > 0\}$.

For the case where we both have a discrete component and a continuous component we get the same formula, now with

$$\frac{dP}{dP_0}(x) = \begin{cases} \frac{P(X=x)}{P_0(X=x)} & x \in S, \\ \frac{f(x)}{f_0(x)} & \text{otherwise.} \end{cases}$$

A.6 Convergence in probability

For a sequence of random variables X_n we want a quantification of the statement that X_n tends to zero as $n \to \infty$, $X_n \to 0$. One way of doing this is to require that the *probability* that X_n is close to zero tends to 1 as $n \to \infty$, or, equivalently, that the probability that X_n is not close to zero tends to 0.

Definition A.1
A sequence of random variables X_n tends to zero in probability *as $n \to \infty$ if, for all $\epsilon > 0$, we have that*

$$P(|X_n| > \epsilon) \to 0.$$

If we include the definition of convergence of a sequence of numbers we get: for all $\epsilon > 0$ and all $\delta > 0$ there exists $N(\epsilon, \delta)$ such that

$$P(|X_n| > \epsilon) < \delta \text{ for } n \geq N(\epsilon, \delta).$$

If X_n does *not* tend to zero in probability there exists $\epsilon > 0$ and $\delta > 0$ and a subsequence $t_1, t_2, \ldots$ such that

$$P(|X_{t_j}| > \epsilon) > \delta, \text{ for } j = 1, 2, \ldots$$

Proposition A.2
Consider a sequence of random variables X_n with $Var(X_n) \to 0$ as $n \to \infty$. Then $X_n - E(X_n)$ tends to zero in probability.

Proof. For any $\epsilon > 0$ we have

$$P(|X_n - E(X_n)| > \epsilon) = E\big(1(|X_n - E(X_n)| > \epsilon)\big) \leq E\big(1(|X_n - E(X_n)| > \epsilon)\frac{(X_n - E(X_n))^2}{\epsilon^2}\big)$$
$$\leq \frac{1}{\epsilon^2}\text{Var}(X_n), \qquad \square$$

which tends to zero as $n \to \infty$. This proves the result.

A.7 Convergence in distribution

A sequence X_n of random variables (values in $\mathbf{R}$) converge in distribution to a random variable X if

$$P(X_n \le x) \to P(X \le x) \text{ for all } x \text{ with } P(X = x) = 0.$$

You know this concept from the central limit theorem. For random k-dimensional vectors we consider instead $P(X_{n1} \le x_1, \ldots, X_{nk} \le x_k)$ for all $x = (x_1, \ldots, x_k)$ with zero probability of the boundary of the set $\{u : u_1 \le x_1, \ldots u_k \le x_k\}$. We write convergence in distribution as $X_n \xrightarrow{d} X$.

A very helpful tool is Slutsky's theorem. The theorem says that if X_n converge in distribution to X and if Y_n converge in probability to a constant c, then

$$X_n + Y_n \text{ converge in distribution to } X + c.$$

There are other similar results like $X_n Y_n \xrightarrow{d} Xc$ and $X_n / Y_n \xrightarrow{d} X/c$ when $c \neq 0$.

Slutsky's theorem follows from the *continuous mapping theorem* which states that $g(X_n) \xrightarrow{d} g(X)$ when $X_n \xrightarrow{d} X$ and g is a continuous function.

A.8 Convexity results

A.8.1 Uniform convergence of convex funtions

The formulation here is taken from Pollard (1991). We consider a sequnce of random convex function $A_n(\theta)$ defined on an open convex subset Θ of $\mathbf{R}^d$. Assume that $A_n(\theta)$ converge in probability to a real-valued function $A(\theta)$ for each $\theta \in \Theta$. Then for each compact subset K of Θ we have

$$\sup_{\theta \in K} |A_n(\theta) - A(\theta)| \xrightarrow{P} 0, \tag{A.1}$$

and the limit function $A(\theta)$ is convex. We express the result as *uniform convergence on compact subsets*.

There is a fairly simple proof in Pollard (1991). However, perhaps a slight clarification is needed. Pollard uses that the convexity of $A(\theta)$ (which follows easily from convexity of $A_n(\theta)$) implies continuity, but I believe a stronger result is needed like convexity implies local Lipschitz (see for example http://users.mat.unimi.it/users/libor/AnConvessa/continuity_all.pdf).

I will indicate the principle of the proof using the case $d = 1$. We start by choosing ϵ and next choose δ small such that $|A(x) - A(y)| \le \epsilon$ for $|x - y| \le \delta$ and x, y belonging to a compact set K. This can be done due to the convexity of $A(\theta)$. We next choose a finite set of points, θ_i, separated by δ to cover K. Figure A.1 shows four of these points. From the assumed pointwise convergence we have

$$\max_i \{|A_n(\theta_i) - A(\theta_i)|\} \xrightarrow{P} 0.$$

In particular, we can choose an ϵ_1 and have $|A_n(\theta_i) - A(\theta_i)| \leq \epsilon_1$ for all i with a probability tending to one. Figure A.1 shows the most extreme placement of $A_n(\theta_i)$ within the ϵ_1 distance from $A(\theta_)$, so as to allow for the smallest values of $A_n(\theta)$ in the interval between the two central θ_i values. The figure shows that in this interval

$$A_n(\theta) \geq A(\theta) - (2\epsilon + 3\epsilon_1).$$

Thus, we get a bound that is uniform in θ which gives the required result.

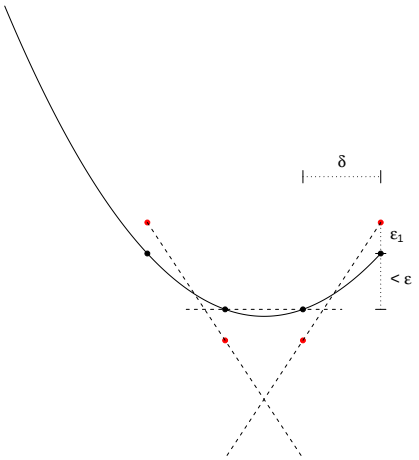

Figure A.1: Illustration for bounding difference between convex functions.

A.8.2 Concergence of argmin

This subsection gives two results from Hjort and Pollard (2011) that shows how convexity can be used in the study of estimators.

As above we have convex functions $A_n(\theta)$, $\theta \in \Theta$. Let $B_n(\theta)$ be a function, that we think of as an approximation to $A_n(\theta)$, amd let $B_n(\theta)$ have a unique minimizer β_n. Let α_n be a minimizer of $A_n(\theta)$.

Lemma A.3
Define for $\delta > 0$

$$\Delta_n(\delta) = \sup_{|\theta-\beta_n|\leq\delta} |A_n(\theta) - B_n(\theta), \quad h_n(\delta) = \inf_{|\theta-\beta_n|=\delta} B_n(\theta) - B_n(\beta_n),$$

then for any $\delta > 0$

$$P(|\alpha_n - \beta_n| \geq \delta) \leq P(\Delta_n(\delta) \geq \frac{1}{2} h_n(\delta)).$$

Proof. Consider θ with $l = |\theta - \beta_n| > \delta$ and write $\theta = \beta_n + lu$, where u then becomes a unit vector. Then $(\delta/l)\theta + (1-\delta/l)\beta_n = \beta_n + \delta u$ and convexity of A_n gives

$$A_n(\beta_n + \delta u) \le (\delta/l)A_n(\theta) + (1-\delta/l)A_n(\beta_n).$$

Rewriting this gives, with $r_n(\theta) = A_n(\theta) - B_n(\theta)$,

$$\begin{aligned}(\delta/l)\big(A_n(\theta) - A_n(\beta_n) &\ge A_n(\beta_n + \delta u) - A_n(\beta_n) && \text{(A.2)}\\ &= B_n(\beta_n + \delta u) + r_n(\beta_n + \delta u) - B_n(\beta_n) - r_n(\beta_n) && \text{(A.3)}\\ &\ge h_n(\delta) - 2\Delta_n(\delta). && \text{(A.4)}\end{aligned}$$

Now, if $\Delta_n(\delta) < \frac{1}{2}h_n(\delta)$ the right hand side is positive, showing that $A_n(\theta) > A_n(\beta_n)$ for all θ with $|\theta - \beta_n| > \delta$. This implies that the the value α_n minimizing $A_n(\theta)$ must be in the set $|\theta - \beta_n| \le \delta$. This proves the lemma. □

Consider the situation where the convex function $A_n(\theta)$ can be written as

$$A_n(\theta = \frac{1}{2}\theta^{\mathrm{T}} V\theta + U_n^{\mathrm{T}}\theta + C_n + r_n(\theta), \quad \theta \in \Theta = \mathbf{R}^d$$

where V is symmetric and positive definite, U_n is stochastically bounded (for any $\epsilon > 0$ we can find C_ϵ such that the probability $P(|U_n > C_\epsilon)$ is less than ϵ in the limit), C_n is arbitrary and $r_n(\theta)$ goes to zero in probability for each θ. Note that we here have $\Theta = \mathbf{R}^d$, which is not stated explicitly in Hjort and Pollard (2011), but seems to be needed for an argument below. Let $B_n(\theta) = \frac{1}{2}\theta^{\mathrm{T}} V\theta + U_n^{\mathrm{T}}\theta + C_n$ which is strictly convex with unique minimizer $\beta_n = -V^{-1}U_n$.

Corollary A.4
Let α_n be the argmin of $A_n(\theta)$. We have that $\alpha_n - (-V^{-1}U_n) \xrightarrow{P} 0$.

If $U_n \xrightarrow{d} U$ we have $\alpha_n \xrightarrow{d} -V^{-1}U$.

The result also holds for the case where V is replaced by V_n being a positive semidefinite symmetric matrix converging in probability to the positive definite matrix V.

Proof. Consider first $\bar{A}_n(\theta) = A_n(\theta) - U_n^{\mathrm{T}}\theta - C_n = \frac{1}{2}\theta^{\mathrm{T}} V\theta + r_n(\theta)$. Since $A_n(\theta)$ is convex so is $\bar{A}_n(\theta)$ and $\bar{A}_n(\theta) \xrightarrow{P} \frac{1}{2}\theta^{\mathrm{T}} V\theta$ for each θ. By (A.1) this convergence is uniform on compact sets, that is, $r_n(\theta)$ converges to zero uniformly on compact sets.

Stochastic boundedness of U_n implies stochastic boundedness of $\beta_n = -V^{-1}U_n$. From this we can deduce that

$$\Delta_n(\delta) = \sup_{|\theta-\beta_n|\le\delta} |A_n(\theta) - B_n(\theta)| = \sup_{|\theta-\beta_n|\le\delta} |r_n(\theta)| \xrightarrow{P} 0,$$

since we first bound β_n to a ball with radius C, say, with sufficient high probability, and next use the uniform convergence of $r_n(\theta)$ on a ball with radius $C + \delta$.

For $h_n(\delta$ from Lemma A.3 we get

$$h_n(\delta) = \inf_{|u|=1} \frac{1}{2}\delta^2 u^{\mathrm{T}} Vu = \frac{1}{2}\delta^2 \lambda_{\min},$$

where $\lambda_{\min}$ is the smallest eigenvalue of V. From Lemma A.3 we therefore have

$$P(|\alpha_n - \beta_n| \geq \delta) \leq P(\Delta_n(\delta) \geq \frac{1}{2}\delta^2 \lambda_{\min}) \to 0,$$

Using the convergence in probability to zero of $\Delta_n(\delta)$. We have ths proved that $\alpha_n - (-V^{-1}U_n)$ goes to zero in probability, and by Slutsky's theorem $\alpha_n \xrightarrow{d} -V^{-1}U$ when $U_n \xrightarrow{d} U$.

When V is replace by V_n we simply write $V_n = V + \eta_n$ and

$$\frac{1}{2}\theta^{\mathrm{T}} V_n \theta = \frac{1}{2}\theta^{\mathrm{T}} V \theta + \theta^{\mathrm{T}} V \eta_n + \frac{1}{2}\eta^{\mathrm{T}} V \eta.$$

The two last terms here can then be joined with $r_n(\theta)$ since $\eta_n \xrightarrow{P} 0$. □

A.9 Exercises

Exercise A.5 (Gaussian tail)
Show that

$$\bar{\Phi}(x) \leq \frac{1}{2} e^{-x^2/2}, \text{ for } x > 0.$$

Hint: write

$$\bar{\Phi}(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy,$$

and use the transformation $z = y - x$. ■

Exercise A.6 (Gaussian difference)
Show that

$$|\bar{\Phi}(x(1+\epsilon)) - \bar{\Phi}(x)| \leq \frac{\epsilon}{2}, \text{ for } x > 0 \text{ and } |\epsilon| < \frac{1}{2}.$$

Hint: the mean value theorem gives that

$$|\bar{\Phi}(x(1+\epsilon)) - \bar{\Phi}(x)| \leq x|\epsilon|\varphi(y),$$

for some y in between x and $x(1+\epsilon)$. Then

$$|\bar{\Phi}(x(1+\epsilon)) - \bar{\Phi}(x)| \leq 2|\epsilon| y\varphi(y).$$

Now show that

$$y\varphi(y) \leq 1 \cdot \varphi(1) \leq \frac{1}{4}$$

for $y > 0$. ■

Exercise A.7 (Lower tail of χ-squared)
Consider $V \sim \chi^2(n)/n$. Then $E(V) = 1$ and $\mathrm{Var}(V) = 2/n$. Also we know that the χ^2-distribution is a Gamma distribution.

1. Using the density of a Gamma-distribution, show that the density of V is
$$f_V(v) = \frac{(n/2)^{n/2} v^{n/2-1}}{\Gamma(n/2)} e^{-\frac{n}{2}v}.$$
Show that
$$E\big(e^{-sV}\big) = \frac{1}{\big(1+2s/n\big)^{n/2}}, \quad s>0,$$
by writing the mean value as an integral and recognize a Gamma density as part of the integrand.

2. Consider now
$$e^{s(1-a)} E\big(e^{-sV}\big) = e^{s(1-a)-\frac{n}{2}\log\left(1+2s/n\right)},$$
and show that the minimal value for $s>0$ is at $1+2s/n = 1/(1-a)$ or $s(1-a) = na/2$, and the minimum is $e^{\frac{n}{2}\left(a+\log(1-a)\right)}$.

3. Show that $a+\log(1-a) \le -a^2/2$ for $0 \le a < 1$.

4. Finally, use that
$$P(V < 1-a) \le \frac{E(e^{-sV})}{e^{-s(1-a)}}$$
to find an upper bound on $P(V \le 1-a)$. ∎

Exercise A.8 (Upper tail of χ-squared)
Consider $V \sim \chi^2(n)/n$. Show that for $a>0$
$$\begin{aligned} P(V \ge 1+a) &\le \frac{E(e^{sV})}{e^{s(1+a)}} = e^{-s(1+a)-\frac{n}{2}\log\left(1-2s/n\right)} \\ &\le e^{-\frac{n}{2}\left(a-\log(1+a)\right)} \le e^{-\frac{n}{4}a^2}. \end{aligned}$$
∎

Exercise A.9 (Upper tail of t-distribution)
Let $t = U/\sqrt{V}$ with $U \sim N(0,1)$ and $V \sim \chi^2(n)/n$, and with U and V independent. Consider, for $a>0$,
$$\begin{aligned} P(t > \sqrt{n}a) &= \int_0^\infty P(U > \sqrt{n}a\sqrt{v}) f_V(v)\,dv \\ &\le P(V < 1-\omega) + \bar{\Phi}\big(\sqrt{n}a(1-\omega)\big). \end{aligned}$$
Take $\omega^2 = a^2/2$ and use the exercises A.5 and A.7 to get
$$P(t > \sqrt{n}a) \le \frac{3}{2} e^{-a^2/8}$$
for $a < 1/\sqrt{2}$. ∎

Exercise A.10 (Non-central t-distribution)
Consider $t = \sqrt{n}(\delta + U/\sqrt{n})/\sqrt{V}$ with $U \sim N(0,1)$ and $V \sim \chi^2(n)/n$, and with U and V independent. Find an upper bound for $P(|t| \leq a\sqrt{n})$, when $|\delta| > 2a$, using ideas similar to the ones in exercise A.9.

Bibliography

Akey, J. M., S. Biswas, J. T. Leek, and J. D. Storey (2007). On the design and analysis of gene expression studies in human populations. *Nature genetics 39*(7), 807–809.

Allen, G. I. and R. Tibshirani (2012). Inference with transposable data: modelling the effects of row and column correlations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 74*(4), 721–743.

Bak, B. A., M. F. Grøn, and J. L. Jensen (2015, March). 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. Eprint: arXiv1405.5989*, 1–5.

Bak, B. A. and J. L. Jensen (2016). High dimensional classifiers in the imbalanced case. *Computational Statitics and Data Analysis 98*, 46–59.

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 (Methodological) 57*(1), pp. 289–300.

Benjamini, Y. and D. Yekutieli (2001). The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics 29*, 1165–1188.

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*. New York, NY, USA: Cambridge University Press.

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

Dyrskjøt, L., T. Thykjær, M. Kruhøffer, J. L. Jensen, N. Marcussen, S. Hamilton-Dutoit, H. Wolf, and T. F. Ørntoft (2003). Identifying distinct classes of bladder carcinoma using microarray. *Nature Genetics 33*, 90–96.

Efron, B. (2010a). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association 105*, 1042–1055.

Efron, B. (2010b). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *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.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics 7*(2), 179–188.

Ge, Y., S. Dudoit, and T. Speed (2003). Resampling-based multiple testing for microarray data analysis. *Test 12*(1), 1–77.

Genovese, C. R. and L. Wasserman (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association 101*(476), 1408–1417.

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.

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

Hastie, T., R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. New York: Chapman and Hall/CRC.

Hjort, N. L. and D. Pollard (2011). Asymptotics for minimisers of convex processes.

Hochberg, Y. and A. C. Tamhane (1987). *Multiple Comparison Procedures*. New York, NY, USA: John Wiley.

Holm, S. (1979). A simple sequential rejective multiple test procedure. *Scandinavian Journal of Statistics 6*, 65–70.

Huang, J., S. Ma, and C.-H. Zhang (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica 18*, 1603–1618.

Jogdeo, K. (1977). Association and probability inequalities. *The Annals of Statistics 5*(3), 495–504.

Korn, E. L., J. F. Troendle, L. M. McShane, and R. Simon (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference 124*(2), 379–398.

Kruhøffer, M., J. Jensen, P. Laiho, L. Dyrskjøt, R. Salovaara, D. Arango, K. Birkenkamp-Demtroder, F. Sørensen, L. Christensen, L. Buhl, J.-P. Mecklin, H. Järvinen, T. Thykjaer, F. Wikman, F. Bech-Knudsen, M. Juhola, N. Nupponen, S.Laurberg, C. Andersen, L. Aaltonen, and T. Ørntoft (2005). Gene expression signatures for colorectal cancer microsatellite status and hnpcc. *British Journal of Cancer 92*(12), pp. 2240–2248.

Lehmann, E. L. and J. P. Romano (2005, 06). Generalizations of the familywise error rate. *Ann. Statist. 33*(3), 1138–1154.

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

Mojiri, A., A. Khalili, and A. Zeinal Hamadani (2022). New hard-thresholding rules based on data splitting in high-dimensional imbalanced classification. *Electronic Journal of Statistics 16*, 814–861.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory 7*(2), 186–199.

Rosset, S. and J. Zhu (2007). Piecewise linear regularized solution paths. *The Annals of Statistics 35*(3), 1012 – 1030.

Sarkar, S. K. (2008). *On the Simes inequality and its generalization*, Volume Volume 1 of *Collections*, pp. 231–242. Beachwood, Ohio, USA: Institute of Mathematical Statistics.

Sidak, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association 62*(318), pp. 626–633.

Sidak, Z. (1971, 02). On probabilities of rectangles in multivariate student distributions: Their dependence on correlations. *Ann. Math. Statist. 42*(1), 169–175.

Singh, D., P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell 42*, 203–209.

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.

Spielman, R. S., L. A. Bastone, J. T. Burdick, M. Morley, W. J. Ewens, and V. G. Cheung (2007). Common genetic variants account for differences in gene expression among ethnic groups. *Nature Genetics 39*, pages 226–231.

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

Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America 98*(9), 5116–5121.

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Westfall, P. H. and S. S. Young (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley.

Wright, S. P. (1992). Adjusted p-values for simultaneous inference. *Biometrics 48*(4), 1005–1013.

Wu, M. C., L. Zhang, and X. Lin (2008). Two-group classfication using sparse linear discriminants analysis. Technical report, Department of Biostatistics, Harvard School of Public Health.

Wu, M. C., L. Zhang, Z. Wang, D. C. Christiani, and X. Lin (2009). Sparse linear discriminant analysis for simultaneous testing for the significance of a gene set/pathway and gene selection. *Bioinformatics 25*(9), 1145–1151.

Young, S. S., H. Bang, and K. Oktay (2009). Cereal-induced gender selection? most likely a multiple testing false positive. *Proceedings. Biological sciences 18*, 1211¿1214.

Young, S. S. and A. Karr (2011). Deming, data and observational studies. *Significance 8*(3), 116–120.

Zhou, S. (2010). Thresholded lasso for high dimensional variable selection and statistical estimation.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association 101*(476), 1418–1429.