universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

"Testing and model selection for prediction in large sets of variables."

Verfasserin

Alexandra Graf

angestrebter akademischer Grad

Magistra der Naturwissenschaften (Mag.rer.nat.)

Wien, im November 2008

Studienkennzahl lt. Studienblatt:	A 405
Studienrichtung lt. Studienblatt:	Mathematik
Betreuer:	o.Univ.Prof. Dr. Peter Bauer

Contents

List of Figures

List of Tables

1 Preface

1.1 Introduction

In gene expression or proteomic studies large numbers of variables are investigated. We generally can not assume that a few of the investigated variables show noticeable effects. Instead we often hope that there is at least a combination of several variables which, e.g., allow a prediction of the response of an individual patient to a particular therapy. The task of selecting useful variables with rather moderate effects from a very large number of candidates and estimating suitable scores to be used for the prediction of a clinical outcome (e.g. success of a specific therapy) in future patients is a hard exercise. Moreover, due to limited resources generally small sample sizes per variable are available which makes the problem even less tractable. For medical research reported in this field there is not always sufficient awareness of the statistical properties of the resulting prognostic scores. For instance, Ntzani and Ioannidis (2003) showed for prediction of cancer outcome that the constructed scores are poorly performing in external validation samples.

Subset selection procedures (e.g. Shao (1993), Miller (2002)) are widely used for such type of problem. However, there is a general problem of how to quantify the probability for falsely selecting variables not related to the clinical outcome. There have been proposals of estimating the positive false discovery rate in case that a nonzero model has been selected (Li and Hui (2007)). It is known that model selection by multiple testing of individual model parameters under fairly general conditions asymptotically is a consistent selection procedure: for increasing sample size the critical boundary for the univariate test statistics (the parameter estimate divided by its standard error) has to approach infinity at a smaller order than the inverse of the standard error (Bauer et al. (1988)). Asymptotic relationships between model selection procedures and multiple tests controlling the false discovery rate, i.e. the expected proportion of type I errors among all rejected hypotheses (FDR, see Benjamini and Hochberg (1995)), have been shown (Abramovich et al. (2006)). However such asymptotic results do not help how to tune the multiple test procedure in a specific sample in order to achieve good prediction of the outcome of a future patient.

1.2 Investigated problem

In this thesis we consider the following scenario: we want to search for predictors of a binary outcome (e.g. success of a therapy) among a large set of candidate variables (e.g. genes, proteins). Independent samples of patients responding and non-responding to the therapy are available (case-control study). Based on the given samples variables have to be selected and a score has to be constructed which will be used to predict response to therapy in future patients. The candidate variables are assumed to follow normal distributions.

We consider multiple tests controlling the false discovery rate (FDR) for the selection of variables. A linear score is estimated from the selected variables and its performance is assessed in terms of the statistical properties of the resulting receiver operating characteristic curve (ROC-curve). The area under the ROC-curve (AUC), a widely used measure how well a score can predict the clinical outcome is calculated varying the FDR level for selection, the number of candidate variables, per-group sample sizes and the number of prognostic variables related with the clinical outcome (alternatives).

We demonstrate that the threshold for the FDR which achieves the maximal AUC largely varies between different parameter constellations. Therefore we propose that cross validation is used to determine the FDR for the test based selection procedure optimal with regard to the AUC. It is investigated to what extend this optimization has an impact on the resulting FDR of the multiple test procedure.

A further typical data analytic approach used for such type of problem is the binary logistic regression. For comparison to the multiple test procedure we additionally investigate what can be achieved in terms of the AUC by using a stepwise (forward) binary logistic regression model for selecting variables and building a linear prediction score.

1.3 Outline of the thesis

This work is a continuation of the second part of my doctoral thesis in Statistics (A084 136) where the described procedures were only investigated under the simple assumptions of independence across hypotheses and known variance (compare Goll (2008)). In this diploma thesis, after a repetition of the results assuming the simple assumptions additionally a more sophisticated method to determine the optimal FDR has been used, another form of the cross validation procedure (using the mean difference in the score values) and extensions to the two-sided test situation as well as to the case of unequal effect sizes are discussed. Furthermore, results with respect to deviations from the underlying assumptions as unknown variance and correlation between hypotheses are investigated. Additionally the cross validation procedure is investigated for four real data sets.

First, an introduction to the general methodology is given in Chapter 2. Chapter 3 gives an overview of some results under the simple assumptions of independence across variables and known variance. Section 3.1 describes the basic assumptions, in Section 3.2 the selection methods (multiple test controlling the FDR and the binary logistic regression) are introduced and in Section 3.3 we explain the construction of a simple prediction score based on the selection methods. The results of the simulation studies for the multiple testing procedure for different parameter constellations can be seen in Section 3.5. Selection and prediction using a forward logistic regression model is discussed in Section 3.6. The situation under the global null hypothesis is described in Section 3.7. Using cross validation to determine selection boundaries for the multiple testing procedure by optimizing the AUC is discussed in Section 4 (compare also Goll (2008)). Chapter 5 gives some extensions as the two-sided test situation (Section 5.1) and the situation of unequal effect sizes (Section 5.2) among the alternatives again assuming independence across variables and known variance. Chapter 6 presents the results for the situation of unknown variances. The differences of the selection procedure and the prediction score as compared to the known variance case are discussed in Sections 6.1 and simulation results are given in Section 6.2. The cross validation procedure for the unknown variance case is discussed in Section 6.3. The situation of an autoregressive correlation structure between variables is discussed in Chapter 7. Selection of variables and the corresponding changes in the prediction score are discussed in Section 7.1. The corresponding simulation studies are given in Section 7.2. The cross validation procedure under the assumption of correlated hypotheses is discussed in Section 7.3. In Chapter 8 the investigated cross validation procedure is applied for four example data sets. A short discussion of the results is given in Chapter 9.

1.4 Publications

As mentioned above, this work is a continuation of the second part of my doctoral thesis:

> Goll (2008): Inference on a large number of hypotheses based on limited samples - some points to consider.

This thesis is based on the following submitted paper:

> Goll and Bauer (2008): Model selection based on the false discovery rate optimizing the area under the receiver operating characteristic curve.

A few results have been used and cited in

> Bauer (2008): Adaptive designs: looking for a needle in the haystack - a new challenge in medical research, *Statistics in Medicine*, 27: 1565-1580.

1.5 Availability

An R-program (R (2005)) for the cross validation procedure is available on:

http://statistics.msi.meduniwien.ac.at/index.php?page=page_ag_publications

These R-program can also be seen in the appendix.

2 General Methodology

2.1 Properties of the normal distribution

As mentioned in the preface, in this thesis we assume that the observed candidate variables (e.g. gene expression data or protein volumes) are normal distributed. In the following some important properties of the normal distribution, which may be used later in the thesis, are discussed. The given definitions, theorems and the corresponding proofs can be seen in e.g. Sachs (1999) and Anderson (2003).

Definition 2.1.0.1 *A random variable X is normal distributed with mean value μ and variance σ^2 (standard deviation σ) if the corresponding density is given as:*

$$f(x \mid \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu)}{\sigma}\right]^2\right\}.$$

This distribution is denoted by $N[\mu, \sigma^2]$.

Note that the turning points of $f(x \mid \mu, \sigma)$ are $\mu - \sigma$ and $\mu + \sigma$. Approximately 2/3 of all observations are lying within the two turning points. Note also that the mean value μ and the variance σ^2 are estimated by the sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the empirical variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Theorem 2.1.0.1 *$\bar{X}$ and S^2 are stochastically independent.*

Definition 2.1.0.2 *The normal distribution with mean value $\mu = 0$ and variance $\sigma^2 = 1$ is denoted as standard normal distribution ($N[0,1]$) with density:*

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp(\frac{-z^2}{2})$$

Theorem 2.1.0.2 *Let X be a random variable distributed according to $N[\mu, \sigma^2]$. The random variable*

$$Z = \frac{X - \mu}{\sigma}$$

is standard normal distributed ($Z \sim N[0,1]$).

Theorem 2.1.0.3 *The cumulative distribution function of the standard normal distribution can be calculated as:*

$$\Phi(z) = P[Z \leq z] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(\frac{-v^2}{2}) dv$$

In the thesis we may refer to the following properties of the standard normal distribution:

Theorem 2.1.0.4 *Properties of the standard normal distribution ($N[0,1]$):*

- *Because of the symmetry of the normal distribution:*
  $\Phi(-z) = 1 - \Phi(z)$
- $P[\mid Z \mid \leq z] = 2\Phi(z) - 1$ *and* $P[\mid Z \mid > z] = 2(1 - \Phi(z))$

Typical properties of the normal distribution ($N[\mu, \sigma^2]$) are summarized in the next theorem.

Theorem 2.1.0.5 *Let $X_i \sim N[\mu, \sigma^2]$ for $i = 1, ..., n$, then*

- $\sum_{i=1}^{n} X_i \sim N[n\mu, n\sigma^2]$
- $\frac{1}{n} \sum_{i=1}^{n} X_i \sim N[\mu, \sigma^2/n]$
- $\frac{X_i - \mu}{\sigma} \sim N[0,1]$
- $\frac{\bar{X} - \mu}{\sigma} \sqrt{n} \sim N[0,1]$
- $\sum_{i=1}^{n} (\frac{X_i - \mu}{\sigma})^2 \sim \chi_v^2$ *chi-square distributed with $v = n$ degrees of freedom*
- $S^2/\sigma^2(n-1) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2 \sim \chi_v^2$ *chi-square distributed with $v = n-1$ degrees of freedom*

The normal distribution also plays an important role for the approximation of other distributions.

Theorem 2.1.0.6 ***Central limit theorem:*** *Let $S_n = \sum_{i=1}^{n} X_i$ be the sum of n independent identical distributed random variables with the same expected value μ and the same variance σ^2 and let*

$$S_n^* = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

be the corresponding standardized random variable, then

$$\lim_{n\to\infty} P[S_n^* \leq z] = \Phi(z)$$

(convergence in distribution).

When moving to more than one candidate variable we need the following definition:

Definition 2.1.0.3 *Let $X = (X_1, ..., X_m)$ be random variables. The common distribution is called a m-dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ if the density is of the form:*

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{m/2} det(\sum)^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \sum^{-1}(\mathbf{x} - \boldsymbol{\mu})\right]$$

with Σ positive definite. We denote this distribution by $N[\boldsymbol{\mu}, \Sigma]$.

One important property of the m-dimensional normal distribution used in the thesis is:

Theorem 2.1.0.7 *If the m-dimensional random vector $X = (X_1, ..., X_m)$ is distributed according to $N[\boldsymbol{\mu}, \Sigma]$ then each linear combination with weights $\mathbf{c}^T = (c_1, ..., c_m)$*

$$Y = \mathbf{c}^T X = \sum_{i=1}^{m} c_i X_i$$

is distributed according to $N[\mathbf{c}^T \boldsymbol{\mu}, \mathbf{c}^T \Sigma \mathbf{c}]$.

2.2 Measures of accuracy for binary tests

Classification and prediction are fundamental components of clinical practice, e.g. a patient should be classified as a responder to a specific treatment or not. Classification errors can lead to serious consequences in medicine, e.g. a patient who would in fact response to the therapy but who was erroneously classified as non-responder may not receive the vital treatment. A patient who would in fact not response to the therapy but was erroneously classified as a responder will at a minimum undergo unnecessary medical procedures and emotional stress. The accuracy of such a diagnostic test that classifies a subject as either responder or non-responder can be defined in various ways discussed in the following. The given definitions, theorems and the corresponding proofs can be seen in Pepe (2003).

Let the binary variable R denote the true response status of a patient:

$$R = \begin{cases} 1 & \text{for response} \\ 0 & \text{for non-response} \end{cases} . \tag{2.1}$$

The variable Y denotes the result of a diagnostic test:

$$Y = \begin{cases} 1 & \text{positive for response} \\ 0 & \text{negative for response} \end{cases} . \tag{2.2}$$

If we know the truth, the result of the test can than be classified as true positive, true negative, false positive or false negative. Hence, the test can have two types of errors, false positive errors and false negative errors.

Definition 2.2.0.4 *Classification probabilities:*

- *false positive fraction*= $FPF = P[Y = 1 \mid R = 0]$
- *true negative fraction*= $TNF = P[Y = 0 \mid R = 0] = 1 - FPF$
- *true positive fraction*= $TPF = P[Y = 1 \mid R = 1]$
- *false negative fraction*= $FNF = P[Y = 0 \mid R = 1] = 1 - TPF$

TPF and TNF are also known as ***sensitivity*** *and* ***specificity****.*

Because $FNF = 1 - TPF$ the pair (FPF, TPF) defines the probabilities with which errors occur when using the given test. An ideal test clearly has $FPF = 0$ and $TPF = 1$. For a useless test on the other had, i.e. if response to therapy has no relation to the test outcome, $TPF = FPF$. Note that in context of statistical hypothesis testing of a null hypothesis ($R = 0$) versus an alternative hypothesis ($R = 1$) the terms significance level ($\alpha = FPF$) and statistical power ($1 - \beta = TPF$) are used (see next Section 2.3).

As an alternative, accuracy can be quantified by how well a test result predicts true response status.

Definition 2.2.0.5 *The predictive values are:*

- *positive predictive value*= $PPV = P[R = 1 \mid Y = 1]$
- *negative predictive value*= $NPV = P[R = 0 \mid Y = 0]$

Thus a perfect test will predict response perfectly with PPV=1 and NPV=1. On the other hand a useless test has no information about true response status and thus $P[R = 1 \mid Y = 1] = P[R = 1]$ and $P[R = 0 \mid Y = 0] = P[R = 0]$. We see that the predictive values depend not only on the performance of the test in responding and non-responding patients, but also on the prevalence of response. A low PPV may simply be a result of low prevalence of response or it may be due to a test that does not reflect the true response status of the patient very well. The classification probabilities, TPF and FPF, are considered more relevant to quantify the inherent accuracy of the test because they quantify how well the test reflects true response status. There is a direct relationship between predictive values and the classification probabilities.

Theorem 2.2.0.8 *Let* $\nu = P[R = 1]$ *and* $\tau = P[Y = 1]$ *then*

- $PPV = \nu TPF/(\nu TPF + (1 - \nu)FPF)$
- $NPV = (1 - \nu)(1 - FPF)/((1 - \nu)(1 - FPF) + \nu(1 - TPF))$
- $\tau = \nu TPF + (1 - \nu)FPF$
- $TPF = \tau PPV/(\tau PPV + (1 - \tau)(1 - NPV))$
- $FPF = \tau(1 - PPV)/(\tau(1 - PPV) + (1 - \tau)NPV)$
- $\nu = \tau PPV + (1 - \tau)(1 - NPV)$

Likelihood ratios are a further way of describing the prognostic or diagnostic value of a test.

Definition 2.2.0.6 *Diagnostic likelihood ratios (DLR) are defined as:*

- *positive DLR=DLR*$^{+} = P[Y = 1 \mid R = 1]/P[Y = 1 \mid R = 0]$
- *negative DLR=DLR*$^{-} = P[Y = 0 \mid R = 1]/P[Y = 0 \mid R = 0]$

They are the ratios of the likelihood of the observed test result in the responding versus the non-responding populations. An uninformative test having no relation to response status has DLRs of unity. On the other hand, a perfect test, for which $Y = R$ with probability 1 has DLR parameters $DLR^{+} = \infty$ and $DLR^{-} = 0$. A $DLR^{+} > 1$ indicates that a positive test is more likely in a responding subject than in a non-responding subject. Similarly with $DLR^{-} \leq 1$.

Consider now the odds that a subject response to therapy before the test is performed, i.e. in absence of the test result.

Definition 2.2.0.7 *The pre-test odds are defined as:*

$$pre - test \quad odds = \frac{P[R = 1]}{1 - P[R = 1]} = \frac{P[R = 1]}{P[R = 0]}$$

After the test is performed, i.e. with knowledge of the test results, the odds of response are:

Definition 2.2.0.8 *The post-test odds are defined as:*

$$post - test \quad odds(Y = y) = \frac{P[R = 1 \mid Y = y]}{P[R = 0 \mid Y = y]}$$

where $y = 0$ or 1.

Some relationships between DLRs, predictive values, classification probabilities and odds are discussed in the following .

Theorem 2.2.0.9 *The following results hold:*

- *post-test odds (Y=1)= $DLR^{+}\times$ (pre-test odds)*
- *post-test odds (Y=0)= $DLR^{-}\times$ (pre-test odds)*

Theorem 2.2.0.10 *The following results hold:*

- *post-test odds (Y=1)=PPV /(1-PPV)*
- *post-test odds (Y=0)=(1-NPV)/ NPV*

Theorem 2.2.0.11 *The following results hold:*

- $DLR^{+} = TPF/FPF$
- $DLR^{-} = (1 - TPF)/(1 - FPF)$

A single index of classifier performance commonly used in medicine is the odds ratio (ratio of post-test odds).

Theorem 2.2.0.12 *The odds ratio can be written as:*

$$OR = \frac{post - test \quad odds(Y=1)}{post - test \quad odds(Y=0)} = DLR^{+}\frac{1}{DLR^{-}} =$$
$$= \frac{PPV}{(1-PPV)}\frac{NPV}{(1-NPV)} = \frac{TPF}{FPF}\frac{(1-FPF)}{(1-TPF)}$$

A single odds ratio value can result from a wide variety of classification performances (FPF, TPF). For example an odds ratio of 36 results from $(FPF = 0.1, TPF = 0.8)$ which might be considered "good" classification or from $(FPF = 0.5, TPF = 0.973)$ which is likely considered "poor" classification (see e.g. Pepe and Thompson (2002), Pepe et al. (2004)).

2.3 Multiple tests controlling the false discovery rate

2.3.1 Error rates for multiple testing

Binary responses are commonly studied in medical and epidemiologic research, e.g. the response to a particular therapy. To find the variables related to the clinical outcome among a large set of candidate genes one may apply a multiple test procedure.

As mentioned in the last section, when a single null hypotheses H is tested, a type I error, that is rejecting the hypotheses, when it is in fact true (a false positive decision) may occur. A standard approach is to specify an acceptable level α for the probability of a type I error (significance level). Let $H = 0$ if the null hypotheses is in fact true, and $H = 1$ if the alternative holds. The control of a specified type I error probability α can be achieved by choosing a critical value c_α such that $P[T \geq c_\alpha \mid H = 0] \leq \alpha$, where T is the corresponding test statistic for hypothesis H. The hypothesis H is rejected if $T \geq c_\alpha$.

If the hypothesis is accepted, although in fact the alternative holds, a type II error occurs (a false negative decision). The probability of a type II error is: $\beta = P[T < c_\alpha \mid H = 1]$.

Table 2.1: Possible outcomes after a multiple testing procedure

Number of	not rejected	rejected	Total
True null hypotheses	TN	FP	$m\pi_0 = m - m_e$
False null hypotheses	FN	TP	$m(1-\pi_0) = m_e$
Total	m-R	R	m

Multiple testing refers to the testing of more than one hypothesis at the same time. For example in gene expression or proteomic studies thousands of hypotheses are tested simultaneously. Since the probability of at least one type I error increases with the number of hypotheses, in such studies large multiplicity problems occur. Table 2.1 shows the possible outcome after a multiple testing procedure. Consider the problem of testing simultaneously m null hypotheses H_i, $i = 1, ..., m$ and denote by R the number of rejected hypotheses among the m hypotheses. Assume that there are $m\pi_0 (= m - m_e)$ true null hypotheses among all m hypotheses. The proportion of true null hypotheses π_0 is an unknown parameter. The number of rejected hypotheses R is an observed random variable and TP (number of true positive decisions), FN (number of false negative decisions), TN (number of true negative decisions) and FP (number of false positive decisions) are unobservable random variables.

Two common error rates used to control the type I error are:

Definition 2.3.1.1 *The **Family Wise Error Rate (FWER)** is defined as the probability of at least one type I error:*

$$FWER = P[FP \geq 1],$$

were FP is the number of rejected true null hypotheses (false positives).

Definition 2.3.1.2 *The **False Discovery Rate (FDR)** is the expected proportion of type I errors among the rejected hypotheses:*

$$FDR = \boldsymbol{E}\left[\frac{FP}{R} \mid R > 0\right] P(R > 0) = \boldsymbol{E}\left[\frac{V}{\max(R,1)}\right]$$

where FP is the again the number of false positives and R denotes the number of rejected hypotheses. The effect $\max(R,1)$ *in the denominator is to set* $FP/R = 0$ *if* $R = 0$ *(compare Benjamini and Hochberg (1995)).*

A multiple testing procedure is said to control a particular type I error rate at level α if this error rate is less than or equal to α. There is a distinction between strong and weak control of a type I error rate. Strong control refers to the control of the type I error rate under any combination of true and false null hypotheses. In contrast, weak control refers to the control of the type I error rate only under the global null hypothesis, that is when all null hypotheses are in fact true. Weak control is unsatisfactory, because in reality, some null hypotheses may be true and others false, but the subset of true null hypotheses is unknown. Strong control ensures that the type I error rate is controlled under the unknown combination of true and false null hypotheses.

The following properties of the FDR were shown in Benjamini and Hochberg (1995):

Theorem 2.3.1.1 *Properties of the FDR:*

- *Under the complete null hypotheses (if all null hypotheses are true: $m_e = 0$), the FDR is equivalent to the FWER. Therefore control of the FDR implies control of the FWER in the weak sense.*

- *If $\pi_0 < 1$, the FDR is smaller than or equal to the FWER.*

As a result of theorem 2.3.1.1, any procedure that controls the FWER also controls the FDR. Procedures that control the FWER are more conservative, that is, lead to fewer rejections than those controlling the FDR. If a procedure only controls the FDR, more type I errors but less type II errors occur and thus, the power of the procedure may be increased. In the long run there is always a fraction of at most α true null hypotheses among the rejected hypotheses.

Within the class of multiple testing procedures that control a given type I error rate at an acceptable level α, one seeks for test procedures that maximize the power $(1 - \beta)$, that is, minimize the type II error rate (β). As with type I error rates, the concept of power can be generalized when moving from single to multiple hypotheses testing.

Definition 2.3.1.3 *Under the assumption of a common alternative (as considered in the* *following) the power is the expected fraction of null hypotheses correctly rejected*

$$1 - \beta = \frac{E[TP]}{m_e} = \frac{E[TP]}{m(1 - \pi_0)}.$$

2.3.2 Procedures controlling the FDR

In the genomic or proteomic setting, where thousands of tests are performed simultaneously and only a small number of genes or proteins are expected to be differentially expressed, FDR controlling procedures present a promising alternative to FWER approaches (as e.g. the Bonferroni correction or the Bonferroni-Holm procedure). In such situations, controlling the FWER can lead to unduly conservative procedures. One may tolerate some type I errors, provided their number is small in comparison to the number of rejected hypotheses. The FDR offers a less strict multiple testing criterion than the FWER.

Two approaches to provide FDR controlling procedures are the following: One is to fix the acceptable FDR level beforehand, and find a data-dependent thresholding rule so that the FDR of this rule is less than or equal to the pre-chosen level. This is the approach taken by Benjamini and Hochberg (1995). Another is to fix the thresholding rule and form an estimate of the FDR whose expectation is greater than or equal to the true FDR over that significance region. This is the approach taken by Storey (2002). These two procedures are discussed in the following.

The Benjamini-Hochberg procedure

Benjamini and Hochberg (1995) derived the following step-up procedure for strong control of the FDR for independent test statistics. In contrast to step-down procedures, step-up procedures begin with the largest p-value. Benjamini and Hochberg proved that the following procedure controls the FDR at a pre-chosen level α when the p-values following the null distribution are independent and uniformly distributed.

Definition 2.3.2.1 *The method of Benjamini and Hochberg proceeds as follows:*

1. *Let $p_1 \leq ... \leq p_m$ denote the observed ordered p-values corresponding to the hypotheses $H_1, ..., H_m$.*

2. *For the control of the FDR at level α calculate* $\hat{k} = \max\{1 \leq k \leq m : p_k \leq \frac{k}{m}\alpha\}$.

3. *If $\hat{k}$ exists, then reject the null hypotheses H_j for $j = 1, ..., \hat{k}$ corresponding to $p_1 \leq ... \leq p_{\hat{k}}$. Otherwise, reject nothing.*

Theorem 2.3.2.1 *For independent test statistics and for any configuration of false null hypotheses, the above procedure controls the FDR at level α.*

The proof of theorem 2.3.2.1 can be found in Benjamini and Hochberg (1995). It was also shown by Storey, Taylor and Siegmund (2004) that the Benjamini-Hochberg procedure controls the FDR in the strong sense. Benjamini and Yekutieli (2001) proved that this procedure also controls the FDR when the test statistics have positive dependency on each of the test statistics corresponding to the true null hypothesis. They also proposed, referring to Hommel (1988), a simple conservative modification of the procedure, replacing $\alpha k/m$ with $\alpha k/(m \sum_{j=1}^{m} \frac{1}{j})$ in the second step, which provides FDR control under arbitrary dependence structures (see also Dudoit et al. (2003)).

The Benjamini-Hochberg procedure was originally introduced by Simes (1986) to weakly control the FWER when all p-values are independent, although it happens to provide strong control of the FDR.

Storey's procedure

As mentioned before, instead of fixing α and estimating the rejection region, Storey (2002) fixed the rejection region and then estimated the FDR. Storey's method uses information about π_0, which yields a less stringent procedure and more power, while maintaining strong control. Typically the power of a multiple test procedure decreases with increasing m. But the larger m, the more information about π_0 is obtained.

Again m identical hypothesis tests $H_1, ..., H_m$ are performed with independent test statistics $T_1, ..., T_m$. Let $H_i = 0$ when the null hypothesis i is true and $H_i = 1$ otherwise. It is assumed that the test statistics under the true null $T_i|(H_i = 0)$ and under the alternative hypothesis $T_i|(H_i = 1)$ are identically distributed. It is further assumed that the same rejection region is used for each test. Finally it is assumed, that the H_i are independent Bernoulli random variables with $P[H_i = 0] = \pi_0$ and $P[H_i = 1] = 1 - \pi_0 = \pi_1$. Let Γ be the common rejection region for each hypothesis test.

Theorem 2.3.2.2 *Under the above assumptions the FDR can be written as:*

$$\begin{aligned} FDR = P[H = 0 \mid T \in \Gamma] &= \frac{\pi_0 P[T \in \Gamma \mid H = 0]}{\pi_0 P[T \in \Gamma \mid H = 0] + \pi_1 P[T \in \Gamma \mid H = 1]} \\ &= \frac{\pi_0 P[T \in \Gamma \mid H = 0]}{P[T \in \Gamma]} \end{aligned} \tag{2.3}$$

In the following hypotheses are rejected on the basis of independent p-values. For rejections based on p-values, all rejection regions are of the form $[0, \gamma]$ for some $\gamma \geq 0$.

Theorem 2.3.2.3 *In terms of p-values the above result can be written as:*

$$FDR(\gamma) = \frac{\pi_0 P[p \leq \gamma \mid H = 0]}{P[p \leq \gamma]} = \frac{\pi_0 \gamma}{P[p \leq \gamma]} \tag{2.4}$$

where p is the random p-value resulting from any test.

Since π_0 is an unknown parameter, it has to be estimated. Storey (2002) proposed the following conservative estimate of π_0:

Definition 2.3.2.2 *The proportion of true null hypotheses π_0 is estimated by:*

$$\hat{\pi}_0(\lambda) = \frac{\sharp\{p_i > \lambda\}}{(1 - \lambda)m} = \frac{W(\lambda)}{(1 - \lambda)m} \tag{2.5}$$

for some well-chosen λ, where $p_1, ..., p_m$ are the observed p-values, and $W(\lambda) = \sharp\{p_i > \lambda\}$ is the number of observed p-values exceeding λ. For a small proportion of null hypotheses this estimator can be larger than 1, thus in this cases it is set to 1.

The argument for the choice of the estimator $\hat{\pi}_0(\lambda)$ he explained as follows: As long as each test has reasonable power the large p-values are most likely to come from the true null hypothesis. Therefore for a well chosen λ, it is expected, that $\pi_0(1 - \lambda)$ of the p-values lie in the interval $(\lambda, 1]$, because the p-values under the true null hypotheses are uniformly distributed. Therefore $W(\lambda)/m \approx \pi_0(1 - \lambda)$, where $\mathbf{E}[\hat{\pi}_0(\lambda)] \geq \pi_0$ when the p-values corresponding to the true null hypotheses are uniformly distributed.

There is an inherent bias-variance trade off in the choice of λ. When λ gets smaller, the bias of $\hat{\pi}_0$ gets larger, but the variance gets smaller. Choosing a larger λ reduces the bias at the cost of higher variance (Storey et al. (2004)). Therefore, λ can be chosen to try to balance this trade-off. Storey (2002) optimized the value for λ to minimize the mean squared error of the estimate with bootstrap methods. However, simulations showed that when choosing a non-optimal λ the difference in their true mean-squared errors is not very drastic. For his calculations he used $\lambda = 0.5$. For our calculations we will also use $\lambda = 0.5$

It is now assumed that λ is fixed.

Definition 2.3.2.3 *An estimate of $P[p \leq \gamma]$ is:*

$$\widehat{P}[p \leq \gamma] = \frac{\sharp\{p_i \leq \gamma\}}{m} = \frac{R(\gamma)}{m}$$

where $R(\gamma) = \sharp\{p_i \leq \gamma\}$.

Theorem 2.3.2.4 *The estimate for the FDR can be calculated as:*

$$\widehat{FDR}_\lambda(\gamma) = \frac{\hat{\pi}_0(\lambda)\gamma}{\widehat{P}(p \leq \gamma)} = \frac{W(\lambda)\gamma}{(1-\lambda)\max\{R(\gamma), 1\}} \tag{2.6}$$

If $\widehat{FDR}_\lambda(\gamma) > 1$ Storey suggest setting $\widehat{FDR}_\lambda(\gamma) = 1$.

The following important result was proven by Storey, Taylor and Siegmund (2004).

Theorem 2.3.2.5 *Suppose that the p-values corresponding to the true null hypotheses are independent and uniformly distributed. Then for fixed $\lambda \in [0, 1)$:*

$$\boldsymbol{E}[\widehat{FDR}_\lambda(\gamma)] \geq FDR(\gamma)$$

for all γ and $\pi_0 < 1$.

Note that Storey (2002) fixed a rejection boundary γ and proposed an estimator for the FDR. To perform a test controlling a pre-chosen FDR α, the largest γ has to be determined, such that $\widehat{FDR}_\lambda(\gamma) \leq \alpha$. For $\lambda = 0$ Storey's procedure for a pre-chosen FDR is equivalent to the Benjamini-Hochberg procedure. For $\lambda > 0$, the rejection boundary γ is larger compared to the Bejamini-Hochberg method and thus it may be more powerful.

The following theorem may also be used in the thesis:

Theorem 2.3.2.6 *Asymptotically, for $m \to \infty$ the FDR can be written as:*

$$\alpha = \frac{\pi_0\gamma}{\pi_0\gamma + (1-\pi_0)(1-\beta(\gamma))}$$

where $(1-\beta(\gamma))$ is the power and $\beta(\gamma)$ is the type II error as a function of the rejection boundary γ.

2.4 Linear methods for classification

2.4.1 Binary logistic regression

In medical research it is often studied how a set of predictor variables $X = (X_1, X_2, ..., X_m)$ is related to a dichotomous response variable R. Note that the true response is defined by $R = 0$ if a patient does not response or 1 if a patient responds to a specific therapy. The statistical model that is generally preferred for the analysis of binary responses is the binary logistic regression model (see e.g. Harrel (2001), Hastie et al. (2001)).

Definition 2.4.1.1 *The binary logistic regression model is stated in terms of the probability that $R = 1$ given X, the values of the predictors:*

$$P[R = 1 \mid X = \mathbf{x}] = [1 + \exp(-\mathbf{x}\boldsymbol{\beta})]^{-1}$$

where $\mathbf{x}\boldsymbol{\beta} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m$.

Definition 2.4.1.2 *The function $q = [1+\exp(-x)]^{-1}$ is called the logistic function. Note that this function has an unlimited range for x but q is restricted to range from 0 to 1.*

Solving the equation above for x by using

$$1 - q = \exp(-x)/[1 + \exp(-x)]$$

yields the inverse of the logistic function

$$x = \log[\frac{q}{1-q}] = logit(R = 1 \mid X = \mathbf{x})$$

Note that $\frac{q}{1-q}$ is the odds ratio that $R = 1$ occurs. Since the logistic model is stated in terms of $q = P[R = 1 \mid X = \mathbf{x}]$ its only assumptions relate to the form of the regression equation. We transform $P[R = 1]$ to make a model that is linear in $\mathbf{x}\boldsymbol{\beta}$:

$$logit(R = 1 \mid X = \mathbf{x}) = logit(q) = \log[\frac{q}{1-q}] = \mathbf{x}\boldsymbol{\beta}$$

where $q = P[R = 1 \mid X = \mathbf{x}]$. Thus the model is a linear regression model in the log odds that $R = 1$ since $logit(q)$ is a weighted sum of $\mathbf{x}$.

The parameter β_j is then the change in the log odds per unit change in x_j if x_j represents a single factor that is linear and does not interact with other factors and if all other factors are held constant. Instead of writing this relation ship in terms of log odds, it could just as easily be written in terms of the odds that $R = 1$.

$$odds\{R = 1 \mid X = \mathbf{x}\} = exp(\mathbf{x}\boldsymbol{\beta})$$

The effect of increasing x_j by d is to increase the odds that $R = 1$ by a factor of $\exp(\beta_j d)$ or to increase the log odds that $R = 1$ by an increment $\beta_j d$. The logistic model quantifies the effect of a predictor in terms of an odds ratio or log odds ratio.

The parameters in the logistic regression model are estimated using the maximum likelihood method. Denoting the response and vector of predictors of response of the ith subject by R_i and $\mathbf{x}_i$, respectively, the model states that $q_i = P[R_i = 1 \mid X = \mathbf{x}_i] = [1 + \exp(-\mathbf{x}_i\boldsymbol{\beta})]^{-1}$. The likelihood of an observed responder R_i given predictors $\mathbf{x}_i$ and the unknown parameter vector $\boldsymbol{\beta}$ is

$$q_i^{R_i}[1 - q_i]^{1-R_i}.$$

The joint likelihood of all responses R_i, $i = 1, ..., n$ is the product of theses likelihoods:

$$L(\beta) = \prod_{i=1}^{n} q_i^{R_i}[1 - q_i]^{1-R_i}$$

Thus, the log likelihood is:

$$\begin{aligned} \log(L(\beta)) &= \sum_{i=1}^{n} R_i \log(q_i) + (1 - R_i)\log(1 - q_i) \\ &= \sum_{i=1}^{n} R_i \log(\frac{q_i}{1 - q_i}) + \log(1 - q_i) \\ &= \sum_{i=1}^{n} R_i logit(q_i) - log(1 + \exp(logit(q_i))). \end{aligned}$$

The likelihood and log likelihood functions are rewritten by using the definition of q_i above to allow them to be recognized as a function of the unknown parameters $\boldsymbol{\beta}$. Note that $\hat{\boldsymbol{\beta}}$ cannot be written explicitly. The Newton-Raphson method is usually used to salve iteratively for the list of values $\boldsymbol{\beta}$ that maximize the log likelihood.

Since the maximum likelihood estimate of a function of a parameter is the same function of the maximum likelihood estimate if the parameter:

$$\hat{q}_i = [1 + \exp(-\mathbf{x}_i \hat{\beta})]^{-1}.$$

To test the null hypotheses $H_0 : \beta_i = 0$ against the alternative $H_1 : \beta_i \neq 0$ the Wald test statistic is used.

2.4.2 Linear discriminant analysis

Another linear method for classification is the linear discriminant analysis (see e.g. Hastie et al. (2001)). Again a set of predictor variables $X = (X_1, X_2, ..., X_m)$ is given. We assume that the population can be split into $l = 1, ..., K$ sub-populations (e.g. for 2 sub-populations: responder and non-responder). Assume that for a training sample the corresponding true class $R = k$ is known (e.g. 0 for non-responders and 1 for responders in the case of 2 sub-populations). We are now searching for decision functions that discriminates between two classes respectively. Therefore we need the class posteriors $P[R = k \mid X = \mathbf{x}]$. Let $f_k(\mathbf{x})$ be the class-conditional density of $\mathbf{x}$ in class $R = k$ and $q_k = P[R = k]$ the a priori probability of class k, where $\sum_{k=1}^{K} q_k = 1$. From Bayes theorem we get:

$$P[R = k \mid X = \mathbf{x}] = \frac{f_k(\mathbf{x}) q_k}{\sum_{l=1}^{K} f_l(\mathbf{x}) q_l}.$$

$P[R = k \mid X = \mathbf{x}]$ is the a posteriori probability that an observation (patient) with predictor vector $\mathbf{x}$ belongs to class k. To estimate the a posteriori probability we have to estimate $f_k(\mathbf{x})$ and q_k from the sample.

Definition 2.4.2.1 ***Bayes decision rule:*** *An observation with predictor vector* $\mathbf{x}$ *will be allocated to the class which has the largest a posteriori probability* $P[R = k \mid X = \mathbf{x}]$*:*

$$\hat{k} \text{ such that } P[R = \hat{k} \mid X = \mathbf{x}] \geq P[R = l \mid X = \mathbf{x}] \text{ for } l = 1, ...K$$

Definition 2.4.2.2 *For a decision function* $\hat{k} = e(\mathbf{x})$ *the conditional error rate is the probability that an observation with true class* k *and predictors* $\mathbf{x}$ *is allocated to the wrong class:*

$$\varepsilon(e(\mathbf{x})) = P[e(\mathbf{x}) \neq k \mid X = \mathbf{x}].$$

Definition 2.4.2.3 *For a decision function $\hat{k} = e(\mathbf{x})$ the total error rate is the probability that an observation with predictors $\mathbf{x}$ is allocated to the wrong class:*

$$\varepsilon(e) = P[e(\mathbf{x}) \neq k].$$

Theorem 2.4.2.1 *The Bayes rule minimizes the conditional error rate and thus also the total error rate.*

Linear discriminant analysis arises in the special case when we assume that the classes have a common covariance matrix $\sum_k = \sum \; \forall k$ and each class density is multivariate normal. When comparing 2 classes k and l we have to look at the log-ratio of the two a posteriori probabilities and we see that:

$$\log(\frac{P[R = k \mid X = \mathbf{x}]}{P[R = l \mid X = \mathbf{x}]}) = \log(\frac{f_k(\mathbf{x})}{f_l(\mathbf{x})}) + \log(\frac{q_k}{q_l})$$
$$= \log(\frac{q_k}{q_l}) - \frac{1}{2}(\mu_k - \mu_l)^T \sum^{-1}(\mu_k - \mu_l) + \mathbf{x}^T \sum^{-1}(\mu_k + \mu_l)$$

an equation linear in $\mathbf{x}$. The linear log-odds function implies that the decision boundary between classes k and l, the set were $P[R = k \mid X = \mathbf{x}] = P[R = l \mid X = \mathbf{x}]$, is linear in x (in m dimensions a hyperplane).

Definition 2.4.2.4 *Linear discriminant function:*

$$LDF(\mathbf{x}) = (\mu_k - \mu_l)^T \sum^{-1} \mathbf{x} - \frac{1}{2}(\mu_k - \mu_l)^T \sum^{-1}(\mu_k + \mu_l)$$

Thus, one will decide for k if

$$(\mu_k - \mu_l)^T \sum^{-1} \mathbf{x} - \frac{1}{2}(\mu_k - \mu_l)^T \sum^{-1}(\mu_k + \mu_l) > \log(\frac{q_k}{q_l})$$

and class l otherwise. In practice the parameters of the normal distribution are unknown and we will have to estimate proportions, means and covariance matrices from the given sample.

For a observed patient with predictor vector $\mathbf{x}$ we can now calculate K discriminant functions. The patient will allocate to that class $\hat{k}$ which has maximal $LDF_{\hat{k}}(\mathbf{x})$.

2.5 The receiver operating characteristic curve

Various measures have been proposed to capture discrimination, but the receiver operating characteristic curve (ROC-curve) has become the standard description of classification accuracy for scalar-used classifiers. It is a measure of the predictive ability of a score if the score is used for different thresholds with varying values of sensitivity and specificity. The area under the ROC-curve (AUC) is the widely used measure to summarize the ROC. The ROC-curve is currently the best-developed statistical tool for describing the performance of tests with results that are not simply positive or negative but that are measured on continuous scales. The following definitions, theorems and their corresponding proofs can be found in Pepe (2003).

2.5.1 ROC-curve for continuous tests

Let R denote again the true response status, $R = 1$ if the patient is responding and $R = 0$ if he is not responding to a particular therapy.

Definition 2.5.1.1 *Using a threshold c, a binary test from a continuous result from a diagnostic test Y is defined as positive if $Y \geq c$ and negative if $Y < c$. Let the the corresponding true and false positive fractions at the threshold c be*

$$TPF(c) = P[Y \geq c \mid R = 1] \text{ and } FPF(c) = P[Y < c \mid R = 0].$$

Note again that TPF and FPF are also known as ***sensitivity*** *and* ***1-specificity****.*

Definition 2.5.1.2 *The ROC curve is the entire set of possible true and false positive fractions attainable by dichotomizing Y with different thresholds. Thus the ROC-curve is*

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}.$$

As the threshold c increases, both $FPF(c)$ and $TPF(c)$ decrease. At one extreme, $c = \infty$, we have $\lim_{c\to\infty} TPF(c) = 0$ and $\lim_{c\to\infty} FPF(c) = 0$. At the other extreme, $c = -\infty$, we have $\lim_{c\to-\infty} TPF(c) = 1$ and $\lim_{c\to-\infty} FPF(c) = 1$. Thus, the ROC-curve is a monotone increasing function in the positive quadrant.

Definition 2.5.1.3 *Because of the above discussed properties, the ROC-curve can also be written as*

$$ROC(\cdot) = \{(t, ROC(t)), \quad t \in (0,1)\},$$

where the ROC function maps t to $TPF(c)$ and c is the threshold corresponding to $FPF(c) = t$.

The ROC-curve is a monotone increasing function mapping $(0,1)$ onto $(0,1)$. An uninformative test is one such that Y is unrelated to the response status R. The probability distributions for Y are assumed to be the same in the responding and non-responding populations and therefore for any threshold c we have $TPF(c) = FPF(c)$. The ROC-curve for a useless test is therefore $ROC(t) = t$. A perfect test on the other hand completely separates responding and non-responding subjects. Thus, for some threshold c, we have $TPF(c) = 1$ and $FPF(c) = 0$. Note that most tests have ROC-curves that lie between those of the perfect and useless tests. Better tests have ROC-curves closer to the upper left corner. If we choose thresholds c_A and c_B for which $TPF_A(c_A) = TPF_B(c_B)$, the corresponding false positive fractions are ordered in favor of test A, so that $FPF_A(c_A) < FPF_B(c_B)$.

An important property of the ROC-curve is the following:

Theorem 2.5.1.1 *The ROC-curve is invariant to strictly increasing transformations of Y.*

The ROC-curve for evaluating diagnostic tests provides a complete description of potential performance, facilitates comparing and combining information across studies of the same test, guides the choice of threshold in applications and provides a mechanism for relevant comparisons between different non-binary tests (see Pepe (2003)).

2.5.2 Area under the ROC-curve

The most widely used summary measure for the ROC-curve is the area under the ROC-curve (AUC).

Definition 2.5.2.1 *The area under the ROC-curve (AUC) is defined as:*

$$\int_0^1 ROC(t)dt$$

A perfect test, one with the perfect ROC-curve, has the value $AUC = 1$, in contrast, an uninformative test with $ROC(t) = t$ has $AUC = 0.5$. Most tests have values falling in between. Clearly, if two tests are ordered with test A uniformly better than test B, in the sense that

$$ROC_A(t) \geq ROC_B(t) \qquad \forall t \in (0, 1),$$

then their AUC statistics are also ordered

$$AUC_A \geq AUC_B.$$

However, the converse is not necessarily true.

Let Y_r denotes the test result for a (true) responding and Y_{nr} for a (true) non-responding patient respectively. The area under the ROC-curve can be interpreted as the probability that in a randomly selected pair for responders and non-responders, the score value of the non-responder is smaller than the score value of the responder:

Theorem 2.5.2.1 *The following result holds:*

$$AUC = P[Y_r > Y_{nr}]$$

where Y_r and Y_{nr} correspond to independent and randomly chosen test results from the responding and non-responding populations, respectively.

This theorem has been shown by Bamber (1975). A proof of this theorem can also be seen in e.g. Pepe (2003).

2.5.3 Binormal ROC-curve and AUC

To derive the functional form of the binormal ROC-curve, suppose that test results are normally distributed in the responding and non-responding populations.

Theorem 2.5.3.1 *If $Y_r \sim N(\mu_r, \sigma_r^2)$ and $Y_{nr} \sim N(\mu_{nr}, \sigma_{nr}^2)$ then*

$$ROC(t) = \Phi(a + bz(t))$$

where $a = (\mu_r - \mu_{nr})/\sigma_r$ and $b = \sigma_{nr}/\sigma_r$ and $z(t) = \Phi^{-1}(t)$ denotes the t-quantile of the standard normal distribution.

Proof: For any threshold c,

$$FPF(c) = P[Y_{nr} > c] = \Phi(\frac{\mu_{nr} - c}{\sigma_{nr}}),$$

$$TPF(c) = P[Y_r > c] = \Phi(\frac{\mu_r - c}{\sigma_r}).$$

For a false positive fraction t, we see that $c = \mu_{nr} - \sigma_{nr} z(t)$ is the corresponding threshold for the test positivity criterion. Hence,

$$ROC(t) = TPF(c) = \Phi(\frac{\mu_r - c}{\sigma_r}) = \Phi(\frac{\mu_r - \mu_{nr} + \sigma_{nr} z(t)}{\sigma_r}) = \Phi(a + bz(t))$$

We call a the intercept and b the slope for the binormal ROC curve.

Theorem 2.5.3.2 *The AUC for the binormal ROC curve is*

$$AUC = \Phi(\frac{a}{\sqrt{1 - b^2}}).$$

Proof: Recall that $AUC = P[Y_r > Y_{nr}] = P[Y_r - Y_{nr} > 0]$. Let $W = Y_r - Y_{nr}$. Then

$$W \sim N(\mu_r - \mu_{nr}, \sigma_r^2 + \sigma_{nr}^2)$$

and

$$P[W > 0] = 1 - \Phi\left(\frac{-\mu_r + \mu_{nr}}{\sqrt{\sigma_r^2 + \sigma_{nr}^2}}\right) = \Phi\left(\frac{-\mu_r + \mu_{nr}}{\sigma_r^2} / \sqrt{1 + \frac{\sigma_{nr}^2}{\sigma_r^2}}\right) = \Phi\left(\frac{a}{\sqrt{1 - b^2}}\right)$$

The AUC is an increasing function of a and a decreasing function of b.

2.5.4 Estimating the ROC-curve

We assume that the data can be represented as test results for n_r cases and n_{nr} controls: $Y_{r,i}, i = 1, ..., n_r$ and $Y_{nr,i}, i = 1, ..., n_{nr}$. We assume that $Y_{r,i}$ and $Y_{nr,i}$ are selected randomly from the populations of test results associated with responding and non-responding states, respectively. The empirical estimator of the ROC-curve simply applies the definition of the ROC-curve to the observed data.

Definition 2.5.4.1 *For each possible cut-point c, the empirical true and false positive fractions are calculated as follows:*

$$\widehat{TPF}(c) = \sum_{i=1}^{n_r} I[Y_{r,i} \geq c]/n_r,$$

$$\widehat{FPF}(c) = \sum_{i=1}^{n_{nr}} I[Y_{nr,i} \geq c]/n_{nr}.$$

The empirical ROC curve, $\widehat{ROC}(t)$, is a plot of $\widehat{TPF}(c)$ versus $\widehat{FPF}(c)$ for all $c \in (-\infty, \infty)$.

Clearly the empirical AUC ($\widehat{AUC}$) can be calculated by applying there definition to the empirical ROC. Note that the empirical $\widehat{ROC}(t)$ is generally a step function.

A rank-based estimate of the AUC is the Mann-Whitney U Statistic introduced by Mann and Whitney (1947). The following theorem was proven by Hanley and McNeil (1982).

Theorem 2.5.4.1 *The area under the empirical ROC curve is the Mann-Whitney U-statistic:*

$$\widehat{AUC} = \sum_{j=1}^{n_r} \sum_{i=1}^{n_{nr}} \left\{ I[Y_{r,i} > Y_{nr,i}] + \frac{1}{2} I[Y_{r,i} = Y_{nr,i}] \right\} / (n_r n_{nr})$$

2.5.5 ROC-curve of a linear score

Let $X = (X_1, ..., X_m)$ again be a set of predictor variables. We now consider a linear combination of the test result: $S(\boldsymbol{\beta}, X) = \sum_{i=1}^{m} \beta_i X_i$.

Definition 2.5.5.1 *The ROC-curve for a score S, is then defined as the set of points*

$$ROC(\cdot) = \{(FPF(c), TPF(c)), c \in (-\infty, \infty)\}$$

where $TPF(c) = P[S_i > c \mid R_i = 1]$ which is interpreted as the true positive rate associated with the positivity criterion $S > c$ and $FPF(c) = P[S_i > c \mid R_i = 0]$ which is the false positive rate at threshold c.

We now want to find $\boldsymbol{\beta}_{opt}$ which is the $(\beta_1, ..., \beta_m)$ that maximizes the area under the ROC-curve associated with S.

Theorem 2.5.5.1 *If $X_1, ..., X_m$ has a multivariate normal distribution in each of the responding ($N[\boldsymbol{\mu}_r, \Sigma]$) and non-responding populations ($N[\boldsymbol{\mu}_{nr}, \Sigma]$), then the score defined by the linear discriminant function maximizes the area under the ROC-curve:*

$$\boldsymbol{\beta}_{opt} = (\boldsymbol{\mu}_r - \boldsymbol{\mu}_{nr})^T \Sigma^{-1}$$

This theorem was proven by Su and Liu (1993) (see also e.g. Pepe and Thompson (2002)).

It is well known that in the multivariate binormal setting the linear discriminant and logistic scores are equal if the covariance matrices are proportional. The linear discriminant procedure has been shown to be statistically more efficient when the model is correct. Logistic regression, however, can be applied outside of the multivariate binormal setting. It relies only on an assumption about the form of the conditional probability for response given $X_1, ..., X_m$ and does not require specification of the much more complex joint distribution of $X_1, ..., X_m$. However, the logistic regression is not motivated as a procedure which maximizes the area under the ROC-curve for a linear score. In logistic regression analysis, the coefficients are chosen to maximize the logistic likelihood. It is not clear if the logistic likelihood relates to any natural measure of the discriminatory capacity of the linear score. Hence, in general, the logistic regression linear score is not easily motivated as an optimal discriminator of responding and non-responding populations except in the multivariate binormal setting. It has been shown, however, that, if complete discrimination is possible, the logistic regression will estimate the linear combination which separates the populations (compare Pepe and Thompson (2002), Pepe et al. (2004)).

3 Model selection for prediction of a clinical outcome

3.1 Assumptions

We want to search for predictors of a binary outcome (e.g. success of a therapy) among a large set of m candidate variables (e.g. genes, proteins). Independent samples of patients responding (n_r) and non-responding (n_{nr}) to the therapy are available. Based on these samples variables have to be selected and a score has to be constructed which will be used to predict response to therapy in future patients. We will aim at a score which optimizes the AUC.

To simplify the problem we first assume that the variables follow normal distributions with common known variance $\sigma^2 = 1$ with means $\mu_{r,i}$, $i = 1, ..., m$, for responders and means $\mu_{nr,i}$, $i = 1, ..., m$, for non-responders. Furthermore we assume that among the m candidates there is a set E, $m_e = \sharp\{E\}$, of prognostic variables related to the clinical outcome (alternatives). We also assume that these prognostic variables have a common mean $\mu_{r,i} = \mu_r$, $i \in E$ in the responding patients and also a common mean $\mu_{nr,i} = \mu_{nr}$, $i \in E$ in the non-responding patients. Hence a common effect size $\mu_r - \mu_{nr} = \Delta$ is assumed for the prognostic variables. For the non-prognostic variables without loss of generality the difference in means between responders and non-responders is assumed to be zero, $\mu_{r,j} - \mu_{nr,j} = 0$, $j \in (1, 2, .., m) \setminus E$ (true null hypotheses).

3.2 Selection of variables for future prediction

We investigate two methods for the selection of promising variables to build a prediction score for a clinical outcome of a future patient.

3.2.1 Selection based on a multiple test controlling the FDR

For the selection of variables for the prediction score we test the following set of one-sided null hypotheses:

$$H_{0i} : \mu_{r,i} - \mu_{nr,i} = 0 \text{ against } H_{1i} : \mu_{r,i} - \mu_{nr,i} > 0 \text{ for } i = 1, \ldots, m.$$

The standardized mean differences between responder and non-responder

$$z_i = (\bar{x}_{r,i} - \bar{x}_{nr,i})\sqrt{n/2},\ i = 1, \ldots, m$$

are calculated, where for simplicity we assume equal sample sizes per variable and group $n = n_{r,i} = n_{nr,i}$ for $i = 1, \ldots, m$. The test decisions are based on the one-sided p-values

$$p_i = 1 - \Phi(z_i),$$

where Φ denotes the cumulative distribution function of the standard normal distribution (see theorem 2.1.0.3). Note that the two-sided case is considered later.

To adjust for multiplicity, i.e. not to include too many nuisance variables without any predictive ability in the score, we use Storeys approach (Storey (2002)) to control the FDR (see Section 2.3.2). Due to formula (2.6) the critical boundary is determined from the sample such that the estimated FDR never exceeds the targeted value α. Note that this method adapts to the estimated proportion of true null hypotheses.

The variables whose p-values fall below the critical boundary γ ($p_i < \gamma$) corresponding to the targeted threshold α are selected to build a score in order to predict whether a future patient will respond or not respond to the treatment.

3.2.2 Selection based on stepwise forward logistic regression

Whereas the first approach is only based on individual selection criteria for the variables, here we use a multiple logistic regression approach (see Section 2.4.1) with stepwise forward selection to assess the contribution of the individual variables to predict response to therapy in the training sample. We again use a fixed threshold γ for the p-values, this time calculated from the final model evolving in the multivariate logistic regression. The selection is done in such a way that selected variables can again be removed from the model when their p-values in the aggregated model fall above the threshold γ. The stepwise procedure ends with a final model when further variables fail to meet the selection criterion.

3.3 Prediction of the clinical outcome

Based on our assumptions we simply use the linear score of the selected variables. This would be the optimal solution in the case of no selection for a given set of variables which follow independent and identical normal distributions with unknown means and known variance $\sigma^2 = 1$ (as in the classical linear discriminant analysis, see theorem 2.5.5.1). Thus in our case we use the following prediction score:

Definition 3.3.0.1 *Let $\bar{x}_{r,i}$ and $\bar{x}_{nr,i}$ denote the sample means of the ith variable of patients responding and not responding to therapy respectively and $\mathbf{x} = (x_1, ..., x_m)$ the values of the variables of a future patient. The prediction score is a linear combination of $\mathbf{x}$:*

$$\hat{f}(\mathbf{x}; \gamma) = \hat{\mathbf{c}}^T \mathbf{x} = \sum_{i=1}^{m} \hat{c}_i x_i. \tag{3.1}$$

where

$$\hat{c}_i = \begin{cases} \bar{x}_{r,i} - \bar{x}_{nr,i} & \text{if } p_i < \gamma \\ 0 & \text{else} \end{cases} . \tag{3.2}$$

I.e. $k \leq m$ variables for which $p_i < \gamma$ are selected to build a linear score and $m - k$ variables are not selected (their weights in the score are set to 0).

If $\hat{f}(\mathbf{x}; \gamma) > b$ we predict a response, otherwise a non-response. To measure the predictive ability of such a score we use the ROC-curve resulting from varying threshold values for the score, where sensitivity (TPF), further denoted by v, is plotted against FPF=(1-specificity), further denoted by w, as a function of b (see Section 2.5). Since we are interested in ROC-curves the following results are invariant to any strictly monotonic transformation of this score (see theorem 2.5.1.1).

In case of forward selection we simply use the estimated linear predictor for the log odds from the final model in the forward logistic regression, which is of the same form as (3.1) but uses the parameter estimates (β_i) from the model instead of the difference in the sample means of the selected variables.

Theorem 3.3.0.1 *Given the selection threshold α for the FDR (and thus the corresponding selection boundary γ for the individual p-values) and the estimated weights (3.2) from the samples the prognostic score (3.1) follows two normal distributions:*

$$\hat{f}(\mathbf{x};\gamma) \sim N[\mu_a, \sigma_a^2] = N[\hat{\mathbf{c}}^T \boldsymbol{\mu}_a, \hat{\mathbf{c}}^T \hat{\mathbf{c}}]$$

where $\boldsymbol{\mu}_a$, $a = r$ or nr is the true mean vector in a future responder or non-responder, respectively. Note that because of the independence between variables the true covariance matrix $\Sigma = I$.

Proof: This result can be simply calculated using theorem 2.1.0.7 in Section 2.1.

Theorem 3.3.0.2 *Fixing the appropriate $\boldsymbol{\mu}_a$ for the populations of responders and non-responders, respectively, the AUC for future independent populations can be calculated as:*

$$AUC(\alpha) = \int_0^1 \left\{1 - \Phi\left[z(1-w) - \frac{\hat{\mathbf{c}}^T(\boldsymbol{\mu}_r - \boldsymbol{\mu}_{nr})}{\sqrt{\hat{\mathbf{c}}^T\hat{\mathbf{c}}}}\right]\right\} dw \tag{3.3}$$

where $z(q)$ is the q-quantile of the standard normal distribution.

Proof: This can be calculated from:

$$Sensitivity = v = 1 - \Phi_{\hat{\mathbf{c}}^T\boldsymbol{\mu}_r, \hat{\mathbf{c}}^T\hat{\mathbf{c}}}(b) = 1 - \Phi(\frac{b - \hat{\mathbf{c}}^T\boldsymbol{\mu}_r}{\sqrt{\hat{\mathbf{c}}^T\hat{\mathbf{c}}}}) \tag{3.4}$$

and

$$Specificity = 1 - w = \Phi_{\hat{\mathbf{c}}^T\boldsymbol{\mu}_{nr}, \hat{\mathbf{c}}^T\hat{\mathbf{c}}}(b) = \Phi(\frac{b - \hat{\mathbf{c}}^T\boldsymbol{\mu}_{nr}}{\sqrt{\hat{\mathbf{c}}^T\hat{\mathbf{c}}}}) \tag{3.5}$$

where Φ_{μ,σ^2} denotes the cumulative distribution function of the normal distribution with mean value μ and variance σ^2. Calculating b from formula (3.5) results in:

$$b = z(1-w)\sqrt{\hat{\mathbf{c}}^T\hat{\mathbf{c}}} + \hat{\mathbf{c}}^T\boldsymbol{\mu}_{nr}.$$

Inserting this into formula (3.4) results in (3.3).

For the prognostic score based on the logistic regression similar results can be derived.

3.4 Assumptions on the effect size Δ

To get a benchmark let us assume that the optimal linear score built from the m_e prognostic variables is known. We now will ask, depending on the number m_e of prognostic variables, what minimal common effect size Δ is required to achieve a ROC-curve crossing through the point where both sensitivity (v) and specificity $(1-w)$ have a certain pre-specified values?

Theorem 3.4.0.3 *Under the assumption of equal effect sizes among the alternatives, the effect size required for a ROC-curve crossing through the fixed point $(v, 1-w)$ can be calculated as:*

$$\Delta = \frac{z(1-w) - z(1-v)}{\sqrt{m_e}} \tag{3.6}$$

Proof: Clearly we get the best prognostic score if all m_e prognostic variables and no non-prognostic variables are selected, i.e. we know the true score:

$$f(\mathbf{x}) = \Sigma_{i\in E}x_i.$$

Note that for equal effect sizes Δ for the alternatives, the constant true weights (Δ) can be ignored in the score. From theorem 2.1.0.7 we know that this score follows a normal distribution:

$$f(\mathbf{x}) \sim N[\mu_f, \sigma_f^2] = N[m_e\Delta, m_e]$$

Hence the sensitivity for the theoretically best score for a future patient can be easily calculated as follows:

$$v = 1 - \Phi(z(1-w) - \frac{m_e\Delta}{\sqrt{m_e}}) = 1 - \Phi(z(1-w) - \sqrt{m_e}\Delta).$$

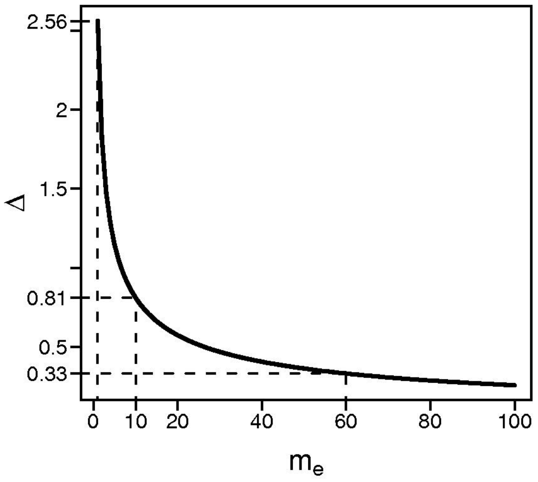

Figure 3.1: Minimal effect size Δ required to achieve a ROC-curve crossing through the point where sensitivity and specificity are equal to 0.9 as a function of the number of prognostic variables m_e. (Figure from Goll (2008)).

By solving the equation it turns out that the effect size required to cross the point $(v, 1-w)$ can be calculated as in formula (3.6).

In the following we choose $v = 1 - w = 0.9$, i.e. a theoretically best achievable AUC of $AUC_* = 0.965$ can be achieved. Figure 3.1 (Figure from Goll (2008)) shows the minimal effect size Δ depending on the number of prognostic variables related with the clinical outcome (m_e) if the best ROC-curve is assumed to cross the point where $v = 1-w = 0.9$. Two examples are marked which will be considered more closely in the simulation studies. For $m_e = 60$ an effect size of $\Delta = 0.331$ is required to achieve such an ROC-curve. For $m_e = 10$ an effect size of $\Delta = 0.811$ is needed to get a ROC-curve with such a property. Note that if there is only a single prognostic variable an effect size of $\Delta = 2.563$ is required. This demonstrates the crucial problem for gene expression or proteomic studies. If many prognostic variables work together they may show a large common effect even if there are only marginal individual effect sizes. Thus, the process of selection of such variables with only marginal effects among a large number of candidates in relatively small samples will be a formidable task. However, in case of a single or few prognostic variables the effect size to achieve good prognostic properties has to be pretty large, so that already small samples may be sufficient to select those very influential variables.

3.5 Variable selection using the FDR approach: simulation studies

Similar results of the simulation studies can also be seen in my doctorial thesis (see Goll (2008)). However, for this thesis we use a more sophisticated optimization procedure to determine the optimal selection threshold. Thus, there may be slight differences of the following results as compared to Goll (2008).

We investigate the selection procedure using a multiple test for constructing a linear score discussed in Section 3.3 by simulation, assuming that two samples of patients responding to a particular treatment and of patients not responding to the treatment are available. For a fixed FDR threshold α we can now calculate in a specific sample $AUC(\alpha)$ using formula (3.3). For a grid of α values with interval 0.01 $AUC(\alpha)$ is evaluated by simulation (10000 simulation runs). The optimal FDR level α_{opt} to achieve the best prediction with a linear score in terms of the $AUC(\alpha)$ in a specific scenario is hard to determine analytically in finite samples. It has to be kept in mind that $AUC(\alpha)$ is a random variable. Thus, optimization of α is based on the averages of the simulated $AUC(\alpha)$ values. For the simulated mean values of $AUC(\alpha)$ for the grid of α values we interpolate a function using splines. To determine α_{opt} which optimizes the average $AUC(\alpha)$, we optimize this interpolated function. Note again that this optimization procedure is a further investigation of the results in Goll (2008) where only a grid with interval 0.05 was investigated and no interpolation was done.

Different parameter constellations are investigated: we vary the number of prognostic variables related with the clinical outcome to be $m_e = 10$ or 60. We fix the group sample sizes to $n_r = n_{nr} = 50$, 100 and 500 and the number of candidate variables to $m = 1000$ and 6000. We also investigate the situation under the global null hypothesis ($m_e = 0$). As discussed in Section 3.4, the effect size Δ is triggered by forcing the optimal ROC-curve through the benchmark point $v = 1 - w = 0.9$, thus for $m_e = 10$, $\Delta = 0.811$ and for $m_e = 60$, $\Delta = 0.331$.

3.5.1 Searching among m=1000 hypotheses

Figure 3.2 shows the interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) as a function of α chosen a priori for the selection of variables for future prediction assuming a sample size of $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line) per group. $m_e = 10$ prognostic variables are searched within $m = 1000$ candidate genes. Note that if no score is selected in a specific sample $AUC(\alpha)$ is set to 0.5. The best achievable $AUC_* = 0.965$ is shown as solid horizontal line.

The figure shows that if a larger α is chosen, more non-prognostic variables are tolerated in the score but also more prognostic variables are selected so that the score still performs well. However, if a too large α is chosen for selection, too many non-prognostic variables are added so that the score gets worse if small sample sizes are applied. When increasing the sample size per group clearly the mean values of $AUC(\alpha)$ of the selected scores also increase. It can be seen that the resulting scores perform well for a wide range of α values for larger sample sizes. This may be due to the better estimate of $\mu_{r,i}$ and $\mu_{nr,i}$ for $i = 1, ..., m$ and thus to a better estimate of the weights in the score function when the sample size increases. It seems that for $n = 500$ it does not really mind which FDR threshold is chosen as selection criteria, the resulting score always performs good. The weights of selected true null hypotheses are nearly null, although their p-values are significant for larger FDR levels. See a detailed summary of the results in Table 3.1. For fixed m from simple consistency arguments it follows:

Theorem 3.5.1.1 *Given any positive α for the selection threshold:*

$$\lim_{n\to\infty} \hat{c}_i = \Delta \qquad \forall i \in E \qquad \text{and}$$

$$\lim_{n\to\infty} \hat{c}_j = 0 \qquad \forall j \in (1, ..., m)\backslash E$$

if the selected model is too large and contains non-prognostic variables. Therefore

$$\lim_{n\to\infty} AUC(\alpha) = AUC_* \qquad \forall\alpha.$$

If we assume $m_e = 60$ alternatives among the $m = 1000$ tested candidate variables, the effect size Δ to achieve the theoretical benchmark ROC-curve now is 0.331. Figure 3.3

again shows the interpolated functions of the mean values (over the simulated samples) of $AUC(\alpha)$ varying the FDR threshold α. Again it is better to tolerate more non-prognostic variables and thus find more prognostic ones, however for small sample sizes (see e.g. the dotted line for $n = 50$) it would be superior to choose an unrealistically large α_{opt}. Again increasing the sample size per group clearly also increases the $AUC(\alpha)$ values of the selected scores. A good performance in terms of future AUC values can be seen over all investigated FDR thresholds α if the sample size is increased to $n = 500$. See a detailed summary of the results in Table 3.1.

Over all investigated examples α_{opt} is decreasing and $AUC(\alpha_{opt})$ is increasing with increasing n. α_{opt} is increasing and $AUC(\alpha_{opt})$ is decreasing with increasing m_e .

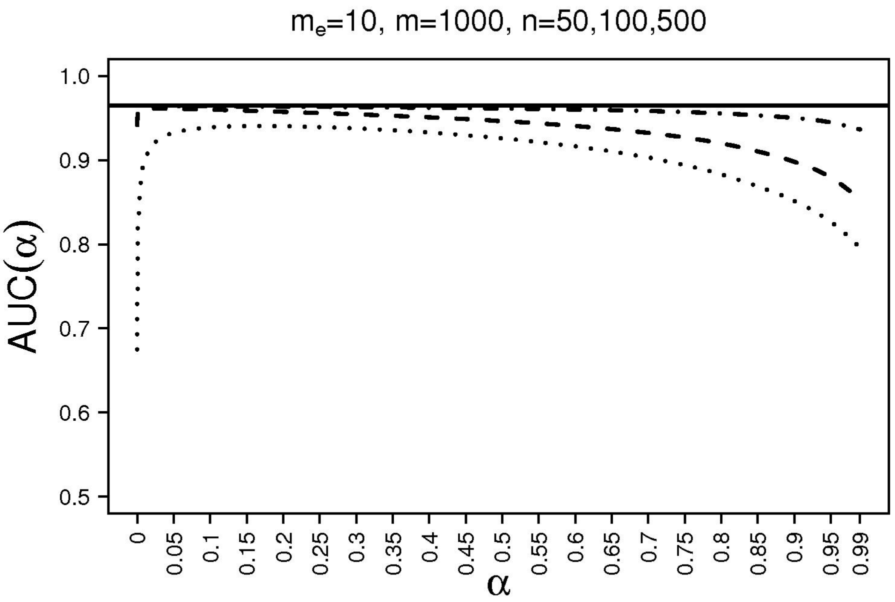

Figure 3.2: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 10$ prognostic variables among $m = 1000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.965$ is given as solid horizontal line. The effect size $\Delta = 0.811$.

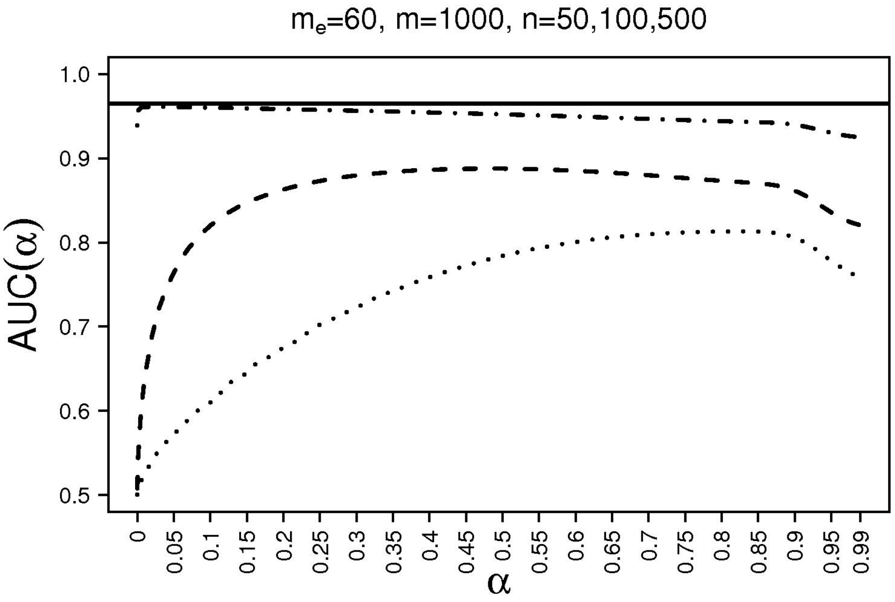

Figure 3.3: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 60$ prognostic variables among $m = 1000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.965$ is given as solid horizontal line. The effect size $\Delta = 0.331$.

3.5.2 Searching among m=6000 hypotheses

Let us furthermore have a look at the situation assuming $m_e = 10$ and $m_e = 60$ prognostic variables (alternatives) among $m = 6000$ tested hypotheses. Because of the larger number of candidate variables the problem to find the alternatives becomes harder. The effect size Δ remains the same, thus $\Delta = 0.811$ for $m_e = 10$ since it only depends on the number m_e of alternatives and not on the number of tested hypotheses. Figure 3.4 shows the situation assuming $m_e = 10$ fixing the sample size per group to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). Clearly the $AUC(\alpha)$ values are smaller as compared to the scenario with m only equal to 1000. However, for $n = 50$, due to the large effect size, selection may lead to good prediction scores if the right threshold is chosen for selection, although selecting out of 6000 hypotheses. Increasing the sample size to $n = 100$ per group again increases $AUC(\alpha_{opt})$ and decreases α_{opt}. A further increase of the sample size per group to $n = 500$ results, as in the case of $m = 1000$, in a good performance for a wide range of α values. For small values of α, $AUC(\alpha_{opt})$ for future prediction is almost equal to $AUC_* = 0.965$. For larger α values the performance is only slightly smaller.

Assuming $m_e = 60$ among $m = 6000$ hypotheses, the situation gets worse. Because of the small effect size ($\Delta = 0.331$) and the larger number of hypotheses to test no good prediction score can be selected if the sample size per group is small. Figure 3.5 shows the results. Applying a sample size of $n = 50$ the best choice of the selection threshold would be an unrealistically large α_{opt}, which applied as selection criterion would lead to prediction scores achieving only an average $AUC(\alpha_{opt})$ smaller than 0.7. This indicates a poor performance of the resulting scores as compared to $AUC_* = 0.965$. This again describes the problem of such studies. If only a few prognostic variables with a large effect size exist it may be possible to find good prediction scores if the right selection criterium is used, but if there are many variables with low effect sizes working together, searching for prediction scores with rather small sample sizes becomes a formidable problem. Increasing the sample size to $n = 100$ per group the situation improves a little as compared to $n = 50$. A further increase of the sample size per group to $n = 500$ changes the situation completely. Again, for a wide range of α values the future performance remains good (see Table 3.1).

A detailed summary of the results can be seen in Table 3.1. For $m = 6000$ the same tendencies can be seen as for $m = 1000$. Over all investigated examples α_{opt} is decreasing and $AUC(\alpha_{opt})$ is increasing with increasing n. α_{opt} is increasing and $AUC(\alpha_{opt})$ is decreasing with increasing m_e. For increasing m, α_{opt} is increasing and $AUC(\alpha_{opt})$ is decreasing.

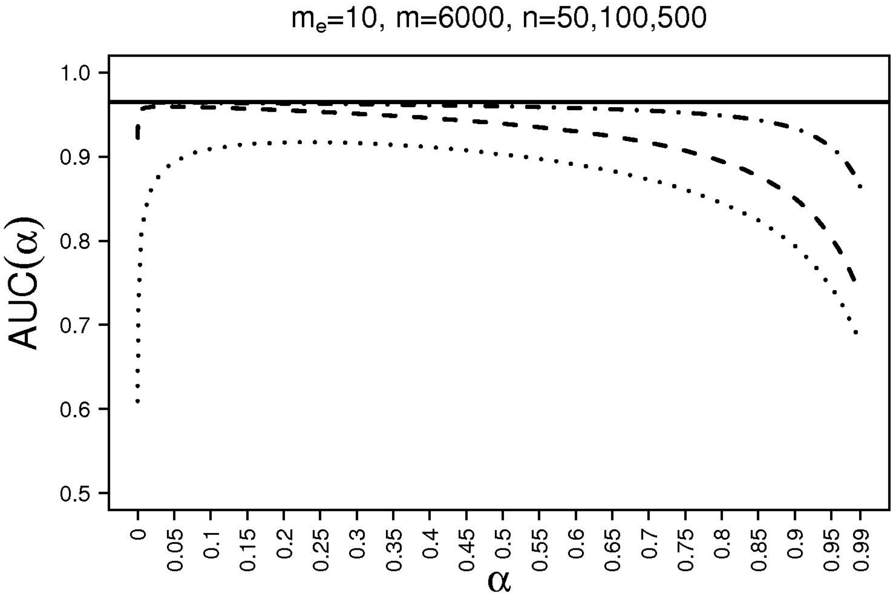

Figure 3.4: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 10$ prognostic variables among $m = 6000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.965$ is given as solid horizontal line. The effect size $\Delta = 0.811$.

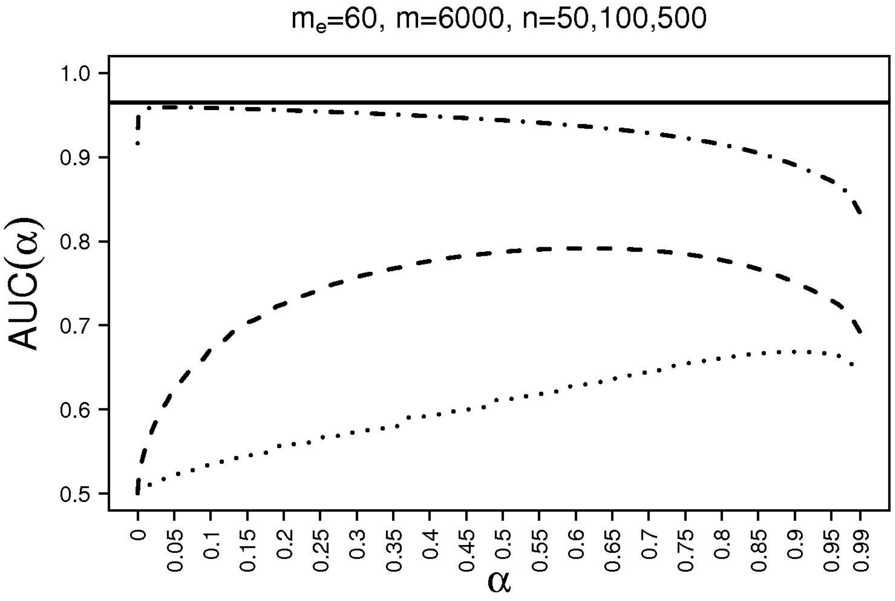

Figure 3.5: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 60$ prognostic variables among $m = 6000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.965$ is given as solid horizontal line. The effect size $\Delta = 0.331$.

Table 3.1: **Simulation results for selection using the FDR approach:** The best choice of the FDR threshold (α_{opt}), the corresponding $AUC(\alpha_{opt})$ as well as the number of selected non-prognostic variables (m_0^s) and the number of selected prognostic variables (m_e^s) for varying number of prognostic variables (m_e), tested hypotheses (m) and per group sample sizes (n). Note that $AUC_* = 0.965$.

m	m_e	Δ	n	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
1000	10	0.811	50	0.170	0.941	1.92	8.69
		0.811	100	0.014	0.963	0.14	9.90
		0.811	500	0.001	0.965	0.01	10.00
	60	0.331	50	0.824	0.813	254.02	48.98
		0.331	100	0.475	0.888	40.86	43.66
		0.331	500	0.033	0.961	2.04	59.53
6000	10	0.811	50	0.250	0.917	2.81	7.59
		0.811	100	0.034	0.960	0.36	9.70
		0.811	500	0.001	0.965	0.11	10.00
	60	0.331	50	0.895	0.669	323.04	27.81
		0.331	100	0.611	0.792	46.07	27.31
		0.331	500	0.034	0.959	0.83	57.50

3.5.3 Variable Selection expecting a small AUC_*

In the last examples we assumed that the optimal linear prediction score of future patients, if known, would lead to a ROC-curve crossing through the benchmark point where sensitivity and specificity are 0.9, which corresponds to a theoretically achievable AUC_* of 0.965 indicating a (in truth) very good discrimination between responders and non-responders. However in medical research often there is no such a good discrimination between two groups. Thus, in the following we will investigate scenarios, where AUC_* is assumed to be 0.8, which may be more realistic in medical research. To achieve a future performance of $AUC_* = 0.8$ the benchmark point is at $v = 1 - w = 0.724$ (assuming a ROC-curve crossing through a point, where sensitivity and specificity are the same). The minimal Δ required to obtain this ROC is 0.376 assuming $m_e = 10$ alternatives and 0.154 for $m_e = 60$. Figure 3.6 (Figure from Goll (2008)) shows the minimal required Δ for different values for AUC_*. Results assuming 10 (dashed line) and 60 (solid line) prognostic variables are shown. The results for assuming a theoretically achievable AUC_* of 0.8 and 0.965 (as assumed in the previous sections) are marked.

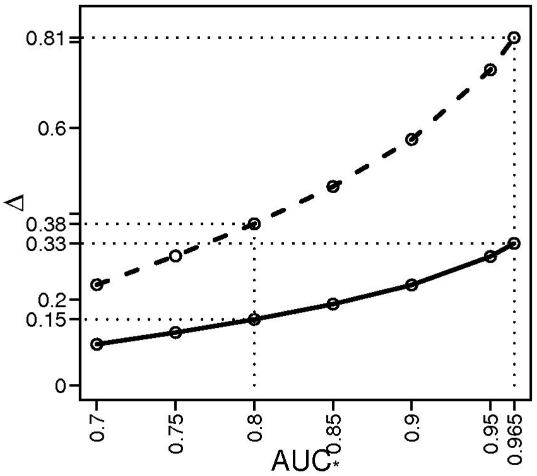

Figure 3.6: Minimal required effect size Δ as a function of the theoretically best possible AUC_* assuming the number of prognostic variables $m_e = 10$ (dashed line) and 60 (solid line) (Figure from Goll (2008)).

We again performed simulations (10000 simulation runs) assuming $m_e = 10$ and 60 alternatives among $m = 1000$ and 6000 hypotheses for a grid of α values with interval 0.01. The optimization of α_{opt} is again based on the interpolated functions from the mean values of the simulated $AUC(\alpha)$ values for each point of the γ-grid. The sample size per group is set to $n = 50$, 100 and 500.

Figure 3.7 shows the resulting mean $AUC(\alpha)$ values assuming $m_e = 10$ among 1000 hypotheses. Applying a small sample size per group we will not be able to find good prediction scores whatever FDR threshold α is used for selection. The mean $AUC(\alpha)$ values for future prediction are smaller than 0.6 over the whole range of α values. Larger sample sizes are needed to detect the alternatives with their only small effect sizes required to achieve $AUC_* = 0.8$. However, when doubling the sample size to $n = 100$ per group only a small increase in values of $AUC(\alpha_{opt})$ can be seen. Fixing the sample size to $n = 500$ leads to good prediction scores for small α_{opt} values.

Increasing the number of prognostic variables to $m_e = 60$ hypotheses (Figure 3.8) again a very large sample size is needed to achieve good prediction scores. However, fixing the sample size to $n = 500$ per group on average $AUC(\alpha_{opt})$ is only 0.720. For $n = 50$ the

AUC values do not exceed 0.6 over the whole range of investigated α values. The results of the simulated examples are summarized in Table 3.2.

If the prognostic variables are searched among 6000 hypotheses the situation gets extremely worse if smaller sample sizes are considered (Figures 3.9 and 3.10). Increasing the sample size to $n = 500$ per group helps if only a small number of alternatives with large effect sizes is assumed. Whereas for $m_e = 10$ (Figure 3.9) large $AUC(\alpha_{opt})$ values can be obtained for small α_{opt} values, for $m_e = 60$ (Figure 3.10), the average $AUC(\alpha_{opt})$ does not exceed 0.7. Thus the conclusion is that if there is only a moderate true discrimination between responders and non-responders very large sample sizes are required to get good prediction scores. Studies with small sizes will mostly produce useless prediction scores (see the summary in Table 3.2).

The following result can be determined:

Theorem 3.5.3.1 *Let Δ_1 be the required effect size to achieve $AUC_{*,1}$ and Δ_2 the required effect size to achieve $AUC_{*,2}$. The per-group sample size n_2 to achieve the same selection procedure as from the other sample (with per-group sample size n_1) can be calculated by:*

$$n_2 = \left(\frac{\Delta_1}{\Delta_2}\right)^2 n_1.$$

Proof: Given m and m_e, scenarios with the same value of $\Delta\sqrt{n}$ lead to identical selection procedures, i.e. to the same test-statistics and thus to the same selected prognostic and non-prognostic variables. Thus to get the same test statistic:

$$\Delta_1\sqrt{n_1/2} = \Delta_2\sqrt{n_2/2}.$$

By solving the equation it can easily be calculated that $n_2 = \left(\frac{\Delta_1}{\Delta_2}\right)^2 n_1$.

For example, to get the same selection procedure in the situation of $AUC_* = 0.8$, we need a $(0.81/0.38)^2 \approx 4.6$ times larger sample size as compared to the situation of $AUC_* = 0.965$. However, because of the smaller effect sizes the AUC achieved in this case may be relatively smaller. E.g. for $\alpha = 0.17$ and applying a sample size of $n = 232$ we get an average $AUC(\alpha)$ of 0.767 (95.9% of $AUC_* = 0.8$) as compared to 0.941 (97.5% of $AUC_* = 0.965$)

for $n = 50$. For $\alpha = 0.6$ the numbers are 0.741 (92.6% of $AUC_* = 0.8$) versus 0.916 (94.9% of $AUC_* = 0.965$).

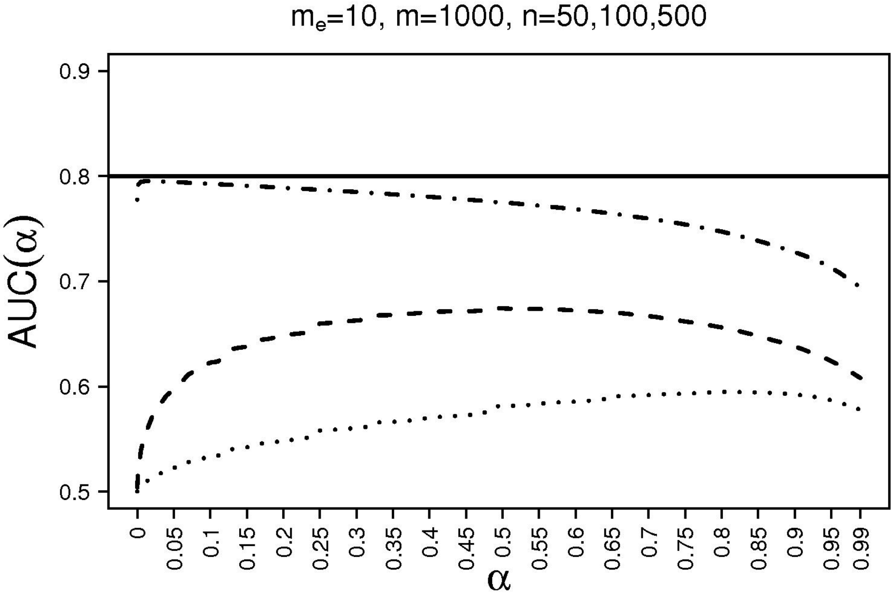

Figure 3.7: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 10$ prognostic variables among $m = 1000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.8$ is given as solid horizontal line. The effect size $\Delta = 0.376$.

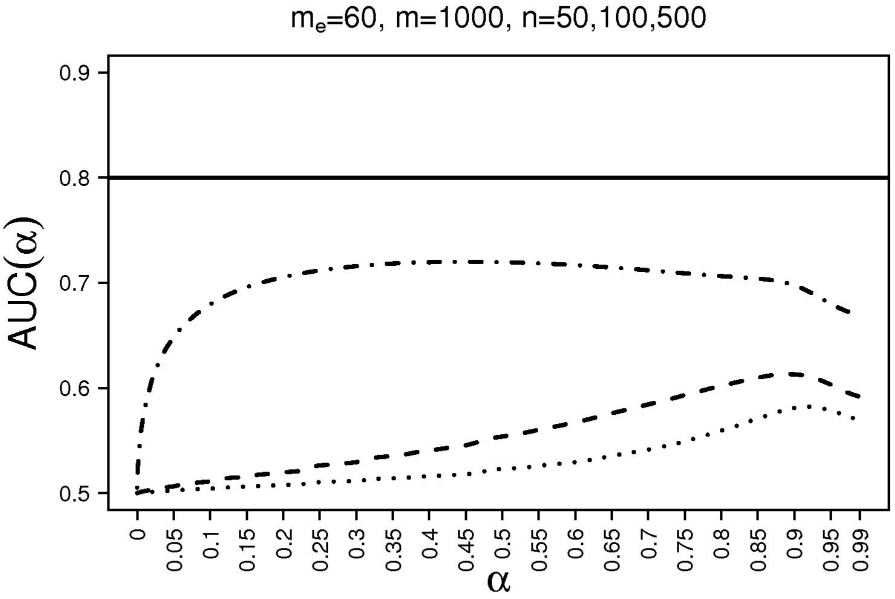

Figure 3.8: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 60$ prognostic variables among $m = 1000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.8$ is given as solid horizontal line. The effect size $\Delta = 0.154$.

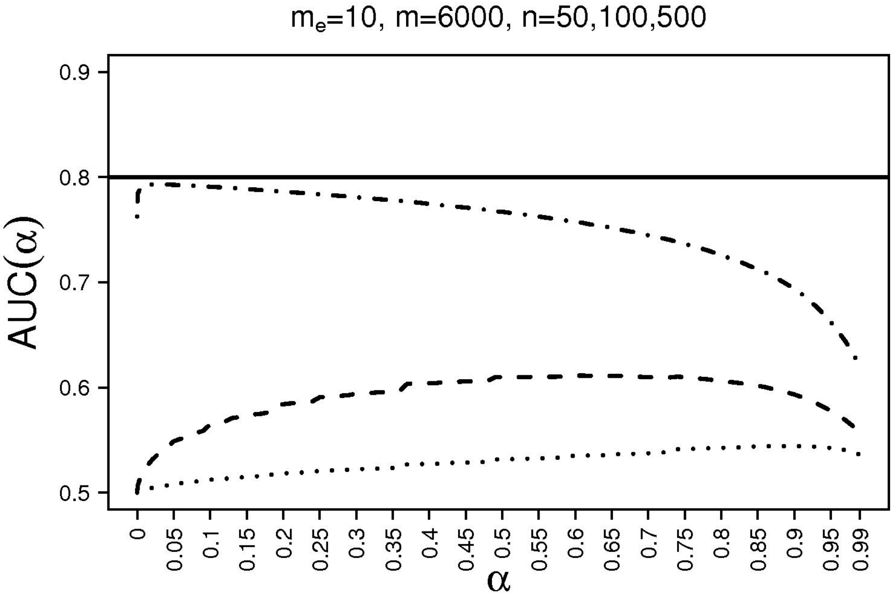

Figure 3.9: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 10$ prognostic variables among $m = 6000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.8$ is given as solid horizontal line. The effect size $\Delta = 0.376$.

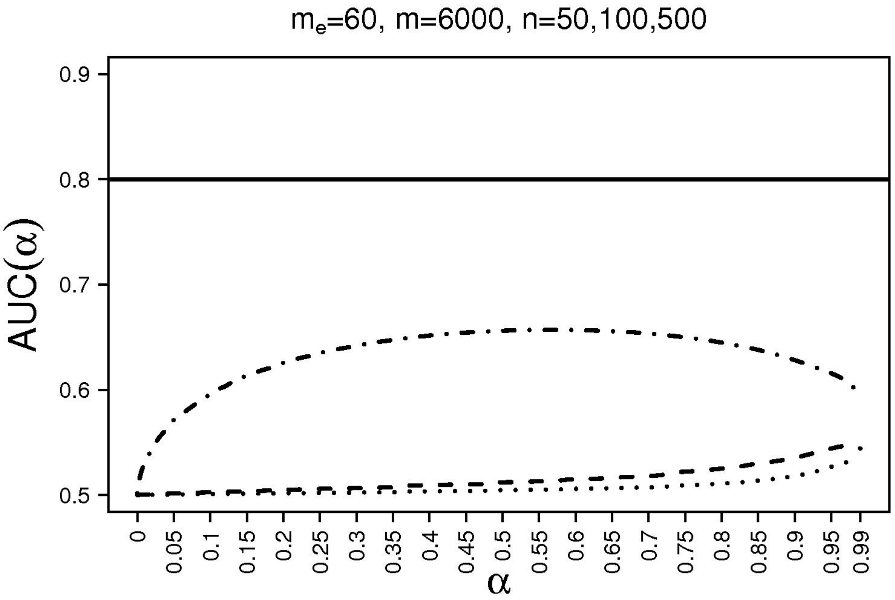

Figure 3.10: Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation runs) for a varying FDR threshold α for selection assuming $m_e = 60$ prognostic variables among $m = 6000$ tested variables. The sample size per group is set to $n = 50$ (dotted line), 100 (dashed line) and 500 (dotdashed line). $AUC_* = 0.8$ is given as solid horizontal line. The effect size $\Delta = 0.154$.

Table 3.2: **Simulation results for selection using the FDR approach assuming a smaller** AUC_***:** The best choice of the FDR threshold (α_{opt}), the corresponding $AUC(\alpha_{opt})$ as well as the number of the selected non-prognostic variables (m_0^s) and the number of selected prognostic variables (m_e^s) for varying number of prognostic variables (m_e), tested hypotheses (m) and per group sample sizes (n). $AUC_* = 0.8$.

m	m_e	Δ	n	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
1000	10	0.376	50	0.794	0.595	36.47	4.61
		0.376	100	0.493	0.675	7.03	5.50
		0.376	500	0.014	0.796	0.15	9.97
	60	0.154	50	0.919	0.582	405.62	37.75
		0.154	100	0.893	0.613	373.63	42.99
		0.154	500	0.442	0.720	36.13	44.27
6000	10	0.376	50	0.875	0.544	69.29	2.86
		0.376	100	0.594	0.612	8.25	3.45
		0.376	500	0.034	0.793	0.35	9.82
	60	0.154	50	0.981	0.532	2522.14	34.09
		0.154	100	0.972	0.547	2055.65	36.87
		0.154	500	0.561	0.657	37.68	27.87

3.6 Variable selection using forward logistic regression

For a comparison also forward logistic regression has been investigated as a method of variable selection. Again Δ was set to 0.811 in the case of $m_e = 10$ and to 0.331 for $m_e = 60$. Because of run-time problems and problems with memory capacity when testing a very large number of hypotheses m, simulations (1000 simulation steps) were only performed for $m = 1000$. Different thresholds γ are applied for the individual p-values in the stepwise selection based on the multiple logistic regression. The simulations for the logistic regression were done using the SAS 9.1. system.

The simulation results of the forward logistic regression show that this selection method for the investigated scenarios performs poor in terms of AUC values for prediction of the outcome of future patients as compared to the selection procedure using the FDR approach. The poor result may be due to the following reason: in a small training sample the forward logistic regression generally leads to a complete separation of data points, i.e. responders and non-responders of the validation data set can be fully separated with the found regression model. Only a few prognostic variables are selected for future prediction using the forward logistic regression which leads to the worse performances of the prediction scores.

For the independent case, the best performance for the situation of $m_e = 10$ occurs for $\gamma_{opt} = 0.0005$ with $AUC(\gamma_{opt}) = 0.812$ for $n = 50$ and again at $\gamma_{opt} = 0.0005$ with $AUC(\gamma_{opt}) = 0.900$ for $n = 100$. The forward logistic regression applying a larger sample size clearly performs better. However, $AUC(\gamma_{opt})$ for the selection procedure using the FDR was 0.941 for $n = 50$ and 0.961 for $n = 100$. Note that for $n = 50$ in average only 2.881 and for $n = 100$ only 5.644 out of the 10 alternatives are selected using the logistic regression.

In the case of $m_e = 60$ for $\gamma = 0.0073$ up to 0.05 the mean $AUC(\gamma)$ on average is 0.613 for $n = 50$. Because of complete separation almost the same performance is achieved for $\gamma \geq 0.0073$. Setting $n = 100$ per group a similar result can be seen. Again choosing $\gamma \geq 0.0073$ lead to the same performance achieving an average $AUC(\gamma) = 0.696$. Note that $AUC(\alpha_{opt})$ for selection using the FDR was 0.813 for $n = 50$ and 0.888 for $n = 100$ per group. However, the theoretically achievable AUC_* is 0.965, which is not achieved with both selection methods.

Note that if we assume correlation between hypotheses, the forward logistic regression model also results in complete separation after selecting a few variables. Furthermore considering a large positive autoregressive correlation structure between alternatives and the neighboring true null hypotheses (as considered later) leads to numerical problems and the model results in implausible estimates for the selected variables. This may also be due to the fact that the underlying data is not generated based on a logistic regression model. Due to the mentioned reasons the logistic regression is no good procedure to determine a prediction score for the given data structure and will not be considered in further applications.

3.7 Situation under the global null hypothesis

Under the global null hypothesis of no existing prognostic variable at all ($m_e = 0$) the ROC curve is always the diagonal (AUC$= 0.5$). Whatever selection procedure is used, if variables are selected, they are always non-prognostic variables (true null hypotheses) and thus the prediction score is useless. For selection using a multiple test controlling the FDR, by definition, the probability to end with a selection of variables and building a score is targeted at the pre-chosen FDR threshold α. Hence in the case of the global null hypothesis it would have been better to choose a small FDR. However, if a large number of alternatives are expected with rather small effect sizes a very large FDR should be chosen as selection criterion.

If we select variables using the forward logistic regression, we may have problems to evaluate the level of false discoveries. However, as for the FDR selection method, the results depend on the boundary γ chosen a priori. E.g. if we decided for Bonferroni corrected boundaries ($\gamma = 0.00005$) in only 2.9% of the 1000 simulated cases at least one variable was identified for future prediction but when increasing γ to 0.0025 in 98.3%

useless prediction scores are produced.

This again demonstrates the dilemma of the task we are faced with. It may be possible to improve the selection and estimation procedures but a contradiction will remain: being cautious may help not to produce too many nuisance results if the postulated relationships do not exist. Being more optimistic and liberal may improve the results if in fact variables related with the clinical outcome exist, but under the global null useless scores may be produced.

4 A cross validation method to estimate an appropriate selection criterion

We have seen from the previous sections that under the given assumptions the forward logistic regression may not be a good selection procedure if small sample sizes are given. There are problems to quantify the number of false positives in a score built with the logistic regression model. Furthermore, there may be numerical problems in the calculations of the estimates. The simple method using a multiple test controlling the FDR and just building a weighted sum of the selected hypotheses (as in the classical discriminant analysis) leads, if the right selection threshold is chosen, to a better performance in terms of AUC values for prediction of the outcome of a future patient. However, it remains the question, how to choose the "optimal" selection boundary, because depending on the parameter constellation (varying number of tested hypotheses, varying proportion of prognostic variables and varying sample size), different boundaries are required in order to achieve a large AUC for future independent patients.

4.1 The cross validation procedure

To estimate an appropriate selection threshold $\hat{\alpha}_{opt}$ for selection using a multiple test controlling the FDR, we investigated in Goll (2008) and Goll and Bauer (2008) a cross validation procedure. Note that therefore we again first assume that the variable levels follow independent normal distributions with mean vector $\boldsymbol{\mu}_r$ for responder and $\boldsymbol{\mu}_{nr}$ for non-responder and known variance $\sigma^2 = 1$. Deviations from these simple assumptions are considered later.

For this cross validation procedure we decided to search for the optimal $\hat{\gamma}_{opt}$ for the individual p-values instead of $\hat{\alpha}_{opt}$ because of the extremely longer runtime needed to search for the corresponding γ values in each training set. For the final selection boundary $\hat{\gamma}_{opt}$ the corresponding FDR threshold $\hat{\alpha}_{opt}$ can be estimated with Storey's estimator (see Section 2.3.2) in the total sample. Storey et al. (2004) showed that searching for the optimal FDR and γ asymptotically leads to the same. However, despite the long runtime, for some scenarios we performed simulations of the procedure searching for $\hat{\alpha}_{opt}$. The results were very similar (data not shown).

The cross validation procedure works as follows:

From a given data set with n_r responders and n_{nr} non-responders, a pair of a single responder and a single non-responder respectively is left out. Note that there are $n_r n_{nr}$ possibilities ($= n^2$ if $n_r = n_{nr} = n$ as assumed in the previous sections) for leaving out a pair of one responder and one non-responder. The remaining $(n_r + n_{nr} - 2)$ patients ($n_r - 1$ responder and $n_{nr} - 1$ non-responder) in each of the $n_r n_{nr}$ "training" samples respectively are used to estimate prediction scores applying a grid of values γ for the selection boundary. As discussed in Section 3.2, the variables, whose one sided p-values lie below the selection boundary γ are selected to build a score for future prediction. For the left out responder and non-responder respectively we now calculate for each γ, the value of the corresponding prediction score:

$$\hat{f}_{(ij)}(\mathbf{x}_{r,i}; \gamma) = \hat{\mathbf{c}}_{(ij)}(\gamma)^T \mathbf{x}_{r,i} \text{ and } \hat{f}_{(ij)}(\mathbf{x}_{nr,j}; \gamma) = \hat{\mathbf{c}}_{(ij)}(\gamma)^T \mathbf{x}_{nr,j}$$

where $\hat{\mathbf{c}}_{(ij)}(\gamma)$ is the vector of the weights of the score calculated from the training sample leaving out the ith responder and the jth non-responder, using γ as selection boundary. $\mathbf{x}_{r,i}$ and $\mathbf{x}_{nr,j}$ denote the corresponding values of the (selected) variables of the single responder and non-responder respectively left out in the construction of the score. Now for each investigated γ the following cross validation function CF_{ij} is calculated.

Definition 4.1.0.1 *For a given selection threshold γ, the cross validation function is calculated by:*

$$CF_{ij}(\mathbf{x}_{r,i}, \mathbf{x}_{nr,j}; \gamma) = \begin{cases} 0 & \textit{if } \hat{f}_{(ij)}(\mathbf{x}_{r,i}; \gamma) < \hat{f}_{(ij)}(\mathbf{x}_{nr,j}; \gamma) \\ 1 & \textit{if } \hat{f}_{(ij)}(\mathbf{x}_{r,i}; \gamma) > \hat{f}_{(ij)}(\mathbf{x}_{nr,j}; \gamma) \\ 0.5 & \textit{if } \hat{f}_{(ij)}(\mathbf{x}_{r,i}; \gamma) = \hat{f}_{(ij)}(\mathbf{x}_{nr,j}; \gamma) \end{cases} . \tag{4.1}$$

If no prediction score is selected from the data using the given γ, $CF_{ij}(\mathbf{x}_{r,i}, \mathbf{x}_{nr,j}; \gamma) = 0.5$.

The values of $CF_{ij}(\mathbf{x}_{r,i}, \mathbf{x}_{nr,j}; \gamma)$ for each γ are now calculated for all $n_r n_{nr}$ training samples. Note that in our balanced scenario overall we use $n_r n_{nr}$ pairs of a single responder and non-responder as validation sample. For each γ we can calculate a "cross validation based" area under the ROC-curve.

Definition 4.1.0.2 *For a given γ, the cross validation based $\widehat{AUC}(\gamma)$ is calculated by:*

$$\widehat{AUC}(\gamma) = \frac{1}{n_r n_{nr}} \sum_{i=1}^{n_r} \sum_{j=1}^{n_{nr}} CF_{ij}(\mathbf{x}_{r,i}, \mathbf{x}_{nr,j}; \gamma). \tag{4.2}$$

It can be shown that the Mann-Whitney-U statistic (4.2) is the AUC of the empirical ROC-curve in the independent sample case (see theorem 2.5.4.1 and e.g. Hanley and McNeil (1982), Pepe (2003), Pepe et al. (2006)). Finally we choose the selection boundary $\hat{\gamma}_{opt}$ such that it maximizes $\widehat{AUC}(\gamma)$.

Definition 4.1.0.3 *The best choice of the selection boundary, $\hat{\gamma}_{opt}$, is calculated by:*

$$\hat{\gamma}_{opt} = \arg\max_{\gamma} \left[\sum_{i=1}^{n_r} \sum_{j=1}^{n_{nr}} CF_{ij}(\mathbf{x}_{r,i}, \mathbf{x}_{nr,j}; \gamma) \right] \tag{4.3}$$

This is a special case of the maximum rank correlation estimator known to be consistent and asymptotically normal when used for the parameters of a the generalized linear model with a given set of predictors (see Han (1987), Sherman (1993), Pepe et al. (2006)).

As already mentioned before the corresponding FDR threshold $\hat{\alpha}_{opt}$ is calculated as follows:

Definition 4.1.0.4 *The best choice of the threshold α for the FDR, $\hat{\alpha}_{opt}$, can be calculated as function of $\hat{\gamma}_{opt}$ using Storey's estimator:*

$$\hat{\alpha}_{opt}(\hat{\gamma}_{opt}) = \frac{\frac{\sharp\{p_i > \lambda\}}{(1-\lambda)} \hat{\gamma}_{opt}}{\max(\sharp\{p_i < \hat{\gamma}_{opt}\}, 1)}$$

where p_i, $i = 1, ..., m$ are the p-values from the individual z-tests calculated from the total sample of n_r responder and n_{nr} non-responder and λ is a constant chosen a priori (compare Section 2.3.2). $\hat{\alpha}_{opt}(\hat{\gamma}_{opt})$ is further on only denoted by $\hat{\alpha}_{opt}$.

Note that in the following λ is set to 0.5 as in Storey (2002) (compare Section 2.3.2).

Below we will investigate by simulation how this estimator of the best FDR threshold α behaves for increasing sample sizes when the set of predictors is chosen from model selection. Generally a large number of weights in the score is set to zero by selection based on an α which is chosen in a data driven way by optimizing $\widehat{AUC}(\gamma)$.

Note also that if more than one γ fulfills the cross validation criterion (4.3), the minimum of these γ values is chosen as final selection boundary.

Since we are working with simulations we can also calculate the asymptotic FDR:

Definition 4.1.0.5 *Assuming that π_0 and Δ are known the best choice of the FDR threshold can be calculated directly from $\hat{\gamma}_{opt}$ by:*

$$\hat{\alpha}_{opt,\infty}(\hat{\gamma}_{opt}, \Delta, \pi_0) = \frac{\pi_0 \hat{\gamma}_{opt}}{\pi_0 \hat{\gamma}_{opt} + (1 - \pi_0)(1 - \beta(\hat{\gamma}_{opt}))}$$

where $(1 - \beta(\hat{\gamma}_{opt})) = 1 - \Phi_{\sqrt{\frac{n}{2}}\Delta,1}(c_{1-\hat{\gamma}_{opt}})$ is the power of the performed one-sided two-sample z-tests (compare Section 2.3.2). $\hat{\alpha}_{opt,\infty}(\hat{\gamma}_{opt}, \Delta, \pi_0)$ is further on only denoted by $\hat{\alpha}_{opt,\infty}$.

It may be also interesting to look at the FWER, which is calculated numerically in the simulations below which turns out to be close to the value $(1 - (1 - \hat{\gamma}_{opt})^{m\pi_0})$ ignoring the random nature of $\hat{\gamma}_{opt}$.

4.2 Cross validation under the alternative

To investigate the cross validation method discussed above we performed simulations for the scenarios assuming $m_e = 10$ and 60 prognostic variables (alternatives) among $m = 1000$ and 6000 tested variables (hypotheses) setting the sample size to $n = 50$ per group. For the scenarios testing $m = 1000$ hypotheses we also investigate the cross validation procedure fixing the per-group sample size to $n = 100$.

Table 4.1 shows the results of the cross validation procedure for the investigated examples (mean values (standard deviations) and *medians* over 500 simulation runs) assuming that the best achievable $AUC_* = 0.965$. The α_{opt} and $AUC(\alpha_{opt})$ values evaluated in the last chapter are also given. If we compare the estimated $\hat{\alpha}_{opt}$ determined by the cross validation procedure to the true α_{opt} we see that there may on average be large differences. Therefore it has to be considered that the optima of the interpolated functions in the different scenarios are generally flat, i.e. we get similar performances among a wide range of α values. Furthermore, because of the skew distribution of $\hat{\alpha}_{opt}$ the medians over the simulations are always closer to the α_{opt} values (see Table 4.1).

The crucial finding is that the true FDR (determined from the simulations) in the selection procedure is always close to the threshold $\hat{\alpha}_{opt}$ determined by cross validation from the data. This behavior is related to theoretical results on the convergence of the FDR simultaneously for different thresholds (Genovese and Wasserman (2004)).

Despite the differences between $\hat{\alpha}_{opt}$ and α_{opt}, the determined $\hat{\gamma}_{opt}$ and accordingly calculated $\hat{\alpha}_{opt}$ values are on average leading to true $AUC(\hat{\gamma}_{opt})$ values for independent future patients which are only slightly smaller than the evaluated $AUC(\alpha_{opt})$ over the whole investigated examples. Note that the true $AUC(\hat{\gamma}_{opt})$ is calculated using formula (3.3) given $\hat{\gamma}_{opt}$ for selection. Increasing the samples size generally leads to smaller differences between $\hat{\alpha}_{opt}$ and α_{opt} and between $AUC(\hat{\gamma}_{opt})$ and $AUC(\alpha_{opt})$. Thus, this cross validation method seems to work under the alternative producing prognostic scores with a true AUC in future patients close to the (for the different scenarios) best possible $AUC(\alpha_{opt})$ when using a multiple testing procedure controlling the FDR for selection.

From the cross validation procedure we also get a positively biased estimate $\widehat{AUC}(\hat{\gamma}_{opt})$ (compare formula (4.2)) of the true $AUC(\alpha_{opt})$ which is closer to the truth the larger the effect and sample sizes (Table 4.1). However, e.g. in the situation assuming $m_e = 60$ alternatives among $m = 6000$ tested hypotheses $\widehat{AUC}(\hat{\gamma}_{opt}) = 0.802$ largely overestimates the true $AUC(\hat{\gamma}_{opt}) = 0.669$.

A further important fact can be seen from the simulation results: the family wise type I error rate (FWER) determined from the simulations is generally very large in selections achieving a good prediction score (Table 4.1). This reflects the fact that allowing at least one non-prognostic variable in the prediction score in order to detect more prognostic variables leads to better performances in terms of the AUC.

In the situation of a small AUC_* close to 0.8 (see results of the cross validation procedure in Table 4.2) the cross validation procedure leads to smaller cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ values indicating a poorer prediction as compared to the situation of $AUC_* = 0.965$. For the resulting prediction scores, the same tendencies of the procedure are found as for $AUC_* = 0.965$. The procedure is again resulting in $\hat{\gamma}_{opt}$ and corresponding $\hat{\alpha}_{opt}$ values

Table 4.1: **Results using the cross validation procedure:** The true best choice of the FDR (α_{opt}) and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure (means (standard deviations) and *medians* over 500 simulation runs): the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true future $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e, per-group sample sizes n and number of tested hypotheses m assuming $AUC_* = 0.965$.

m	1000				6000	
m_e	10		60		10	60
n	50	100	50	100	50	50
α_{opt}	0.170	0.014	0.824	0.475	0.250	0.896
$AUC(\alpha_{opt})$	0.941	0.963	0.813	0.888	0.917	0.669
$\hat{\gamma}_{opt}$	0.005 (0.01)	0.003 (0.01)	0.097 (0.09)	0.053 (0.05)	0.001 (0.001)	0.034 (0.05)
	0.002	*0.001*	*0.060*	*0.039*	*0.0005*	*0.013*
$\hat{\alpha}_{opt}$	0.243 (0.19)	0.149 (0.19)	0.615 (0.18)	0.451 (0.17)	0.286 (0.19)	0.812 (0.13)
	0.190	*0.061*	*0.643*	*0.459*	*0.250*	*0.835*
FDR	0.254 (0.23)	0.165 (0.21)	0.604 (0.20)	0.444 (0.19)	0.297 (0.27)	0.783 (0.18)
	0.200	*0.091*	*0.645*	*0.462*	*0.226*	*0.817*
$\hat{\alpha}_{opt,\infty}$	0.256 (0.20)	0.154 (0.19)	0.613 (0.18)	0.452 (0.17)	0.311 (0.22)	0.801 (0.12)
	0.191	*0.065*	*0.636*	*0.460*	*0.260*	*0.820*
$AUC(\hat{\gamma}_{opt})$	0.934 (0.02)	0.957 (0.01)	0.796 (0.04)	0.881 (0.02)	0.910 (0.03)	0.667 (0.03)
	0.938	*0.960*	*0.803*	*0.883*	*0.918*	*0.669*
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.956 (0.02)	0.966 (0.01)	0.868 (0.05)	0.916 (0.03)	0.945 (0.03)	0.802 (0.07)
	0.959	*0.967*	*0.873*	*0.918*	*0.948*	*0.810*
FWER	0.760	0.586	0.984	0.998	0.720	0.980

Table 4.2: **Results using the cross validation procedure assuming a smaller AUC_*:** The true best choice of the FDR, α_{opt} and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure (means (standard deviations) and *medians* over 500 simulation runs): the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true future $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e, per-group sample sizes n fixing the number of tested hypotheses $m = 1000$ assuming $AUC_* = 0.80$.

m	1000			
m_e	10		60	
n	50	100	50	100
α_{opt}	0.794	0.493	0.919	0.893
$AUC(\alpha_{opt})$	0.595	0.675	0.582	0.613
$\hat{\gamma}_{opt}$	0.042 (0.06)	0.020 (0.03)	0.136 (0.23)	0.146 (0.23)
	0.013	*0.008*	*0.039*	*0.042*
$\hat{\alpha}_{opt}$	0.713 (0.22)	0.546 (0.24)	0.815 (0.16)	0.734 (0.17)
	0.763	*0.560*	*0.852*	*0.757*
FDR	0.711 (0.28)	0.521 (0.29)	0.779 (0.20)	0.700 (0.20)
	0.813	*0.556*	*0.833*	*0.752*
$\hat{\alpha}_{opt,\infty})$	0.735 (0.18)	0.552 (0.23)	0.781 (0.09)	0.712 (0.14)
	0.772	*0.553*	*0.797*	*0.723*
$AUC(\hat{\gamma}_{opt})$	0.598 (0.08)	0.676 (0.04)	0.551 (0.10)	0.587 (0.03)
	0.600	*0.678*	*0.559*	*0.593*
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.741 (0.07)	0.765 (0.05)	0.701 (0.07)	0.700 (0.05)
	0.745	*0.767*	*0.702*	*0.700*
FWER	0.920	0.890	0.962	0.980 (0.07)

leading to prediction scores with $AUC(\hat{\gamma}_{opt})$ values close to $AUC(\alpha_{opt})$.

4.3 Cross validation under the global null hypothesis

Under the global null hypothesis the cross validation procedure searching for decision boundaries resulting in the "best" cross validated ROC-curve will generally produce a score which is always useless for prediction of future outcomes. Note that only in a few cases the FDR threshold determined by cross validation will not lead to selection of any variable in the total sample. In this cases the AUC is set to 0.5 in the following. Note that the true FDR is then 0.

Figures 4.1 show the distributions of the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ of the simulated samples for the examples assuming $m_e = 10$ (first row), 60 (second row) prognostic variables as well as under the global null hypotheses ($m_e = 0$, third row) searching among $m = 1000$ tested hypotheses (Figure from Goll (2008)). A sample size of $n = 50$ (first column) and $n = 100$ (second column) per group is applied. The histograms show that in case of $m_e = 10$, $\widehat{AUC}(\hat{\gamma}_{opt})$ is generally larger than 0.8 whereas under the global null $\widehat{AUC}(\hat{\gamma}_{opt})$ is generally below 0.8. Thus, one criterion could be the following: if the $\widehat{AUC}(\hat{\gamma}_{opt})$ resulting from the cross validation procedure is smaller than a value relevant for practicable purposes then it seems to be preferable not to construct any score at all. Note that applying a sample size of $n = 100$ per group (second column) there is no overlap between the distributions under the alternative and under the global null hypothesis in the simulated samples. Assuming $m_e = 60$ prognostic variables and fixing $n = 50$ per group only a small overlap can be seen if we are searching among 1000 hypotheses.

A further observation may be used for the decision to construct a score or not: under the global null hypothesis generally $\hat{\alpha}_{opt}$ takes very large values exceeding 0.9. The histograms of $\hat{\alpha}_{opt}$ under the alternative (first row: $m_e = 10$, second row: $m_e = 60$) and under the global null hypotheses (third row) are shown in Figure 4.2 for $n = 50$ (first column) and $n = 100$ (second column). Under the alternative $\hat{\alpha}_{opt}$ is varying largely, although the average generally being much smaller than under the global null. Therefore, a large $\hat{\alpha}_{opt}$ found in a real data set may be a good reason to decide against the score because this may signal that we are under the global null hypotheses or that the sample size is to small to detect the given effects. Moreover in such a situation we have to expect that most of the selected variables will not contribute to prediction anyway.

It also has to be mentioned that under the global null the mean estimate $\hat{\alpha}_{opt}$ calculated from $\hat{\gamma}_{opt}$ may be much smaller than the true FDR. Table 4.3 summarizes the results of the cross validation procedure under the global null hypothesis. Note that $\hat{\alpha}_{opt,\infty}$ is always 1 and the true $AUC(\hat{\gamma}_{opt})$ is always 0.5. For comparison to the results under the alternative see Table 4.1.

Looking at both criteria may help to decide for or against the prediction scores. The Scatterplots for $\widehat{AUC}(\hat{\gamma}_{opt})$ versus $\hat{\alpha}_{opt}$ in Figure 4.3 give an overview over the combination of both criteria. It can be seen that the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ is varying less than the estimated $\hat{\alpha}_{opt}$. If both criteria are large this may be a reason to decide against the sore despite the large area under the curve because one may be confident that a large number of non-prognostic variables are included in the score.

Increasing the number of tested hypotheses to $m = 6000$ in the situation of $m_e = 10$ (Figure 4.4: left plot) again $\widehat{AUC}(\hat{\gamma}_{opt})$ is generally larger than 0.8. However, when expecting a larger number of prognostic variables with rather small effect sizes ($m_e = 60$: Figure 4.4 central plot) the distribution of $\widehat{AUC}(\hat{\gamma}_{opt})$ largely overlaps the distribution under the global null hypothesis (right plot). E.g., deciding to construct a score only if the $\widehat{AUC}(\hat{\gamma}_{opt})$ is larger than 0.8 would lead to a false negative decision in more than one half of the cases. Constructing scores only if $\widehat{AUC}(\hat{\gamma}_{opt})$ exceeds 0.7 would reduce the false negative decisions, however increase the false positive decisions to 30% under the global null hypotheses. Again, if the estimated $\widehat{AUC}(\hat{\gamma}_{opt})$ is small this may be an indication that in a specific sample we are close to the global null hypotheses or the sample size is too small to detect the prognostic variables with their rather small effects. $\hat{\alpha}_{opt}$ is again varying largely under the alternative and under the global null (see histograms in Figure 4.5).

Summing up the results for $m = 6000$ it can be seen that if a small number of prognostic variables with large effect sizes is assumed (Figure 4.6, left plot), $\widehat{AUC}(\hat{\gamma}_{opt})$ tends to be large and $\hat{\alpha}_{opt}$ (despite the larger variation) tends to be small suggesting that there may be a good prediction of the response of future patient to a specific therapy. For $m_e = 60$ (see Figure 4.6, central plot) and under the global null (Figure 4.6, right plot) $\widehat{AUC}(\hat{\gamma}_{opt})$ tends to be small and $\hat{\alpha}_{opt}$ tends to be large which indicates that no good prediction score for future patients can be determined from the given data. See also Table 4.3 for the results under the global null and Table 4.1 for the results under the alternative.

The histograms of $\widehat{AUC}(\hat{\gamma}_{opt})$ in Figure 4.7, of $\hat{\alpha}_{opt}$ in Figure 4.8 and the scatterplots in Figure 4.9 show the situation assuming a smaller AUC_* of 0.8. In all three Figures the situations of $m_e = 10$ (first row), $m_e = 60$ (second row) and $m_e = 0$ (third row) are considered for $n = 50$ (first column) and $n = 100$ (second column). In the situation of a small AUC_* it may become difficult to distinguish between the situations under the alternative and under the global null if only small samples are used to search for a prediction score. Under the alternative $\hat{\alpha}_{opt}$ varies largely and $\widehat{AUC}(\hat{\gamma}_{opt})$ tends to be small. Thus, larger sample sizes are needed to detect good prognostic scores. However, increasing the sample size to $n = 100$ only slightly increases the performances of the determined prediction scores. However again, the results of the cross validation procedure are reflecting the poor performance of the detected prediction scores under the alternative and under the global null (see Table 4.3). For comparison to the results of the cross validation procedure under the alternative see Table 4.2.

Table 4.3: **Results using the cross validation procedure under the global null hypotheses:** means (standard deviations) and *medians* over 500 simulation runs of the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR, the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of per-group sample sizes n and tested hypotheses m. Note that $\hat{\alpha}_{opt,\infty}$ is always 1 and $AUC(\hat{\gamma}_{opt})$ is always 0.5.

m	1000		6000
n	50	100	50
$\hat{\gamma}_{opt}$	0.023 (0.03) *0.010*	0.020 (0.03) *0.009*	0.003 (0.003) *0.004*
$\hat{\alpha}_{opt}$	0.855 (0.18) *0.919*	0.857 (0.18) *0.924*	0.818 (0.21) *1.000*
FDR	0.950 (0.22) *1.000*	0.986 (0.18) *1.000*	0.920 (0.27) *1.000*
$\hat{\alpha}_{opt,\infty}$	1.000	1.000	1.000
$AUC(\hat{\gamma}_{opt})$	0.500	0.500	0.500
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.657 (0.08) *0.662*	0.642 (0.06) *0.644*	0.657 (0.08) *0.725*
FWER	0.950	0.986	0.920

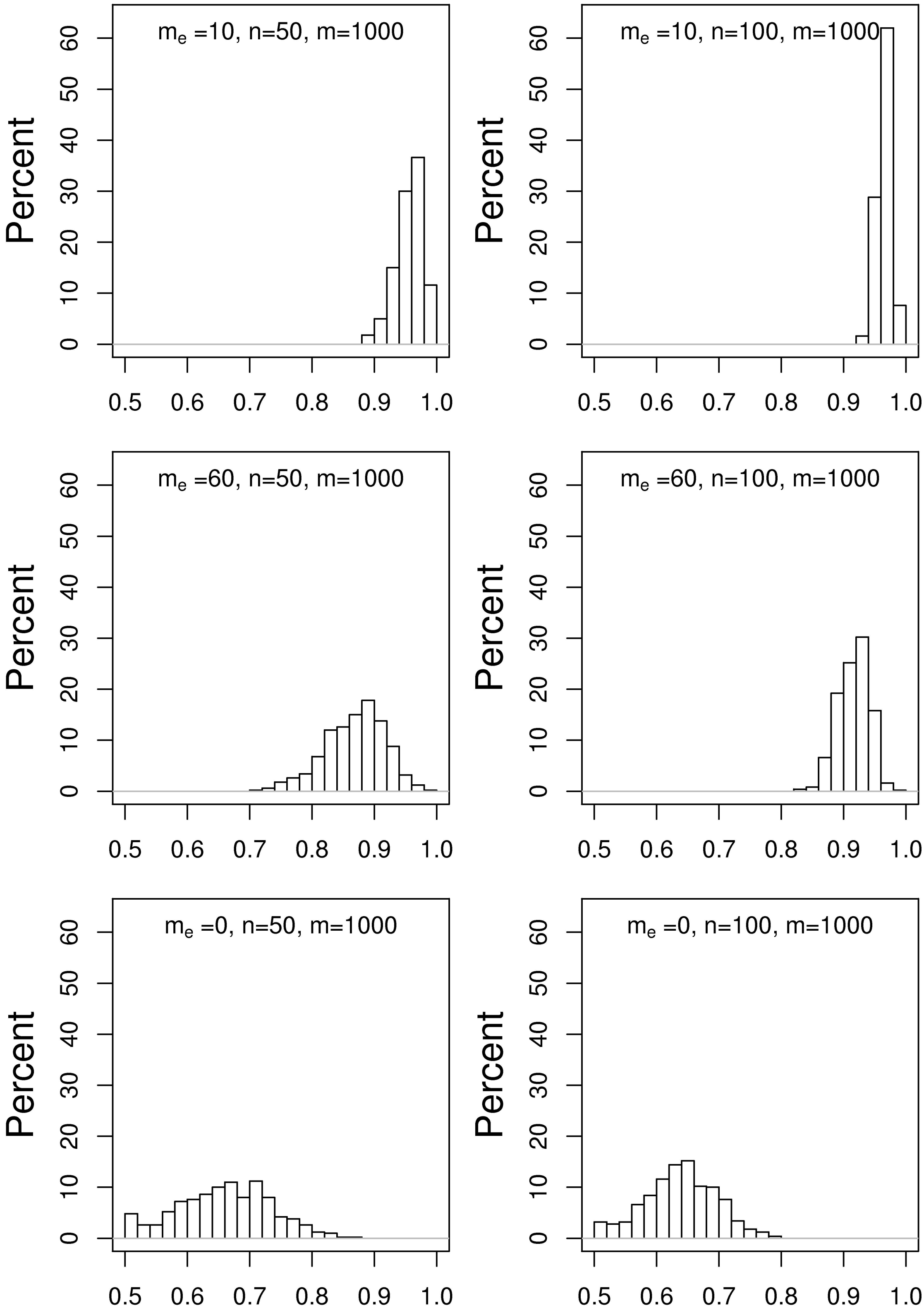

Figure 4.1: Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.965$ (Figure from Goll (2008)).

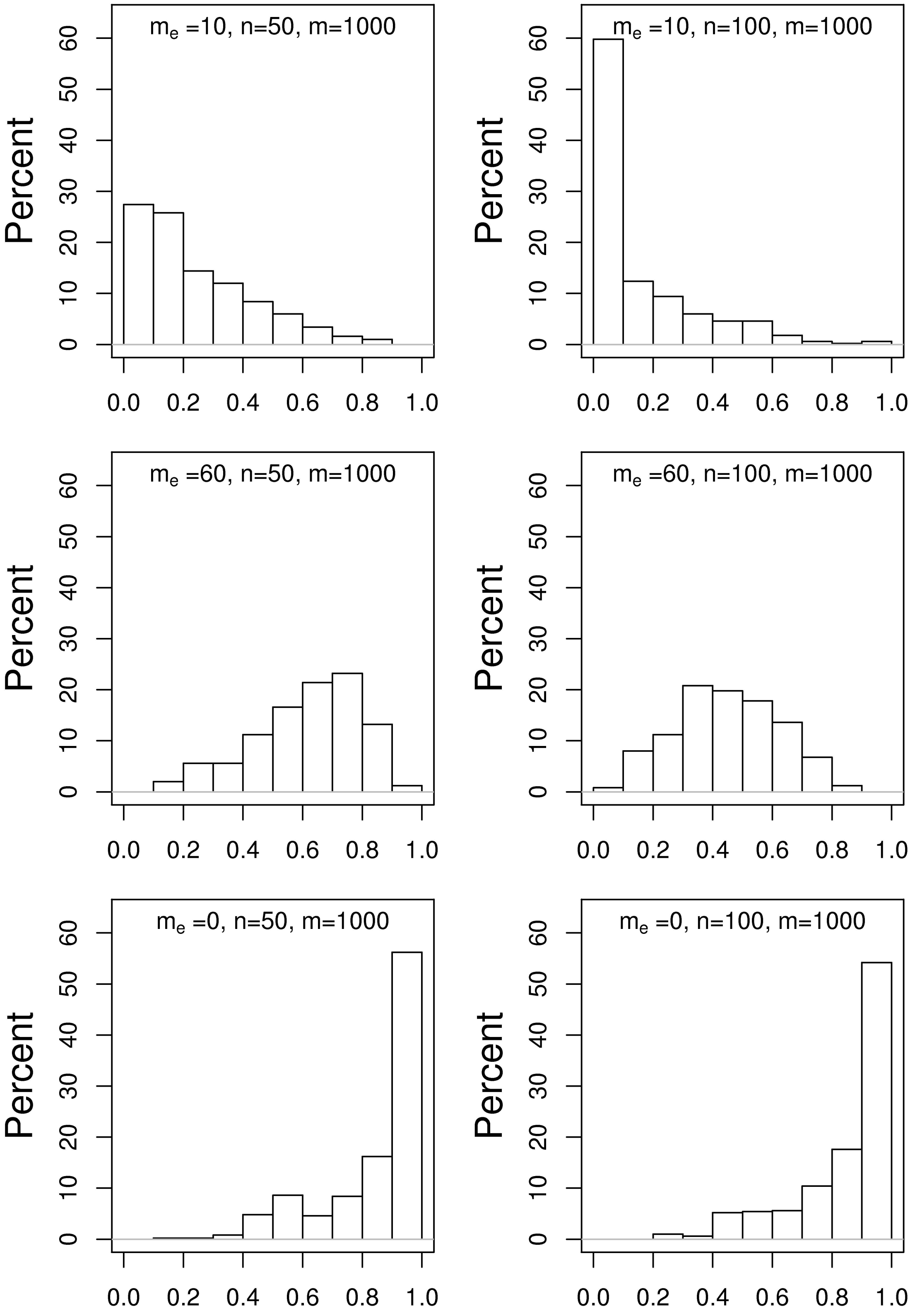

Figure 4.2: Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.965$.

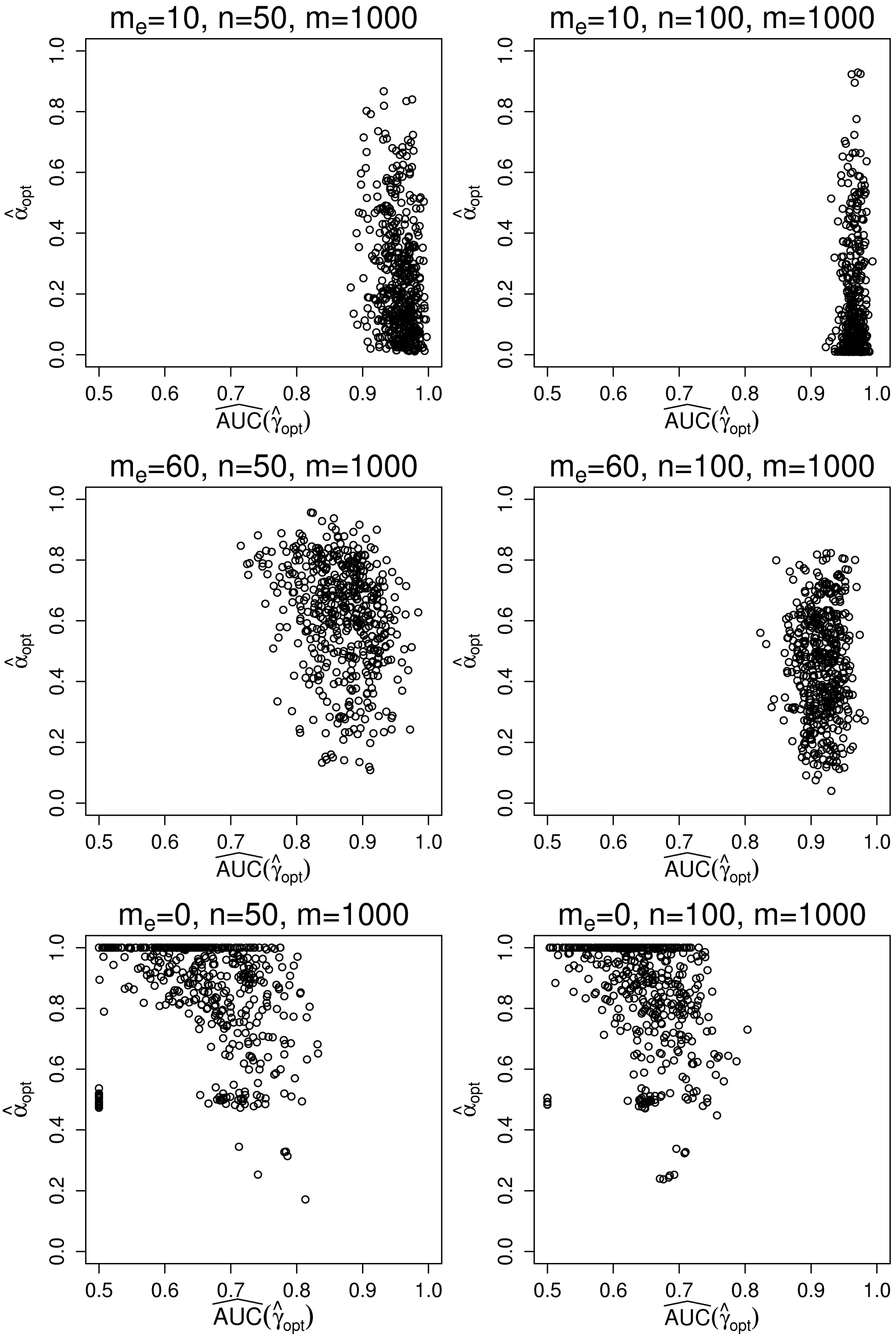

Figure 4.3: Scatterplots of $\widehat{AUC}(\gamma)$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.965$.

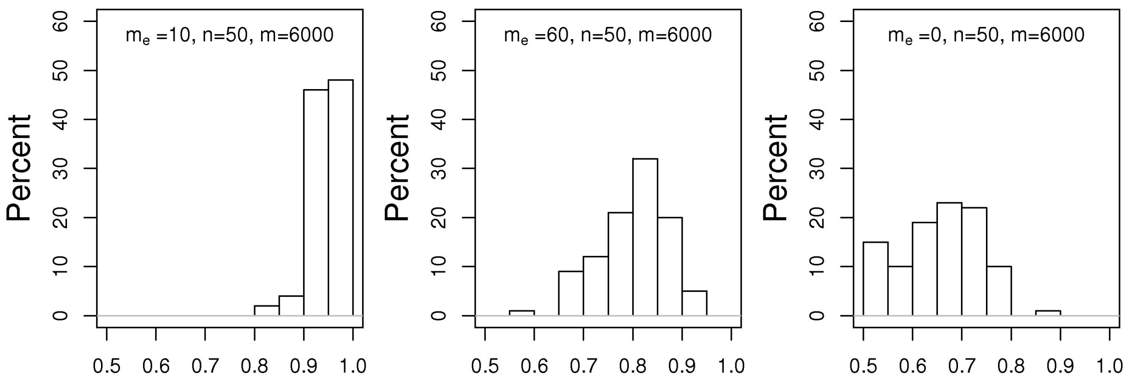

Figure 4.4: Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ for selection using (100 simulated steps): $m_e = 10$ (left plot), 60 (central plot) or 0 among $m = 6000$ (right plot)hypotheses. The sample size was set to $n = 50$ per group. $AUC_* = 0.965$.

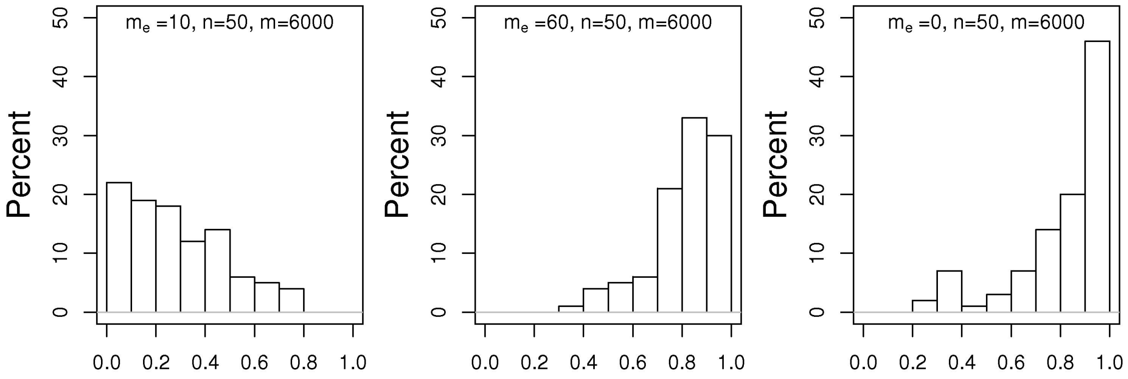

Figure 4.5: Distribution of the cross validation based $\hat{\alpha}_{opt}$ (100 simulation runs): $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 6000$ hypotheses. The sample size was set to $n = 50$ per group. $AUC_* = 0.965$.

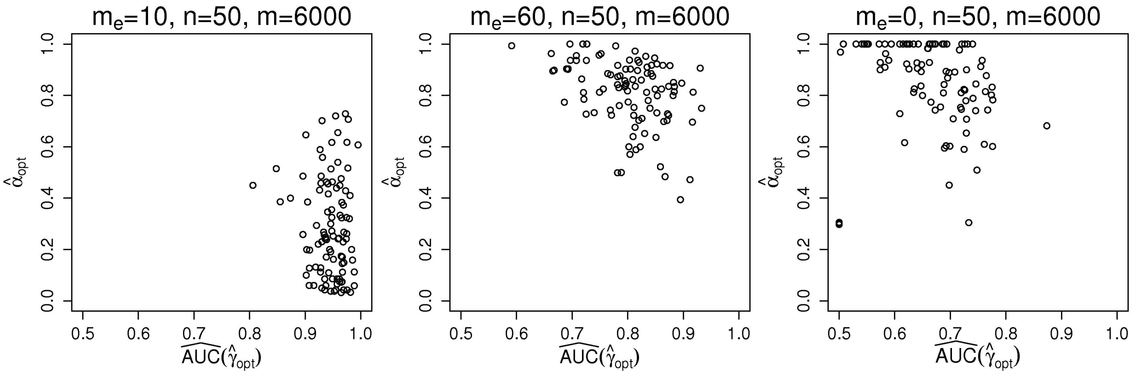

Figure 4.6: Scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (100 simulation runs): $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 6000$ hypotheses. The sample size was set to $n = 50$ per group. $AUC_* = 0.965$.

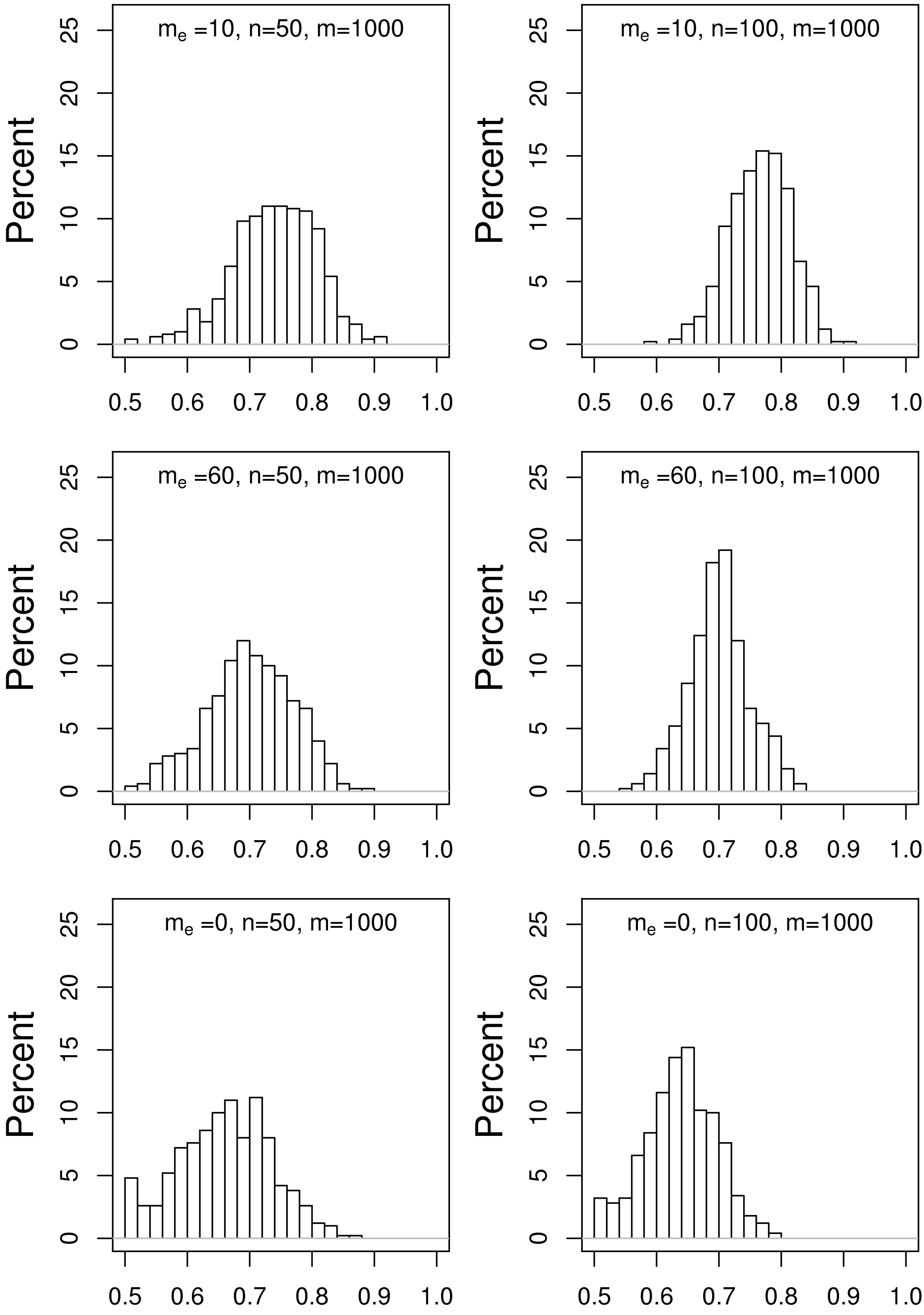

Figure 4.7: **Cross validation assuming a smaller** AUC_***:** Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.8$

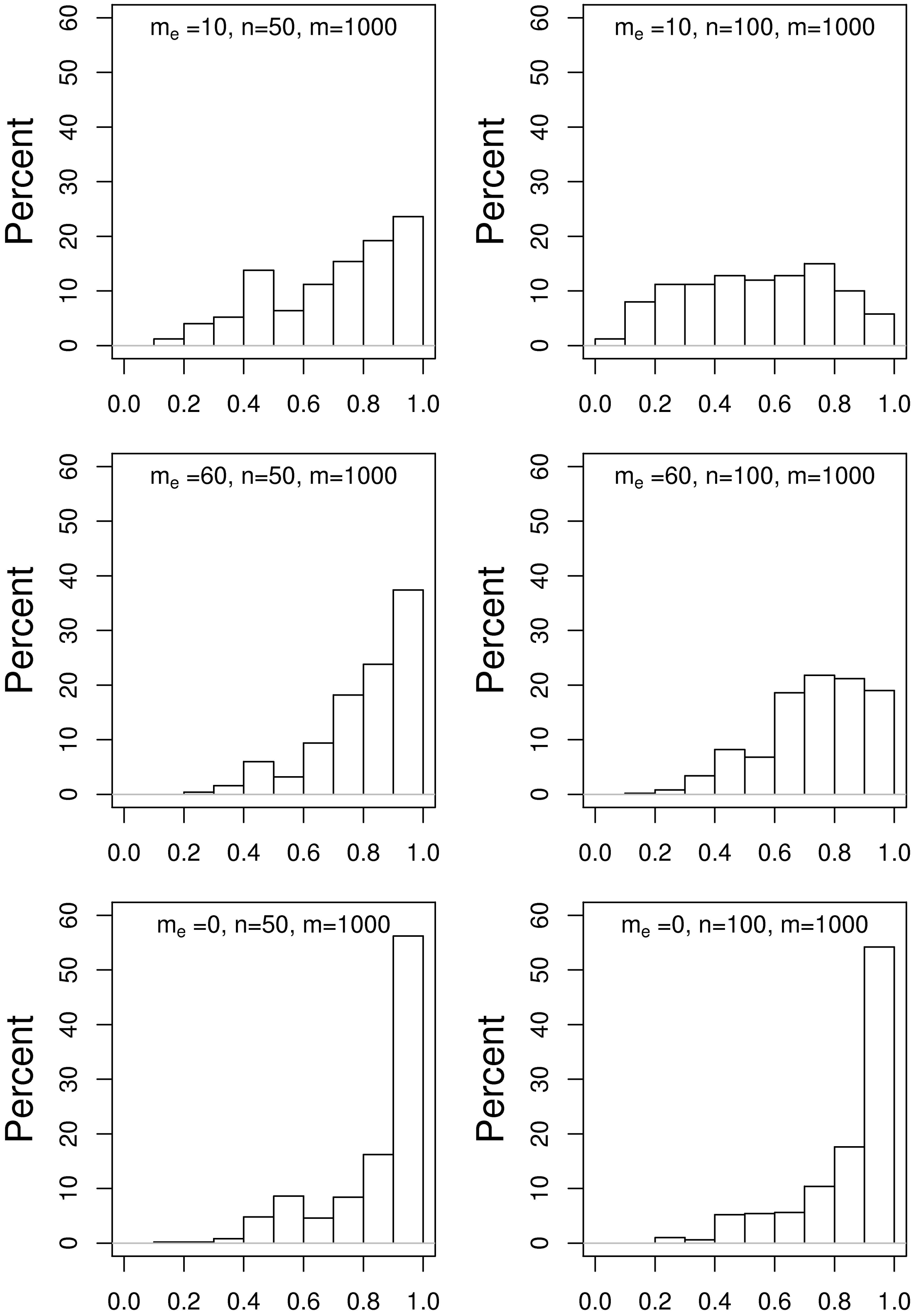

Figure 4.8: **Cross validation assuming a smaller** AUC_***:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.8$.

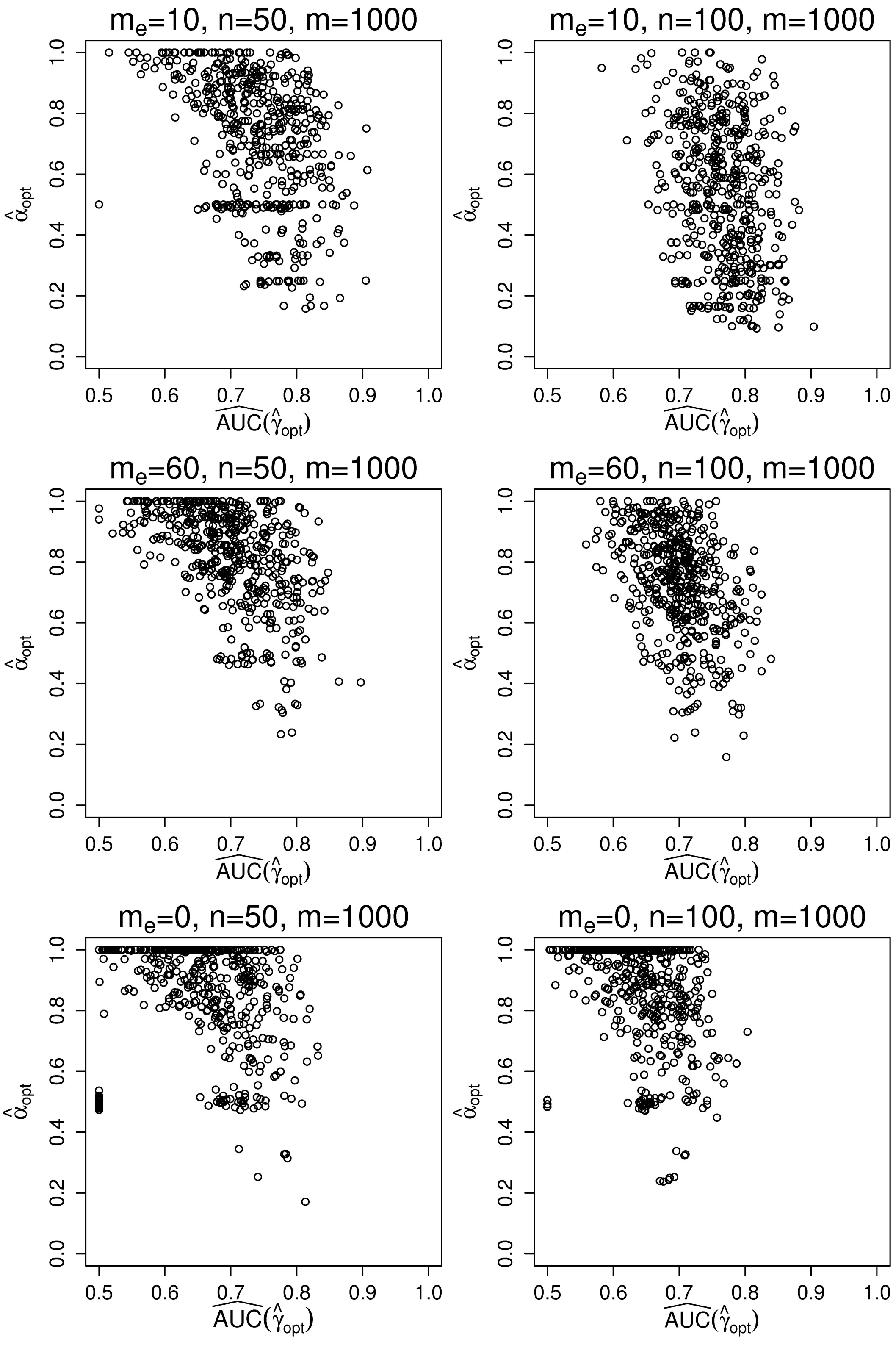

Figure 4.9: **Cross validation assuming a smaller** AUC_***:** Scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column). $AUC_* = 0.8$.

4.4 Cross validation using the mean difference in score values

In the following we will investigate another type of the cross validation method, where the difference between the score values of the leaved out responder and the non-responder is used instead of the cross validation function CF_{ij}. The final selection boundary $\hat{\gamma}_{opt}$ is than chosen such that it maximizes the mean value of the score differences over all $n_r n_{nr}$ possibilities of leaving out one responder and non-responder respectively.

Definition 4.4.0.6 *Let $\hat{f}_{(ij)}(\mathbf{x}_{r,i};\gamma)$ and $\hat{f}_{(ij)}(\mathbf{x}_{nr,j};\gamma)$ be the score values of the left out responder and non-responder respectively. $\hat{f}_{(ij)}$ was determined from the training sample where the ith responder and the jth non-responder was left out and γ was used as selection boundary. The mean difference between the score values is calculated as function of γ by:*

$$\widehat{MD}(\gamma) = \frac{1}{n_r n_{nr}} \sum_{i=1}^{n_r} \sum_{j=1}^{n_{nr}} (\hat{f}_{(ij)}(\mathbf{x}_{r,i};\gamma) - \hat{f}_{(ij)}(\mathbf{x}_{nr,j};\gamma)) \tag{4.4}$$

Definition 4.4.0.7 *The best choice of the final selection boundary, $\hat{\gamma}_{opt}$, is then calculated by:*

$$\hat{\gamma}_{opt} = \arg\max_{\gamma} \left[\frac{1}{n_r n_{nr}} \sum_{i=1}^{n_r} \sum_{j=1}^{n_r} (\hat{f}_{(ij)}(\mathbf{x}_{r,i};\gamma) - \hat{f}_{(ij)}(\mathbf{x}_{nr,j};\gamma)) \right] \tag{4.5}$$

Table 4.4 shows the results applying the cross validation procedure using the difference in score values for a varying number of prognostic variables $m_e = 10$, 60 or 0 and per-group sample sizes $n = 50$ or 100, fixing $m = 1000$ and assuming $AUC_* = 0.965$. The mean difference $\widehat{MD}(\hat{\gamma}_{opt})$ tends to be small under the global null hypotheses and to be large under the alternative. The histograms of $\widehat{MD}(\hat{\gamma}_{opt})$ of the different scenarios can be seen in Figure 4.10. The figure shows a large overlap of the distributions under the alternatives of $m_e = 10$ and 60 (first and second row).

It can also be seen from Table 4.4 that the differences between $\hat{\alpha}_{opt}$ and α_{opt} are much larger than when using the Mann-Whitney U statistic. Thus, a smaller true $AUC(\hat{\gamma}_{opt})$ is achieved. $\hat{\alpha}_{opt}$ varies largely under the alternative, however, it also tends to be rather large for $m_e = 10$ unlike applying the Mann-Whitney U statistic where small values of

$\hat{\alpha}_{opt}$ were found. Figure 4.11 shows the distributions of $\hat{\alpha}_{opt}$ for the different scenarios. Under the global null $\hat{\alpha}_{opt}$ is generally larger than 0.9, similar to using the Mann-Whitney U statistic. Figure 4.12 shows scatterplots of $\widehat{MD}(\hat{\gamma}_{opt})$ versus $\hat{\alpha}_{opt}$ for the combination of both criteria when using the mean difference in score values in the cross validation procedure. For comparison to the cross validation procedure using the Mann-Whitney U Statistic see Table 4.1 and Figure 4.3.

According to the given results one may conclude that using the Mann-Whitney U statistic seems to work better than using the mean difference between the score values and we furthermore get an estimate of the true AUC for future prediction (despite positively biased) which seems to be a good criteria for the basic decision whether a prognostic score should be constructed from the data or not. $\widehat{MD}(\hat{\gamma}_{opt})$ has been found to be no good criterion to reflect the performance of the determined prediction scores.

Table 4.4: **Results using the cross validation procedure using the mean difference in score values:** The true best choice of the FDR, α_{opt} and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure (means (standard deviations) and *medians* over 500 simulation runs): the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true future $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{MD}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e and varying per-group sample sizes n, fixing $m = 1000$ and assuming $AUC_* = 0.965$.

m	1000					
m_e	10		60		0	
n	50	100	50	100	50	100
α_{opt}	0.170	0.014	0.824	0.475		
$AUC(\alpha_{opt})$	0.941	0.963	0.813	0.888	0.500	0.500
$\hat{\gamma}_{opt}$	0.044 (0.03)	0.041 (0.03)	0.241 (0.11)	0.130 (0.05)	0.040 (0.04)	0.041 (0.03)
	0.036	*0.032*	*0.241*	*0.131*	*0.030*	*0.033*
$\hat{\alpha}_{opt}$	0.702 (0.21)	0.668 (0.23)	0.811 (0.08)	0.682 (0.10)	0.910 (0.14)	0.925 (0.11)
	0.762	*0.731*	*0.818*	*0.690*	*0.974*	*0.977*
FDR	0.684 (0.25)	0.667 (0.25)	0.794 (0.09)	0.669 (0.01)	0.940 (0.24)	0.978 (0.15)
	0.778	*0.762*	*0.818*	*0.696*	*1.000*	*1.000*
$\hat{\alpha}_{opt,\infty}$	0.699 (0.22)	0.674 (0.24)	0.798 (0.08)	0.675 (0.09)	1.000	1.000
	0.783	*0.762*	*0.820*	*0.697*		
$AUC(\hat{\gamma}_{opt})$	0.892 (0.04)	0.927 (0.02)	0.812 (0.02)	0.881 (0.01)	0.500	0.500
	0.890	*0.929*	*0.814*	*0.880*		
$\widehat{MD}(\hat{\gamma}_{opt})$	7.591 (1.44)	7.323 (0.89)	7.124 (1.60)	6.934 (0.97)	1.169 (0.84)	0.743 (0.47)
	7.529	*7.301*	*7.026*	*6.953*	*1.028*	*0.678*
FWER	0.974	0.980	0.998	0.998	0.940	0.978

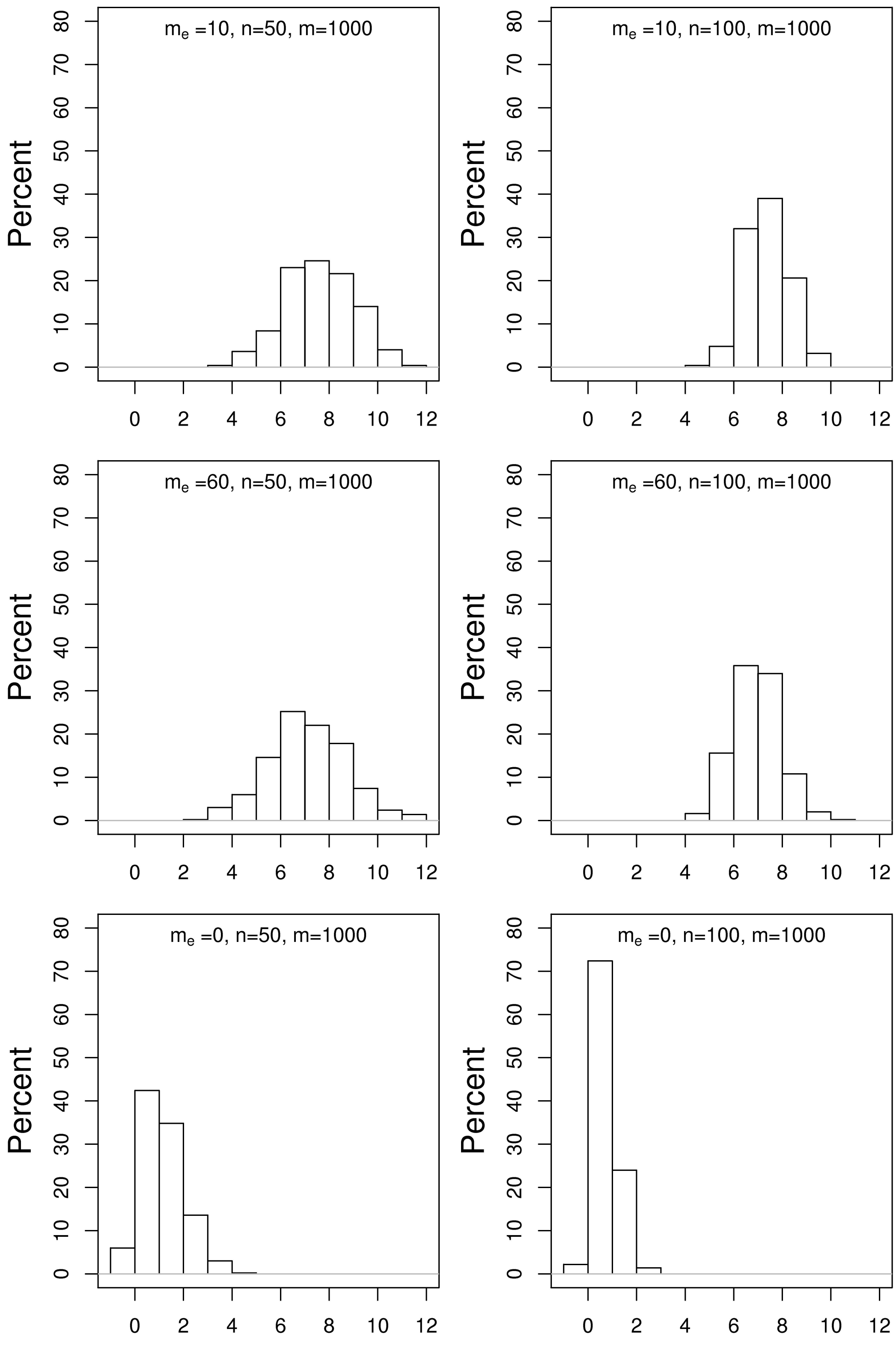

Figure 4.10: **Cross validation using the mean difference in score values:** Distribution of the cross validation based $\widehat{MD}(\hat{\gamma}_{opt})$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column).

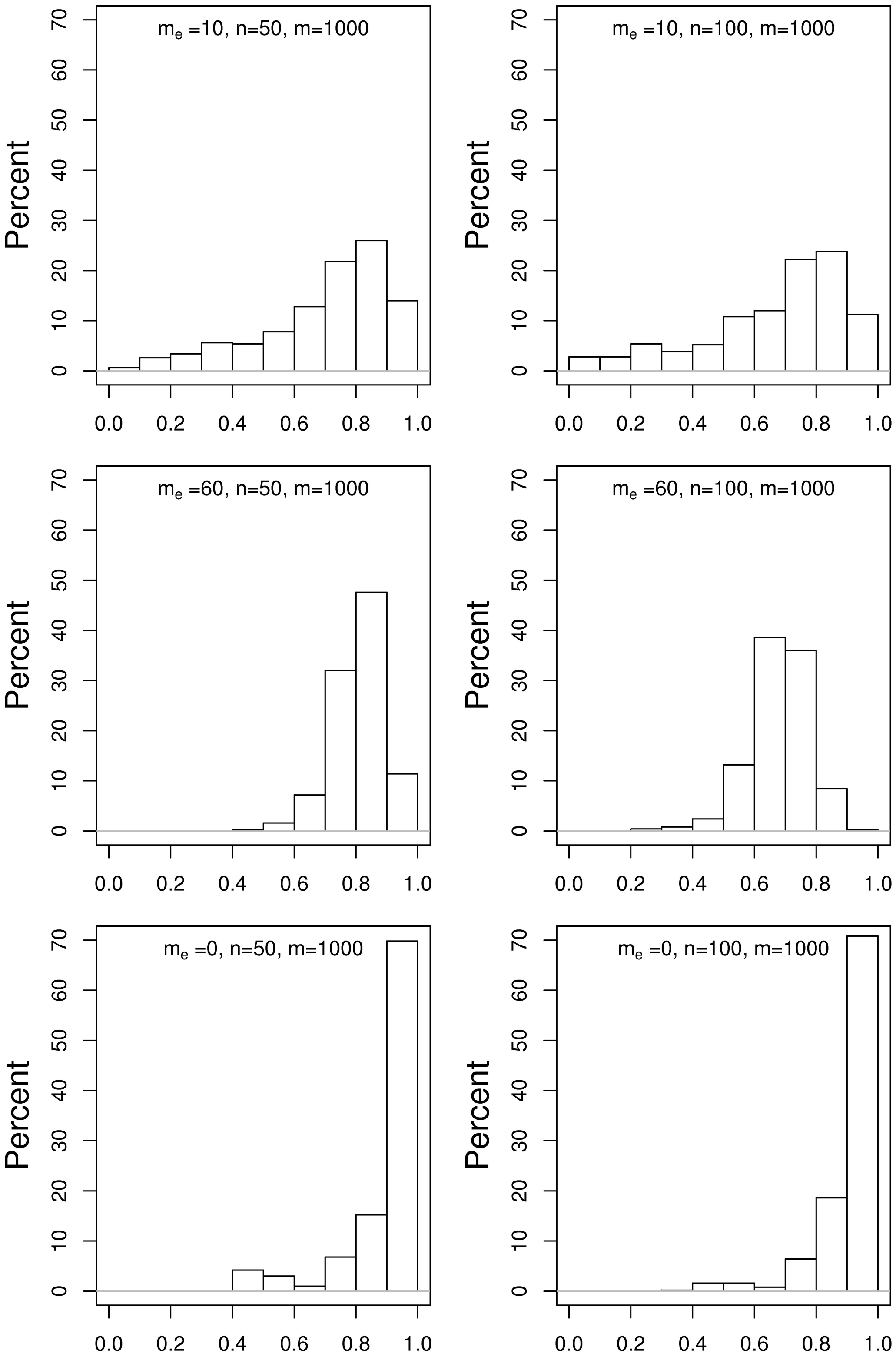

Figure 4.11: **Cross validation using the mean difference in score values:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column).

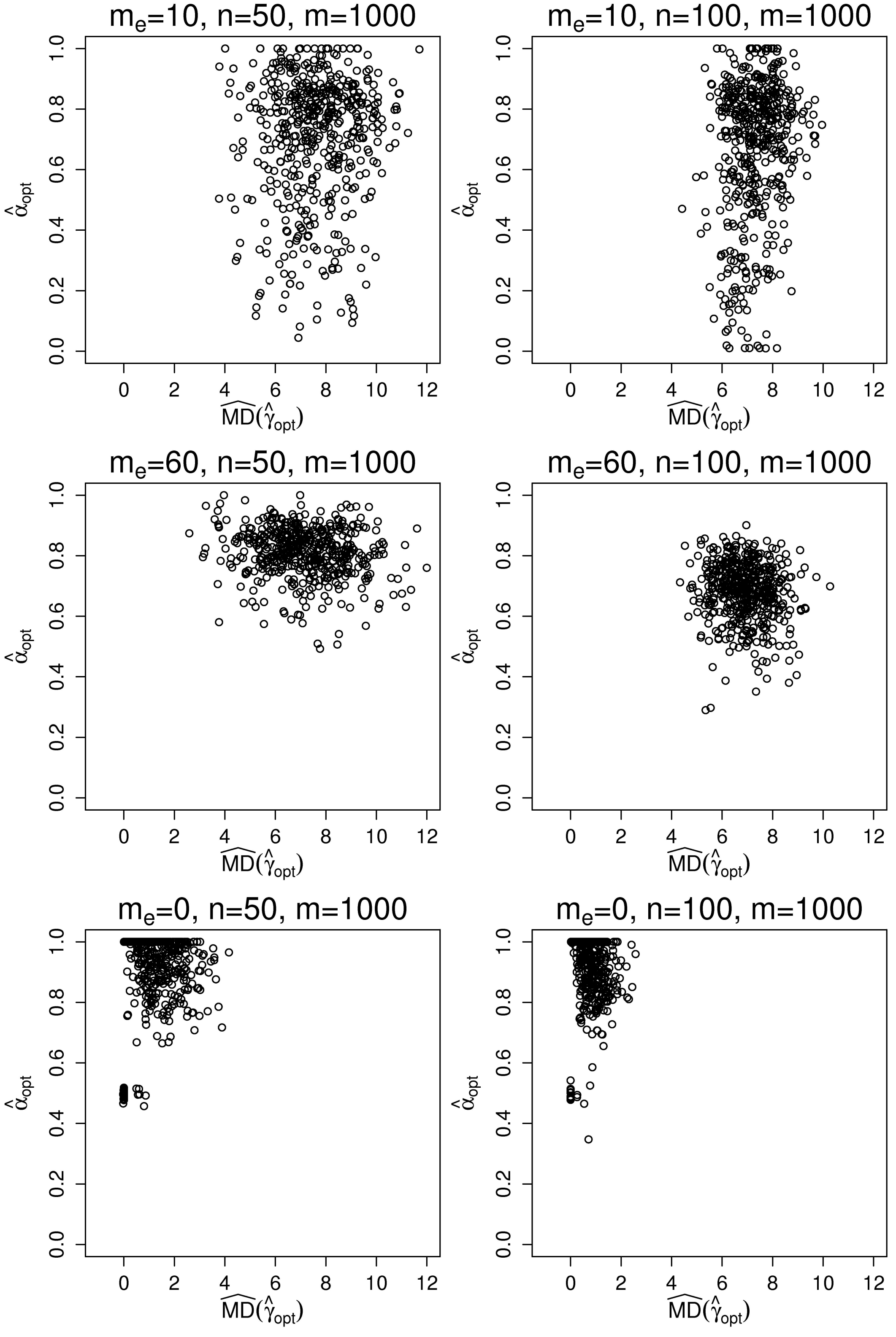

Figure 4.12: **Cross validation using the mean difference in score values:** Scatterplots of $\widehat{MD}(\gamma)$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs): $m_e = 10$ (first row), 60 (second row) or 0 (third row) among $m = 1000$ hypotheses. The sample size was set to $n = 50$ per group (first column) and $n = 100$ (second column).

5 Extensions

5.1 Applying two-sided tests for selection

Up to now we performed one-sided two-sample z-tests for the selection of variables. However, in practice generally two-sided tests are performed. The two-sided p-values $p_i = 2(1 - \Phi(|z_i|))$ are compared to the critical boundary γ which is equivalent to compare the p-values $p_i = 1 - \Phi(|z_i|)$ to the the critical boundary $\gamma/2$.

Theorem 5.1.0.2 *In the two-sided case, for the p-values $p_i = 1 - \Phi(|z_i|)$ the procedure applying the critical boundary $\gamma/2$ leads to the same results as the one-sided test under the global null hypotheses. If under the alternative we assume constant effect size among the prognostic variables the two-sided tests lead to the same results if the effect size (Δ_2) is calculated by*

$$\Delta_2 = \frac{z(1-\frac{\gamma}{2}) - z(1-\gamma)}{\sqrt{\frac{n}{2}}} + \Delta_1.$$

(ignoring directional errors under the alternative). $z(1-\gamma)$ denotes the $(1-\gamma)$-quantile of the standard normal distribution and Δ_1 is the corresponding effect size in the one-sided case.

Proof: Applying the same selection procedure the same power is achieved in the one and two-sided test situation. Ignoring directional errors under the alternative:

$$1 - \beta(\gamma) = 1 - \Phi_{\sqrt{\frac{n}{2}}\Delta_1,1}(z(1-\gamma)) = 1 - \Phi_{\sqrt{\frac{n}{2}}\Delta_2,1}(z(1-\frac{\gamma}{2})) = 1 - \beta(\gamma/2)$$

Thus, the effect size Δ_2 can be easily calculated by solving the above equation.

5.1.1 Simulation studies

Figure 5.1 shows the interpolated functions of the mean values of $AUC(\alpha)$ as a function of α assuming $m_e = 10$ (dashed line) and 60 (dotted line) among $m = 1000$ hypotheses. The sample size per group is set to $n = 50$. Note that $\Delta = 0.811$ for $m_e = 10$ and $\Delta = 0.331$ for $m_e = 60$. The grey curves show the results using one-sided tests and the black curves applying two-sided tests for selection. The best choice expecting $m_e = 10$ among $m = 1000$ tested hypotheses in the two-sided test situation would be a slightly larger $\alpha_{opt} = 0.174$ as compared to 0.170 in the one-sided test situation achieving on average a slightly smaller $AUC(\alpha_{opt})$ of 0.933 as compared to 0.941 in the one-sided situation. If a larger number of prognostic variables with small effects is assumed, the difference in $AUC(\alpha_{opt})$ between the one and two-sided test situation is larger. Assuming 60 alternatives among the $m = 1000$ tested hypotheses the values are $\alpha_{opt} = 0.850$ achieving an average $AUC(\alpha_{opt}) = 0.756$ for applying two-sided tests as compared to $\alpha_{opt} = 0.824$ and $AUC(\alpha_{opt}) = 0.813$ for the one-sided test situation.

The same tendencies can also be seen for the situation where the prognostic variables are searched within $m = 6000$ variables (see Figure 5.2). Trough the whole examples considered, in the two-sided case a slightly larger α_{opt} is determined from the cross validation procedure achieving a slightly smaller $AUC(\alpha_{opt})$. A summary of the results for the two-sided case can be seen in Table 5.1. For comparison to the one-sided case see Table 3.1.

Table 5.1: **Two-sided test situation:** Best choice of the FDR threshold α_{opt}, the corresponding true $AUC(\alpha_{opt})$ as well as the number of non-prognostic variables (m_0^s) and the number of prognostic variables (m_e^s) included in the prediction score.

m	m_e	Δ	n	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
1000	10	0.811	50	0.174	0.933	1.87	8.56
	60	0.331	50	0.850	0.756	242.88	39.39
6000	10	0.811	50	0.252	0.904	2.64	6.89
	60	0.331	50	0.936	0.625	631.37	26.34

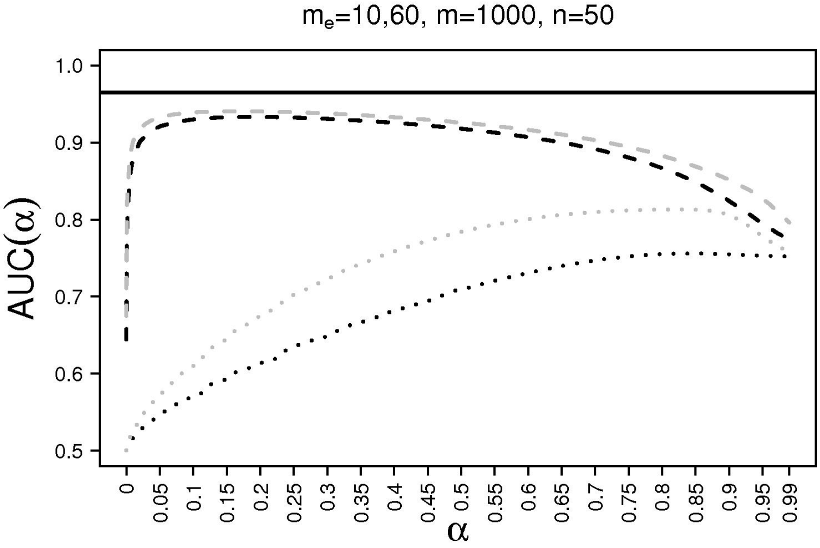

Figure 5.1: **Two-sided test situation:** Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation run) as a function of the FDR threshold α for selection using a two-sided test assuming $m_e = 10$ (black dashed curve) and 60 (black dotted curve) alternatives among $m = 1000$ tested variables. The corresponding one-sided results are given as grey curves. The sample size per group was set to $n = 50$. $AUC_* = 0.965$ is given as solid horizontal line.

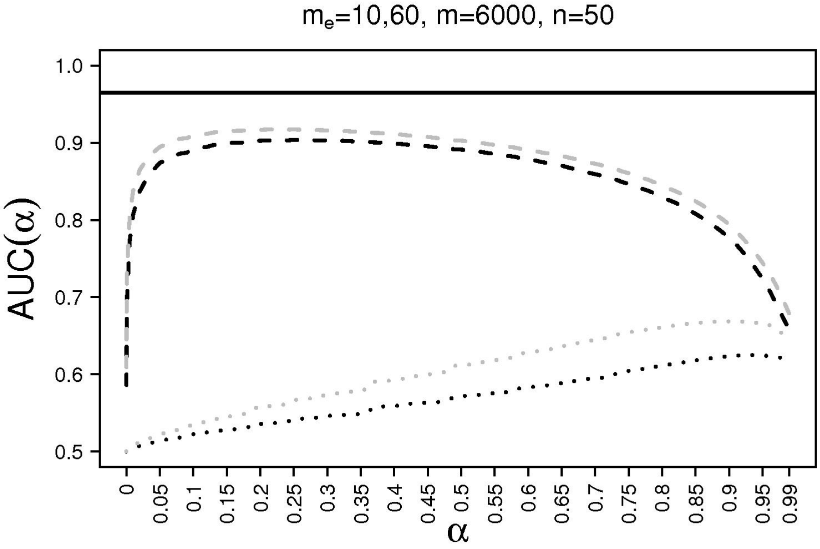

Figure 5.2: **Two-sided test situation:** Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation run) as a function of the FDR threshold α for selection using a two-sided test assuming $m_e = 10$ (black dashed curve) and 60 (black dotted curve) alternatives among $m = 6000$ tested variables. The corresponding one-sided results are given as grey curves. The sample size per group was set to $n = 50$. $AUC_* = 0.965$ is given as solid horizontal line.

5.1.2 Cross validation using two-sided tests

The cross validation procedure applying two-sided tests was investigated by simulation only for the situations testing $m = 1000$ hypotheses. The sample size per group is fixed to $n = 50$. Simulations were performed for $m_e = 10$, 60 and under the global null.

The results of the cross validation procedure are shown in Table 5.2. Similar tendencies for the determined prediction scores as compared to applying one-sided tests are found. The cross validation procedure results in $\hat{\gamma}_{opt}$ and corresponding $\hat{\alpha}_{opt}$ values achieving a future performance with mean $AUC(\hat{\gamma}_{opt})$ values close to the true $AUC(\alpha_{opt})$. However, in the two-sided case the difference between $AUC(\hat{\gamma}_{opt})$ and $AUC(\alpha_{opt})$ is slightly larger as when using one-sided tests. For the difference between $\hat{\alpha}_{opt}$ and α_{opt} no such tendency can be seen. For comparison to the one-sided case see Table 4.1.

Generally, as in the one-sided situation, values of the estimated $\widehat{AUC}(\hat{\gamma}_{opt})$ are large under the alternative and small under the global null. Figure 5.3 shows the histograms of $\widehat{AUC}(\hat{\gamma}_{opt})$ assuming $m_e = 10$ (left plot), $m_e = 60$ (central plot) and under the global null (right plot). As compared to the one-sided situation a only slightly larger variation of $\widehat{AUC}(\hat{\gamma}_{opt})$ can be seen from the histograms (compare Figure 4.1).

Values of the determined $\hat{\alpha}_{opt}$ are as in the one-sided case largely varying under the alternative and are generally larger than 0.9 under the global null (see histograms in Figure 5.4 and Figure 4.2 for the one-sided case). Figures 5.5 show scatterplots for the combination of both arguments for the investigated examples. Thus again we can conclude that both criteria, $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$, should be used for the basic decision for or against building a prediction score from a given sample.

The results can be summarized similar to the one-sided case. The results of the cross validation procedure are reflecting the performance of the determined scores leading to larger $\widehat{AUC}(\hat{\alpha}_{opt})$ and smaller $\hat{\alpha}_{opt}$ values under the alternative and to small $\widehat{AUC}(\hat{\alpha}_{opt})$ and large $\hat{\alpha}_{opt}$ values under the global null. Therefore, applying two-sided tests only slightly reduces the performance of the determined scores.

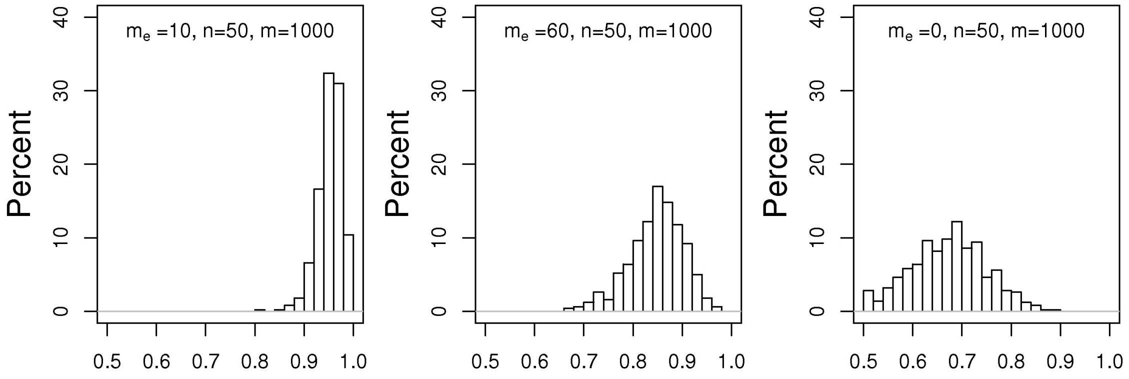

Figure 5.3: **Cross validation using two-sided tests:** Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs) using a two-sided test: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) are assumed among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

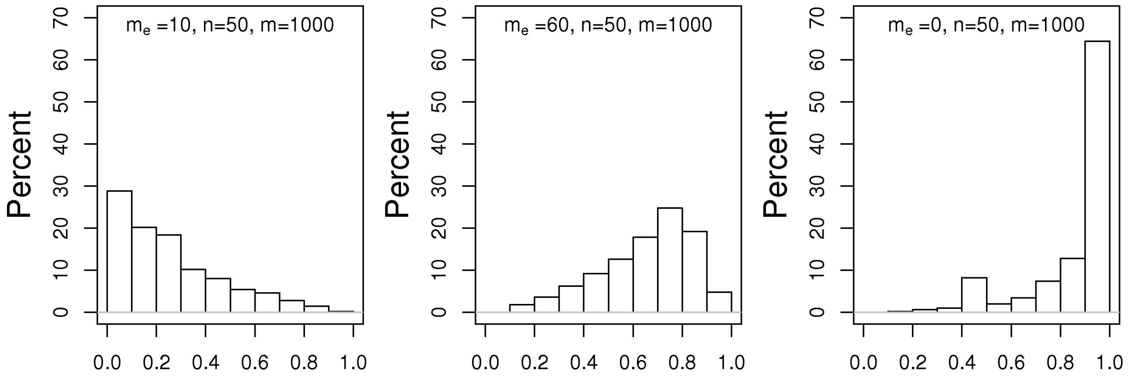

Figure 5.4: **Cross validation using two-sided tests:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs) using a two-sided test: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) are assumed among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

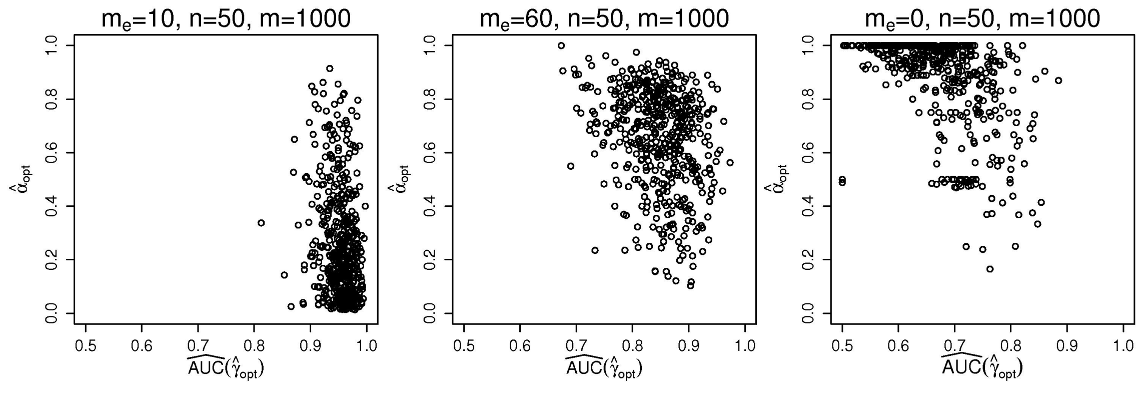

Figure 5.5: **Cross validation using two-sided tests:** Scatterplot of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs) using a two-sided test: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) are assumed among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

Table 5.2: **Cross validation using two-sided tests:** The true best choice of the FDR, α_{opt} and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure: the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e. The per-group sample sizes is fixed to $n = 50$ and $m = 1000$. Under the alternative $AUC_* = 0.965$.

m_e	10	60	0
α_{opt}	0.174	0.850	
$AUC(\alpha_{opt})$	0.933	0.756	0.500
$\hat{\gamma}_{opt}$	0.007 (0.01)	0.103 (0.11)	0.071 (0.11)
	0.002	*0.050*	*0.018*
$\hat{\alpha}_{opt}$	0.260 (0.21)	0.656 (0.19)	0.878 (0.17)
	0.208	*0.696*	*0.958*
FDR	0.266 (0.25)	0.644 (0.21)	0.984 (0.13)
	0.200	*0.682*	*1.000*
$\hat{\alpha}_{opt,\infty}$	0.266 (0.22)	0.600 (0.19)	1.000
	0.201	*0.609*	
$AUC(\hat{\gamma}_{opt})$	0.924 (0.03)	0.743 (0.04)	0.500
	0.931	*0.747*	
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.953 (0.02)	0.848 (0.05)	0.673 (0.08)
	0.957	*0.853*	*0.677*
FWER	0.730	0.980	0.984

5.2 Unequal effect sizes

Up to this point we assumed that all prognostic variables have the same effect size. To investigate situation under unequal effect sizes, we now assume that, if the alternatives holds, the effects follow a uniform distribution ($\Delta \in (0, \tilde{\Delta}]$). The parameter $\tilde{\Delta}$ for the uniform distribution is searched such that the optimal linear prediction score, if known, would lead to the benchmark AUC_* (0.965). Thus, $\tilde{\Delta} = 1.306$ for $m_e = 10$ and $\tilde{\Delta} = 0.566$ for $m_e = 60$.

Secondly we assume that the effects follow an exponential distribution. Again the parameter of the exponential distribution (λ_e) is searched such that the optimal linear prediction score would lead to the benchmark AUC_*. Thus, $\lambda_e = 1.374$ for $m_e = 10$ and $\lambda_e = 3.468$ for $m_e = 60$. In the situation of exponential distributed effect sizes and uniform distributed effect sizes, similar performances in terms of the true $AUC(\alpha)$ can be seen (see Figure 5.6 and Table 5.3), since the effect sizes to achieve AUC_* do not differ largely between the two distributions.

Detecting the rather large effects among the distributed variables results in a slightly better performance of the determined scores as compared to the situation of equal effect sizes. Thus, α_{opt} values are smaller and $AUC(\alpha_{opt})$ values are slightly larger as compared to the case where equal effect sizes are assumed.

However, because of the similarities to the situation of equal effect sizes, we expect similar tendencies for the cross validation procedure to determine $\hat{\gamma}_{opt}$ values leading on average to $AUC(\hat{\gamma}_{opt})$ values close to $AUC(\alpha_{opt})$ as well as giving a good reflection of the performance of the determined scores by the estimates $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$.

Table 5.3: **Unequal effect sizes:** Best choice of the FDR threshold α_{opt}, the corresponding true $AUC(\alpha_{opt})$ as well as the number of non-prognostic variables (m_0^s) and the number of prognostic variables (m_e^s) included in the prediction score assuming unequal effect sizes among the prognostic variables. The parameters of the distributions to achieve AUC_* are given. $m = 1000$, $n = 50$, $AUC_* = 0.965$.

Distribution	m	m_e	n	$\tilde{\Delta}$ or λ_e	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
uniform	1000	10	50	1.306	0.117	0.948	0.83	5.67
	1000	60	50	0.566	0.588	0.834	34.42	22.98
exponential	1000	10	50	1.374	0.117	0.946	0.81	5.62
	1000	60	50	3.468	0.575	0.834	32.04	22.48

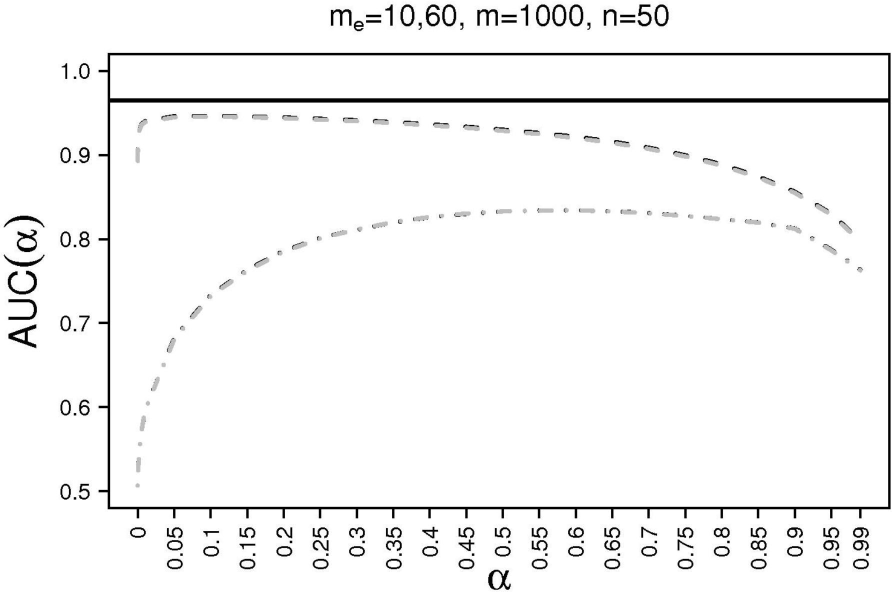

Figure 5.6: **Unequal effect sizes:** Interpolated functions of mean values of $AUC(\alpha)$ over the simulated samples (10000 simulation run) as a function of the FDR threshold α for $m_e = 10$ (dashed curves) and 60 (dotted curves) alternatives among $m = 1000$ tested variables. The effect sizes are assumed to follow uniform distributions (grey curves) or exponential distributions (black curves). The sample size per group is set to $n = 50$. $AUC_* = 0.965$ is given as solid horizontal line.

6 Situation of unknown variances

6.1 Selection and prediction

Up to now we assumed that the variance ($\sigma^2 = 1$) is known. However, in practice, the variance is unknown and has to be estimated from the given data set. We now will investigate the impact of estimating variances on the performance of the resulting prediction scores by simulation. The selection method applying a multiple test with threshold α for the FDR is now based on a one-sided two-sample t-test. Thus, the test statistics, assuming that the unknown within-group variances are equal ($\sigma^2_{r,i} = \sigma^2_{nr,i} = \sigma^2_i = 1$ for $i = 1, ..., m$) is:

$$t_i = (\bar{x}_{r,i} - \bar{x}_{nr,i})/(\sqrt{(s^2_{r,i} + s^2_{nr,i})/n}), \qquad i = 1, \ldots, m \tag{6.1}$$

where we again assume equal sample sizes per variable and group. $s^2_{r,i}$ and $s^2_{nr,i}$ are the estimated variances from the samples of responders and non-responders respectively. The decision is then based on the one-sided p-values

$$p_i = 1 - F_{2n-2}(t_i)$$

where F_{2n-2} is the central t-distribution with $2n - 2$ degrees of freedom.

To calculate the score to predict a clinical outcome we now have to consider the estimated variances in the prediction score (as in the classical discriminant function). The weights of the selected variables in the prediction score have to be divided by the common within groups variance estimate applied in the t-test:

$$s^2_i = \frac{(n_{r,i} - 1)s^2_{r,i} + (n_{nr,i} - 1)s^2_{nr,i}}{n_{r,i} + n_{nr,i} - 2}.$$

Thus in our case of $n_{r,i} = n_{nr,i} = n$ for $i = 1, ..., m$, s_i^2 can be calculated as the simple mean value of the within group variances of the group of responder and non-responder.

Definition 6.1.0.1 *Assume that $k \leq m$ variables are selected to build a prediction score ($p_j \leq \gamma$ for $j = 1, .., k$). Let $\bar{x}_{r,j}$ and $\bar{x}_{nr,j}$ denote the sample means of the jth selected variable of patients responding and not responding to therapy, respectively, and $\mathbf{x} = (x_1, ..., x_k)$ the corresponding values of the selected variables in a future patient. The prediction score is calculated as follows:*

$$\hat{f}(\mathbf{x}; \gamma) = \hat{\mathbf{c}}^T \mathbf{x} = \sum_{j=1}^{k} \hat{c}_j x_j. \tag{6.2}$$

where

$$\hat{c}_j = \frac{\bar{x}_{r,j} - \bar{x}_{nr,j}}{s_j^2} \tag{6.3}$$

for the k selected variables with $p_j \leq \gamma$. All other variables are not included in the prediction score (the weights in the score are set to 0).

If $\hat{f}(\mathbf{x}; \gamma) > b$ we predict a response, otherwise a non-response. Let the diagonal matrix of the estimated variances of the $k \leq m$ selected variables be denoted by

$$\hat{\Sigma}_k = \begin{pmatrix} s_1^2 & 0 & \dots & 0 \\ 0 & s_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & s_k^2 \end{pmatrix} \tag{6.4}$$

Note that we do not estimate the covariances (they are set to 0). Note also that the true covariance matrix Σ_k of the k selected variables under our assumptions of independence and (unknown) variance $\sigma^2 = 1$ is equal to I.

Theorem 6.1.0.1 *Let $\boldsymbol{\mu}_a$, $a = r$ or nr denote the true mean vector of the k selected variables in a future responder or non-responder, respectively. Given the selection threshold α and the estimated weights from the samples, the prognostic score follows two normal distributions:*

$$\begin{aligned} \hat{f}(\mathbf{x}; \gamma) &\sim N[\mu_a, \sigma_a^2] = N[\hat{\mathbf{c}}^T \boldsymbol{\mu}_a, \hat{\mathbf{c}}^T \hat{\mathbf{c}}] \\ &= N[(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \boldsymbol{\mu}_a, (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \Sigma_k (\hat{\Sigma}_k^{-1})^T (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})] \\ &= N[(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \boldsymbol{\mu}_a, (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \hat{\Sigma}_k^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})] \end{aligned}$$

Proof: The results can be derived by using theorem 2.1.0.7.

Theorem 6.1.0.2 *Fixing the appropriate $\boldsymbol{\mu}_a$, $a = r$ or nr, for the populations of responders and non-responders, respectively, the AUC for future independent populations is calculated by:*

$$AUC(\alpha) = \int_0^1 \left\{ 1 - \Phi \left[z(1-w) - \frac{\hat{\mathbf{c}}^T(\boldsymbol{\mu}_r - \boldsymbol{\mu}_{nr})}{\sqrt{\hat{\mathbf{c}}^T\hat{\mathbf{c}}}} \right] \right\} dw$$

$$= \int_0^1 \left\{ 1 - \Phi \left[z(1-w) - \frac{(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_{nr})}{\sqrt{(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \hat{\Sigma}_k^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})}} \right] \right\} dw.$$

Proof: The results can be derived as in theorem 3.3.0.2.

6.2 Simulation studies

The situation of unknown variance is investigated by simulations for the scenarios searching for $m_e = 10$ and 60 alternatives among $m = 1000$ and 6000 hypotheses. The per-group sample size in all investigated scenarios is set to $n = 50$. The effect size is again triggered by forcing the optimal ROC-curve through the benchmark point $v = 1 - w = 0.9$. Thus Δ remains the same as in the known variance case, for $m_e = 10$, $\Delta = 0.811$ and for $m_e = 60$, $\Delta = 0.331$ to achieve $AUC_* = 0.965$ of the ROC-curve crossing through the benchmark point $(v, 1 - w) = (0.9, 0.9)$. Again simulated mean values of $AUC(\alpha)$ for a grid of α values with interval 0.01 are interpolated using splines. α_{opt} is again determined by optimizing the interpolated function.

Figure 6.1 shows the interpolated functions of the mean values of $AUC(\alpha)$ (10000 simulation runs) for the unknown variance case expecting $m_e = 10$ (black dotted line) and 60 (black dashed line) among $m = 1000$ tested hypotheses. The corresponding results of the known variance case are shown as grey lines. $AUC_* = 0.965$ is shown as solid horizontal line. The figure shows that the score using the estimated variances achieves only slightly smaller performances in terms of $AUC(\alpha)$ as compared to the known variance case. For $m_e = 10$, $\alpha_{opt} = 0.174$ on average achieves a performance of $AUC(\alpha_{opt}) = 0.936$. For $m_e = 60$ the optimal threshold $\alpha_{opt} = 0.832$ achieves on average $AUC(\alpha_{opt}) = 0.810$. Note that in the known variance case the values were $AUC(\alpha_{opt}) = 0.941$ and $\alpha_{opt} = 0.17$

for $m_e = 10$ and $AUC(\alpha_{opt}) = 0.813$ and $\alpha_{opt} = 0.824$ for $m_e = 60$ (see results in Table 6.1 for the unknown variance case and Table 3.1 for comparison to the known variance case). Thus, estimating the variances in the unknown case does only slightly decrease the performance of the resulting scores. The selection threshold α_{opt} slightly increases when moving from the known to the unknown variance case.

Figure 6.2 shows the results for $m = 6000$ tested hypotheses. Again, estimating the variances only slightly decreases the performance of the resulting score as compared to the known variance case. Testing $m = 6000$ hypotheses an optimal threshold of $\alpha_{opt} = 0.226$ leads on average to prediction scores with a mean performance of $AUC(\alpha_{opt}) = 0.906$ if $m_e = 10$ prognostic variables are assumed. For $m_e = 60$ the values are $\alpha_{opt} = 0.918$ and $AUC(\alpha_{opt}) = 0.664$. For the known variance case the values were $AUC(\alpha_{opt}) = 0.917$ and $\alpha_{opt} = 0.25$ for $m_e = 10$ and $AUC(\alpha_{opt}) = 0.669$ and $\alpha_{opt} = 0.895$ for $m_e = 60$ (see also Table 6.1 and 3.1 for more details).

Table 6.1: **Unknown variance case:** The best choice of the FDR threshold α_{opt}, the corresponding true $AUC(\alpha_{opt})$ as well as the number of non-prognostic variables (m_0^s) and the number of prognostic variables (m_e^s) included in the prediction score for a varying number of and tested hypotheses m and prognostic variables m_e. $n = 50$ and $AUC_* = 0.965$.

m	m_e	Δ	n	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
1000	10	0.811	50	0.174	0.936	1.93	8.49
	60	0.331	50	0.832	0.810	274.60	49.71
6000	10	0.811	50	0.253	0.912	2.86	7.60
	60	0.331	50	0.918	0.664	516.28	32.30

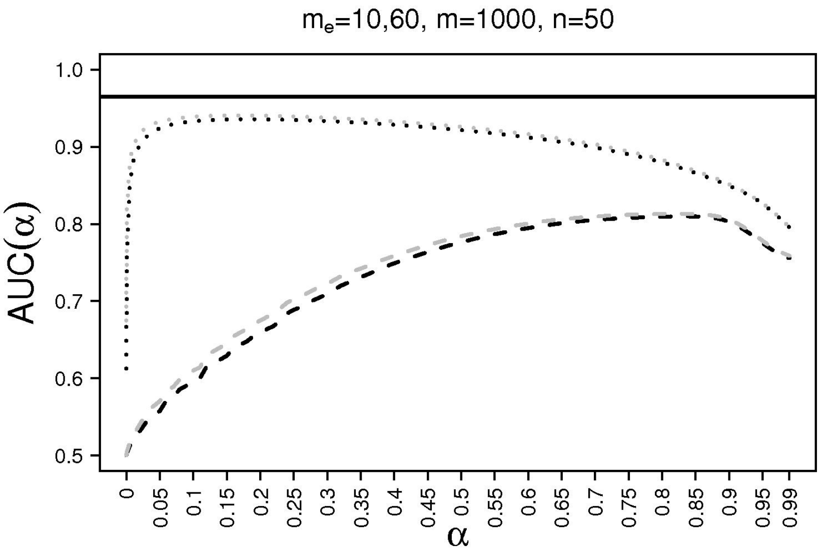

Figure 6.1: **Unknown variance case:** Interpolated functions of mean values of $AUC(\alpha)$ (over 10000 simulation runs) for a varying FDR selection threshold α assuming $m_e = 10$ (black dashed line) alternatives or 60 (black dotted line) among $m = 1000$ tested variables. The sample size per group is set to $n = 50$. The corresponding known case is shown in grey lines. $AUC_* = 0.965$ is given as solid horizontal line.

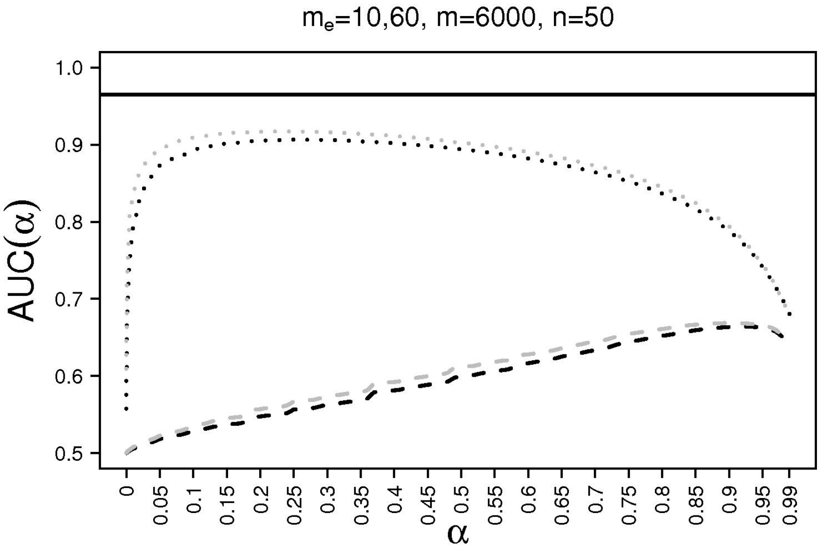

Figure 6.2: **Unknown variance case:** Interpolated functions of mean values of $AUC(\alpha)$ (over 10000 simulation runs) for a varying FDR selection threshold α assuming $m_e = 10$ (black dashed line) alternatives or 60 (black dotted line) among $m = 6000$ tested variables. The sample size per group is set to $n = 50$. The corresponding known case is shown in grey lines. $AUC_* = 0.965$ is given as solid horizontal line.

6.3 Cross validation

We investigated the cross validation procedure applying t-tests for the examples assuming $m_e = 10$ and 60 among $m = 1000$ and 6000 candidate variables. The sample size per group in the four examples is set to $n = 50$.

The general tendencies of the prognostic scores determined from the cross validation procedure in the unknown variance case (results see Table 6.2) are similar as compared to applying z-tests in the known variance case. The cross validation procedure determines $\hat{\gamma}_{opt}$ and corresponding $\hat{\alpha}_{opt}$ values leading to scores with an average performance of $AUC(\hat{\gamma}_{opt})$ values close to $AUC(\alpha_{opt})$.

Figure 6.3 shows the histograms of the estimated $\widehat{AUC}(\hat{\gamma}_{opt})$ for the simulated samples assuming $m_e = 10$ (left plot), 60 (central plot) and under the global null hypothesis ($m_e = 0$: right plot) for $m = 1000$. As in the situation of known variances the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ is small under the global null and large under the alternative indicating that good scores can be constructed from the data if we are only searching within $m = 1000$ hypotheses. Again because of the skew distribution of $\hat{\alpha}_{opt}$ (see histograms in Figure 6.4) and the generally flat optimum of the interpolated functions differences between $\hat{\alpha}_{opt}$ and α_{opt} can be seen. The medians are again closer to the true optimum. However, the differences between $\hat{\alpha}_{opt}$ and α_{opt} are slightly larger then in the known variance case. No such tendency can be seen for the differences between $\widehat{AUC}(\hat{\gamma}_{opt})$ and $AUC(\alpha_{opt})$. As in the known variance case, $\hat{\alpha}_{opt}$ is varying largely under the alternative and under the global null it is generally larger than 0.9 (see Figure 6.4 for histograms of $\hat{\alpha}_{opt}$). However, estimated $\hat{\alpha}_{opt}$ values are again close to the true FDR (refer Table 6.2).

Summing up the results under the unknown variance case it may again be useful to look at the both criteria, the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$, to decide, whether a score should be constructed from a given sample sample or not. Figures 6.5 show scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ versus $\hat{\alpha}_{opt}$ for $m = 1000$. Thus, if a selected score has a small $\widehat{AUC}(\hat{\gamma}_{opt})$ and a large $\hat{\alpha}_{opt}$ one should decide against the determined score. If a selected score has a small $\widehat{AUC}(\hat{\gamma}_{opt})$ and a small $\hat{\alpha}_{opt}$ one may conclude that the effect sizes of the selected prognostic variables are too small for a good prediction of response of a patient to a particular therapy. If $\widehat{AUC}(\hat{\gamma}_{opt})$ is large and $\hat{\alpha}_{opt}$ is small the determined score may have good prognostic abilities to predict the clinical outcome of a future patient. However, if $\widehat{AUC}(\hat{\gamma}_{opt})$ is large and $\hat{\alpha}_{opt}$ is large too one may be careful because despite the good prognostic ability an unrealistic large number of non-prognostic variables may be included in the score. Thus, one may conclude that the sample size is to small to detect the prognostic variables.

Figure 6.6 shows the histograms of $\widehat{AUC}(\hat{\gamma}_{opt})$, Figure 6.7 the histograms of $\hat{\alpha}_{opt}$ and Figure 6.8 scatterplots for $\widehat{AUC}(\hat{\gamma}_{opt})$ versus $\hat{\alpha}_{opt}$ for the investigated scenarios searching the prognostic variables among $m = 6000$ tested genes. The scores that can be determined from given samples by using a FDR-based selection procedures are performing good if $m_e = 10$ prognostic variables are assumed and worse when $m_e = 60$ prognostic variables are searched within the 6000 tested variables (see previous Section 6.2). The cross validation again mirrors this performances by leading to small $\widehat{AUC}(\hat{\gamma}_{opt})$ values and to large $\hat{\alpha}_{opt}$ values under the global null hypotheses and under the alternative of $m_e = 60$. In the situation of $m_e = 10$, where rather good scores can be determined from the underlying samples, the cross validation procedure ends in large $\widehat{AUC}(\hat{\gamma}_{opt})$ values and in small $\hat{\alpha}_{opt}$ values. For example, an average $\widehat{AUC}(\hat{\gamma}_{opt}) = 0.793$ and $\hat{\alpha}_{opt} = 0.803$ for $m_e = 60$ is indicating a poor performance of the evaluated scores. An average $\widehat{AUC}(\hat{\gamma}_{opt}) = 0.993$ and $\hat{\alpha}_{opt} = 0.375$ for $m_e = 10$ is indicating a rather good performance of the evaluated scores if we tolerate the fact that approximately 38% true null hypotheses are included in the prediction score.

A summary of the results of the cross validation procedure under the alternative can be seen in Table 6.2. A summary of the results of the cross validation procedure under the global null hypothesis can be seen in Table 6.3. Note again that under the global null $AUC(\hat{\gamma}_{opt})$ is always equal to 0.5 and $\hat{\alpha}_{opt,\infty}$ is always equal to 1. Note also that again under the global null larger differences between $\hat{\alpha}_{opt}$ and the true FDR can be seen as under the alternative. Looking at the results from the simulations one may conclude that estimating the variances only slightly decreases the quality of the determined prediction scores as compared to the known variance case.

Table 6.2: **Cross Validation in the unknown variance case:** The true best choice of the FDR, α_{opt} and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure (means (standard deviations) and ***medians*** over 500 simulation runs): the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e and tested variables m. The per-group sample sizes is fixed to $n = 50$. $AUC_* = 0.965$.

m	1000		6000	
m_e	10	60	10	60
n	50	50	50	50
α_{opt}	0.174	0.832	0.253	0.918
$AUC(\alpha_{opt})$	0.936	0.810	0.912	0.664
$\hat{\gamma}_{opt}$	0.005 (0.01)	0.113 (0.12)	0.002 (0.004)	0.055 (0.07)
	0.003	*0.065*	*0.0007*	*0.025*
$\hat{\alpha}_{opt}$	0.258 (0.18)	0.632 (0.18)	0.375 (0.25)	0.803 (0.15)
	0.213	*0.663*	*0.313*	*0.858*
FDR	0.249 (0.23)	0.615 (0.20)	0.383 (0.29)	0.811 (0.18)
	0.182	*0.649*	*0.348*	*0.868*
$\hat{\alpha}_{opt,\infty}$	0.264 (0.19)	0.626 (0.18)	0.388 (0.26)	0.820 (0.14)
	0.220	*0.648*	*0.344*	*0.869*
$AUC(\hat{\gamma}_{opt})$	0.931 (0.02)	0.792 (0.04)	0.883 (0.05)	0.658 (0.03)
	0.936	*0.798*	*0.894*	*0.661*
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.952 (0.02)	0.864 (0.05)	0.933 (0.04)	0.793 (0.07)
	0.955	*0.867*	*0.938*	*0.797*
FWER	0.744	0.988	0.820	0.980

Table 6.3: **Cross Validation in the unknown variance case under the global null:** Results determined from the cross validation procedure (means (standard deviations) and *medians* over 500 simulation runs): the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$ (always 1) using the determined selection threshold, the true $AUC(\hat{\gamma}_{opt})$ (always 0.5), the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e and tested variables m. The per-group sample sizes was fixed to $n = 50$. $AUC_* = 0.965$.

m	1000	6000
$\hat{\gamma}_{opt}$	0.046 (0.08)	0.049 (0.07)
	0.012	*0.014*
$\hat{\alpha}_{opt}$	0.909 (0.14)	0.961 (0.07)
	0.978	*0.994*
FDR	0.980 (0.14)	1.000 (0.00)
	1.000	*1.000*
$\hat{\alpha}_{opt,\infty}$	1.000	1.000
$AUC(\hat{\gamma}_{opt})$	0.500	0.500
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.659 (0.07)	0.667 (0.08)
	0.663	*0.675*
FWER	0.980	0.998

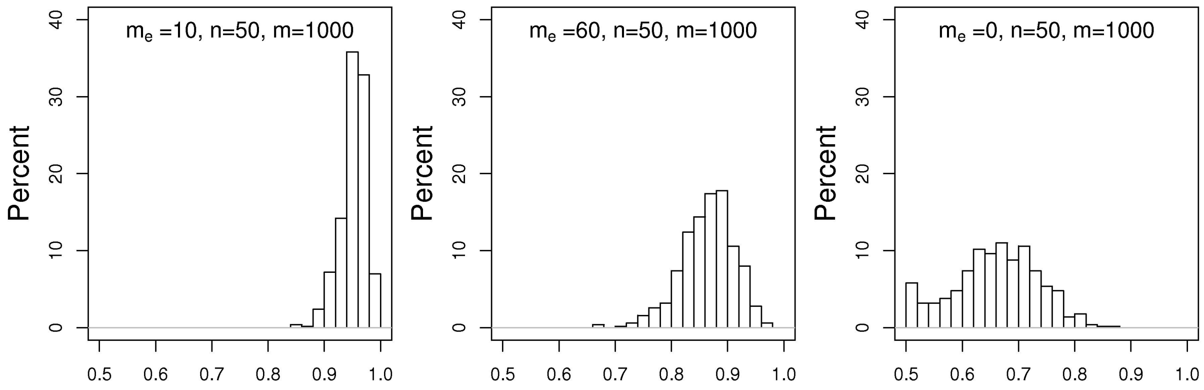

Figure 6.3: **Unknown variance case:** Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

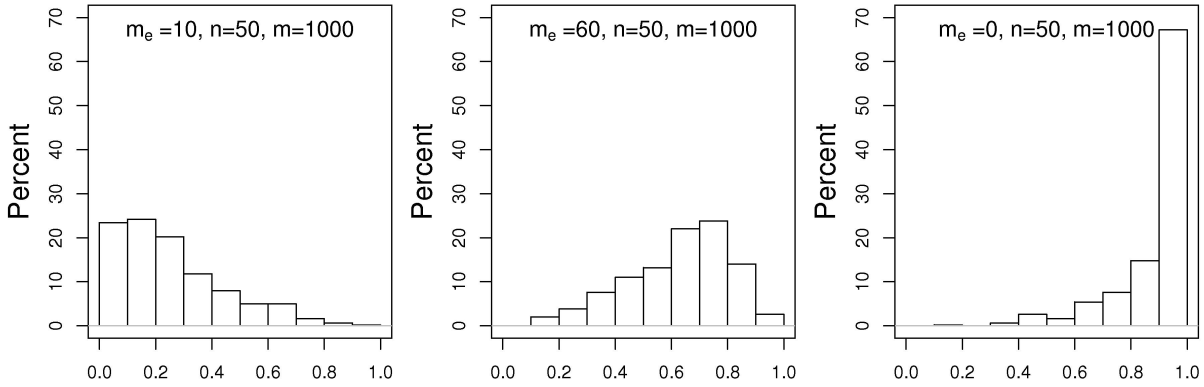

Figure 6.4: **Unknown variance case:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

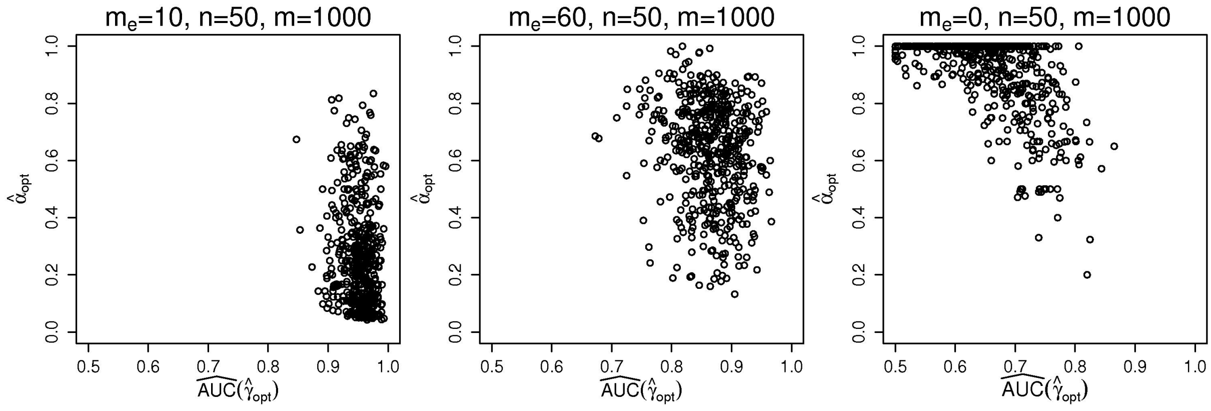

Figure 6.5: **Unknown variance case:** Scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 1000$ hypotheses. The sample size is set to $n = 50$.

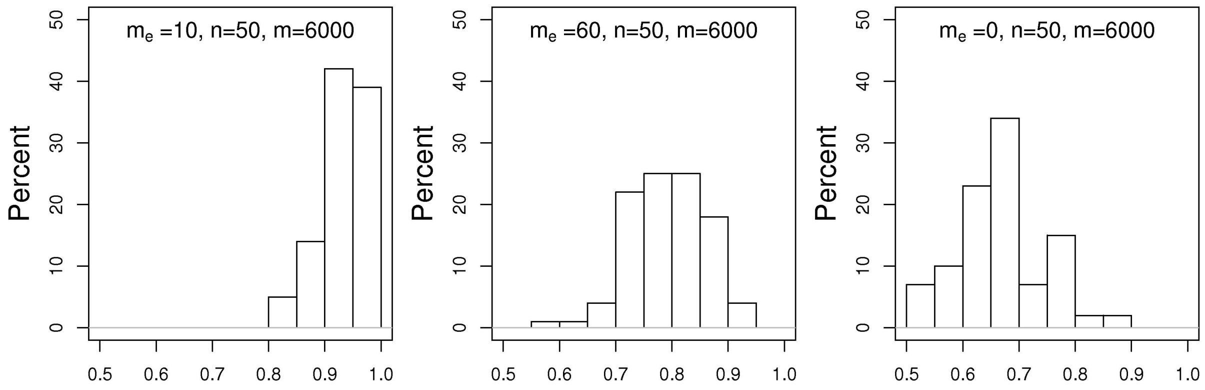

Figure 6.6: **Unknown variance case:** Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 6000$ hypotheses. The sample size is set to $n = 50$.

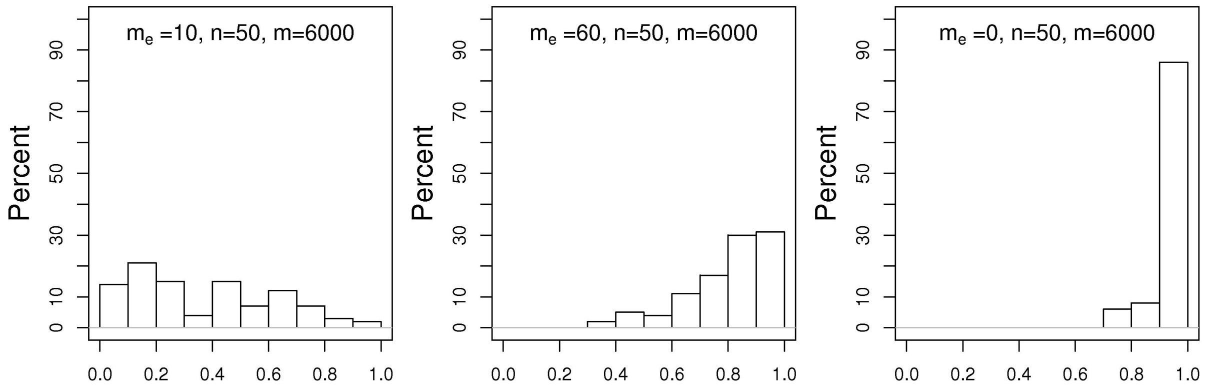

Figure 6.7: **Unknown variance case:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 6000$ hypotheses. The sample size is set to $n = 50$.

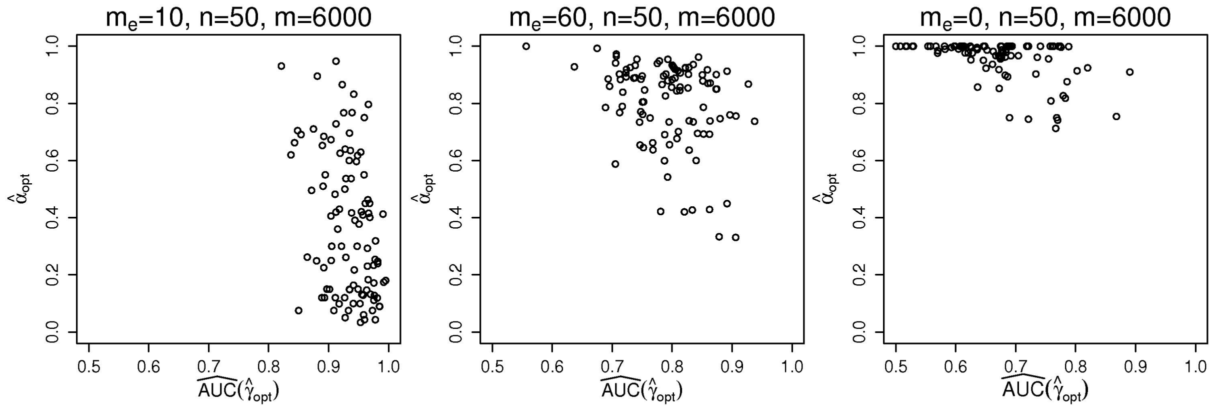

Figure 6.8: **Unknown variance case:** Scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs) assuming unknown variance: $m_e = 10$ (left plot), 60 (central plot) or 0 (right plot) among $m = 6000$ hypotheses. The sample size is set to $n = 50$.

7 Situation of correlation between variables

7.1 Selection and prediction

Up to now we assumed independence across the candidate variables. In practice, one may not always be confident that the tested variables are independent. However, assuming correlation between variables (hypotheses) we use the same test statistics (6.1) for the individual variables as in the unknown variance case for selection using the FDR approach. Thus, the decision is again based on the one-sided p-values $p_i = 1 - F_{2n-2}(t_i)$ (one-sided two-sample t-test) where t_i is calculated using formula (6.1).

The effect size Δ of the prognostic variables is now depending on the correlation structure in order to achieve the benchmark $AUC_* = 0.965$.

Theorem 7.1.0.3 *Let the prognostic variables be distributed according to* $N[\mathbf{\Delta}, \Sigma_{m_e}]$*. We assume equal effect sizes* Δ *for the* m_e *prognostic variables. The required* Δ *to achieve a ROC-curve crossing through a benchmark point with fixed values* v *for sensitivity and* $1 - w$ *for specificity can be calculated by:*

$$\Delta = \frac{z(w) - z(1-v)}{\sqrt{\mathbf{1}^T \Sigma_{m_e}^{-1} \mathbf{1}}} \tag{7.1}$$

Proof: Under the assumption of equal effect sizes among the prognostic variables, the optimal score in this situation is again a linear score of all prognostic variables considering the covariance matrix (as in the discriminant analysis):

$$\hat{f}(\mathbf{x};\gamma) = \mathbf{1}^T\Sigma_{m_e}^{-1}\mathbf{x}.$$

Thus, the sensitivity can be calculated by

$$v = 1 - \Phi\left(z(1-w) - \frac{\mathbf{1}^T\Sigma_{m_e}^{-1}\boldsymbol{\Delta}}{\sqrt{\mathbf{1}^T\Sigma_{m_e}^{-1}\mathbf{1}}}\right) = 1 - \Phi(z(1-w) - \Delta\sqrt{\mathbf{1}^T\Sigma_{m_e}^{-1}\mathbf{1}}) \tag{7.2}$$

Solving equation (7.2) results in (7.1). Note again that Σ_{m_e} here denotes the true covariance matrix of the m_e prognostic variables.

To investigate the impact of correlation between variables (hypotheses) on the performance of the resulting linear prognostic scores we assume an autoregressive correlation structure to exist among the variables, i.e. the correlation between hypothesis i and j is given by

$$\rho^{|i-j|} \text{ for some } \rho \in (0,1).$$

Thus, the m tested variables are distributed according to a m-dimensional normal distribution where the covariance matrix Σ has an autoregressive correlation structure:

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \dots & \rho^{m-1} \\ \rho & 1 & \rho & \rho^2 & \dots & \rho^{m-2} \\ \rho^2 & \rho & 1 & \rho & \dots & \rho^{m-3} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \rho^{m-1} & \dots & \dots & \dots & \dots & 1 \end{pmatrix}. \tag{7.3}$$

We furthermore assume that the first m_e variables are alternatives, i.e. the prognostic variables are lying close to each other and thus are high correlated. The correlations between the alternatives and the true null hypotheses, depending on the distance, may be rather small. However, there is a large correlation between alternatives and true null hypothesis lying close to the alternatives.

Note that we also assumed random distributed alternatives among the m tested variables. In this scenario the selected variables are nearly independent and thus results are close to them determined for the independent case (data not shown).

As mentioned before the effect size is depending on the correlation structure. Figure

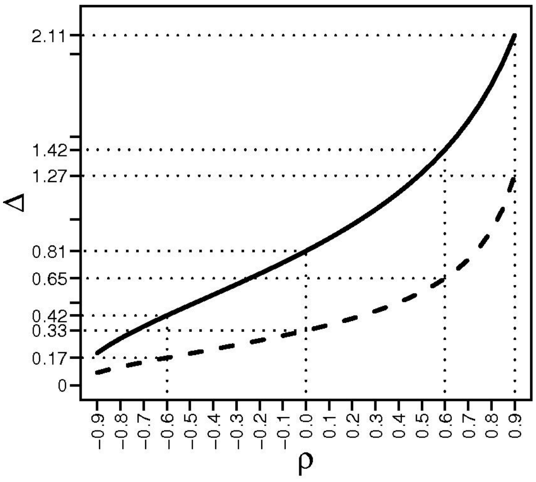

Figure 7.1: **Correlated hypotheses:** Minimal effect size Δ required to achieve a ROC-curve crossing through the point where sensitivity and specificity are equal to 0.9 as a function of the number of prognostic variables m_e.

7.1 shows the required Δ to achieve the benchmark $AUC_* = 0.965$ as a function of the parameter ρ for the situation of an autoregressive correlation structure. For a negative parameter ρ we get positive and negative correlations between the variables and thus also between the alternatives. Therefore, only small effect sizes are required to achieve AUC_*. For a large positive ρ large effect sizes are needed to achieve the benchmark AUC_*. In the following simulation studies the parameter ρ is set to 0.6, −0.6 and 0.9. To achieve a ROC-curve for future prediction that crosses the benchmark point $(0.9, 0.9)$, a minimal Δ of 1.422, 0.421 and 2.111 is required expecting 0.6, −0.6 and 0.9 if $m_e = 10$ prognostic variables are assumed. For $m_e = 60$ a minimal Δ of 0.646, 0.166 and 1.265 respectively is required.

Definition 7.1.0.2 *Assume that $k \leq m$ variables are selected for the construction of the prognostic score whose p-values from the one-sided two-sample t-test were smaller than γ ($p_j < \gamma$). Let $\bar{\mathbf{x}}_r$ and $\bar{\mathbf{x}}_{nr}$ denote the sample means of the $j = 1, ..., k$ selected variables of patients responding and not responding to therapy respectively and $\mathbf{x} = (x_1, ..., x_k)$ are the values of the corresponding variables of a future patient. To calculate the predictive outcome, the estimated covariance matrix $\hat{\Sigma}_k$ of the k selected variables is considered in* *the the prediction score (as in the classical discriminant analysis). The score value than is calculated by:*

$$\hat{f}(\mathbf{x};\gamma) = (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \mathbf{x} \tag{7.4}$$

Note that if the true covariance matrix would be known, the weight of one selected variable in the score also depends on the effect sizes of the two neighboring selected variables and the distance to the two neighbors. The secondary diagonals of the inverse of the true covariance matrix of the selected variables can be calculated by $\rho^{|i-j|}/(\rho^{2|i-j|}-1)$ where $\mid i-j \mid$ is the distance between the ith and jth variable. If $\mid i-j \mid$ is large, this term is close to zero.

According to the unknown variance case, the following results hold:

Theorem 7.1.0.4 *Given a FDR selection threshold α (corresponding to the selection boundary γ) and the estimated weights from the samples, the prognostic score follows two normal distributions:*

$$\hat{f}(\mathbf{x};\gamma) \sim N[\mu_a, \sigma_a^2] =$$
$$N[(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \boldsymbol{\mu}_a, (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \Sigma_k (\hat{\Sigma}_k^{-1})^T (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})]$$

where $\boldsymbol{\mu}_a^T$, $a = r$ or nr is the true mean vector of the k selected variables in a future responder or non-responder, respectively, $\hat{\Sigma}_k$ is the estimated covariance matrix and Σ_k is the true covariance matrix of the k selected hypotheses.

Proof: The results can be again derived by using theorem 2.1.0.7.

Theorem 7.1.0.5 *Fixing the appropriate $\boldsymbol{\mu}_a$ for the populations of responders and non-responders, respectively, it is easy to get the AUC for future independent populations:*

$$AUC(\alpha) = \int_0^1 \left\{ 1 - \Phi\left[z(1-w) - \frac{(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} (\boldsymbol{\mu}_r - \boldsymbol{\mu}_{nr})}{\sqrt{(\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})^T \hat{\Sigma}_k^{-1} \Sigma_k \hat{\Sigma}_k^{-1} (\bar{\mathbf{x}}_r - \bar{\mathbf{x}}_{nr})}} \right] \right\} dw. \tag{7.5}$$

Proof: The results can again be derived as in theorem 3.3.0.2.

Note that because of the equal sample sizes in both groups $\hat{\Sigma}_k = (\hat{\Sigma}_r + \hat{\Sigma}_{nr})/2$ can again be calculated as the mean value over the two within group covariance matrices of the selected k hypotheses. Note also that in the correlated case the whole covariance matrix of the k selected variables is used whereas in the unknown variance case assuming independence across hypotheses, only the estimated variances are included in the score.

However, for the prediction score we need the inverse of $\hat{\Sigma}_k$. If too many hypotheses are selected to build a prediction score, there may be situations where $\hat{\Sigma}_k$ can not be inverted. Thus, our second approach is to ignore the underlying correlation structure in the construction of the prediction score estimating only the variances and setting the covariances to 0. $\hat{\Sigma}_k$ is than estimated as in the unknown variance case assuming independence across variables.

7.2 Simulation studies

The mean values of $AUC(\alpha)$ are evaluated within a grid of α values with interval 0.05 by simulation (5000 simulation runs). As in the previous sections, these mean values are interpolated using splines. α_{opt} is then again determined by maximizing the interpolated function.

Over all investigated examples the sample size is set to $n = 50$ and m is set to 1000. Figure 7.2 shows the example where $m_e = 10$ prognostic variables are searched within $m = 1000$ candidate variables. The black curves show the interpolated functions for the score applying only variance estimates in the weights whereas the grey curves show the interpolated functions for the score for using the whole estimated covariance matrices for the weights. The parameter ρ for the autoregressive correlation structure is assumed to be 0.6 (dashed lines), -0.6 (dotdashed lines) and 0.9 (dotted lines). One can see from the figure that for large α and large positive ρ no good estimate of the covariance matrix can be achieved. The grey curves are falling below the black curves. For $\rho = -0.6$ (positive and negative correlations between the hypotheses) the covariance matrix may also be invertible for larger values of α. The scores are achieving a slightly larger performance for large α as compared to scores only estimating the variance. However, because of the small effect size, both scores result in generally small performances as compared to AUC_*.

Note that the given grey curves can be only calculated from the simulated samples where the estimated covariance matrix $\hat{\Sigma}_k$ is regular. For large positive ρ this is the case only in a few simulated samples. Figure 7.4 shows the proportion of singular covariance matrices among the simulated samples (5000 runs) applying $\rho = 0.6$ (dashed curves), -0.6 (dot-dashed curves) and 0.9 (dotted curves) assuming $m_e = 10$ (Figure (A)) and 60 (Figure (B)) alternatives among $m = 1000$ hypotheses. It can be seen that the larger the α values the larger the probability that $\hat{\Sigma}_k$ is singular. Clearly the larger α the larger the number of selected variables and the worse the estimation of the covariance matrix. For more than 100 selected variables (more variables than samples), no regular estimate of the covariance matrix can be applied.

Assuming $m_e = 60$ among $m = 1000$ hypotheses (Figure 7.3, Figure 7.4 (B)) already for smaller α, large numbers of variables are selected to construct a prediction score and thus worse estimates of the covariance matrix are determined from the samples. The grey curves in Figure 7.3 show a much smaller performance for the scores applying the whole estimated covariance matrix as compared to the score only applying the variance estimates (black curves). For negative ρ the difference between the performances of both scores is smaller. However, over the whole investigated α values no good prediction score can be constructed from the data. The average performance is always smaller than 0.6.

It seems that the second approach only applying variance estimates works better. Table 7.1 shows the results for the best choice of the threshold α if we construct a score based on the unknown variance assumption despite the underlying correlation between the hypotheses. As discussed before if the correlations between prognostic variables are large and positive, the effect size Δ has to be very large. As a consequence of the large effect sizes required to achieve AUC_*, a good prediction score can also be constructed from the data in this case if only small sample sizes are available. If the correlation between prognostic variables is either positive or negative the effect sizes of each prognostic variable to achieve AUC_* can be very small. In such cases no good prediction scores can be achieved over the whole range of investigated FDR thresholds (see also Figures 7.2, 7.3).

Table 7.1: **Correlated hypotheses:** The best choice of the FDR threshold α_{opt}, the corresponding true $AUC(\alpha_{opt})$ as well as the number of non-prognostic variables (m_0^s) and the number of prognostic variables (m_e^s) included in the prediction score assuming a number of $m_e = 10$ and 60 prognostic variables among $m = 1000$ tested hypotheses. The parameter for the autoregressive correlation between hypotheses was set to $\rho = 0.6$, -0.6 and 0.9. The per group sample size is set to $n = 50$ and the effect size Δ is calculated to achieve $AUC_* = 0.965$.

m	m_e	Δ	n	ρ	α_{opt}	$AUC(\alpha_{opt})$	m_0^s	m_e^s
1000	10	1.422	50	0.6	0.024	0.961	0.27	10.00
		0.421	50	-0.6	0.793	0.621	34.43	5.31
		2.111	50	0.9	0.001	0.960	0.02	10.00
1000	60	0.646	50	0.6	0.225	0.950	15.52	50.94
		0.166	50	-0.6	0.908	0.588	340.22	37.07
		1.265	50	0.9	0.025	0.958	1.66	60.00

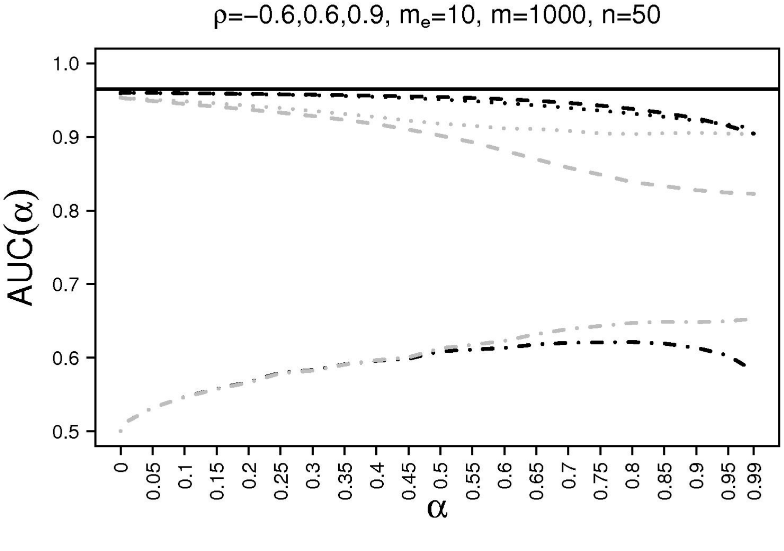

Figure 7.2: **Correlated hypotheses:** Interpolated functions of mean values of $AUC(\alpha)$ (over 5000 simulation runs) for a varying FDR selection threshold α assuming $m_e = 10$ prognostic variables among $m = 1000$ tested variables. The parameter ρ for the autoregressive correlation structure was set to 0.6 (dashed lines), -0.6 (dotdashed lines) and 0.9 (dotted lines). Functions for scores using the whole covariance matrix (grey curves) or only variance estimates (black curves) are given. The sample size is set to $n = 50$. $AUC_* = 0.965$ is given as solid horizontal line.

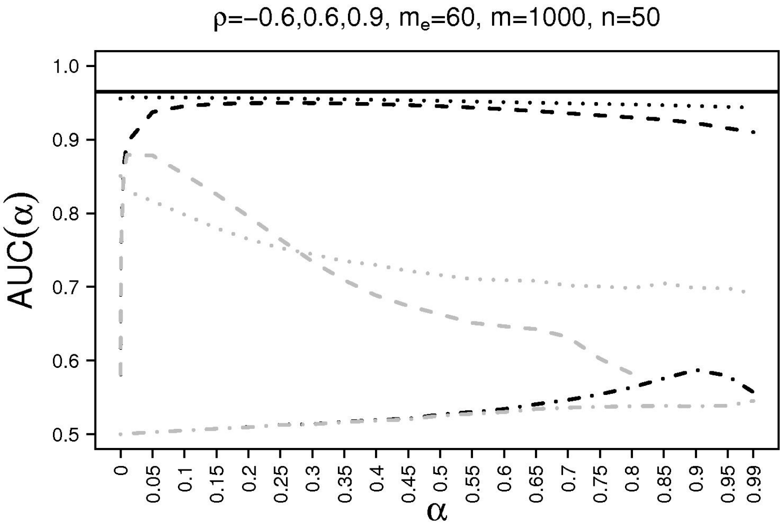

Figure 7.3: **Correlated hypotheses:** Interpolated functions of mean values of $AUC(\alpha)$ (over 5000 simulation runs) for a varying FDR selection threshold α assuming $m_e = 60$ prognostic variables among $m = 1000$ tested variables. The parameter ρ for the autoregressive correlation structure was set to 0.6 (dashed lines), -0.6 (dotdashed lines) and 0.9 (dotted lines). Functions for scores using the whole covariance matrix (grey curves) or only variance estimates (black curves) are given. The sample size is set to $n = 50$. $AUC_* = 0.965$ is given as solid horizontal line.

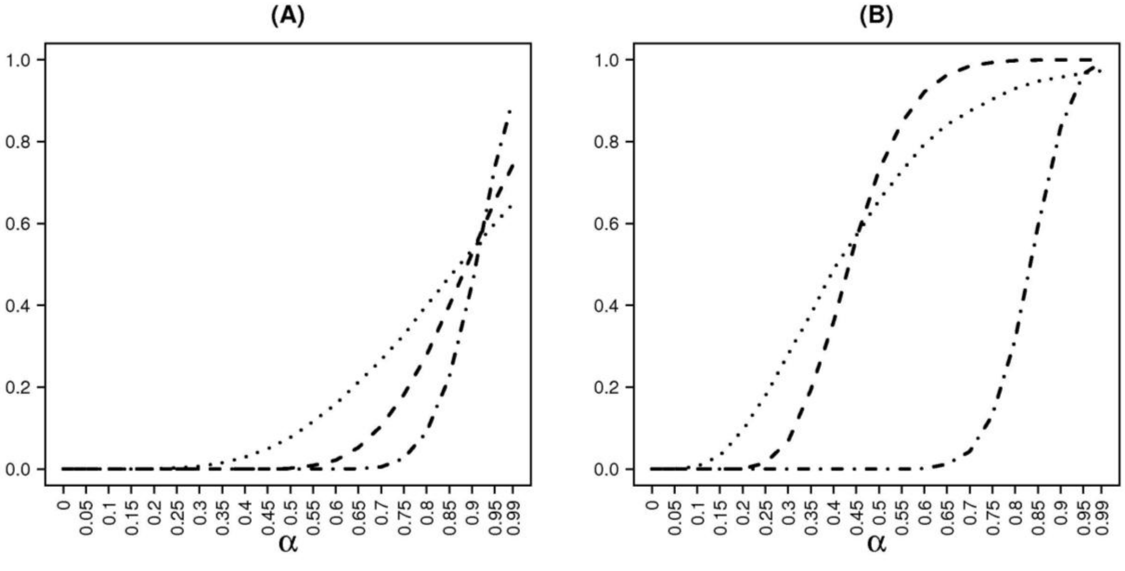

Figure 7.4: Proportion of singular covariance matrices among the simulation runs for a varying FDR threshold α. The parameter ρ for the autoregressive correlation structure was set to 0.6 (dashed line), -0.6 (dotdashed line) and 0.9 (dotted line). $m_e = 10$ (Figure (A)) and 60 (Figure (B)) prognostic variables were assumed among $m = 1000$ tested hypotheses

7.3 Cross Validation

Note again that in the case of correlation between hypotheses we propose the unknown variance assumption standardizing the weights in the score by respective variance estimates. If a large number of variables are included in the score and a small sample size is applied, $\hat{\Sigma}_k$ can not always be inverted. Choosing the simple additive score with standardized weights leads to only slightly poorer performances in terms of the AUC than considering the whole estimated covariance matrix (if possible). This is a reason why the simple additive score has attracted a lot of attention in applications.

In the situation of correlated variables, generally the same tendencies for the prognostic scores determined by the cross validation procedure can be found as in the situation of independence between hypotheses. Because of the flat optimum of the interpolated functions for large positive ρ (see Figures 7.2, 7.3) again there may be large differences between $\hat{\alpha}_{opt}$ determined from the cross validation procedure and α_{opt}. However, again the determined $\hat{\gamma}_{opt}$ and corresponding $\hat{\alpha}_{opt}$ values are leading to a mean future performance in terms of $AUC(\hat{\gamma}_{opt})$ which is close to $AUC(\alpha_{opt})$ (see Table 7.2 for the cross validation results of the investigated examples).

Despite the underlying correlation structure the cross validation procedure (only considering the estimated variances in the score) seem to work well ending in larger cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ values and small $\hat{\alpha}_{opt}$ values if the alternative holds and in small cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ values and large $\hat{\alpha}_{opt}$ values if the global null hypothesis is true (see histograms of $\widehat{AUC}(\hat{\gamma}_{opt})$ in Figure 7.5 and of $\hat{\alpha}_{opt}$ in Figure 7.6). Assuming $\rho = 0.6$ and 0.9 the effect sizes have to be very large to achieve the benchmark AUC_*. Thus the cross validation procedure ends in $\widehat{AUC}(\hat{\gamma}_{opt})$ values larger than 0.9 (Figure 7.5 first and third row). For $\rho = -0.6$ the effect size Δ is very small in order to achieve AUC_*. Fortunately, values of $\widehat{AUC}(\hat{\gamma}_{opt})$ are also small under the alternative indicating that no good prediction score can be determined with the given effect and sample sizes (Figure 7.5 second row).

$\hat{\alpha}_{opt}$ is varying large, however, being on average small for $\rho = 0.6$ and 0.9 due to the large effect sizes. For example, for $\rho = 0.9$, $\hat{\alpha}_{opt}$ on average is 0.003 assuming $m_e = 10$ prognostic variables. For $\rho = -0.6$ and assuming $m_e = 10$ an average $\hat{\alpha}_{opt} = 0.588$ is indicating that the resulting score includes more than 50% non-prognostic variables. For $m_e = 60$ more than 80% non-prognostic variables can be expected in the prediction score. Table 7.2 summarizes the results under the alternative.

Table 7.3 shows the results of the cross validation procedure under the global null hypothesis. Under the global null, $\widehat{AUC}(\hat{\gamma}_{opt})$ is generally smaller than 0.7 and $\hat{\alpha}_{opt}$ is generally larger than 0.9 (see medians in table Table 7.3). This indicates that we can expect that more than 90% variables without prognostic ability are included in the score and that the performance of the selected score is very poor. In this situation one may conclude that no prediction score should be selected from the given data.

The cross validation procedure seems to work also in the situation of correlated hypotheses ending at $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$ values giving a good evaluation of the underlying prediction score. Again the conclusion is that both criteria, $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$ should be considered to decide whether a score should be constructed from a given data set or not. Figure 7.7 shows scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ for all investigated examples. However, one contradiction remains, the differences between the estimated $\hat{\alpha}_{opt}$ and the true FDR are slightly larger under the correlated case than under the independent case (see Tables 7.2 and 7.3).

Table 7.2: **Cross Validation in the correlated case:** The true best choice of the FDR, α_{opt}, and the corresponding $AUC(\alpha_{opt})$ as well as results determined from the cross validation procedure: the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR and $\hat{\alpha}_{opt,\infty}$, the true $AUC(\hat{\gamma}_{opt})$, the cross validated $\widehat{MD}(\hat{\gamma}_{opt})$ and the FWER for a varying number of prognostic variables m_e. The number of tested variables is set to $m = 1000$. The per-group sample size is fixed to $n = 50$. $\rho = 0.6$, -0.6 and 0.9. $AUC_* = 0.965$.

m_e	10			60		
ρ	0.6	-0.6	0.9	0.6	-0.6	0.9
Δ	1.422	0.421	2.111	0.646	0.166	1.265
α_{opt}	0.024	0.793	0.001	0.225	0.908	0.025
$AUC(\alpha_{opt})$	0.961	0.621	0.960	0.950	0.588	0.958
$\hat{\gamma}_{opt}$	0.007 (0.01)	0.013 (0.01)	0.00004 (0.0001)	0.030 (0.03)	0.292 (0.27)	0.0096 (0.01)
	0.002	*0.007*	*0.0000005*	*0.019*	*0.167*	*0.0018*
$\hat{\alpha}_{opt}$	0.220 (0.26)	0.588 (0.25)	0.003 (0.009)	0.281 (0.17)	0.888 (0.07)	0.103 (0.14)
	0.128	*0.634*	*0.0000505*	*0.254*	*0.891*	*0.024*
FDR	0.243 (0.26)	0.579 (0.31)	0.021 (0.07)	0.278 (0.19)	0.857 (0.06)	0.110 (0.16)
	0.167	*0.692*	*0.000*	*0.250*	*0.861*	*0.016*
$\hat{\alpha}_{opt,\infty}$	0.226 (0.26)	0.607 (0.21)	0.004 (0.01)	0.283 (0.17)	0.858 (0.05)	0.105 (0.14)
	0.131	*0.658*	*0.0000495*	*0.255*	*0.854*	*0.027*
$AUC(\hat{\gamma}_{opt})$	0.957 (0.01)	0.618 (0.11)	0.960 (0.001)	0.947 (0.01)	0.584 (0.01)	0.957 (0.002)
	0.959	*0.624*	*0.960*	*0.948*	*0.586*	*0.957*
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.964 (0.02)	0.744 (0.07)	0.960 (0.01)	0.957 (0.02)	0.661 (0.07)	0.960 (0.02)
	0.966	*0.751*	*0.960*	*0.959*	*0.667*	*0.960*
FWER	0.648	0.826	0.110	0.986	0.998	0.542

Table 7.3: **Cross Validation in the correlated case under the global null:** Results determined from the cross validation procedure: the selection boundary $\hat{\gamma}_{opt}$ and the corresponding $\hat{\alpha}_{opt}$, the true FDR using the determined selection threshold, $\hat{\alpha}_{opt,\infty}$ (always 1) the true $AUC(\hat{\gamma}_{opt})$ (always 0.5), the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the FWER for a varying number of parameters for the autoregressive correlation ρ. The number of tested variables is set to $m = 1000$. The per-group sample size is fixed to $n = 50$. $AUC_* = 0.965$.

ρ	0.6	-0.6	0.6
$\hat{\gamma}_{opt}$	0.034 (0.06)	0.044 (0.08)	0.040 (0.08)
	0.006	*0.009*	*0.007*
$\hat{\alpha}_{opt}$	0.812 (0.22)	0.841 (0.20)	0.770 (0.27)
	0.916	*0.935*	*0.904*
FDR	0.972 (0.17)	0.964 (0.19)	0.886 (0.32)
	1.000	*1.000*	*1.000*
$\hat{\alpha}_{opt,\infty}$	1.000	1.000	1.000
$AUC(\hat{\gamma}_{opt})$	0.500	0.500	0.500
$\widehat{AUC}(\hat{\gamma}_{opt})$	0.650 (0.07)	0.662 (0.07)	0.604 (0.07)
	0.659	*0.670*	*0.607*
FWER	0.972	0.964	0.886

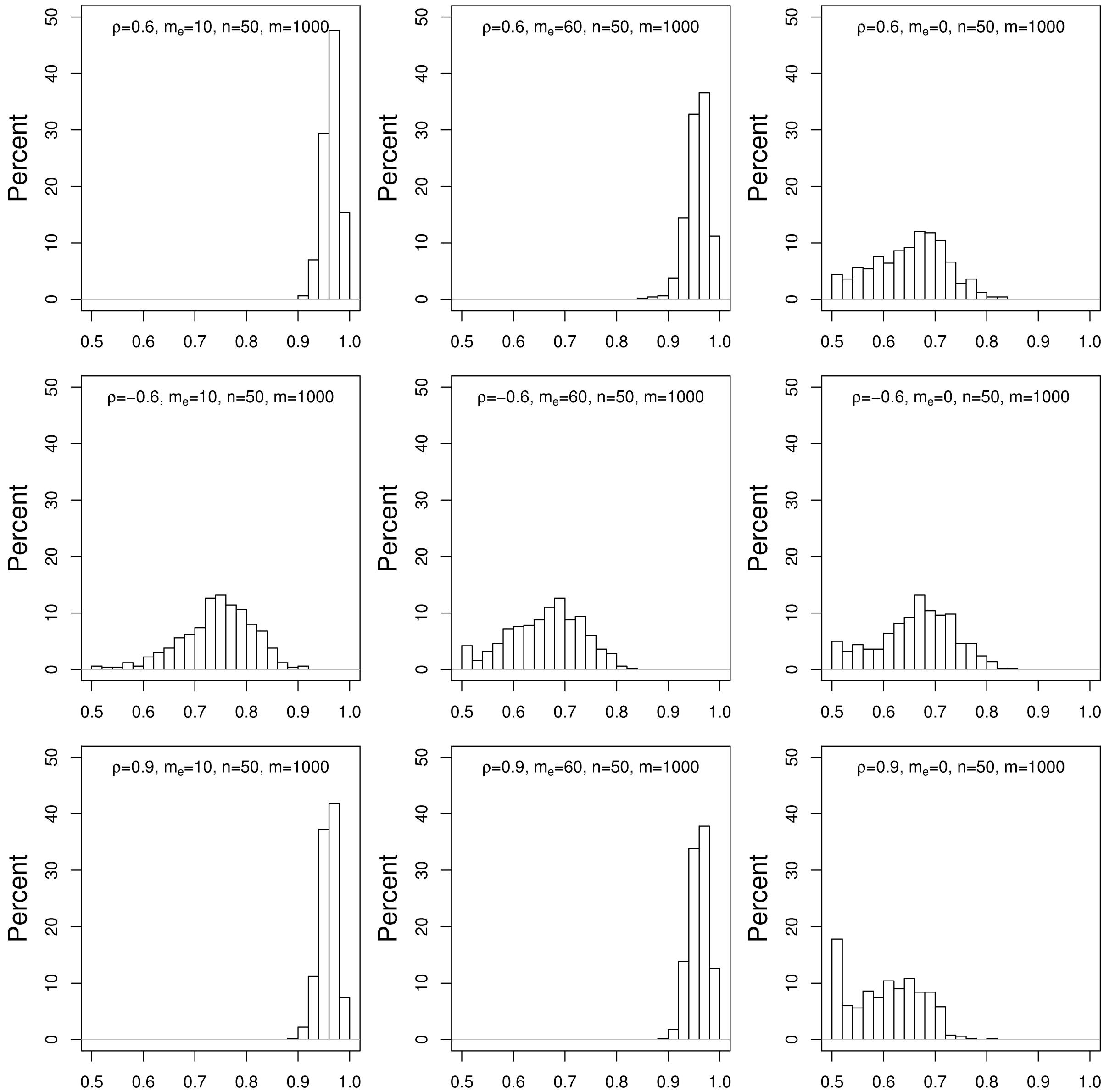

Figure 7.5: **Correlated hypotheses:** Distribution of the cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ (500 simulation runs) assuming correlation between variables: $m_e = 10$ (first column), 60 (second column) or 0 (third column) are assumed among $m = 1000$ hypotheses. $\rho = 0.6$ (first row) -0.6 (second row) and 0.9 (third row) is assumed. The sample size is set to $n = 50$.

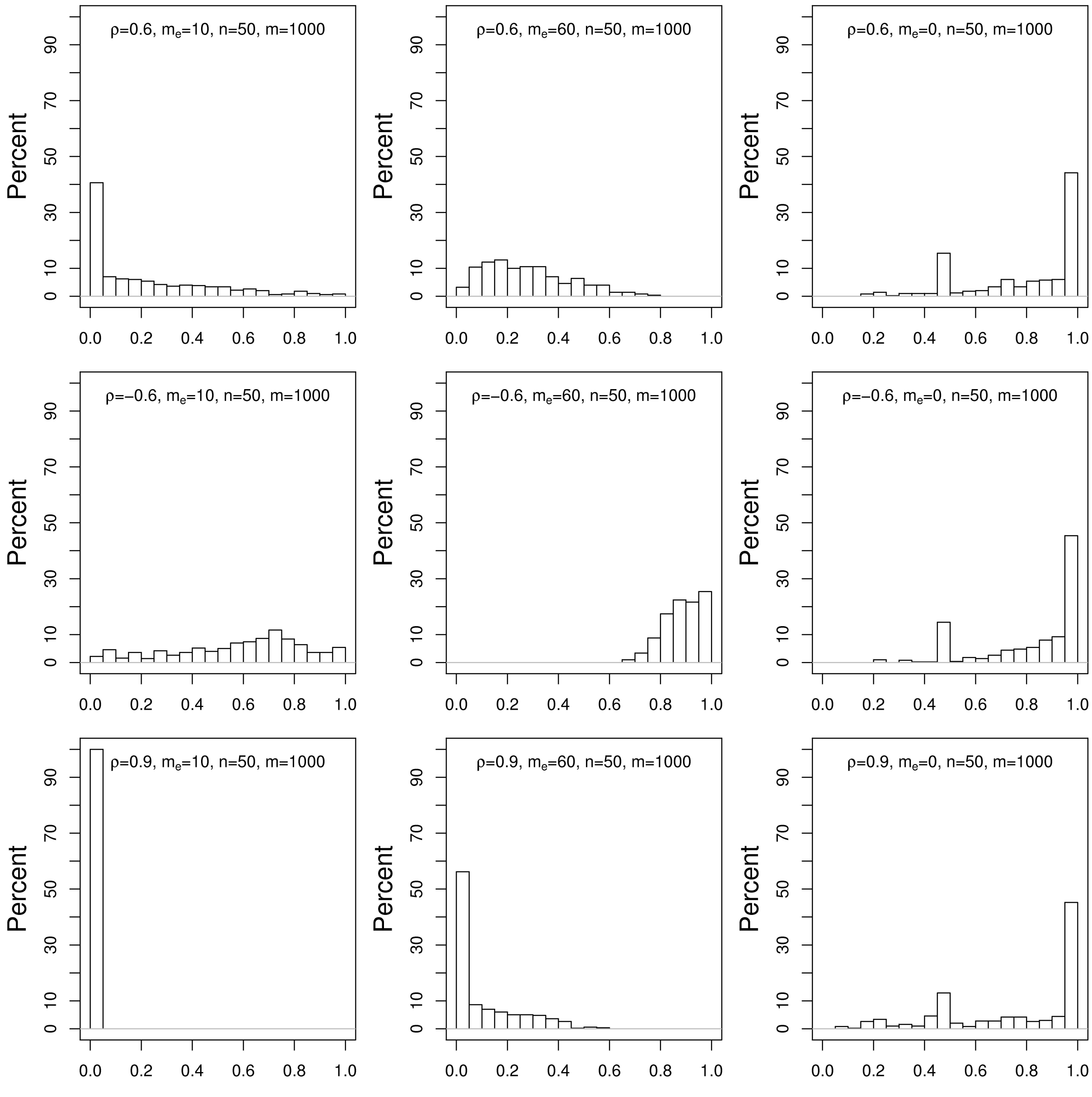

Figure 7.6: **Correlated hypotheses:** Distribution of the cross validation based $\hat{\alpha}_{opt}$ (500 simulation runs) assuming correlation between variables: $m_e = 10$ (first column), 60 (second column) or 0 (third column) are assumed among $m = 1000$ hypotheses. $\rho = 0.6$ (first row) -0.6 (second row) and 0.9 (third row) is assumed. The sample size is set to $n = 50$.

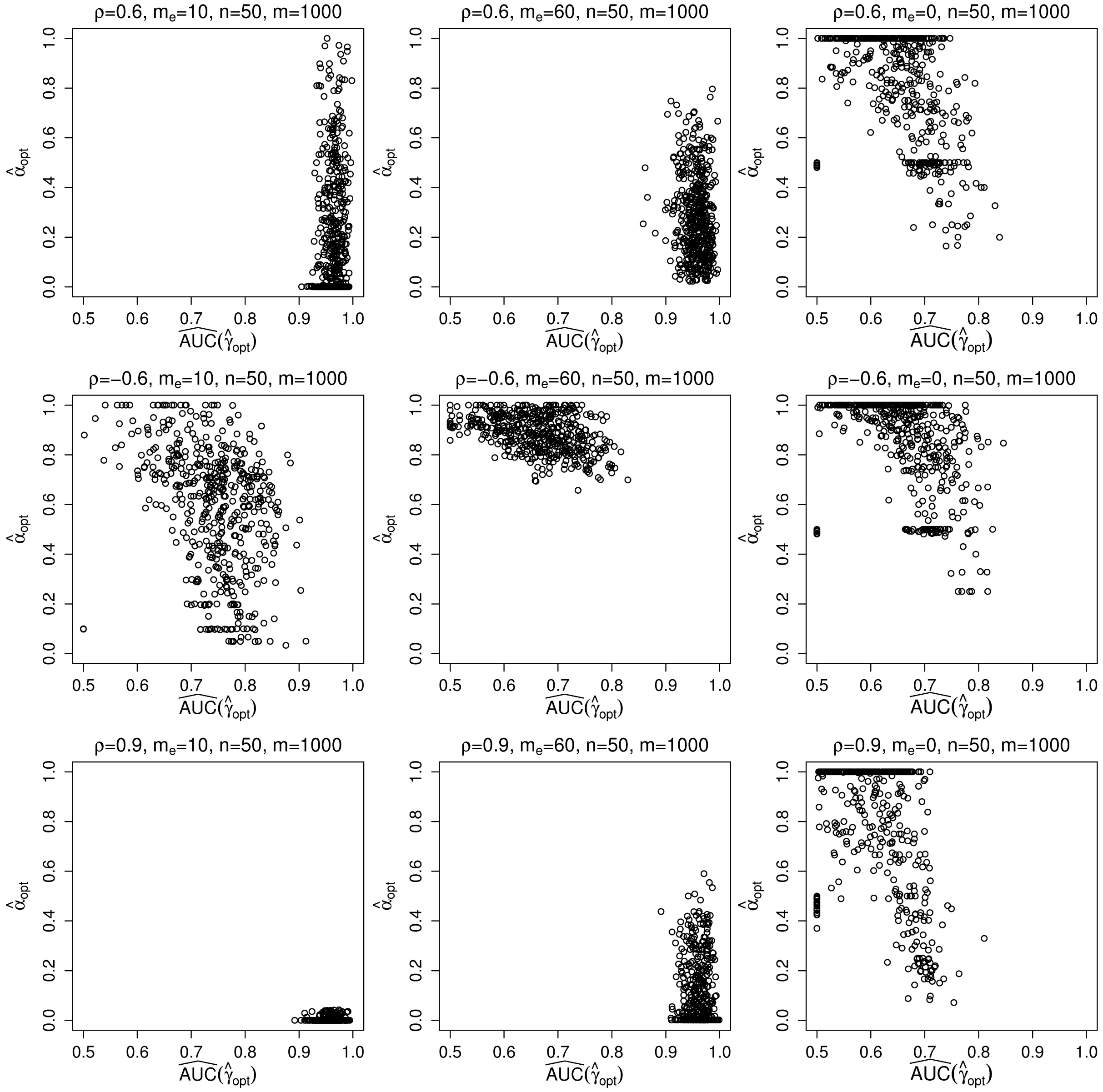

Figure 7.7: **Correlated hypotheses:** Scatterplots of $\widehat{AUC}(\hat{\gamma}_{opt})$ vs. $\hat{\alpha}_{opt}$ (500 simulation runs) assuming correlation between variables: $m_e = 10$ (first column), 60 (second column) or 0 (third column) are assumed among $m = 1000$ hypotheses. $\rho = 0.6$ (first row) -0.6 (second row) and 0.9 (third row) is assumed. The sample size is set to $n = 50$.

8 Real data applications

To evaluate the cross validation procedure we used three data sets summarized and pre-processed by Pavlidis et al. (2003) and a data set investigated in Tian et al. (2003). Note that because of the large number of investigated genes in the four investigated data sets (> 6000) it was not possible to apply the forward logistic regression in SAS 9.1. due to lack of memory space.

8.1 Data set: Tian et al. (2003)

First we investigate the data set taken from Tian et al. (2003) and pre-processed by Jeffery et al. (2006). In this study, patients with multiple myeloma were investigated. 36 patients in whom focal lesions of bone could not be detected were compared to 137 patients with such lesions. They subjected purified plasma cells from the bone marrow of patients with newly diagnosed multiple myeloma to oligonucleotide microarray profiling. The data was generated using Affymetrix human U95A. 12625 probe sets were investigated. In order to construct a prediction score we compare the two independent groups using the p-values of the two-sided t-tests.

The cross validation procedure determines an $\widehat{\alpha}_{opt}$ of 0.0124 leading to a score including 101 probe sets. Figure 8.1 (A) shows $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the selection boundary γ for the individual p-values. Figure (B) shows a histogram of the 12625 two-sided p-values. $\widehat{AUC}(\hat{\alpha}_{opt}) = 0.786$ is indicating a rather limited performance for a future independent patient. However, the result indicates that in the example one may be confident that the selected score will not contain a noticeable

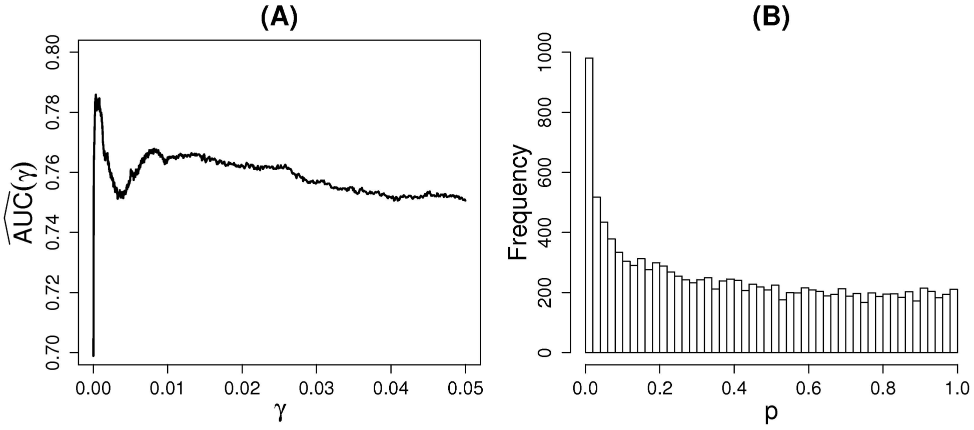

Figure 8.1: Results of the cross validation procedure for the data set investigated in Tian et al. (2003). $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the threshold γ for the individual p-values is shown in Figure (A). Figure (B) shows a histogram of the individual p-values.

fraction of non-prognostic genes. A summary of the results can be seen in Table 8.1.

8.2 Data set: Golub et al. (1999)

In the study by Golub et al. (1999) gene expression profiles of two types of leukaemia were compared. Samples were derived from 47 patients with acute lymphoblastic leucemia (ALL) and 25 patients with acute myeloblastic leucemia (MLL). RNA prepared from bone marrow mononuclear cells was hybridized to high-density oligonucleotide microarrays, produced by Affymetrix. 7129 probe sets were investigated.

To construct a prediction score we again compare the two independent groups (ALL vs. MLL) using the p-values of the two-sided t-tests. A summary of the results of the cross validation procedure can be seen in Table 8.1. Figure 8.2 (A) shows $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the threshold γ for the individual p-values. Figure (B) shows a histogram of the two-sided p-values. From the histogram one can see that for approximately 1500 probe sets the corresponding p-values are smaller than 0.02.

The cross validation procedure determines a very small $\hat{\alpha}_{opt} = 0.0001$ achieving a cross validation based $\widehat{AUC}(\hat{\gamma}_{opt}) = 0.988$. This result may be an indication that with the se-

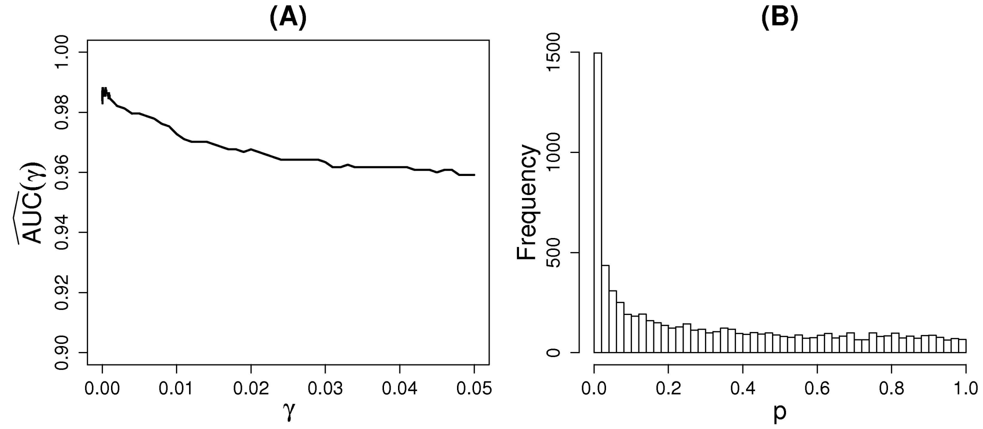

Figure 8.2: Results of the cross validation procedure for the data set investigated in Golub et al. (1999). $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the threshold γ for the individual p-values is shown in Figure (A). Figure (B) shows a histogram of the individual p-values.

lected genes we may get a very good determination between the two investigated groups. The determined prediction score includes 103 genes and the small $\hat{\alpha}_{opt}$ indicates that the selected score may not contain a large fraction of non-prognostic genes.

8.3 Data set: Eaves et al. (2002)

In the study of Eaves et al. (2002) they used high-density oligonucleotide arrays to measure the relative expression levels of 39114 genes of mouse spleen and thymus. We investigated a distinction (spleen vs. thymus) that was not examined in the original publication but have been already discussed by Pavlidis et al. (2003). We used the data set preprocessed by Pavlidis et al. (2003).

Again we performe a two-sided t-test to determine candidate variables for the construction of a prognostic score that distinguishes between genes corresponding to spleen or thymus of mice. Figure 8.3 (B) shows a histogram of the two-sided p-values. The distribution of the p-values shows that only a few p-values are very small and that a large number p-values is larger than 0.8. There is no explanation for the strange distribution of p-values. Despite the strange distribution of the p-values we investigated the cross validation structure for this data set. Looking at the results of the cross validation procedure (see Figure

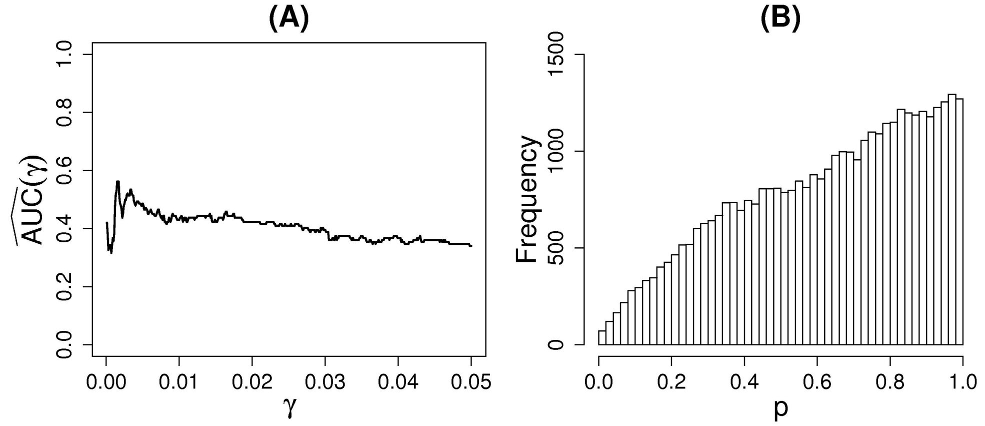

Figure 8.3: Results of the cross validation procedure for the data set investigated in Eaves et al. (2002). $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the threshold γ for the individual p-values is shown in Figure (A). Figure (B) shows a histogram of the individual p-values.

8.3 (A) and Table 8.1) one may get an indication that there is no good discrimination between the two groups. $\hat{\alpha}_{opt}$ is estimated with 1 leading to $\widehat{AUC}(\hat{\gamma}_{opt}) = 0.562$. However, the sample size in both groups is 12 such that the bad result may also be an indication that the sample size is too small to detect the prognostic variables. Note also that only 5 genes were included in the resulting prediction score.

8.4 Data set: Callow et al. (2000)

Callow et al. (2000) tried to identify genes with altered expression levels in knockout mice compared to control mice. Based on the assumption that severe alterations in the expression of genes known to be involved in high-density lipoprotein (HDL) metabolism may affect the expression of other genes, they screened an array of 6384 mouse expressed sequence tags for altered gene expression in the livers of one line of mice with dramatic decreases in HDL plasma concentrations. Labeled cDNA from livers of apoAI-knockout mice and control mice were cohybridized to microarrays. A very small sample of 8 knockout and 8 control mice was used.

Again two-sided two-sample t-tests are used to determine candidates to construct a prediction score (see histogram of the two-sided p-values in Figure 8.4 (B)). The results of the cross validation procedure (see Figure 8.4 (A) and Table 8.1) give an indication that a very good discrimination between the two groups can be achieved with a few genes. The prognostic score determined by using the cross validation procedure includes only 3 genes in the determined score. The cross validation based $\widehat{AUC}(\hat{\gamma}_{opt})$ is set to 1 indicating complete discrimination of the two groups. $\hat{\alpha}_{opt} = 0.0011$ also indicates a vary small proportion of non-prognostic genes.

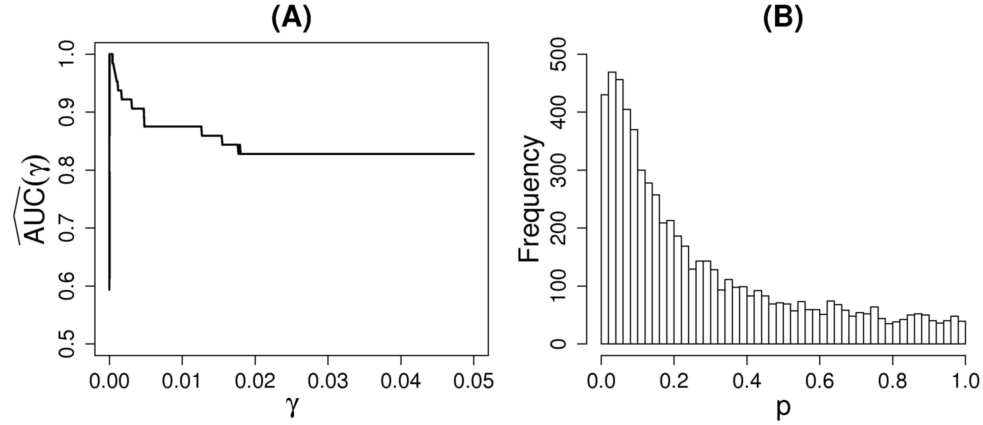

Figure 8.4: Results of the cross validation procedure for the data set investigated in Callow et al. (2000). $\widehat{AUC}(\gamma)$ determined from cross validation as a function of the threshold γ for the individual p-values is shown in Figure (A). Figure (B) shows a histogram of the individual p-values.

Table 8.1: **Real data applications:** Results of the cross validation procedure determined for real data sets are shown. The reference of the corresponding paper, the group sample sizes (n_1/n_2), the number of investigated genes and the type of array which was used for the study, either cDNA for data that was collected using two-color "cDNA" microarrays or "oligo" for Affymetrix-type oligonucleodtide arrays. The best selection boundary $\hat{\gamma}_{opt}$, the corresponding estimated FDR $\hat{\alpha}_{opt}$ and $\hat{\pi}_0$ as well as the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ and the number of genes included in the prediction score (♯) determined by the cross validation procedure are given.

Data Description				Cross validation				
Reference	(n_1/n_2)	Genes	Type	$\hat{\gamma}_{opt}$	$\hat{\alpha}_{opt}$	$\hat{\pi}_0$	$\widehat{AUC}(\hat{\gamma}_{opt})$	♯
Tian *et.al.* (2003)	137/36	12625	oligo	0.00034	0.0125	0.294	0.786	101
Golub *et.al.* (1999)	47/25	7129	oligo	0.000002	0.0001	0.213	0.988	103
Eveas *et.al.* (2002)	12/12	39114	oligo	0.00150	1.0000	0.545	0.562	5
Callow *et.al.* (2000)	8/8	6384	cDNA	0.000005	0.0015	0.143	1.000	3

9 Conclusions

To get a good prediction (in terms of the ROC-Curve) of a clinical outcome with a single prognostic variable the effect size has to be very large. Therefore only small samples are sufficient to identify this single variable among a large number of candidates by a statistical comparison between responder and non-responder. However, this is not the typical situation we are faced with. Generally we are confronted with large numbers of candidates, few of them being related with a clinical outcome having rather small effect sizes. Selection and estimation are often based on samples dramatically smaller than the number of candidates so that the asymptotic of model selection procedures does not apply. The estimates of the selected weights (and the ROC-curves) are biased and highly variable.

We performed simulations using multiple tests controlling the FDR for selection of variables for future prediction. We additionally performed simulations using the binary logistic regression model for the investigated problem. In the situation where the sample sizes are much smaller than the number of tested variables the simple method of additive scores following a selection of variables by multiple testing based on a FDR threshold general outperforms selection by forward stepwise logistic regression. The appearing problem of complete separation of data points results in selecting only a few alternatives for future prediction and thus to a poor performance. If the number of prognostic variables is rather small and they have sufficiently large effects, which, if all would be known, would lead to a large AUC, then the selected scores may have good properties over a range of different FDR values used for selection. Under the alternative in general it seems to be preferable to use rather liberal selection criteria accepting that a certain number of non-prognostic variables is contained in a score to get the advantage of catching more effective ones. For large samples the predictive ability of the estimated score does not depend strongly on the number of selected variables. For large sample sizes the weights for non-prognostic variables contained in the score are estimated more precisely and, despite of the selection procedure, will tend to be the ones closed to zero with a small contribution overall. Hence the performance of the score will not vary much for different numbers of non-prognostic variables contained in the score: More liberal selection criteria will lead to scores containing more nuisance variables (with low weights) but also more variables related with the clinical outcome (with large weights). This mirrors the fact that asymptotically for large sample sizes, multiple test based selection procedures may be consistent procedures for model selection. It also has to be mentioned, that if a very large number of prognostic variables is expected working together with rather small individual effect sizes and only small sample sizes are available, the selection methods based on univariate tests also perform worse, although leading to a larger AUC for future prediction as using the forward logistic regression.

The cruical scenario in the small sample case is the global null hypothesis: There are no prognostic variables at all and hence any selection will lead to completely uninformative prediction scores. To protect against erroneous selection in this situation the FDR applied for selection should be rather small. Under the global null hypothesis control of the FDR also controls the probability of the selection of any variables. Under the alternative, however, we found that rather larger FDR values should be used for selection.

One way to determine the FDR-value to be applied for selection in a concrete sample in order to achieve good prediction by a prognostic score in terms of the AUC is to estimate the selection boundaries by cross validation. The discussed method seems to work if we really are in a situation that we deal with variables with a high prognostic potential which leads to rather large cross validation estimates of the AUC, whereas under the global null hypothesis these estimates are rather small.

However the situation gets much worse if we look at situation when the prognostic variables are not sufficiently large that the optimal score (if known) would lead to a ROC-curve crossing the point with sensitivity and specificity of 0.9 (with a theoretically best achievable $AUC_* = 0.965$ which was the benchmark in most of our investigations). For AUC_*

close to values of 0.8, which are values, e.g., achievable in predicting hospital mortality from scores based on a set of variable measured in patients at admission to an intensive care unit (and constructed in large training samples), the selection procedure will lead to poor scores. However, the cross validation procedure still seem to work well by identifying scores with an estimated AUC close to the best AUC achievable by a FDR-based selection procedure in the sample.

With this cross validation method we may achieve several goals:

1. We determine an optimal selection threshold for selection of variables to be used in a prediction score for future sample units which provides a good performance in terms of the ROC-curve.

2. We get a positively biased estimate of the AUC which is closer to the true AUC for prediction the larger the effect and sample sizes.

3. If the estimate of the AUC is small this may be an indication that in a specific sample we are close to the global null hypotheses or the effect sizes are to small for the given sample size.

4. We also get an estimate of the FDR among the selected variables which is close to the true FDR (with a direction depending on the magnitude of the FDR).

The specific contribution of this diploma thesis was to investigate the properties of the proposed procedure in case deviations from the simple assumptions of one-sided tests for independent normally distributed variables with common known variance: two-sided tests, unknown variance, distributed alternatives and correlation between variables. The general tendencies found for the simple one-sided known variance case still apply under more general model assumptions. The performances of the determined prognostic scores only slightly decrease as compared to the known variance case. Assuming an autoregressive correlation structure between the candidate variables, the effect size of the alternatives and thus also the performance of the determined scores depend on the parameter ρ. Assuming a large positive ρ the effect size has to be very large in order to achieve AUC_*. Thus, a good prediction score can be constructed from the data in this case if only small sample sizes are available. If the correlation is either positive or negative (negative ρ), the effect size to achieve AUC_* can be very small. Thus, no good prediction scores can be achieved if small sample sizes are applied. However, also in the correlated case the cross validation procedure is reflecting the performance of the prediction scores with the estimated $\widehat{AUC}(\hat{\gamma}_{opt})$ and $\hat{\alpha}_{opt}$ values.

Our findings show that simple method can lead to well performing prediction scores even in rather small samples, given that we deal with a problem where prognostic variables with noticeable effects are involved. However, they also tell us that there is no such thing as a free lunch in a statistically odd problem of dealing with large numbers of variables considerably exceeding sample sizes.

Acknowledgement

I am grateful to my supervisor Peter Bauer for giving me the opportunity to carry out this work and for his support and assistance during the years.

I also want to thank Franz König for his advice regarding SAS codes and Martin Posch for valuable commands. Furthermore I would like to express my gratitude to my parents Johann and Maria Goll as well as to Franklin, Ferdinand and my husband Andreas Graf for their support and their continuous encouragement to remain defiant.

This work was supported by the Austrian Science Fond (FWF) No. P18698-n15.

A Abstract

Multiple testing has been applied for selecting prognostic variables related with a clinical outcome (response to therapy) from a large number of candidates in small samples of "responding" or "non-responding" patients which are then used to estimate a score for prediction in future patients. We evaluated selection based on control of the false discovery rate (FDR) to build a linear score by considering the resulting receiver operating characteristic (ROC) for independent prediction of future patients. We simulated different scenarios with varying number of tested candidates, proportion of prognostic variables and sample sizes. Underlying effect sizes were determined such that optimal prediction, if known, would lead to a ROC-curve crossing through a benchmark point with pre-fixed values of sensitivity and specificity. We show that the "best" FDR-threshold which provides the ROC-curve with the largest area under the curve (AUC) varies largely over the different parameter constellation not known in advance.

Hence, cross validation is proposed to determine the optimal selection threshold in a specific sample. This procedure (i) allows to choose an appropriate selection criterion, (ii) results in an estimate of the AUC for future prediction (though positively biased) and (iii) provides an estimate of the FDR close to the true FDR. Moreover, low estimates of the cross validated AUC and large estimates of the cross validated FDR may indicate a lack of sufficiently prognostic variables and/or too small sample sizes.

Keywords: Variable Selection; False Discovery Rate; Receiver Operating Characteristic Curve; Cross Validation

B Kurzfassung

In dieser Arbeit wird die Selektion von Variablen die (in Wahrheit) einen Einfluss auf einen klinischen Endpunkt (z.B. den Ausgang einer bestimmten Therapie) haben aus einer großen Menge von Kandidatenvariablen mit Hilfe von nur kleinen Stichproben von Patienten, die auf die Therapie reagieren bzw. nicht reagieren, behandelt. Die Selektion basiert auf einer multiplen Testprozedur die die False Discovery Rate (FDR) einhält. Mit jenen, mit Hilfe der multiplen Testprozedur selektierten, Variablen soll ein prognostischer Score konstruiert werden, mit dem man den klinischen Endpunkt eines zukünftigen Patienten vorhersagen kann. Dieser lineare Score wird aufgrund der resultierenden Receiver Operating Characteristic Curve (ROC) bewertet. Die Selektionsgrenze für die FDR, welche die beste Fläche unter der ROC-Kurve (AUC) liefert ist allerdings von unbekannten Parametern wie z.B. der Effektgröße oder der Anzahl der Variablen, die tatsächlich einen Einfluss auf den klinischen Endpunkt haben stark abhängig.

Um in einem spezifischen Datensatz nach der optimalen Selektionsschranke zu suchen wird die Verwendung einer Prozedur zur Kreuzvalidierung vorgeschlagen. Diese Prozedur (i) ermittelt ein adäquates Selektionskriterium für die multiple Testprozedur, (ii) berechnet einen (positiv verzerrten) Schätzer für die AUC für zukünftige Prognosen und (iii) liefert einen Schätzer für die FDR, der nahe der wahren FDR ist. Darüber hinaus geben niedrige Werte der ermittelten kreuzvalidierten AUC und große Werte der kreuzvalidierten FDR einen Hinweis darauf, dass der Einfluss der Variablen auf den klinischen Endpunkt zu gering ist und/oder dass die gegebene Stichprobengröße zu gering ist um die gegebenen Effekte zu finden.

Sichwörter: Variablenselektion; False Discovery Rate; Receiver Operating Characteristic Curve; Kreuzvalidierung

C Curriculum Vitae

Personal data:

- Name: Alexandra Christine Graf
- Name before Marriage: Alexandra Christine Goll
- Academic Degree: Mag.rer.soc.oec. Dr.rer.soc.oec
- Date of Birth: 20.08.1979
- Place of Birth: Vienna
- Nationality: Austria
- Parents: Ing. Johann and Maria Goll
- Date of Marriage: 20.09.2008
- Name of Husband: Andreas Graf

Education:

- 1985-1989 Elementary school in Langenzersdorf, Niederösterreich
- 1989-1993 Grammar school in Stockerau, Niederösterreich
- 1993-1998 Commercial academy, Korneuburg, Niederösterreich
  Graduation: June, 1998
- 1998-2003 Academic studies in Statistics, Graduation: October, 2003
  Title of Diploma Thesis: "Multikriterielle Tourenplanung für mobile Gesundheitseinrichtungen."
  Supervisor: Ao.Univ.Prof. Mag. Dr. Walter Gutjahr, ISDS, University of Vienna
- 2005-2008: Doctoral studies in Statistics, Graduation: April, 2008
  Title of Doctoral Thesis: "Inference on a large number of hypotheses based on limited samples - some points to consider."
  Supervisor: o.Univ.Prof. Dr. Peter Bauer, IMS, Medical University of Vienna
- since WT 2000: Academic studies in Mathematics, completion first part: March, 2004

Work experience:

- Summer 1997-2002: several vacation jobs at CA, Mobilcom Austria and Statistik Austria
- October 2003 - September 2008: Part-time job at the Institute of Medical Statistics, Core Unit for Medical Statistics and Informatics, Medical University of Vienna
- May 2005 - April 2008: additional part-time job as doctoral student at the Institute of Medical Statistics: FWF-Project Nr.: P18698-n15

- May 2008 - September 2008: additional part-time post doctoral job at the Institute of Medical Statistics: FWF-Project Nr.: P18698-n15
- since September 2008: Univ.-Ass. at the Institute of Medical Statistics
- ST 2005: Lecturer at the Medical University Vienna: "SSM2: Methoden der Medizininischen Wissenschaften"
- ST 2006: Lecturer at the Medical University Vienna: "SSM3: Methodenseminar Statistik"
- ST 2006: External lecturer at the University of Vienna: "Übungen zu Mathematik 2 A für Statistiker und Volkswirtschaftler"
- ST 2006: External lecturer at the University of Vienna: "Übungen zu Mathematik 2 B für Statistiker"

Talks about the thesis:

- Statistics and Life Sciences: Perspectives and Challenges (LIFESTAT), March 2008 in Munich, Germany

Other talks:

- International Conference on Multiple Comparison Procedures (MCP), Juli 2007 in Vienna, Austria
- ROeS Seminar, September 2007 in Bern, Switzerland

Papers concerning the thesis:

- Goll, A. and Bauer, P. (2008). Model Selection based on the false discovery rate optimizing the area under the receiver operating characteristic curve. Submitted.

Other papers and clinical cooperations:

- Goll, A. and Bauer, P. (2007). Two-stage designs applying methods differing in costs, *Bioinformatics*, 23: 1519-1526.
- Zehetmayer, S., Goll, A., Bauer, P. and Posch, M. (2007). Step by Step: mehr Effizienz mit neuen Studiendesigns, *Biospektrum*, 7: 754-755.

- Krebs, I., Binder, S., Stolba, U., Kellner, L., Glittenberg, C., Goll, A. (2008). Subretinal surgery and transplantation of autologous pigment epithelial cells in retinal angiomatous proliferation. *Acta Ophthalmol*, 86: 504-509.
- Sipos, W., Holzer, M., Bayegan, K., Janata, A., Unterweger, Ch., Goll, A., Weihs, W., Bauer, P., Sterz, F. and Behringer, W. (2008). A novel highly observer-independent neurologic examination porcedure for pigs in a model for cardiac arrest resuscitation. *Vet.Med. Austria*, 95: 28-38.
- Sabeti-Aschraf, M., Serek, M., Pachtner, T., Auner, K., Machinek, M., Geisler, M. and Goll, A. (2008). The Enduro motorcyclist's wrist and other overuse injuries in competitive Enduro motorcyclists: a prospective study. *Scand J Med Sci Sports*, to appear

- Sabeti, M., Dorotka, R., Goll, A., Funovics, P., Schmidt, M., Schatz, K. and Kotz, R. (2008). Focussed extracorporeal shockwave therapy for tennis elbow. *Phys Med Rehab Kuror*, 18: 83-86.

- Krebs, I., Ansari-Shahrezaei, S., Goll, A. and Binder, S. (2008). Activity of neovascular lesions treated with bevacizumab: comparison between optical coherence tomography and fluorescein angiography. *Graefes Arch Clin Exp Ophthalmol*, 246: 811-815.

- Rabenlehner, D., Stanzel, B., Krebs, I., Binder, S. and Goll, A. (2008). Reduction of iatrogenic RPE lesions in AMD patients: evidence for wound healing? *Graefes Arch Clin Exp Ophthalmol*, 246: 345-352.

- Krebs, I., Krepler, K., Stolba, U., Goll, A. and Binder, S. (2008). Retinal angiomatous proliferation: combined therapy of intravitreal triamcinolone acetonide and PDT versus PDT alone. *Graefes Arch Clin Exp Ophthalmol*, 246: 237-243.

- Winkler, W., Zellner, M., Diestinger, M., Babeluk, R., Marchetti, M., Goll, A., Zehetmayer, S., Bauer, P., Rappold, E., Miller, I., Roth, E., Allmaier, G. and Oehler, R. (2008). Biological variation of the platelet proteome in the elderly population and its implication for biomarker research. *Mol Cell Proteomics*, 7: 193-203.

- Sabeti, M., Dorotka, R., Goll, A., Gruber, M. and Schatz, K.D. (2007). A comparison of two different treatments with navigated extracorporeal shock-wave therapy for calcifying tendinitis - a randomized controlled trial. *Wien Klin Wochenschr*, 119: 124-128.

- Dorotka, R., Sabeti, M., Jimenez-Boj, E., Goll, A., Schubert, S. and Trieb, K. (2006). Location modalities for focused extracorporeal shock wave application in the treatment of chronic plantar fasciitis. *Foot Ankle Int*, 27: 943-947.

- Stacher, G., Lenglinger, J., Eisler, M., Hoffmann, M., Goll, A., Bergmann, H. and Stacher-Janotta, G. (2006). Esophageal acid exposure in upright and recumbent postures: roles of lower esophageal sphincter, esophageal contractile and transport function, hiatal hernia, age, sex, and body mass. *Dig Dis Sci*, 51: 1896-1903.

- Zehetgruber, H., Grübl, A., Goll, A., Schwameis, E., Wurnig, C. and Giurea, A. (2005). Prevention of heterotopic ossification after THA with indomethacin: analysis of risk factors. *Z Orthop Ihre Grenzgeb*, 143: 631-637.

- Krebs, I., Binder, S., Stolba, U., Glittenberg, C., Brannath, W. and Goll, A. (2005) Choroidal neovascularization in pathologic myopia: three-year results after photodynamic therapy. *Am J Ophthalmol*, 140: 416-425.

- Krebs, I., Binder, S., Stolba, U., Schmid, K., Glittenberg, C., Brannath, W. and Goll, A. (2005). Optical coherence tomography guided retreatment of photodynamic therapy. *Br J Ophthalmol*, 89: 1184-1187.

- Sabeti-Aschraf, M., Dorotka, R., Goll, A. and Trieb, K. (2005). Extracorporeal shock wave therapy in the treatment of calcific tendinitis of the rotator cuff. *Am J Sport Med*, 33: 1365-1368.

- Windberger, U., Grohmann, K., Goll, A., Plasenzotti, R. and Losert, U. (2005). Fetal and juvenile animal hemorheology. *Clin Hemorheol Microcirc.* 32: 191-197.

D R-Code

This is a R-program for the cross validation procedure to construct a score for future prediction. The program calculates the optimal choice of the FDR ($\hat{\alpha}_{opt}$), the corresponding selection boundary ($\hat{\gamma}_{opt}$), the estimated number of true null hypotheses ($\hat{\pi}_0$), the cross validated $\widehat{AUC}(\hat{\alpha}_{opt})$ as well as the weights of the determined prediction score and the corresponding identification number of the variables included in the prediction score.

Note that with this R-program we search for the optimal estimate of the selection boundary γ instead of the optimal FDR because of the extremely longer runtime needed to search for the corresponding γ values in each training set. For the finally chosen selection boundary the FDR can be estimated with Storey's estimator in the total sample. Note that searching for the optimal FDR and γ asymptotically leads to the same (see Storey et al. (2004)), the similarity of outcome being affirmed by simulations also in our finite case.

Functions:

crossvalsub ... subroutine of crossvalfun: calculation of function CF_{ij} (see Section 4.1.)

fdrestt ... Function to compute the estimate of FDR and the estimate of the number of true null hypotheses π_0

crossvalfun ... Function to compute cross validated results for a given data set

Parameters:

daten ... data set: one column for each patient, one row for each gene/protein

group ... vector containing 0 or 1 identifying each patient either as responder or non-responder

gamma1 ... grid of γ values in which the optimal $\hat{\gamma}_{opt}$ should be searched. Defalt= $seq(0.005, 0.5, 0.005)$

sided ... if sided=1 a one-sided test is performed, if sided=2, a two-sided test will be performed for selection of variables for the prediction score. Defalt=1

known ... if known=0 the variance will be assumed as unknown. If the variance is known, the input is a vector containing the within-group variances for each gene. Defalt=0

lambda ... Parameter λ for Storeys estimate (see Storey (2002)). Defalt=0.5

Output:

A list of 3 Items:

CrossValAUC ... includes the cross validated $\widehat{AUC}(\gamma)$ for each of the investigated γ values in the grid

SelectedHyp ... gives the identification numbers of variables included in the score and the corresponding weights

CrossValResult ... gives the estimate of the optimal choice of the selection boundary, $\hat{\gamma}_{opt}$, the corresponding FDR, $\hat{\alpha}_{opt}$, and the estimated proportion of true null hypotheses (π_0) as well as the cross validated $\widehat{AUC}(\hat{\gamma}_{opt})$ of the estimated selection boundary and the number of variables selected for the score.

R-Code:

```
crossvalsub<-function(parms,daten,rimat,nrimat,tcrit,resp,nresp,n1,n2,known,m,sided)
{
i<-parms[1]
j<-parms[2]
meanr<-rimat[1:m,i]
meannr<-nrimat[1:m,j]
meandiff<-meanr-meannr

if(length(known)==1)
  {
  varr<-rimat[(m+1):(2*m),i]
  varnr<-nrimat[(m+1):(2*m),j]
  ssq<-((n1-2)*varr+(n2-2)*varnr)/(n1+n2-4)    }
else
  {
   ssq<-known
  }

tstat<-meandiff/sqrt(ssq*(1/(n1-1)+1/(n2-1)))
if(sided==2)tstat<-abs(tstat)

rloi<-daten[,resp[i]]
nrloi<-daten[,nresp[j]]
weight<-outer(tstat,tcrit,">")*(meandiff/ssq)
scorer<-rloi%*%weight
scorenr<-nrloi%*%weight
(scorer>scorenr)+(scorer==scorenr)*0.5
}
```

```
fdrestt<-function(tst,tcrit,gamma1,lambda,ilambda,n1,n2)
{
pi0estim<-min(sum(tst<ilambda)/((1-lambda)*length(tst)),1)
c(min(pi0estim*gamma1*length(tst)/(max(sum(tst>tcrit),1)),1),pi0estim)
}

crossvalfun<-function(daten,group,gamma1=seq(0.005,0.5,0.005),sided=1,known=0,lambda=0.5)

m<-nrow(daten)
fact<-ifelse(sided==1,1,1/2)
resp<-c(1:length(group))[group==0]
nresp<-c(1:length(group))[group==1]
n1<-length(resp)
n2<-length(nresp)
allcomb<-cbind(rep(c(1:n1),n2),sort(rep(c(1:n2),n1)))
tsumr<-apply(daten[,resp],1,sum)
tsumnr<-apply(daten[,nresp],1,sum)
rimat<-(tsumr-daten[,resp])/(n1-1)
nrimat<-(tsumnr-daten[,nresp])/(n2-1)

if(sum(known)==0)
  {
  tcrit<-qt(1-gamma1*fact,df=(n1+n2-2))
  ilambda<-qt(lambda,df=(n1+n2-2))
  tqsumr<-apply(daten[,resp]*daten[,resp],1,sum)
  tqsumnr<-apply(daten[,nresp]*daten[,nresp],1,sum)
  rimat<-rbind(rimat,(tqsumr-daten[,resp]*daten[,resp]-rimat*(n1-1)*rimat)/(n1-2))
  nrimat<-rbind(nrimat,(tqsumnr-daten[,nresp]*daten[,nresp]-nrimat*(n2-1)*nrimat)/(n2-2))
  }
else
  {
  tcrit<-qnorm(1-gamma1*fact)
  ilambda<-qnorm(lambda)
  }
}

crossvalresult<-apply(allcomb,1,crossvalsub,daten,rimat,nrimat,tcrit,resp,nresp,n1,n2,known,m,sided)
{
resultj<-apply(crossvalresult,1,sum)
tmax<-tcrit[resultj==max(resultj)]
tmax<-max(tmax)
gammamax<-gamma1[tcrit==tmax]
resultj<-resultj/(n1*n2)
resultj<-rbind(gamma1,resultj)
rownames(resultj)<-c("gamma","jackknife AUC")
meanr<-tsumr/n1
meannr<-tsumnr/n2

if(sum(known)==0)
```

```
  {
  varrt<-(tqsumr-tsumr^2/n1)/(n1-1)
  varnrt<-(tqsumnr-tsumnr^2/n2)/(n2-1)
  ssqt<-((n1-1)*varrt+(n2-1)*varnrt)/(n1+n2-2)
  }
else
  {
  ssqt<-known
  }

meandifft<-meanr-meannr
tstatt<-meandifft/sqrt(ssqt*(1/n1+1/n2))
if(sided==2){tstatt<-abs(tstatt)}
weight<-meandifft/ssqt
sel<-c(1:m)[tstatt>tmax]
weightr<-weight[sel]
result1<-rbind(sel,weightr)
rownames(result1)<-c("Nr. sel. Hyp.","weight")

result2<-c(gammamax,fdrestt(tstatt,tmax,gammamax,ilambda,lambda),max(resultj),length(sel))
names(result2)<-c("opt choice gamma","opt choice FDR","opt choice pi0","cross validated AUC(opt
gamma)","Number of selected Hyp")

erg<-list(CrossValAUC=resultj,SelectedHyp=result1,CrossValResult=result2)
erg
}
```

Examples:

Construction of a random data set:

```
n1<-10
n2<-20
m<-1000
delta<-0.4
fhyp<-10
daten<-matrix(rnorm((n1+n2)*m),ncol=(n1+n2),nrow=m)
daten[1:fhyp,1:n1]<-daten[1:fhyp,1:n1]+delta
```

Identification of groups of responders and non-responders:

```
group<-c(rep(0,n1),rep(1,n2))
```

Example 1: sided=1, known=0: Selection using one-sided tests assuming unknown variance:

```
ex1<-crossvalfun(daten,group,gamma1=seq(0.005,0.6,0.005))
plot(ex1[[1]][1,],ex1[[1]][2,],ylim=c(0,1),xlab=expression(hat(gamma)),
    ylab=expression(widehat(AUC)(hat(gamma))))
ex1
```

Example 2: sided=1, known=0: Selection using one-sided tests assuming unknown variance:

```
ex2<-crossvalfun(daten,group,sided=2,gamma1=seq(0.005,0.6,0.005))
plot(ex2[[1]][1,],ex2[[1]][2,],ylim=c(0,1),xlab=expression(hat(gamma)),
    ylab=expression(widehat(AUC)(hat(gamma))))
ex2
```

Example 3: sided=1, known=rep(1,1000): Selection using one-sided tests assuming known variance 1 for each variable:

```
ex3<-crossvalfun(daten,group,sided=1,known=rep(1,1000),gamma1=seq(0.005,0.6,0.005))
plot(ex3[[1]][1,],ex3[[1]][2,],ylim=c(0,1),xlab=expression(hat(gamma)),
    ylab=expression(widehat(AUC)(hat(gamma))))
ex3
```

Example 4: sided=2, known=rep(1,1000): Selection using two-sided tests assuming known variance 1 for each variable:

```
ex4<-crossvalfun(daten,group,sided=2,known=rep(1,1000),gamma1=seq(0.005,0.6,0.005))
plot(ex4[[1]][1,],ex4[[1]][2,],ylim=c(0,1),xlab=expression(hat(gamma)),
    ylab=expression(widehat(AUC)(hat(gamma))))
ex4
```

Bibliography

Abramovich, F., Benjamini, Y., Donoho, D. and Johnstone, I. (2006). Adapting to unknown sparsity by controlling the false discovery rate.
The Annals of Statistics, 34: 584–653.

Anderson, T. (2003). An introduction to multivariate statistical analysis.
Wiley series in probability and statistics, third edition.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph.
Journal of Mathematical Psychology, 12: 387–415.

Bauer, P. (2008). Adaptive designs: looking for a needle in the haystack - a new challenge in medical research. *Statistics in Medicine,* 27:1565–1580.

Bauer, P., Pötscher, B. and Hackl, P. (1988). Model selection by multiple test procedures. *Statistics,* 19: 39–44.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing.
Journal of the royal statistical society, Series B, 57: 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency.
The Annals of Statistics, 29: 1165–1188.

Callow, M., Dudoit, S., Gong, W., Speed, T. and Rubin, E. (2000). Microarray expression profiling identifies genes with altered expression hdl-deficiont mice.
Genomom Research, 10: 2022–2029.

Dudoit, S., Shaffer, J. and Boldrick, J. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science,* 18: 71–103.

Eaves, I., Wicker, L., Ghandour, G., Lyons, P., Peterson, L., Todd, J. and Glynne, R. (2002). Combining mouse congenic strains and microarray gene expression analyses to study a complex trait: the nod model of type 1 diabetes.
Genome Research, 12: 232–243.

Genovese, C. and Wasserman, L. (2004). A stochastic approach to false discovery control. *The Annals of Statistics,* 32: 1035–1061.

Goll, A. (2008). Inference on a large number of hypotheses based on limited samples - some points to consider. *Doctoral Thesis.*

GOLL, A. and BAUER, P. (2008). Model selection based on the false discovery rate optimizing the area under the receiver operating characteristic curve. submitted.

GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M., BLOOMFIELD, C. and LANDER, E. (1999). Molecular classification of cancer: class discovery and glass prediction by gene expression monitoring. *Science* 286: 531–537.

HAN, A. (1987). Non-parametric analysis of generalized regression model. the maximum rank correlation estimator. *Journal of Econometrics,* 35: 303–316.

HANLEY, J. and MCNEIL, B. (1982). The meaning and use of the area under a receiver operating characteristic curve. *Radiology,* 143: 29–36.

HARREL, F. (2001). Regression modeling strategies. *Springer Series in Statistics.*

HASTIE, T., R., T. and FRIEDMAN, J. (2001). The elements of statistical learning. *Springer Series in Statistics.*

HOMMEL, G. (1988). A stage-wise rejective multiple test procedure based on a modified bonferroni test. *Biometrica* 75: 383–386.

JEFFERY, I., HIGGINS, D. and CULHANE, A. (2006). Comparison and evaluation of methods for generating differentially expressed genes lists from microarray data. *BMC Bioinformatics* 7: 359–375.

LI, L. and HUI, S. (2007). Step-wise variable selection and positive false discovery rate estimate in pharmacogenetics studies. *Journal of Biopharmaceutical Statistics,* 17: 883–902.

MANN, H. and WHITNEY, D. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics,* 18: 50–60.

MILLER, A. (2002). Subset selection in regression. *Chapmann and Hall/CRC,* second edition.

NTZANI, E. and IOANNIDIS, J. (2003). Predictive ability of DNA microarrays for cancer outcomes and correlates: an empirical assessment. *Lancet,* 362: 1439–1444.

PAVLIDIS, P., LI, Q. and NOBLE, W. (2003). The effect of replication on gene expression microarray experiments. *Bioinformatics,* 13: 1620–1627.

PEPE, M. (2003). The statistical evaluation of medical tests for classification and prediction. *Oxford University Press.*

PEPE, M., CAI, T. and LONGTON, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics,* 62: 221–229.

Pepe, M., Janes, H., Longton, G., Leisenring, W. and Newcomp, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic or screening marker. *American Journal of Epidemiology,* 159: 882–890.

Pepe, M. and Thompson, M. (2002). Combining diagnostic test results to increase accuracy. *Biostatistics,* 1: 123–140.

R (2005). R development core team: A language and environment for statistical computing. *R Foundation for Statistical Computing, Vienna, Austria.*

Sachs, L. (1999). Angewandte Statistik: Anwendung statistischer Methoden. *Springer Verlag.*

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the american statistical association,* 88: 486–494.

Sherman, R. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrics,* 93: 123–137.

Simes, R. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrica,* 73: 751–754.

Storey, J. (2002). A direct approach to false discovery rate. *Journal of the royal statistical society, Series B,* 64: 479–498.

Storey, J., Taylor, J. and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the royal statistical society, Series B,* 66: 187–205.

Su, J. and Liu, J. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association,* 88: 1350–1355.

Tian, E., Zhan, F., Walker, R., Rasmussen, E., Ma, Y., Barlogie, B. and Shaughnessy, J. (2003). The role of wnt-signaling antagonist dkk1 in the development of osteolytic lesions in multiple myeloma. *New England Journal of Medicine,* 26: 2483–2494.