False Discovery Rates, Higher Criticism and Related Methods in High–Dimensional Multiple Testing

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)

im Fachgebiet
Mathematik

vorgelegt
von Dipl.-Wirtsch.-Math. Bernd Klaus
geboren am 29. Oktober 1984 in Nordhausen

Die Annahme der Dissertation wurde empfohlen von

1. Professor Dr. Anne–Laure Boulesteix (LMU München)

2. Professor Dr. Bernd Kirstein (Universität Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 09.01.2013 mit dem Gesamtprädikat magna cum laude.

Danksagung

Diese Arbeit entstand im Laufe der letzten drei Jahre während meiner Tätigkeit als wissenschaftlicher Mitarbeiter am Institut für Medizinische Informatik, Statistik und Epidemiologie der Universität Leipzig. Teile der Arbeit wurden im Rahmen des BMBF Forschungsprojektes 0315452A (HaematoSys) gefördert.

Allen, die mir in den letzen 3 Jahren zur Seite standen, möchte ich herzlich danken. Zuallererst gilt mein Dank Prof. Dr. Korbinian Strimmer, der meine Dissertation in herausragender Weise betreute. Gleichermaßen bin ich ihm für manches Gespräch über den Tellerand hinaus zu Dank verpflichtet! Mein Dank gilt ebenso meiner Bürokollegin Verena Zuber, die mich in vielerlei Hinsicht unterstützt hat. Sebastian Gibb als "Teilzeitbürokollegen" danke ich für einige Linux–Tipps und interessante Gespräche.

Mein Dank gilt in besonderer Weise auch Prof. Dr. Markus Löffler und den vielen netten Kollegen vom IMISE, die stets für ein angenehmes Arbeitsklima gesorgt haben.

Je remercie Prof. Christophe Ambroise et Catherine Matias de m'avoir si chaleureusement accueilli au sein du laboratoire "Statistique et Génome" à Évry pendant le printemps de 2012. Les discussions avec les membres du laboratoire étaient vraiment enrichissants et j'en ai beaucoup profité! En plus je remercie Stéphane Robin, Tristan Mary-Huard et Marie-Laure Martin-Magniette de l'AgroParisTech pour des discussions fructueuses sur les méthodes presentées dans le chapitre 4 de la thèse.

Je tiens à adresser mes plus sincères remerciements également à toute la famille Tourret chez laquelle j'avais le grand plaisir de habiter pendant mon séjour à Évry.

Prof. Dr. Anne Laure Boulesteix und Prof. Dr. Bernd Kirstein danke ich für Ihre Bereitschaft die Begutachtung der Arbeit zu übernehmen. Prof. Dr. Kirstein danke zudem ich besonders für seine Unterstützung in allen administrativen Belangen, ohne die eine derat schnelle Durchführung des Promotionsverfahrens nicht möglich gewesen wäre.

Jana Neuhaus und Anett Hänisch danke ich für das Korrekturlesen der Arbeit, sie haben viel zur Vollendung der sprachliche Form beigetragen! Lorraine Garchery danke ich für einige Französischtipps.

Zu guter Letzt möchte ich mich bei meinen Eltern für ihre jahrelange Unterstützung bedanken, ohne die diese Arbeit nicht entstanden wäre.

Leipzig, den 10. Januar 2013 Bernd Klaus

Abstract

The technical advancements in genomics, functional magnetic-resonance and other areas of scientific research seen in the last two decades have led to a burst of interest in multiple testing procedures.

A driving factor for innovations in the field of multiple testing has been the problem of large scale simultaneous testing. There, the goal is to uncover lower–dimensional signals from high–dimensional data. Mathematically speaking, this means that the dimension d is usually in the thousands while the sample size n is relatively small (max. 100 in general, often due to cost constraints) — a characteristic commonly abbreviated as $d \gg n$.

In my thesis I look at several multiple testing problems and corresponding procedures from a false discovery rate (FDR) perspective, a methodology originally introduced in a seminal paper by Benjamini and Hochberg (2005).

FDR analysis starts by fitting a two–component mixture model to the observed test statistics. This mixture consists of a null model density and an alternative component density from which the interesting cases are assumed to be drawn.

In the thesis I proposed a new approach called log–FDR to the estimation of false discovery rates. Specifically, my new approach to truncated maximum likelihood estimation yields accurate null model estimates. This is complemented by constrained maximum likelihood estimation for the alternative density using log–concave density estimation.

A recent competitor to the FDR is the method of "Higher Criticism". It has been and strongly advocated in the context of variable selection in classification which is deeply linked to multiple comparisons. Hence, I also looked at variable selection in class prediction which can be viewed as a special signal identification problem. Both FDR methods and Higher Criticism can be highly useful for signal identification. This is discussed in the context of variable selection in linear discriminant analysis (LDA), a popular classification method.

FDR methods are not only useful for multiple testing situations in the strict sense, they are also applicable to related problems. I looked at several kinds of applications of FDR in linear classification. I present and extend statistical techniques related to effect size estimation using false discovery rates and showed how to use these for variable selection. The resulting fdr–effect method proposed for effect size estimation is shown to work as well as competing approaches while being conceptually simple and computationally

inexpensive.
Additionally, I applied the fdr–effect method to variable selection by minimizing the misclassification rate and showed that it works very well and leads to compact and interpretable feature sets.

Contents

List of Figures

List of Tables

List of Symbols and Abbreviations

d	Dimension / number of test statistics 7
d_0	Number of true null hypotheses 7
BH	Benjamini and Hochberg method 7
FWER	Family wise error rate 8
V	Number of falsely rejected hypotheses, additionally the estimated variance funtion in section 4.2.1 8
R	Family wise error rate 8
cdf	Cumulative distribution function 9
Φ	cdf of the standard normal distribution 9
$\varphi(;\mu,\sigma^2)$	Density of the normal distribution with expectation μ and variance σ ...36
f_0 / F_0	Null model density and corresponding distribution 10
f_A / F_A	Alternative / non–null model density and corresponding distribution .. 10
y	General test statistic 11
η_0	The true proportion of the null statistics 12
fdr	Local false discovery rate 12
Fdr	Tail–area–based false discovery rate, a.k.a. q–value 12
Fndr	Tail–area–based false *non* discovery rate 13
fndr	Local false *non* discovery rate 13
F(N)DR	False (non) discovery rates in a general sense 13
θ	Parameter(s) of the null model 15
HND	Half–normal decay model 31
BUM	Beta–uniform mixture model 29
LDA	Linear disciriminant analysis 48
SDA	Shrinkage disciriminant analysis — an LDA variant 62
PAM, NSC	Nearest Shrunken Centroids — a classification method a.k.a. PAM after its software implementation 48
$\boldsymbol{\omega}^{(k,\text{pool})}$	Feature weight vector in LDA 49
$\boldsymbol{\omega}^{(k,l)}$	Effect size vector in LDA 49
$\boldsymbol{\omega}_{\text{cat}}^{(k,l)}$	Cat–score vector between the classes k and l on the population level 50
MR	Misclassification Rate in LDA 50
$\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}$	fdr–effect size estimation (fdr–effect) 57
HCT	Higher Criticism thresholding 61
S_i	Summary statistic for each feature i in LDA 60
MRT	Misclassification rate (MR) based variable thresholding 61

1 Introduction

The technical advancements in genomics, functional magnetic-resonance and other areas of scientific research seen in the last two decades have lead to a burst of interest in multiple testing procedures [Benjamini, 2010a]. Since these procedures mostly deal with multiple comparisons, the terms "multiple testing procedures" and "multiple comparison procedures" (MCP) are used synonymously in general [Shaffer, 1995]. In 2008, almost 8% of all publications in top statistical journals devoted to statistical methodology were concerned with procedures for multiple comparisons as shown in Fig. **1.1**.

A driving factor for innovations in the field of MCP has been the problem of large scale simultaneous testing. There, the goal is to uncover lower–dimensional signals from high dimensional data. Mathematically speaking, this means that the dimension d is usually in the thousands while the sample size n is relatively small (max. 100 in general, often due to cost constraints) — a characteristic commonly abbreviated as $d \gg n$. Observe that this is in stark contrast to traditional applications of multiple testing in e.g. ANOVA post–hoc tests where usually less than 20 group comparisons are performed [Rüger, 2002]. Today, we face very high dimensional problems, where very often conventional statistical tools no longer work satisfactorily. Nonetheless, the high dimension also has its benefits: It allows to estimate certain model parameters from data.

A case in point for high dimensional multiple comparisons are DNA microarray experiments which have revolutionized genomic research in the last twenty years by allowing the measurement of gene activity for thousands of genes simultaneously [Dudoit et al., 2003]. In a microarray experiment, gene activity, which is commonly referred to as "gene expression", is measured by RNA concentration. An important task arising in these kind of experiments is the identification of differentially expressed genes, that is, genes whose expression levels are associated with a response or covariate of interest.

This variable is very often dichotomous, e.g. indicating for each sample whether a sample originates from normal or cancer tissue. The biological question of differential

expression can be recasted as a multiple testing problem: the *simultaneous* test for each gene of the null hypothesis of no association between the expression and the response. This is usually assessed by performing a (regularized) t–test for each gene. In this way, a ranking of the genes according to their discriminatory power is obtained. Multiple testing methodologies then help to find the "right" cutoff, separating signal from noise.

However, the covariate can also have multiple levels. In this case, the biological question at hand is the development of so–called "molecular signatures" characterizing (for example) subtypes of a disease [Alizadeh et al., 2000]. In this scenario, the covariate represents the different subcases. Here multiple testing can help to identify a (compact) set of genes that differ strongly in expression between the different subtypes. This set then characterizes the different classes on a molecular level. From a statistical point of view, the characterization of subtypes using a compact set of genes is equivalent to the problem of variable selection in classification (= class prediction). Here, multiple testing procedures are also extremely helpful.

In this thesis, I will mostly consider multiple testing problems arising from microarray data. Note however that multiple comparisons are also a very important issue in other areas. Examples include recommender systems and even business analytics: In a thought–provoking article, Polson and Scott [2012] use multiple testing approaches which are conceptually similar to the ones presented in this thesis to filter out successful firms from a plethora of several ten thousand candidates. Perhaps surprisingly, only a small fraction of the companies commonly considered as leading organizations appears on the upper echelons of their final results.

Guideline through the Thesis

In the course of this thesis, I will look at several multiple testing problems and procedures from a false discovery rate (FDR) perspective, a methodology originally introduced in a seminal paper by Benjamini and Hochberg [1995]. The FDR approach aims at controlling the expected number of false positives among all null hypotheses rejected. It is generally much less conservative than the traditionally employed family wise error rate (FWER), the probability of performing at least one false rejection. This makes the FDR highly suitable for high dimensional multiple comparisons, where the control of the FWER usually leads to very strict procedures. Most of my considerations in the thesis will be based on a Bayesian perspective on FDR methods, originally introduced by Efron et al. [2001] as well as Storey [2003], and well summarized by Efron [2008].

A recent competitor to the FDR in the field of MCP is the method of "Higher Criticism" (HC), originally introduced by Tukey [1976] in order to perform a multiple testing correction for the original p–values obtained from original test statistics. It has been strongly advocated by Donoho and Jin [2008, 2009] in the context of variable selection

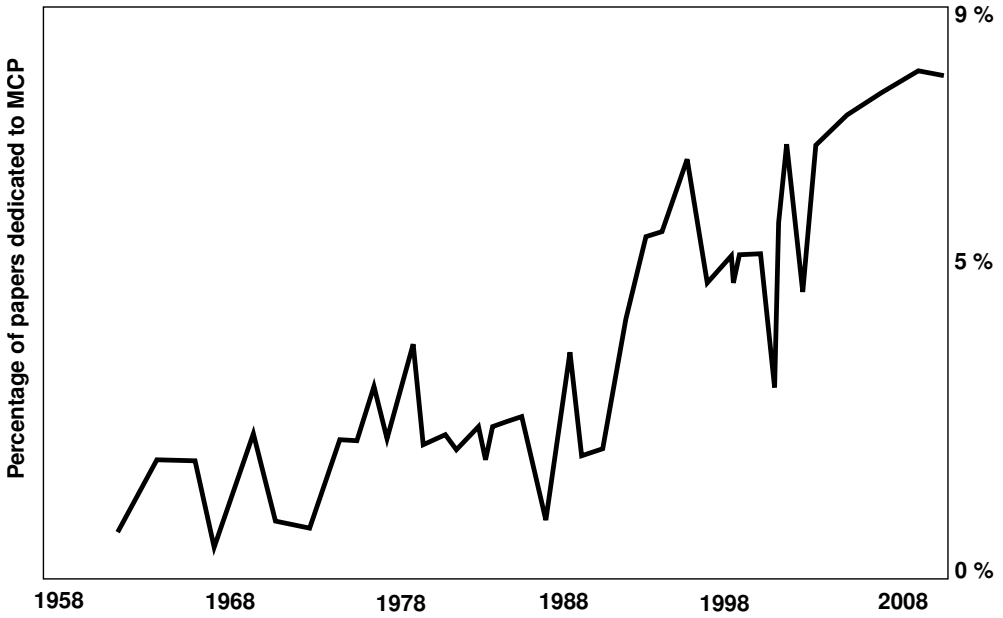


Figure 1.1: Distribution of articles related to multiple comparison procedures (MCP) in top statistical journals, adapted from Benjamini [2010a].

in classification, which is deeply linked to multiple comparisons. Hence, I will also look at variable selection in class prediction, which can be viewed as a special signal identification problem. Both FDR methods and Higher Criticism can be highly useful for signal identification. This will be discussed in the context of variable selection in linear discriminant analysis (LDA), a popular classification method. A specimen of this method called "Nearest Shrunken Centroids" (NSC) has been widely used in the characterization of disease subtypes [Tibshirani et al., 2003].

Chapter Summaries

- An introduction to false discovery rate methods is given in **Chapter 2**. The classical Benjamini–Hochberg procedure is introduced and empirical Bayes analyses of the FDR is treated in some detail.

- **Chapter 3** treats the estimation of false discovery rates in detail. Specifically, truncated maximum likelihood estimation of the null model and non–parametric estimation of the alternative model are discussed. The chapter recapitulates established methods but also presents new techniques for the estimation of the FDR.

- In **Chapter 4**, application areas of the FDR in linear classification are treated. FDR methods are used to estimate effect sizes and to select variables, respectively. Specifically, a simple and heuristic but nonetheless competitive empirical Bayes

approach to the estimation of effect sizes is presented and several thresholding methods for variable selection are compared. The introduced effect size estimates are used to derive a variable selection threshold based on an approximation of a prediction error. This is then compared to thresholding using the method of Higher Criticism (HC) and thresholding based on false discovery rates.

- Signal identification in large–dimensional settings, of which variable selection can be viewed as a special case, is a challenging problem in biostatistics. The method of Higher Criticism (HC) was shown to be an effective means for determining appropriate decision thresholds. In **Chapter 5**, HC is studied from a false discovery rate perspective. It is shown that the HC threshold may be viewed as an approximation to a natural class boundary (CB) in two–class discriminant analysis, which, in turn, is expressible as an FDR threshold.

- In **Chapter 6**, the work is summarized and several important aspects of the preceding chapters are once again highlighted. Furthermore an outlook is given and some directions for future research are indicated.

Contributions of the Thesis

This section gives an overview of the contributions of this thesis.

Contributions of Chapter 3

(i) A new approach to the estimation of the truncation–point necessary for empirical null estimation (section 3.1.2).

(ii) Application of a log–concave density estimator to the estimation of the alternative model, leading to a "smoother" density estimate *without* any additional tuning parameters (part of section 3.2.2).

(iii) A study of FDR estimation using threshold functions (section 3.3).

(iv) A detailed comparison of the variability and accuracy of several FDR estimation methods (section 3.7).

Contributions of Chapter 4

(i) A new conceptually simple and computationally efficient method for effect size estimation in linear discriminant analysis (LDA, section 4.2.1).

(ii) An extended version of missclassification rate based variable selection, allowing *fast* variable selection for *any number of classes* in LDA (section 4.3.2).

(iii) Demonstration of the connection between missclassification rate based variable selection and Higher Criticism thresholding (section 4.4).

Contributions of Chapter 5

(i) A non–technical and accessible introduction to the method of Higher Criticism (HC) on both the sample and the population level (section 5.1).

(ii) A "class boundary" (CB) FDR–based threshold is introduced and compared to the Kolmogorov-Smirnov (KS) and HC thresholds (sections 5.2 and 5.3).

(iii) It is shown that in a so called "rare–weak" setting, if signal identification is possible, the CB and HC thresholds are practicably indistinguishable. Thus HC thresholding is in this case identical to using a simple FDR cutoff (section 5.4).

Publications

Parts of this thesis have already been already published. Section 3.3 is based on the article "**Learning false discovery rates by fitting sigmoidal threshold functions**" [Klaus and Strimmer, 2011].

Chapter 4 is essentially an extended version of the ArXiv preprint "**Effect size estimation and misclassification rate based variable selection in linear discriminant analysis**" [Klaus, 2012]. It additionally contains material from the conference contribution "**Thresholding methods for feature selection in genomics: higher criticism versus false non-discovery rates**" [Klaus and Strimmer, 2010].

The final chapter 5 is based on the article "**Signal identification for rare and weak features: higher criticism or false discovery rates?**" [Klaus and Strimmer, 2012].

Software implementing the approaches presented in this thesis is described in Appendix A. Summaries of all the publications described here can be found in Appendix B.

2 False Discovery Rates (FDR)

2.1 Classical FDR Analysis – the BH Procedure

False discovery rate analysis has its origin in an influential paper by Schweder and Spjøtvoll [1982], see also Benjamini [2010b]. Schweder and Spjøtvoll looked at the classical multiple testing situation of d tested hypotheses of which only a number $d_0 < d$ were true. It is illustrative to look at their method in some detail. Remember that for a collection of continuous null distributions the corresponding p–values are uniform. Let test statistics $s_1, \ldots, s_d$ have continuous null distributions $F_{0,1}, \ldots, F_{0,d}$. Suppose $H_{0,i}$ gets rejected when $s_{0,i}$ is large. Then the corresponding p–values are $p_i = 1 - F_{0,i}(p_i)$. If s_i does not correspond to a true null hypothesis, the corresponding p–value p_i will be very small. Let N_p be the number of p–values that are greater than p. For large p–values which likely will correspond to true null hypotheses it then holds that

$$E(N_p) = d_0(1 - p)\,.$$

Large (probably non–null) p–values will thus be close to a straight line with slope d_0. Accordingly, small p–values (probably null) will deviate from that line. Fig. **2.1** shows such a plot for 200 p–values drawn from the mixture model $0.75N(0,1) + 0.25N(2,1)$, with $d_0 = 150$. Schweder and Spjøtvoll now suggest to reject all p–values that deviate "strongly" from the line. From the inspection of Fig. **2.1** we can infer a cutoff of roughly 175, i.e. we reject the 25 smallest p–values. While being both easy to implement and intuitive, Schweder and Spjøtvoll's method is rather ad–hoc and subjective. Furthermore, the operating characteristics of the procedure are unclear.

Benjamini and Hochberg [1995] — hereafter BH — proposed a more precise procedure to evaluate many p–values simultaneously. Their method aims at controlling the BH false discovery rate (FDR_{BH}), the expected ratio $E(V/R)$ of the number of falsely rejected hypotheses V among all tests R declared significant. If $R = 0, V/R$ is set to 0. The quantity $E(V/R)$ is called BH false discovery rate — FDR_{BH} — here to avoid confusion,

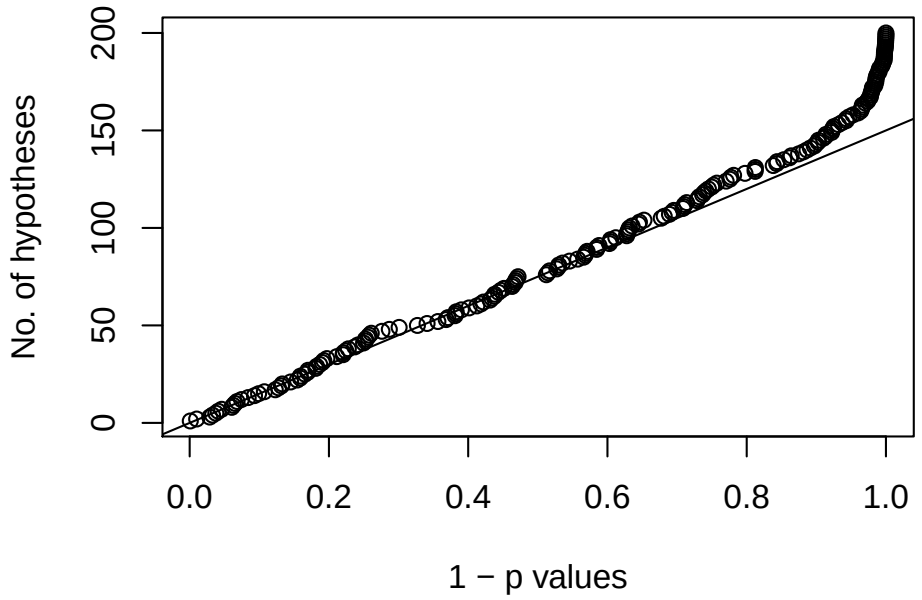


Figure 2.1: The Schweder and Spjøtvoll approach applied for 200 p–values from the mixture model $0.75N(0,1) + 0.25N(2,1)$.

since additional Bayesian false discovery rate definitions will be introduced below. As usually $E(V/R) \leq \text{prob}(V \geq 1) (=$ family wise error rate, FWER) holds, controlling the FDR_{BH} is less conservative than controlling the FWER [Dickhaus, 2008]. BH introduced the following linear step–up procedure to control the FDR_{BH} at a desired level q when the statistics $s_1, \ldots, s_d$ (and therefore also the corresponding p–values $p_1, \ldots, p_d$) are independent:

1. The p–values are ordered so that $p_{(1)} \leq \ldots \leq p_{(d)}$.

2. Each value $p_{(i)}$ is compared with $q\frac{i}{d}$.

3. Setting $k := \max_i p_{(i)} \leq q\frac{i}{d}$ all hypotheses belonging to $p_{(1)}, \ldots, p_{(k)}$ are rejected.

The values $q\frac{i}{d}$ are referred to as Simes' critical values in the literature [Dickhaus, 2008]. Observe that the empirical cumulative density function (ecdf) of the p–values $\hat{F}$ fulfills $\hat{F}(p) = \frac{\text{order}(p_i)}{d}$. Hence the p–values are compared to $q\hat{F}(.)$ rather than to $d_0(1-p)$ as in the Schweder and Spjøtvoll approach. Here $\text{order}(p_i)$ equals one for the smallest and d for the largest p–value, respectively. Interestingly, applied to the simulated p–values of Fig. **2.1** using a FDR_{BH} level of $q = 10\%$ the BH–procedure yields 21 rejections. Both approaches therefore give similar results in this example.

The FDR_{BH} step up procedure can be reformulated into a correction procedure for p–values:

$$p_i^{\text{BH}} = p_i \frac{d}{\text{order}(p_i)}, \quad i = 1, \ldots, d. \tag{2.1}$$

Using a significance level q for these corrected values will control the FDR_{BH} at level q. For comparison, the standard Bonferroni correction [Bonferroni, 1935] is $p_i^{\text{Bf}} = p_i \cdot d$, and hence $p_i \leq p_i^{\text{BH}} \leq p_i^{\text{Bf}}$.

There are a couple of extensions of the BH linear step up procedure, see e.g. Finner et al. [2009]. In this thesis, however, I will mostly discuss empirical Bayes approaches to the estimation of false discovery rates. The focus of the BH and related procedures lies on the control of the false discovery rate. However, when analyzing high dimensional data usually at least several hundred hypotheses tests are performed. This, in fact, allows to estimate rather than only to control the false discovery rate. It will turn out that Bayesian approaches are highly appropriate here. However no "full" Bayesian modeling will be performed. Model parameters are estimated from data rather than inferred from posterior distributions, hence the name for these methodologies is "empirical Bayes". The next section will introduce the key concepts starting with a data example from microarray analysis.

2.2 Empirical Bayes FDR Analysis

A typical scenario encountered in microarray data analysis is a gene–wise comparison of two sample groups. Fig. **2.2** shows a histogram of 2000 two–sample t–scores measuring differential gene expression between cancer (40 samples) and healthy tissue (22 samples) in a colon cancer microarray study [Alon et al., 1999]. For every gene, each sample group is assumed to follow a normal distribution and the null hypothesis is that there is no differential expression between the sample groups. Due to the large number of degrees of freedom, the t–scores can be assumed to be normally distributed under the null hypothesis for this data set. In general, when faced with smaller sample sizes, the two sample t–statistics can easily be transformed to the normal scale. Let n_1 and n_2 be the sizes of group 1 and 2, respectively, and F_{df} the cumulative distribution function (cdf) of a Student's t distribution with df degrees of freedom. Then the t–scores can be transformed to z–scores by the following transformation:

$$z_i := \Phi^{-1}\left[F_{n_1+n_2-2}(t_i)\right], \tag{2.2}$$

where Φ denotes the cdf of the standard normal distribution. Thus we will always assume that our gene–wise comparisons yield z statistics $z_1, \ldots, z_d$. Under the null hypothesis of no differential expression, the statistics should follow a standard normal distribution, i.e. $z_i \sim N(0,1)$.

In microarray experiments, a common and realistic assumption is that for "most" ($\geq 75\%$) genes the null hypothesis is true since it is very unlikely that several thousand genes are involved in the genesis of a disease at the same time. Therefore. the central region of the z–value histogram in Fig. **2.2** should roughly correspond to a standard normal distribution. However, this is clearly not the case. Unfortunately, our theoretical null distribution $N(0,1)$ turns out to be more theoretical than we might have anticipated. Efron [2008] and Benjamini [2008] state a couple of typical reasons for this common phenomenon:

1. The assumption of a normal distribution in each sample group might not be valid for all genes.
2. There might be unobserved covariates, e.g. age, sex and eating habits that affect the expression of particular genes in a study.
3. The samples might not be independent: In microarray experiments so–called "batch effects" are very often observed, i.e. samples coming from the same laboratory are usually correlated with each other.
4. Genes work together in groups so we expect them to be correlated. This can both increase or decrease the variance of the null distribution [Efron, 2007a].
5. Benjamini [2008] indicates that set of statistics used for the final analysis have usually already been preselected from the original set measured. A typical microarray platform can measure up to 20,000 genes. However, typically only a couple of thousand genes are considered in the final analysis. This preselection possibly distorts the center of the histogram, since generally a large number of genes showing no differential activity have already been filtered out before any false discovery rate analysis is performed.

These problems indicate that the theoretical null model is very often not appropriate for high dimensional testing situations. But thanks to the large number of hypothesis tests, we can actually estimate an appropriate null distribution and the FDR. This is best done in the context of a simple two–groups model, which will be introduced next.

2.2.1 The Two–Groups Model and a Generalized Test Statistic

Estimation of FDR typically starts by fitting a two–component mixture model to the observed test statistics [Efron, 2008]. This mixture consists of a null model density f_0 (and corresponding distribution F_0) and an alternative component density f_A (and corresponding distribution F_A) from which the "interesting" or "non–null" (corresponding to the alternative component) cases are assumed to be drawn. In this thesis, I mostly will use a general test statistic $y \geq 0$, with large values of y indicating an "interesting"

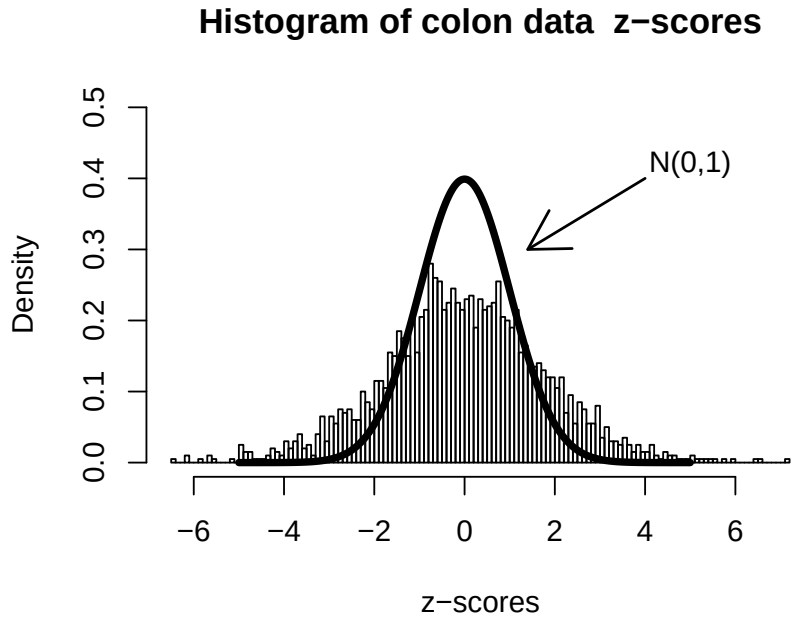


Figure 2.2: A histogram of z–scores obtained from colon cancer data [Alon et al., 1999].

and small values close to zero an "uninteresting" (corresponding to the null component) case. Examples for suitable statistics y include:

- $y = 1 - p$ where p is a p–value,
- $y = |z|$ where z is a normal score,
- $y = |r|$ where r is a correlation, and
- $y = |t|$ where t is a t–score.

See also Strimmer [2008b]. We can write the mixture model in terms of densities as

$$f(y) = \eta_0 f_0(y) + (1 - \eta_0) f_A(y) \tag{2.3}$$

and using distributions as

$$F(y) = \eta_0 F_0(y) + (1 - \eta_0) F_A(y)\,. \tag{2.4}$$

The parameter η_0 is the true proportion of the null statistics. From a given mixture model using Bayes' rule, the so–called local FDR (= fdr) is readily obtained by

$$\begin{aligned} \text{fdr}(y) &= \text{prob}(\text{“null”}|Y = y) \\ &= \eta_0 \frac{f_0(y)}{f(y)} \end{aligned} \tag{2.5}$$

and the tail–area–based FDR (= Fdr), also known as q–value, is defined by

$$\begin{aligned} \text{Fdr}(y) &:= \text{prob}(\text{“null”}|Y \geq y) \\ &= \eta_0 \frac{1 - F_0(y)}{1 - F(y)}. \end{aligned} \tag{2.6}$$

Note that the "Bayesian" Fdr defined in Eq. **2.6** corresponds to estimating the ratio $E(V)/E(R)$, while the BH procedure controls $E(V/R)$. However, these quantities are equivalent under independence assumptions. See Storey [2003], theorem 1 and corollary 1 for details. Specifically, if $\eta_0 < 1$ then the following relations hold approximately:

$$\text{FDR}_{\text{BH}} = E(V/R) = E(V/R|R > 0)\,\text{Pr}(R > 0) \approx E(V)/E(R) = \text{Fdr}.$$

There is another interesting connection between the BH rule Eq. **2.1** and definition Eq. **2.6**. The former can be interpreted as a non–parametric empirical estimator of Fdr:

$$\begin{aligned} \text{Fdr}(1 - p_i) &= \text{Prob}(\text{“null”}|1 - P \geq 1 - p_i) \\ &= \eta_0 \frac{1 - (1 - p_i)}{1 - F(1 - p_i)} = \eta_0 \frac{p_i}{1 - F(1 - p_i)}. \end{aligned}$$

Plugging in the ecdf $\hat{F}(1 - p_i) = 1 - \frac{\text{order}(p_i)}{d}$ as an estimator of $F(1 - p)$ and using the conservative guess $\hat{\eta}_0 = 1$ yields:

$$\widehat{\text{Fdr}}(1 - p_i) = \frac{\hat{\eta}_0 p_i}{1 - \hat{F}(1 - p_i)} = p_i \frac{\hat{\eta}_0 d}{\text{order}(p_i)} = p_i \frac{d}{\text{order}(p_i)}.$$

It follows that using the above estimator and controlling $\widehat{\text{Fdr}}$ at level q is equivalent to the BH–rule Eq. **2.1** for the same level q.

The mixture models of Eq. **2.3** and Eq. **2.4** also allow the definition of the false non discovery rate (FNDR) [Genovese and Wassermann, 2002]. Here, the roles of null and alternative are interchanged. The local false non discovery rate (fndr) is given by:

$$\begin{aligned} \text{fndr}(y) &:= \text{prob}(\text{“alternative”}|Y = y) \\ &= (1 - \eta_0) \frac{f_a(y)}{f(y)} = 1 - \text{fdr}(y) \end{aligned} \tag{2.7}$$

Table 2.1: Definitions of FDR and FNDR on the y scale.

Quantity	Definition
Specificity$(y) =$	prob$(Y \leq y$|"null"$) = F_0(y)$
Power / Sensitivity$(y) =$	prob$(Y \geq y$|"alternative"$) = 1 - F_A(y)$
Fdr$(y) =$	prob$($"null"$\|Y \geq y) = \eta_0 \frac{1-F_0(y)}{1-F(y)}$
Fndr$(y) =$	prob$($"alternative"$\|Y \leq y) = (1-\eta_0) \frac{F_A(y)}{F(y)}$
fdr$(y) =$	prob$($"null"$\|Y = y) = \eta_0 \frac{f_0(y)}{f(y)}$
fndr$(y) =$	prob$($"alternative"$\|Y = y) = (1-\eta_0) \frac{f_a(y)}{f(y)} = 1 - \text{fdr}(y)$

All expressions are defined on the y scale. Fdr and Fndr are based on distributions (capital "F") whereasfdr and fndr are computed from densities (lower case "f").

and the tail area based FNDR (=Fndr), is defined by

$$\begin{aligned} \text{Fndr}(y) &:= \text{prob}(\text{"alternative"}|Y \leq y) \\ &= (1-\eta_0)\frac{F_A(y)}{F(y)}\,. \end{aligned} \tag{2.8}$$

Tab. **2.1** summarizes the definitions given above. In this work, the term F(N)DR will abbreviate false (non) discovery rates in a general sense, including both local and tail area based definitions, while quantities derived from densities will begin with a lower case "f" and correspondingly quantities derived from distributions will begin with a capital "F".

It is instructive to compare the definitions of Fdr and Fndr for a given threshold y with those of sensitivity and specificity — see Tab. **2.1**. The order of conditioning is reversed in the two instances; apart from that, the definitions are very similar. Furthermore, both Fdr–Fndr and sensitivity–specificity can be used as risk measures for a testing procedure. In a conventional testing situation, the threshold y is chosen to maximize both sensitivity and specificity (i.e. typically specificity is fixed and power is maximized). Analogously, in an FDR analysis, one seeks to minimize Fdr and Fndr (e.g; by fixing Fndr and minimizing Fdr). Hence, there is a tradeoff between Fndr and Fdr, just as there is a tradeoff between sensitivity and specificity.

2.2.2 Technical Assumptions and Monotonicity of False Discovery Rates if the Null Model is Known

Most MCPs are based on p–values and hence implicitly assume that a null model is known. In this subsection, I will also make this assumption. The null model will assumed to be known and all calculations will be based on p–values.

Table 2.2: Definitions of FDR and FNDR on the p–value scale using the models of Eq. **2.10** and Eq. **2.11**. Observe that when the null model is known, statistics can always be transformed to p–values, see Eq. **2.9**.

Quantity	Definition
$\text{Specificity}(p) =$	$\text{prob}(P \geq p \vert \text{“null”}) = 1 - F_0(p) = 1 - p$
$\text{Sensitivity}(p) =$	$\text{prob}(P \leq p \vert \text{“alternative”}) = F_A(p)$
$\text{Fdr}(p) =$	$\text{prob}(\text{“null”} \vert P \leq p) = \frac{\eta_0 F_0(p)}{F(p)} = \frac{\eta_0 p}{F(p)}$
$\text{Fndr}(p) =$	$\text{prob}(\text{“alternative”} \vert P \geq p) = (1-\eta_0)\frac{1-F_A(p)}{1-F(p)}$
$\text{fdr}(p) =$	$\text{prob}(\text{“null”} \vert P = p) = \frac{\eta_0 f_0(p)}{f(p)} = \frac{\eta_0}{f(p)}$
$\text{fndr}(p) =$	$\text{prob}(\text{“alternative”} \vert P = p) = \frac{(1-\eta_0) f_A(p)}{f(p)} = 1 - \text{fdr}(p)$

All expressions are defined on the p–value scale. Fdr and Fndr are based on distributions (capital “F”) whereas fdr and fndr are computed from densities (lower case “f”).

Once a (possibly empirical) null model has been specified and f_0 / F_0 is known, we can transform our test statistics into p–values via the transformation

$$p = 1 - F_0(y)\,. \tag{2.9}$$

Since p–values are uniform under the null hypothesis, the null density becomes $f_0(p) = 1$, and the null distribution is $F_0(p) = p$. Note that Eq. **2.9** results in a reformulation of the models Eq. **2.3** and Eq. **2.4** as

$$f(p) = \eta_0 + (1-\eta_0) f_A(p) \tag{2.10}$$

and

$$F(p) = \eta_0 p + (1-\eta_0) F_A(p)\,, \tag{2.11}$$

respectively. Now, small values of p indicate ”interesting“ cases. This does not change the essence of the mixture models introduced but results in slightly different formulae than setting $y = 1 - p$ in Eq. **2.3** and Eq. **2.4**. For convenience, Tab. **2.2** gives Tab. **2.1** on the p–value scale. Assuming the null model as known allows to derive several interesting properties of the FDR estimates.

Firstly, identifiability of the mixture weight (proportion of true null hypotheses) η_0 of the mixture models Eq. **2.10** and Eq. **2.11** is assured if the alternative density is assumed to vanish near 1, i.e. $f_A(p \to 1) = 0$. Then we have from Eq. **2.10** that:

$$f(p \to 1) = \eta_0 + (1-\eta_0) f_A(p \to 1) = \eta_0\,,$$

and hence η_0 can be identified and it follows that $\text{fdr}(1) = 1$. Similarly, we have $\text{Fdr}(1) = \eta_0$.

Secondly, it is desirable to obtain monotone FDR values, since in this way the original ordering of the p–values (and correspondingly the test statistics) remains unchanged, i.e. $\text{FDR}(p_1) \leq \text{FDR}(p_2)$ for $p_1 \leq p_2$ holds. Note that, this is *not* the case for the BH procedure. Therefore concavity constraints are often imposed on the marginal distribution F on the p–value scale. Specifically, if F is concave and sufficiently smooth, such that $f(p)$ is a monotonically decreasing density, $\text{Fdr}(p)$ and $\text{fdr}(p)$ are monotonically increasing with p. Additionally, it holds that $0 \leq \text{fdr}(p) \leq 1$ and $0 \leq \text{Fdr}(p) \leq \eta_0$.

In order to see that $\text{Fdr}(p)$ is monotonically increasing with p, note that since F is concave it holds for any $p \in [0,1]$ that $F(p) \geq pF(1) = p = F_0(p)$. Therefore, $F'(p) = f(p) \geq 1 = f_0(p)$ while $F(0) = 0 = F_0(0)$. Hence, F grows faster than $F_0(p) = p$ (or as fast as $F_0(p)$) while both have the same starting point. Thus, $\frac{p}{F(p)}$ is always smaller than or equal to 1. Since f_A is monotonically decreasing and $f_A(p \to 1) = 0$, the growth speed of F decreases monotonically and approaches that of $F_0(p) = p$ for $p \to 1$. It follows that $F(p) - p \geq 0$ is monotonically decreasing with p. Therefore, $\frac{p}{F(p)}$ is obviously monotonically increasing with p.

In theory, it suffices to impose restrictions on the alternative density only. If we assume sufficient smoothness of $F_A(p)$, then $f_A(p)$ is a monotonically decreasing density, if and only if, $F_A(p)$ is concave. In the literature, usually the concavity assumption is made. This also implies that the alternative and the null model are stochastically ordered with $F_A(p) \geq F_0(p) = p$ for all p [Langaas et al., 2005, Strimmer, 2008b]. However, in practice it is generally easier to impose monotonicity restrictions on the marginal density f.

2.3 Empirical Null Modeling

Efron [2004] has shown that problems with theoretical null model F_0 / f_0 as exemplified in Fig. **2.2** can be elegantly avoided by estimating the parameters of the null model in Eq. **2.3** and Eq. **2.4**. Let these parameters be denoted by θ, i.e. $f_0(y) = f_0(y;\theta)$ and $F_0(y) = F_0(y;\theta)$. In this thesis, θ will denote a variance parameter and centered test statistics will be assumed (unless stated otherwise). Intriguingly, this empirical null modeling is greatly facilitated by high dimensions. Here high–dimensionality is not a curse but a blessing. The truncated maximum likelihood empirical null modeling approach presented in this section follows the presentation of Strimmer [2008b]. It is implemented in the `R`–package [R Development Core Team, 2012] `fdrtool` [Strimmer, 2008a].

For empirical null modeling, suitable estimates of the parameters θ and η_0 are necessary. In other words, the null sub–density $\eta_0 f_0(y;\theta)$ of the two–component model (Eqs. **2.3**

and **2.4**) needs to be fit to the observed test statistics. This is more or less straightforward in fully parametric models such as BUM [Pounds and Morris, 2003] or the mixture models of McLachlan et al. [2006] and Muralidharan [2010]. McLachlan et al. [2006] use a two–component normal mixture including null parameters for location and scale, Muralidharan [2010] uses mixtures of multiple components. Unfortunately, even maximum likelihood estimates of simple mixtures models can exhibit pathological behavior, e.g. an infinite likelihood at some points in the parameter space [Le Cam, 1990, example 1]. There is, however, another, more pronounced problem with mixture models. Their parameters are not necessarily identifiable. This means that several different parameter sets may lead to the same model. The following example illustrates this behavior. Assume that the following normal mixture is given

$$\eta_0 N(\mu_1,1) + (1-\eta_0)N(\mu_2,1)\,.$$

When $\mu_1 = \mu_2 = \mu$, the mixture reduces to the single distribution $N(\mu,1)$. The parameter η_0 has disappeared. Similarly, when $\eta_0 = 1$, the parameter μ_2 disappears. This means that there are subspaces of the parameter space where the family is not identifiable. It indicates in particular that the proportion of the null model is generally not identifiable if the null model is not known beforehand. Additionally, since the null and alternative densities are usually not clearly separated in high dimensional multiple testing problems, classic estimation strategies for mixture models such as the EM–algorithm cannot be directly applied: Estimates have to be penalized or other modifications need to be made [cf. section 3.5.2, Muralidharan, 2010].

Accordingly, it is often preferred to leave f_A unspecified and to estimate it non–parametrically imposing shape constraints such as monotonicity. This will be explained in detail in the next chapter (see section 3.2). Due to the non–parametric estimation of f_A, standard procedures for inferring mixture models cannot be applied. Instead, a truncated maximum likelihood approach is necessary. The data are censored using some threshold y_t, so that only test statistics smaller than the threshold, corresponding to the set $\mathbf{y}_t = \{y_i : y_i < y_t\}$, are retained. The underlying assumption is that for $y_i < y_t$, (nearly) all data points belong to the null part. This is called the "**(strong) zero assumption**" in Turnbull [2007] and Efron [2008]. The truncated null density then can be written as:

$$f_0^t(y;\theta) = \begin{cases} f_0(y;\theta)\,/\,F_0(y_t;\theta) \text{ for } y < y_t, \\ 0 \text{ otherwise.} \end{cases} \tag{2.12}$$

In equation (2.12), $F_0(y_t;\theta)$ plays the role of a normalization factor, insuring that the truncated density integrates to 1. Maximization of the corresponding likelihood function returns $\hat{\theta}$ as well as an estimate of its asymptotic error. Once the null model parameters θ and a suitable cutoff y_t are known, the proportion of null values η_0 can be inferred by

assuming a simple binomial model for $d_t = |\mathbf{y}_t|$, i.e.

$$d_t \sim \text{Binom}\,(d, \eta_0)\ ,$$

which leads to the estimate

$$\hat{\eta}_0 = \min\{1, \frac{d_t}{dF_0(y_t;\hat{\theta})}\} \tag{2.13}$$

plus an associated error. In order to see why this estimate of η_0 makes sense, observe that if the zero assumption indeed holds, it should be possible to separate null and alternative density at least approximately using a truncation point y_t. That is, there is a value y_t for which the alternative distribution function is very small and $F(y_t;\theta) \approx \eta_0 F_0(y_t;\theta)$ holds. Then the equation

$$d_t = F(y_t;\theta) \cdot d \approx \eta_0 F_0(y_t;\theta) \cdot d$$

is valid and Eq. **2.13** is an almost unbiased estimate for η_0. Truncated maximum likelihood is used in the `locfdr` [Efron, 2007b, Turnbull, 2007, section 3.5.1] and the `fdrtool` [Strimmer, 2008a] algorithms. If the test statistics are p–values, the truncated maximum likelihood algorithm is equivalent to the simple cutoff technique used in `qvalue` [Storey and Tibshirani, 2003] and most other p–value based FDR estimation software packages. This cutoff technique works as follows: Let λ be covering the range $[0;1]$, e.g. $\lambda = 0, 0.05, 0.1, \ldots, 0.95$. At first

$$\hat{\eta}_0(\lambda) = \frac{|\{p > \lambda\}|}{d(1-\lambda)} \tag{2.14}$$

is calculated. Observe that Eq. **2.14** is essentially a cutoff dependent version of Eq. **2.13** with $\lambda = 1 - y_t$. Furthermore, if $\{p > \lambda\}$ contained only null statistics, then Eq. **2.14** would give an unbiased estimator of η_0. Since most of the alternative p–values are close to 0, Eq. **2.14** will overestimate η_0 especially for small λ. Having obtained an estimate for each λ value from Eq. **2.14**, there are different ways of furnishing a final estimate of η_0. In Storey and Tibshirani [2003], a spline l is fit through the pairs $(\lambda, \eta_0(\lambda))$. The final estimate of η_0 is then given by $\hat{\eta}_0 = l(1)$. Here λ is 1 and η_0 is estimated for the case of complete truncation, i.e. $y_t = 0$. This will lead to a minimal bias of the estimate: For λ close to 1 in Eq. **2.14** there is little contamination from the alternative model in the set $\{p > \lambda\}$ leading to an almost unbiased estimate of η_0. In Strimmer [2008b], an estimate of η_0 is computed by using the 0.1 empirical quantile of distribution of $\eta_0(\lambda)$ resulting from Eq. **2.14**.

Selection of a Suitable Truncation Point

Estimating the null model parameters θ and the proportion η_0 of true null hypotheses by truncated maximum likelihood requires a suitable truncation point y_t. It has to be small enough to ensure that the zero assumption is met and that there are relatively few observations from the alternative f_A in the set $\mathbf{y}_t$. On the other hand, y_t should not be too small since d_t has to be large enough to allow a reliable estimation of θ and η_0. As a rule of thumb, d_t should be at least 200. There are various ideas on how to compute an optimal cutoff. All of them are closely linked to the overall process of estimation of false discovery rates and hence will be treated in the next chapter.

3 Estimation of FDR

In this chapter, the estimation of false discovery rates will be treated. Specifically, I am going to take a look at several methods of estimating the null model and the marginal density f of the mixture model in Eq. **2.3**. Estimating densities is harder than estimating distributions. Thus, once the density estimation problem is solved, the corresponding distribution functions (Eq. **2.4**) are easily obtained with most of the methods considered in this work. Often, both of them are even estimated simultaneously. I will begin with the treatment of the estimation of the null model. However, it is important to note that the estimation of the null and alternative model can never be completely separated. The techniques presented in the following sections will finally be summarized in a general FDR estimation algorithm in section 3.4.1.

3.1 Estimation of the Null Model

First of all, for estimating the null model parameters θ and the proportion η_0 of true null hypotheses by truncated maximum likelihood (Eq. **2.12**) a suitable truncation point y_t is needed.

3.1.1 The Fndr Approach

Strimmer [2008b] uses a simple procedure that enforces the "zero assumption" by requiring that the tail area based false non discovery rate (Fndr) is minimized. The truncation point y_t is chosen such that $\text{Fndr}(y_t)$ is small. Unfortunately, this leads to the following circular inferential problem: In order to determine a suitable truncation point y_t, the Fndr must be known, yet, to compute Fndr and other FDR quantities, a suitable value for y_t must be specified so that truncated maximum likelihood estimation can be performed. Fortunately, in most situations the location of the truncation point y_t does not need to be known exactly.

The Fndr–strategy employed in the algorithm `fdrtool` [Strimmer, 2008a], implementing the approach of Strimmer [2008b], proceeds in two steps. In the first step, the null model is fit approximately. This is achieved by matching its median $F_0^{-1}(1/2;\theta)$ with that of the observed statistics y_i (Note that the median for the half–normal distribution corresponds to the interquartile range (IQR) of the corresponding normal with mean zero). Subsequently, after converting the test statistics into p–values (via Eq. **2.9**), an estimate of the null proportion is determined by using the 0.1 empirical quantile of the distribution of $\eta_0(\lambda)$ computed according to Eq. **2.14**. This then leads to an approximate Fndr curve from which an optimal y_t is obtained. Finally, truncated maximum likelihood estimation on the basis of this truncation point y_t is used for a refined fit of the null model via Eq. **2.12**. This, in turn, allows to compute FDR quantities of interest after the marginal density has been estimated.

3.1.2 Heuristic Truncation Point Finding by Smoothing

In Fig. **3.1**, the η_0 estimate for the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10)$ $+0.1\text{Unif}(5,10)$, depending on the truncation point y_t, is shown. Unless otherwise noted, the null model parameter θ estimated for a normal null distribution is the standard deviation. In most cases, it is not necessary to estimate its mean since this can easily be estimated by computing the median of the raw test statistics.

It is intuitively clear that η_0 should be monotonically increasing with increasing y_t — at least for large enough y_t, for which the parameter θ of the null density can be reliably estimated. The greater y_t is, the more important will be the number of non–null hypotheses within the set $\mathbf{y_t}$. Therefore, d_t will be significantly larger than $\eta_0 F_0(y_t;\theta) \cdot d$ and hence from Eq. **2.13** it can be seen that η_0 is overestimated. Usually for too small y_t, the variance estimate of the null model is very unstable, and hence also the corresponding η_0 estimate is very variable. However, if the zero assumption is met, there should exist a region where null and alternative distribution are at least approximately separated. In this "stability region", the estimate of η_0 should be roughly constant. One can see from Fig. **3.1** that in our example this is fulfilled for a truncation point between 3 and 5. Indeed all truncation points from this interval yield an $\hat{\eta}_0$ of ≈ 0.8 and a corresponding estimated variance of $\hat{\sigma} \approx 2$. In the "separation" area between null and alternative hypothesis, both slope and curvature will be close to zero. However, without smoothing the $(y_t, \hat{\eta}_0)$ curve, this area is usually hard to identify, and it is impossible to compute the derivative of the $\hat{\eta}_0$ –curve.

In Fig. **3.1**, the curve has been smoothed using a cubic B-spline system with spline breakpoints placed on the 30–99% quantiles of the test statistics in 1% steps. This is done by computing a penalized spline function. The least squares fit is penalized by a term depending on the second derivative:

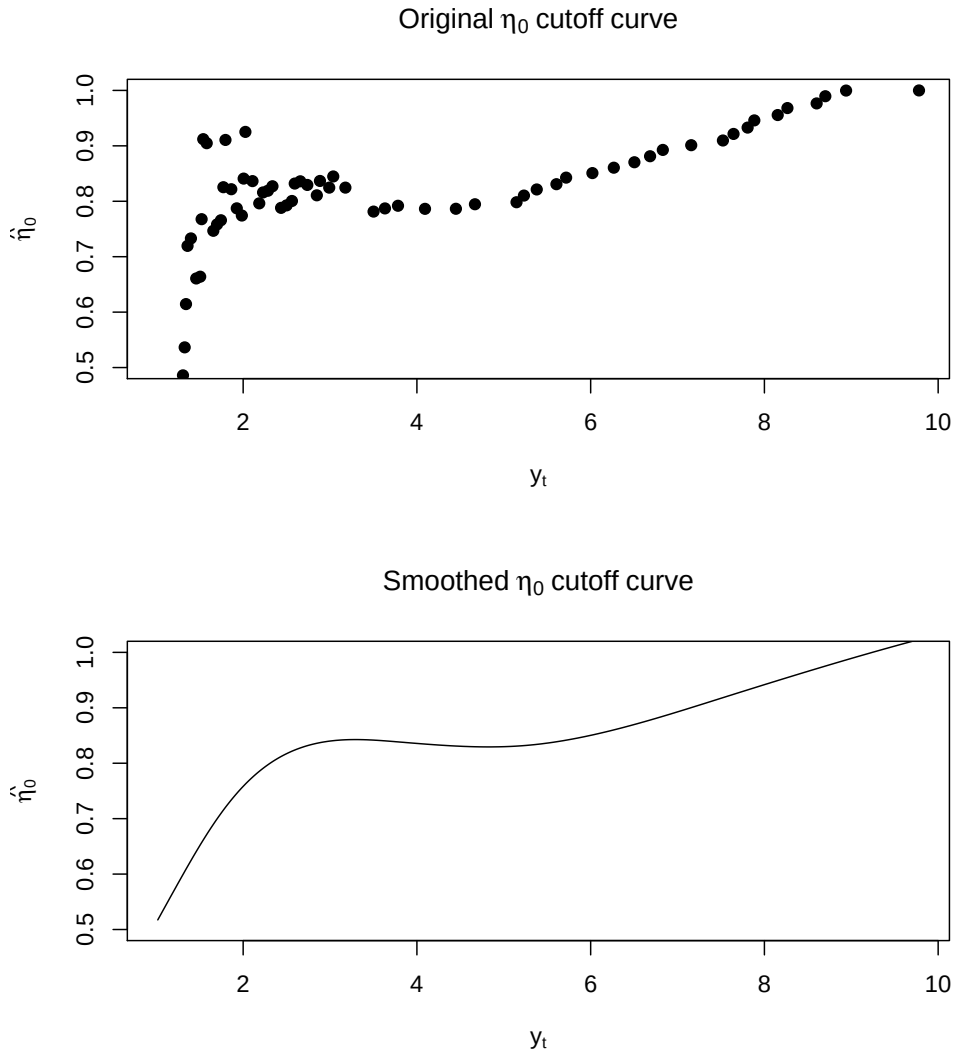


Figure 3.1: Raw and smoothed $\hat{\eta}_0$ curve depending on the truncation point y_t for 200 statistics from the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$.

Let b_m be the cubic B–spline (see e.g. section 3.5 of Ramsay and Silverman [2005]) basis functions with compact support on the sampling interval and a_m the vector of their coefficients. Denote by B the $70 \times K$ matrix of the values of the M basis functions on the 70 sampling points, $\mathbf{b}$ the M–vector of the basis functions b_m and $\mathbf{a}$ the K–vector of the coefficients a_m. Then $\hat{\eta}_0$ is essentially approximated by the B–spline as

$$\hat{\eta}_0(y_t) = \mathbf{a}^T \mathbf{b}(y_t). \tag{3.1}$$

In order to fit this model, the following *penalized* least squared estimate is computed

$$(\hat{\eta}_0 - B\mathbf{a})^T(\hat{\eta}_0 - B\mathbf{a}) + \lambda \int \left[\mathbf{a}^T \frac{\partial \mathbf{b}(y_t)}{\partial^2 y_t}\right]^2 dy_t. \tag{3.2}$$

It uses a penalty based on the second derivative of the basis functions $\mathbf{b}$ and a tuning parameter λ. Especially if the null and the alternative are overlapping, the "stability region" containing a constant $\hat{\eta}_0$ can be very small–sized, so that the tuning parameter λ is best set close to zero, yielding only a mild penalization (Ramsay and Silverman [2005], section 5.2). In conjunction with the densely spaced breakpoints this leads to a smoothed $\hat{\eta}_0$ curve that follows the data points closely. This allows for the identification even of tiny areas of small slope and curvature. Since the second derivative of cubic spline is piecewise constant, the "stability region" is found by identifying local minima of the first derivative of the smoothed η_0 curve. Then their median is used as the truncation point. A detailed description of the algorithm can be found in section 3.4.2.

3.2 Estimation of the Alternative Density

After having introduced truncated maximum estimation of the null model, I now will describe approaches to the estimation of the alternative density and distribution function. Shape constraints such as monotonicity have to be imposed on the overall density (see section 2.2.2). Consequently, it is usually easier to compute the marginal density than to compute the alternative density directly. Taking the estimated null model into account leads to certain constraints on the marginal density derived in the following subsection.

3.2.1 Technical Constraints for the Marginal Density

Before methods for the estimation of the alternative density f_A are considered, I am going to look at some boundaries that the marginal distribution has to observe. They are best described on the p–value scale. That is, the test statistics y_i are transformed in p–values, using the estimated null model and Eq. **2.9**. The key problem can be understood best by going back to Eq. **2.11**, the mixture model for the marginal cdf on the p–value scale. This equation leads to two constraints that any distribution must satisfy in order

to be compatible with the two-component model:

(a) Obviously, the cdf has to fulfill the condition

$$F(p) \geq \eta_0 p, \text{ since } F(p) = \eta_0 p + (1 - \eta_0) F_A(p).$$

(b) Additionally, the inequality

$$1 - F(p) \geq \eta_0 (1 - p)$$

or equivalently

$$F(p) \leq 1 - \eta_0 (1 - p)$$

must be met, due to

$$1 - F(p) = \eta_0 (1 - p) + (1 - \eta_0)(1 - F_A(p)).$$

Conditions (a) and (b) essentially define a "tunnel" that the estimated marginal distribution functions must not leave, where the width of the tunnel depends on the estimated η_0. The second constraint is not particularly obvious but important as Fig. **3.2** illustrates. The model used to generate the data in Fig. **3.2** is $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$ and the cdf is estimated from 200 statistics. The left panel shows a "raw" cdf F. As it can be easily seen, it substantially violates the upper bound. On the right panel, a corrected version is displayed. Both of them were computed using the log–concave estimator that will be discussed in section 3.2.2. Observe that the upper boundary (b) ensures that the *minimum* possible slope equals η_0. This boundary implies that

$$f(p) \leq -\eta_0(-1) = \eta_0$$

is valid.

3.2.2 Estimating the Marginal Density and Distribution with Constrained Maximum Likelihood

After the null model has been fit, p–values can be computed via Eq. **2.9** and the marginal density f can be fit to these p–values. As elucidated in section 2.2.2, we require a monotone marginal density and consequently a concave marginal distribution function in order to obtain monotone FDR values that respect the ordering of the test statistics. A simple way of obtaining monotone density estimates is to impose shape constraints on the maximum likelihood estimation of the density. A straightforward monotone density estimation procedure is provided by the Grenander density estimator [Grenander, 1956].

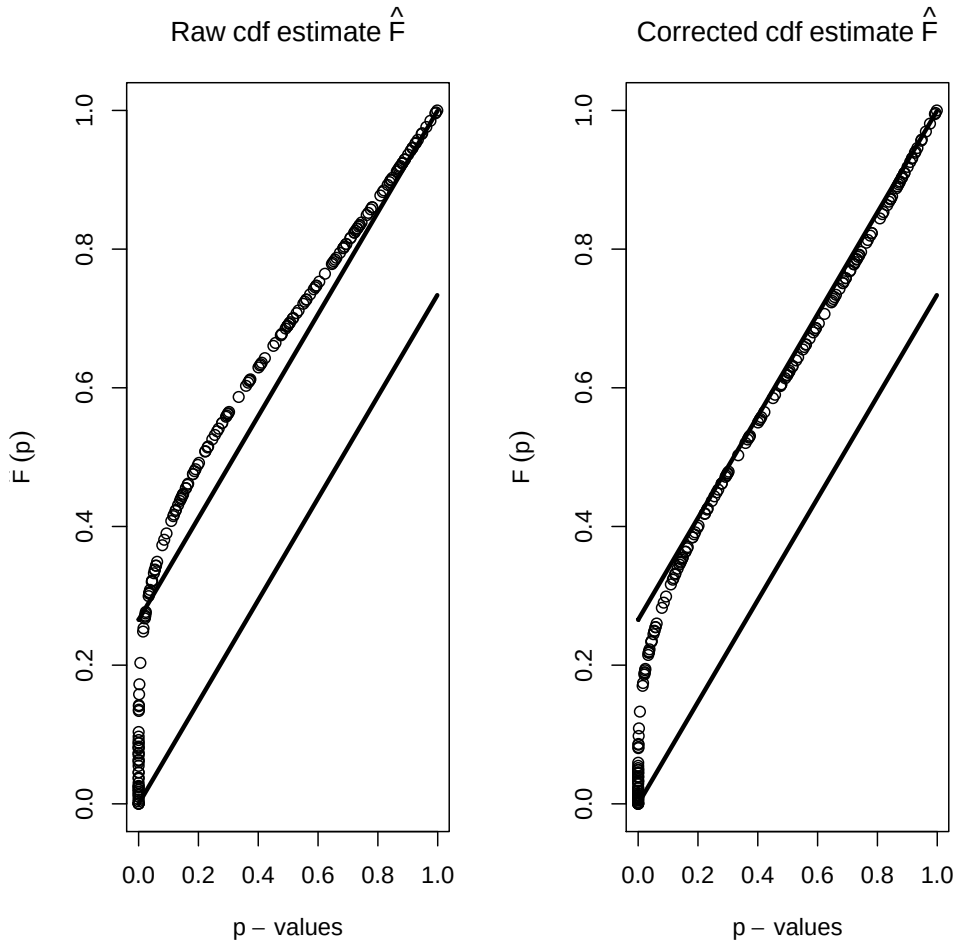


Figure 3.2: Raw and corrected $\hat{F}$ for 200 statistics from the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$.

It not only gives a monotone density but also estimates both f and F simultaneously. For FDR analysis, the Grenander estimator has been first suggested by Langaas et al. [2005] and Broberg [2005]. It is also used by the `fdrtool` algorithm Strimmer [2008a,b]. The Grenander estimator is the non–parametric maximum likelihood estimator of a density under the constraint of monotonicity.

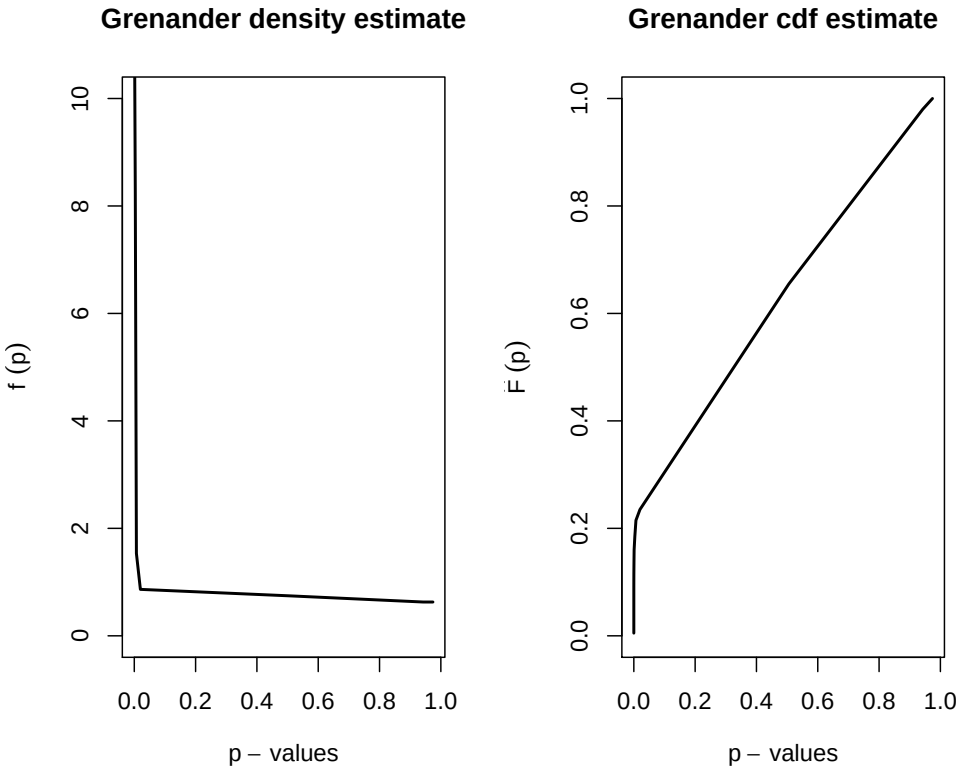


Figure 3.3: Grenander density and distribution function estimate for 200 statistics from the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$.

Fig. **3.3** illustrates the density and distribution function estimate for 200 statistics from the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$ on the p–value scale. Principally, the Grenander density estimator is composed of decreasing piecewise–constant functions. They are equal to the slopes of the least concave majorant (LCM) of the empirical cdf. A detailed derivation of its computation and properties (among them consistency) can be found in the accessible teaching manuscript of Jankowski [2009]. The left part of the figure shows the estimated monotonically decreasing density and the right part the corresponding empirical cumulative distribution. Note that the resulting distribution F is piecewise linear, whereas the density f is piecewise constant. Although the Grenander estimate is easy to obtain and requires no tuning parameters, it is sometimes desirable to have a "smoother" estimate. Since the estimated marginal density f is piecewise constant, using the Grenander estimate obviously leads

to constant fdr–values for certain groups of test statistics. This can be undesirable since these statistics do not necessarily have the same value.

The indicated shortcoming can be elegantly fixed by imposing a log–concavity constraint on the density estimation. More precisely, first the p–values are turned into half–normal (and therefore *positive*) z–values by the transformation

$$z(p) = \Phi^{-1}(1 - p/2)\,. \tag{3.3}$$

Then, these z–values are "mirrored" and the data $\{-z(p)\} \cup \{z(p)\}$ are used as the starting point of a log–concave density estimation. Just as the estimation under monotonicity constraints, log–concave density estimation is fully automatic and does not necessitate any choice of a tuning parameter (penalty, bandwidth etc.). A density f is called log–concave if it may be written as

$$f(x) = \exp[\vartheta(x)], \quad x \in \mathbb{R}\,, \tag{3.4}$$

for some concave function ϑ. Based on data $x_1, \ldots, x_d$ (e.g. our transformed p–values $\{-z(p)\} \cup \{z(p)\}$), this density is estimated by maximizing the log likelihood

$$l(\vartheta | x_i) = d^{-1} \sum_{i=1}^{d} \log f(x_i) = d^{-1} \sum_{i=1}^{d} \vartheta(x_i)$$

over all concave functions ϑ, subject to $\int \exp[\vartheta(x)]dx = 1$. Let $x_{(1)}, \ldots, x_{(d)}$ be the ordered observations. In Dümbgen and Rufibach [2009], it is shown that the maximizer of l is unique, piecewise linear on the interval $[x_{(1)}, x_{(d)}]$ and $\vartheta = -\infty$ elsewhere. Additionally, it has breakpoints (points of changing slope) only at some of the data points $x_{(i)}$ and is consistent. Many parametric families include log–concave densities, e.g. the normal, uniform, Gamma(a,b) (for $a \geq 0$) and Beta(a,b) (for $a,b \geq 1$) distributions. A log–concave density is always unimodal (the reverse is not true) and since the logarithm of the maximum likelihood density is piecewise linear the distribution function is easily computed from the density. A review of log–concave densities can be found in Walther [2009]. Dümbgen and Rufibach [2011] and the references therein discuss optimization algorithms for computing the maximum likelihood estimator and give all important formulae. In this thesis, an iterative convex minorant algorithm will be used to perform the computations. It is implemented in the `R`–package `logcondens` [R Development Core Team, 2012, Dümbgen and Rufibach, 2011].

The log–concave density estimate will usually have its mode at 0. More precisely, the estimate will very often be constant around its maximum, i.e. it will have a "cap" at 0. If the maximum z–value is not at or around 0 the corresponding values will simply be interpolated linearly to preserve monotonicity. This is important since only the estimated density on the negative values $\{-z(p)\}$ will be used to furnish the final estimate. This

corresponds to the "left side" of the density curve depicted in the left panel of Fig. **3.4**. Furthermore, the technical constrains of section 3.2.1 are fulfilled by multiplying $\hat{\vartheta}$ with a factor λ. This gives a density with slightly sharper descent, and hence a more slowly growing distribution function estimate. See Fig. **3.2**. Having obtained the final

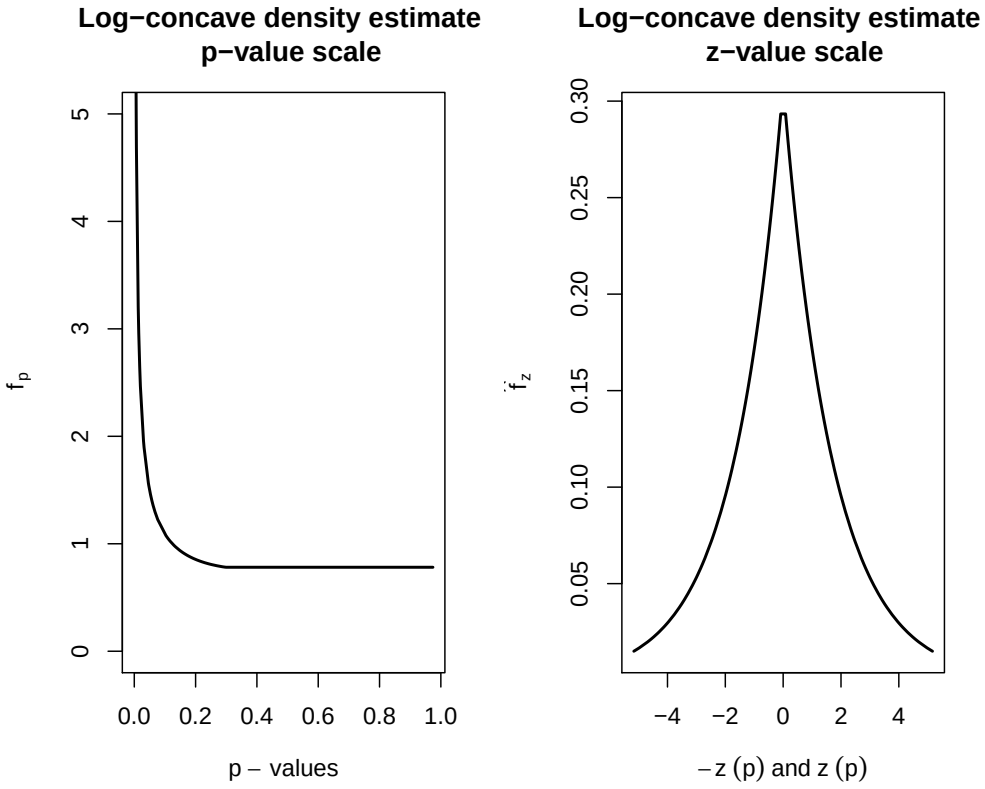


Figure 3.4: Log–concave density on function estimate for 200 statistics from the mixture model $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$ on the z– and p–values scales.

log–concave density estimate $\hat{f}_z$ on the z–value scale for our data $\{-z(p)\} \cup \{z(p)\}$, it can simply be transformed into the p–value scale via the following formula

$$\hat{f}_p(p) = \frac{\hat{f}_z[-z(p)]}{2\varphi(-z(p))}, \tag{3.5}$$

where $[2\varphi(-z(p))]^{-1}$ is the volume element $-\frac{\partial z}{\partial p}$ computed from Eq. **3.3**. Note that the minus sign in Eq. **3.5** is due to the change of the integration direction. Large z–values correspond to small p–values and vice versa, so the inverse transformation of Eq. **3.3** would lead to a p–value density integrating from large to small p–values. In order to reverse this, the minus sign is required.

3.3 FDR Estimation via Threshold Curves

In an interesting comment, Rice and Spiegelhalter [2008] reverse the traditional view on FDR estimation followed in the previous sections. Rather than assuming a two groups model (a null and an alternative model) to derive FDR curves as in section 2.2.1, they proceed by specifying a null model plus a parametric family for the fdr threshold function. The advantage of this procedure is that the alternative model does not need to be specified explicitly, and that at the same time monotonicity of FDR is automatically enforced. In this section, I will investigate the Rice–Spiegelhalter approach by studying two different choices of threshold functions. Two simple models for threshold curves are considered, the beta–uniform mixture (BUM) and the half–normal decay (HND) model.

The approach to FDR estimation presented by Rice and Spiegelhalter [2008] suggests to viewing the null model f_0 plus the fdr curve defined by $\text{fdr}(y)$ as the primary objects, rather than the two densities f_0 and f_A. From Eq. **2.5**, we obtain the marginal distribution as

$$f(y) = \frac{\eta_0 f_0(y)}{\text{fdr}(y)}\,, \tag{3.6}$$

which is here represented as a function of the null model and the fdr. Similarly, the alternative component is given by

$$f_A(y) = \frac{\eta_0}{1-\eta_0}\frac{1-\text{fdr}(y)}{\text{fdr}(y)} f_0(y)\,. \tag{3.7}$$

Furthermore, as $f(y)$ is a density with $\int_0^\infty f(y)dy = 1$ we get the relationship

$$\eta_0 = \left(\int_0^\infty \frac{f_0(y)}{\text{fdr}(y)} dy\right)^{-1}. \tag{3.8}$$

As a consequence, specifying $f_0(y)$ together with $\text{fdr}(y)$ is equivalent to the standard two-component formulation, but with η_0 and $f_A(y)$ viewed as derived rather than primary quantities. Eq. **3.6** also plays an important role in the general algorithm for FDR estimation (cf. section 3.4.1; step G).

3.3.1 Models for fdr Threshold Curves

In this section I study the estimation of FDR using two continuous variants of threshold curves for $\text{fdr}(y)$. Specifically, the half–normal decay (HND) model by Rice and Spiegelhalter [2008] and the beta–uniform mixture (BUM) model of Pounds and Morris [2003] are considered. There are two natural properties for such curves. First, the function should be monotonically decreasing, so that the FDR values lead to the same ranking as the raw statistics y (cf. section 2.2.2). Second, on a $|z|$–score scale ($y = |z|$), the shape of

the curve should be sigmoidal ranging from $\text{fdr}(0) = 1$ onwards to $\text{fdr}(y \to \infty) = 0$. The beta–uniform mixture (BUM) and the half–normal decay (HND) model, as well as their generalizations, satisfy these criteria.

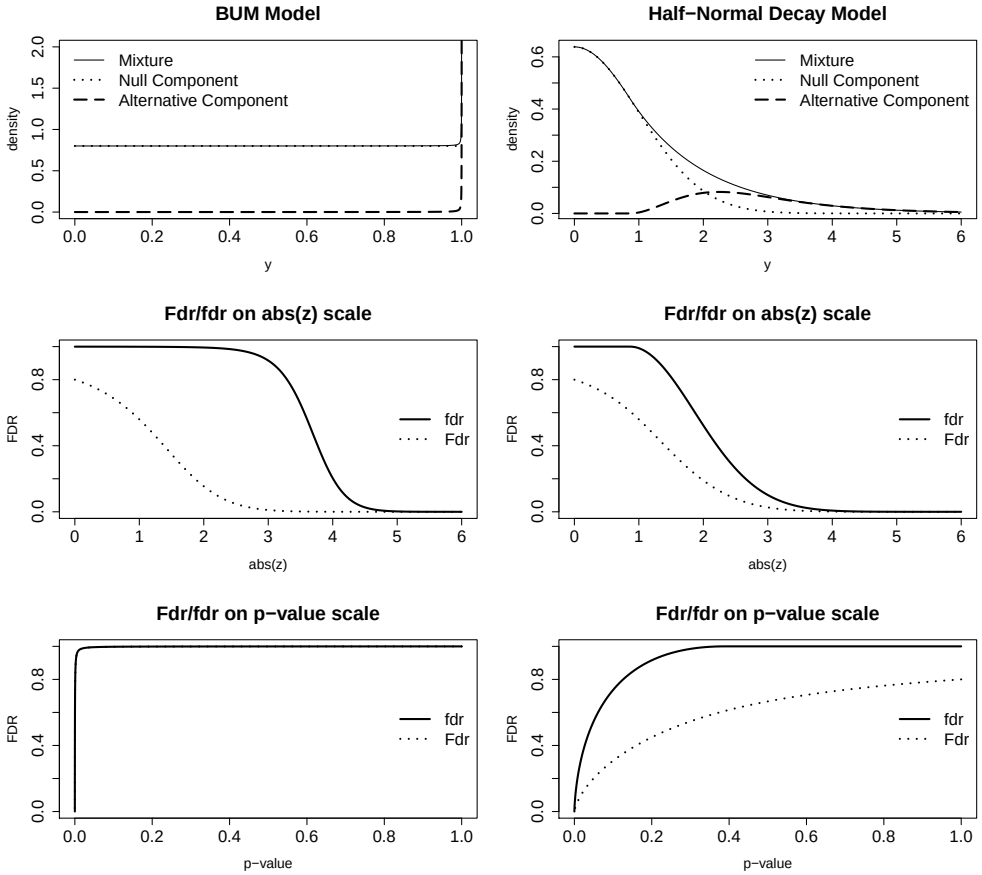


Figure 3.5: Examples of the BUM and HND model for $\eta_0 = 0.8$. The first row shows the corresponding joint, null and alternative densities. The second row displays the fdr– and Fdr–values on the standard normal z–score scale. The third row shows fdr and Fdr values on the p–value scale.

3.3.2 Beta–Uniform Mixture (BUM) Model

The BUM model was proposed in the context of FDR estimation from p–values [Pounds and Morris, 2003]. It is based on a random variable $Y \in [0,1]$ with uniform distribution as null model. The null density is therefore

$$f_0(y) = 1$$

and the corresponding distribution

$$F_0(y) = y\,.$$

The BUM fdr function is given as a one parameter family

$$\mathrm{fdr}^{\mathrm{BUM}}(y|s) = \frac{s}{s + a(1-s)(1-y)^{a-1}}\,.$$

Note that a is not a parameter but a small constant so that approximately $\mathrm{fdr}^{\mathrm{BUM}}(0|s) \approx 1$ ($a = 0.001$ is used throughout). From Eq. **3.8** we find the identity

$$\eta_0 = s\,,$$

which greatly facilitates the interpretation of the parameter s. The marginal density in the BUM model is therefore (Eq. **3.6**)

$$f(y) = \eta_0 + a(1-\eta_0)(1-y)^{a-1}\,.$$

Similarly, the alternative density is

$$f_A(y) = a(1-y)^{a-1}$$

and the alternative distribution

$$F_A(y) = 1 - (1-y)^a\,.$$

The resulting marginal distribution is

$$F(y) = \eta_0 y + (1-\eta_0)(1-(1-y)^a)\,,$$

which leads with Eq. **2.6** to the following expression for the q-value

$$\mathrm{Fdr}(y) = \frac{\eta_0}{\eta_0 + (1-\eta_0)(1-y)^{a-1}}\,,$$

which has $\mathrm{Fdr}(0) = \eta_0$ as required.

The BUM model can also be trivially reformulated using p–values ($y(p) = 1 - p$). Alternatively, as null statistic, one can also use standard normal z–scores with $y(z) = 2\Phi(|z|) - 1$. Observe that this is the inverse transformation of Eq. **3.3**. The Fdr and fdr curves are invariant against reparameterization, i.e. $\mathrm{Fdr}(z) = \mathrm{Fdr}(y(z))$ and $\mathrm{fdr}(z) = \mathrm{fdr}(y(z))$. The marginal density is computed as $f(z) = \eta_0 f_0(z)/\mathrm{fdr}(y(z))$ and thus requires as an additional factor the volume element (which is hidden in the transformation from $f_0(y)$ to $f_0(z)$). In Fig. **3.5**, the BUM model and the associated Fdr and fdr values are shown for $\eta_0 = 0.8$, both on a p–value scale and on a standard normal z–score scale.

3.3.3 Half–Normal decay (HND) model

The half–normal decay model was first described by Rice and Spiegelhalter [2008]. Its starting point is the random variable Y drawn from standard half–normal distribution. Thus, the observations $y \in [0, \infty]$ have the null density

$$f_0(y) = \sqrt{\frac{2}{\pi}} e^{-y^2/2}$$

and corresponding distribution function

$$F_0(y) = 2\Phi(y) - 1\,.$$

The fdr curve is given by a one parameter family

$$\text{fdr}^{\text{HND}}(y|s) = \begin{cases} 1 & \text{, for } y \leq s \\ e^{-(y-s)^2/2} & \text{, for } y > s\,. \end{cases}$$

The parameter s has a natural interpretation as cut-off threshold below which there are no "interesting" cases. This specification of null model and fdr curve results in

$$\eta_0 = \left(2\phi(s) - 1 + \sqrt{\frac{2}{\pi}} e^{-s^2/2 - \log s}\right)^{-1}.$$

This equation is invertible, hence the parameter s has a one-to-one correspondence to the proportion of the null features η_0. In the HND model, the marginal density is

$$f(y) = \begin{cases} \eta_0 \sqrt{\frac{2}{\pi}} e^{-y^2/2} & \text{, for } y \leq s \\ \eta_0 \sqrt{\frac{2}{\pi}} e^{s^2/2 - ys} & \text{, for } y > s \end{cases}$$

and the alternative density

$$f_A(y) = \begin{cases} 0 & \text{, for } y \leq s \\ \frac{\eta_0}{1-\eta_0} \sqrt{\frac{2}{\pi}} (e^{s^2/2 - ys} - e^{-y^2/2}) & \text{, for } y > s\,. \end{cases}$$

Finally, the marginal distribution function is

$$F(y) = \begin{cases} \eta_0 (2\phi(y) - 1) & \text{, for } y \leq s \\ \eta_0 \left(2\phi(s) - 1 + \sqrt{\frac{2}{\pi}} e^{s^2/2 - \log s} (e^{s^2} - e^{-sy})\right) & \text{, for } y > s, \end{cases}$$

which, together with $F_0(y)$, allows to compute the tail-area-based Fdr via Eq. **2.6**.

The HND model can also be expressed in terms of p–values, using the transformation $y = \varphi^{-1}(1 - p/2)$, cf. Eq. **3.3**. In Fig. **3.5**, the HND model for $\eta_0 = 0.8$ (or equivalently $s = 0.862$) is shown and contrasted with the notably different BUM model.

3.3.4 Generalizations and Problem of Confounding

The BUM and HND fdr threshold functions are one parameter families indexed by the parameter s, which in both models has a one–to–one mapping onto the true proportion of null hypotheses η_0. In order to allow for more flexibility, it is useful to introduce additional parameters, either in the null density $f_0(y)$ for empirical null modeling, or in the fdr function $\mathrm{fdr}(y)$. That is, no truncated maximum likelihood estimation is required for empirical null modeling here a priori. For example, both the BUM and HND model can be employed with an additional scale parameter σ in the null model. Specifically, it is assumed that the null density is a normal distribution $N(0, \sigma^2)$ with mean zero and variance σ^2 so that for the HND model $y = |z/\sigma|$ and for BUM $y = 2\varphi(|z/\sigma|) - 1$, where z is the observed test statistic. Additionally, if the alternative density is not flexible enough, this may be fixed by introducing extra parameters into the fdr curve. However, in generalizing null models and fdr functions, particular care is necessary because of potential confounding of parameters, especially if the null model and the fdr threshold function are extended simultaneously.

For example, the fdr curve of the standard HND model has an inflection point at $y_0 = z_0 = s + 1$ with fdr value $e^{-1/2} \approx 0.6$ and slope $-e^{-1/2} \approx -0.6$. The extended HND model with additional scale parameter σ in the null model leads to an fdr curve with inflection point $z_0 = \sigma(s + 1)$ with a corresponding fdr value of $e^{-1/2} \approx 0.6$ and slope $-e^{-1/2}/\sigma \approx -0.6/\sigma$. Thus, the scale parameter of the null model determines directly the slope of the fdr curve at its inflection point, which implies that scale and slope parameters are confounded.

3.4 Workflow of an FDR Estimation Algorithm

In this section, the results of the previous sections will be summarized and embedded in a general workflow of an FDR estimation algorithm. The algorithm includes empirical null modeling and simultaneous estimation of fdr and Fdr. For general test statistics y_i, it can be put together as follows:

3.4.1 The General Algorithm

(A) Determine a suitable truncation point y_t. Possible options are:

(a1) via Fndr optimization as explained in section 3.1.1

(a2) via smoothing as explained in section 3.1.2.

(B) Estimate the null model and its parameters via truncated maximum likelihood, yielding $\hat{\eta}_0$ and $\hat{\theta}$.

(C) Convert test statistics into p–values via Eq. **2.9**: $p_i = 1 - F_0(y|\hat{\theta})$.

(D) Estimate the marginal density $\hat{f}_p(p)$ and cdf $\hat{F}_p(p)$ on the p–value scale using the methods discussed in section 3.2.2:

(d1) the modified Grenander estimator

(d2) the modified log–concave estimator.

Observe that this requires $\hat{\eta}_0$.

(E) Compute estimates of Fdr and fdr values based on p–values:

$$\widehat{\text{fdr}}_p(p) = \frac{\hat{\eta}_0}{\hat{f}_p(p)},$$
$$\widehat{\text{Fdr}}_p(p) = \frac{\hat{\eta}_0 p}{\hat{F}_p(p)}.$$

(F) Compute estimated Fdr and fdr values as a function of the original test statistics y:

$$\widehat{\text{fdr}}(y) = \widehat{\text{fdr}}_p(1 - \hat{F}_0(y)),$$
$$\widehat{\text{Fdr}}(y) = \widehat{\text{Fdr}}_p(1 - \hat{F}_0(y)).$$

(G) Compute cdf and marginal density on the y-scale:

$$\hat{f}(y) = \hat{\eta}_0 \frac{\hat{f}_0(y)}{\widehat{\text{fdr}}(y)},$$
$$\hat{F}(y) = 1 - \hat{\eta}_0 \frac{1 - \hat{F}_0(y)}{\widehat{\text{Fdr}}(y)}.$$

Note that this transformation is directly derived from the definition of fdr and Fdr in **Eqs.** 2.5 and 2.6.

(H) Estimate alternative sub-density:

$$\hat{F}_A(y) = \frac{\hat{F}(y) - \hat{\eta}_0 \hat{F}_0(y)}{1 - \hat{\eta}_0},$$
$$\hat{f}_A(y) = \frac{\hat{f}(y) - \hat{\eta}_0 \hat{f}_0(y)}{1 - \hat{\eta}_0}.$$

Alternatively, steps (C)–(F) can be replaced by the threshold curve methodology presented in 3.3, where fdr curves are directly computed. In step (C), the estimated null

parameters can be used to obtain half normal statistics for the HND model via Eq. **3.3**. Additionally, the null model step (B) can be performed by directly optimizing a threshold curve model with additional parameters as discussed in section 3.3.4. However, empirical null modeling does not work very well for threshold curve models. Especially the BUM model gives disappointing results, see section 3.6 for details. The newly proposed FDR estimation via heuristic truncation point finding and constrained log–concave density estimation (using Options (a2) and (d2) above) will be termed the *log–FDR* approach henceforth. A more detailed description of the technical details of a FDR estimation algorithm implementing the log–FDR approach is given in the next section.

3.4.2 Some Implementation Details of a log–FDR Algorithm

Step (a2) of the general algorithm is performed by computing the null model using truncated maximum likelihood (Eq. **2.12**) for truncation points y_t varying from the 30% to the 99% quantile of the data in 1% steps. Thereafter, models with very large or very small η_0 values are removed. Specifically, all truncation points yielding an estimated η_0 smaller than the 10% quantile or greater than the 90% quantile of all 70 estimated η_0s are removed. If this filters out to many potential truncation points (> 60), this usually means that η_0 is in fact 1, i.e. there is no alternative present. In this case the filtering step is skipped. A smooth η_0 curve as depicted in Fig. **3.1** is obtained by computing a penalized B–spline of order 4 (see Eq. **3.2**) with a penalty parameter of $\lambda = 0.01$ and all the potential truncation points as breakpoints. In order to achieve this, the `R`–package [R Development Core Team, 2012] `fda` is used.

Subsequently, local minima of the smoothed η_0 curve are found using the function `turnpoints` from the `R`–package `pastecs` [R Development Core Team, 2012] implementing the method of finding "significant" turning points of Ibanez [1982]. The default p–value cutoff of 0.05 is used. Finally, the median of the identified minima is used as a truncation point. In step (d2) the log–concave density estimation is performed with the `R`–package [R Development Core Team, 2012] `logcondens` on the mirrored half–normal z–values obtained from the p–values computed in step (C). Then the preliminary estimate $\hat{\vartheta}$ is multiplied with a factor λ so that the Euclidean norm of the deviations from the "tunnel" constraints presented in section 3.2.1 is less than 0.05. An illustration of a raw and a corrected cdf is given in Fig. **3.2**.

3.5 Other State of the Art Estimation Algorithms

In the upcoming section 3.6 the approaches summarized in the general algorithm of section 3.4.1 will be compared to the two current FDR estimation algorithms locfdr and MixFdr. They are described in some detail in this section.

3.5.1 fdr Estimation with locfdr

The algorithm locfdr as implemented in an R–package [R Development Core Team, 2012] of the same name was first described in Efron [2004] and summarized in Efron [2008]. A more thorough discussion is found in Efron [2007b]. Efron's algorithm is based on z–values and hence starts with assuming a (not necessarily standard) normal null

$$f_0(z) \sim N(\mu_0, \sigma^2);$$

locfdr then uses two different approaches to the estimation of the null parameters μ_0, σ^2 and η_0. When using "**Central Matching**", (CM) the logarithm of the estimated marginal density $\log \hat{f}(z)$ is quadratically approximated using a Taylor series around $z = 0$. In principle, $\hat{f}(z)$ itself is estimated using an exponential regression with seven parameters

$$f_{\boldsymbol{\beta}}(z) = c_{\boldsymbol{\beta}} \exp\{\sum_{j=1}^{7} \beta_j z^j\}.$$

In the actual implementation, a natural spline basis with 7 degrees is used as a default instead of the polynomial, but this does not alter the general idea. Therefore, the attention is restricted to the polynomial case here. For the quadratic approximation only, the first three β–coefficients are used. The null subdensity is assumed to be given by this approximation:

$$log(\hat{\eta}_0 \hat{f}_0(z)) = \beta_0 + \beta_1 z + \beta_2 z^2.$$

Since $f_0(z) \sim N(\mu_0, \sigma^2)$, the parameters μ_0, σ^2 and η_0 are easily computed from the above equation. Since with Central Matching the zero assumption is enforced by looking at an approximation around $z = 0$, it will usually work best for large $\eta_0 > 0.90$. Additionally, it tends to overestimate η_0 [Efron, 2007b].

The second and newer method used in the locfdr algorithm is **"MLE fitting"**. It is based on a truncated null model and conceptually identical to the truncated maximum likelihood estimation of the null model parameters introduced in section 2.3. The truncation point z_t, which is necessary in order to construct the set $\mathbf{z}_t = \{z_i : |z_i| < z_t\}$ of z–values used to compute the null model parameters, differs among the versions of the locfdr R–packages. Possible choices of z_t are displayed in Tab. **3.1**. It can be seen that *no adaptive* choice of the threshold is performed. The threshold used in the current version of locfdr is derived from the mixture model $0.9N(0,1) + 0.05N(-3,1) + 0.05N(-3,1)$ on a theoretical basis [Turnbull, 2007]. Nonetheless, it works rather well in most cases. According to Efron [2007b], MLE fitting generally gives more stable parameter estimates than CM. Therefore, it is the default method used in locfdr. Due to its non–parametric spline estimate of the marginal density $f(z)$, locfdr only computes local FDR values.

Table 3.1: Various choices of normal truncation points implemented in locfdr according to Strimmer [2008b].

Version	*Released*	*Truncation Point*	*Reference*
1.1-1	July 2006	$z_t = 2$	
1.1-3	December 2006	$z_t = \hat{\mu} + b\hat{\sigma}$ with $b = 3.55 - 0.44\log_{10}(d)$, $\hat{\mu} = \text{median}(z_i)$ and $\hat{\sigma} = \text{IQR}(z_i)/1.349$.	
1.1-6	November 2007	as in version 1.1-3, but with $b = \max(1, 4.3d^{-0.112966})$	Turnbull [2007]
1.1-7	February 2011	as in version 1.1-6	

Note that $\hat{\mu}$ and $\hat{\sigma}$ are robust location and scale estimates, respectively.

3.5.2 FDR Estimation with MixFdr

The MixFdr algorithm introduced by Muralidharan [2010] estimates both tail–area based FDR and fdr. It is inspired by Efron's effect size model of section 4.2.1. The following hierarchical model is assumed

$$\delta \sim g(\delta) = \sum_{j=0}^{J-1} \pi_j \varphi(\delta; \mu_j, \sigma_j^2),$$
$$z|\delta \sim N(\delta, 1).$$

Here φ denotes the normal distribution density, J is the total number of mixture components and π_j are the mixture proportions with π_0 corresponding to η_0. The assumed model leads to the following marginal density f:

$$f(z) = \sum_{j=0}^{J-1} \pi_j \varphi(z; \mu_j, \sigma_j^2 + 1)\,.$$

If μ_0 and σ_0 are set to 0, a theoretical null is imposed. The false discovery rates are then readily obtained by the formulae

$$\text{Fdr} = \frac{\pi_0 \left(1 - \Phi(z; \mu_0, \sigma_0^2 + 1) + \Phi(-z; \mu_0, \sigma_0^2 + 1)\right)}{1 - F(z) + F(-z)},$$

and

$$\text{fdr} = \frac{\pi_0 \varphi(z; \mu_0, \sigma_0^2 + 1)}{f(z)},$$

respectively. The mixture model is fit using a penalized EM–algorithm. Thereby, the null proportion π_0 is increased by a constant P in the maximization step of the algorithm.

The penalization can be determined by a bootstrap algorithm, its standard value is $P = d/5$. Muralidharan [2010] suggests using $J = 3$ components in general, one for the null model, one for the negative and one for the positive effects.

3.6 Data Analysis and Simulation Studies

In this section, results from the analysis of synthetic data will be shown. Both an overlapping as well as a well separated scenario are considered. In the first case, the null and alternative group are relatively easy to separate. In the second case they are closer to each other, which makes FDR estimation much harder. Muralidharan [2010] notes that truncated maximum likelihood estimation of the null model (via Eq. **2.12**) and constrained maximum likelihood estimation of the marginal density can lead to highly variable fdr estimates. This claim will be investigated using test statistics distributed according to the HND model. In this situation, a generalized HND model including a variance parameter should perform best, giving a "gold-standard" in terms of accuracy and variability of the fdr estimates. Finally, the simulated data analyses will be followed by a reanalysis of four real world data sets from Rice and Spiegelhalter [2008].

3.6.1 Setup of the Overlapping and Well Separated Scenarios

For the data generation of the overlapping and well separated scenarios, I followed the simulation setup for z scores described in Strimmer [2008b]:

- **Well Separated Scenario**
  Data $z_1, \ldots, z_{200}$ were drawn from a mixture of the normal distribution $N(\mu = 0, \sigma^2 = 4)$ with the symmetric uniform alternatives Unif$(-10, -5)$ and Unif$(5, 10)$ and a null proportion of $\eta_0 = 0.8$, i.e. $0.8N(0,4) + 0.1\text{Unif}(-5,10) + 0.1\text{Unif}(5,10)$.
- The sampling was repeated $B = 1000$ times.

Observe that the alternative density of this model does not match the implied alternative density f_A of neither the BUM nor the HND parametrizations. Thus, with this simulation setup it can be investigated how the fdr threshold models perform under misspecification. The data for an overlapping alternative and the null model are:

- **Overlapping Scenario**
  Setup as above, but with Unif$(-10, -2)$ and Unif$(2, 10)$ as alternative distribution, i.e. $0.8N(0,4) + 0.1\text{Unif}(-2,10) + 0.1\text{Unif}(2,10)$.

This scenario leads to a marginal density that is similar in shape to the HND model. In the subsequent step of comparison of resulting FDR values and model parameters, two

different strategies for fitting the parameters for the threshold models BUM and HND are employed.

1. **External estimation**: The parameters of the fdr threshold models σ and η_0 are estimated using heuristic truncation point finding by smoothing as in log–FDR (sections 3.1.2 and 3.4.2). They are then plugged into the corresponding equations of the BUM and HND models. This allows to assess the influence of null model estimation on the threshold models. Furthermore, it allows for a direct comparison with results obtained using the log–FDR approach.

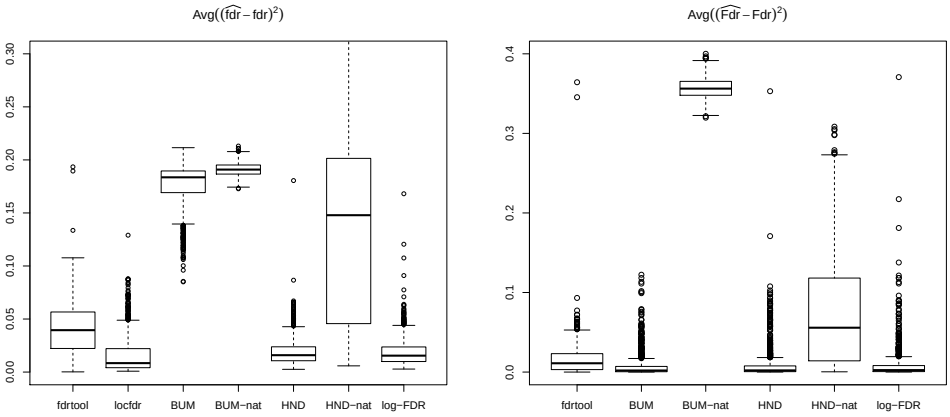
2. **Empirical null model**: The parameters of fdr threshold models are estimated by maximizing the marginal likelihood of the BUM and HND models. I refer to these estimated models as BUM–native and HND–native (abbreviated as *–nat*). This allows to compare the estimated null models of all algorithms considered and to evaluate the effect of misspecification on parameter estimation for the BUM and HND models.

In each case I computed the Fdr– and fdr–values of all $d = 200$ hypotheses for all $B = 1000$ repetitions and compared these estimates with the true Fdr and fdr values as given by the true known mixture model.

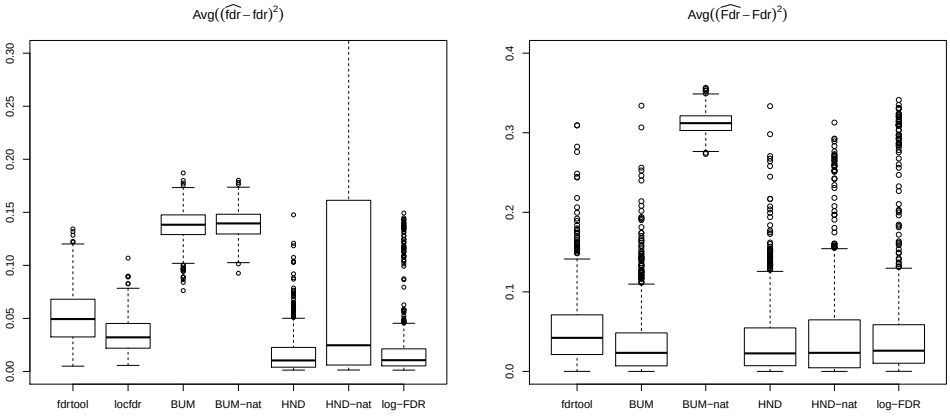
3.6.2 Results for the Overlapping and Well Separated Scenarios

In Fig. **3.6**, the results from the comparison of true and estimated FDR–values are shown using the following abbreviations for the investigated algorithms: fdrtool corresponds to using the `fdrtool` software [Strimmer, 2008a,b], with constrained maximum likelihood estimation for the null model and constrained Grenander density estimation for the marginal density. This corresponds to the options (a1) and (d1) in the general algorithm (section 3.4.1). BUM and HND denote the two fdr threshold methods with the null model given by log–FDR; and BUM–native and HND–native correspond to the two fdr threshold methods with empirical null model (i.e. including an extra standard deviation parameter σ). Furthermore, the state of the art fdr estimation algorithm locfdr as implemented in the `locfdr` `R`–package [R Development Core Team, 2012] is used in the comparisons [Efron, 2004, 2007b, 2008, section 3.5.2)]. This algorithm does not compute Fdr values. So it is not included in the Fdr comparisons.

The results can be summarized as follows: For fdr (first column in Fig. **3.6**) the HND model, locfdr and log–FDR are on top. Intriguingly, however, HND–native exhibits a dramatic reduction of accuracy in fdr estimation if the null and alternative are well separated (upper left image). On the other hand, if the null and the alternative are overlapping, the HND–native approach performs well (albeit with a large variance). The BUM model performs worst, both with and without empirical null model. For



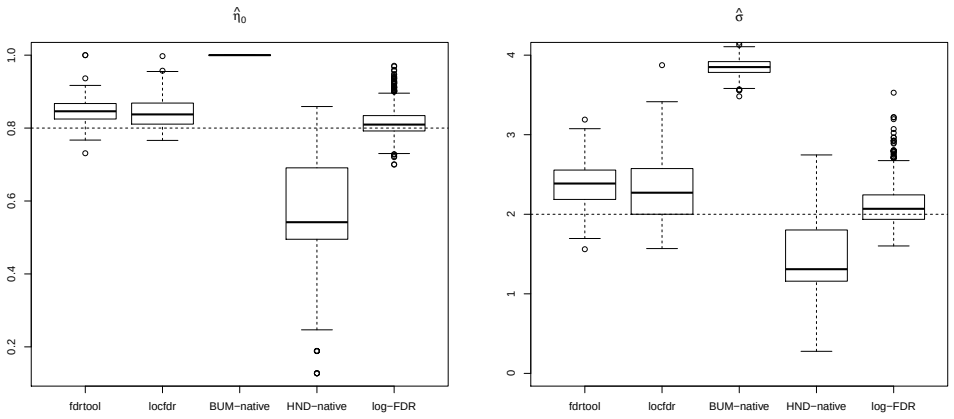
(a) Well separated scenario



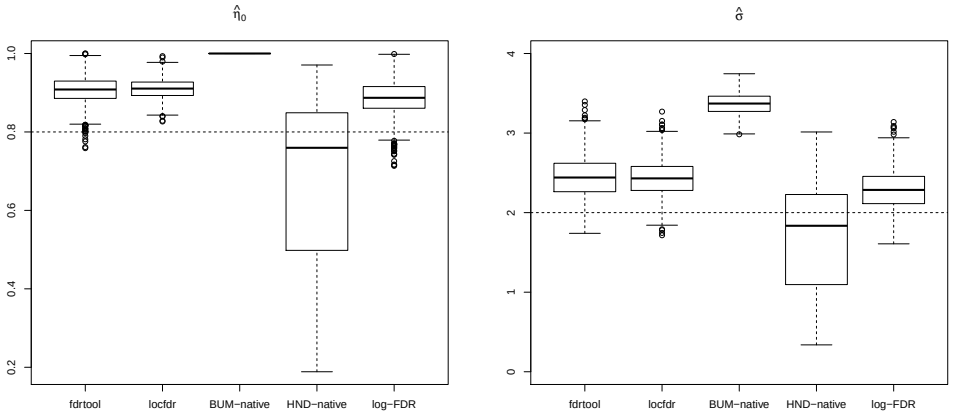
(b) Overlapping scenario

Figure 3.6: Comparison of the accuracy of fdr and Fdr estimates for the simulated data: (a) well separated case, and (b) overlapping scenario.

tail-area based FDR (second column in Fig. **3.6**) both `fdrtool` and log–FDR perform similar to BUM and HND. However, again there is a drastic reduction in accuracy for HND–native and BUM–native in the case of clear separation of null and alternative density (upper right image). Overall, log–FDR performs best, although the number of outliers is relatively large. This may be due to the complicated and thus variable null estimation process. The similar number of outliers of the HND and BUM approaches hints to this since these algorithms use the same null model estimation as log–FDR.



(a) Well separated scenario



(b) Overlapping scenario

Figure 3.7: Comparison of the accuracy of parameter estimates for the simulated data: (a) well separated case, and (b) overlapping scenario.

Fig. **3.7** shows the accuracy of the estimated null models for fdrtool, BUM–native, HND–native, locfdr and log–FDR. In the first column, boxplots for the estimated null proportion $\hat{\eta_0}$ are shown. With the true value of $\eta_0 = 0.8$ it is evident that BUM–native always overestimates η_0, whereas HND–native mostly underestimates η_0. The second column shows that the scale parameter σ is also always overestimated by BUM–native

and mostly underestimated by HND–native. Fdrtool, locfdr and log–FDR show similar results for both η_0 and σ in the overlapping case, while log–FDR is almost unbiased in the well separated case. As in Fig. **3.6**, the impact of the misspecification on HND–native can

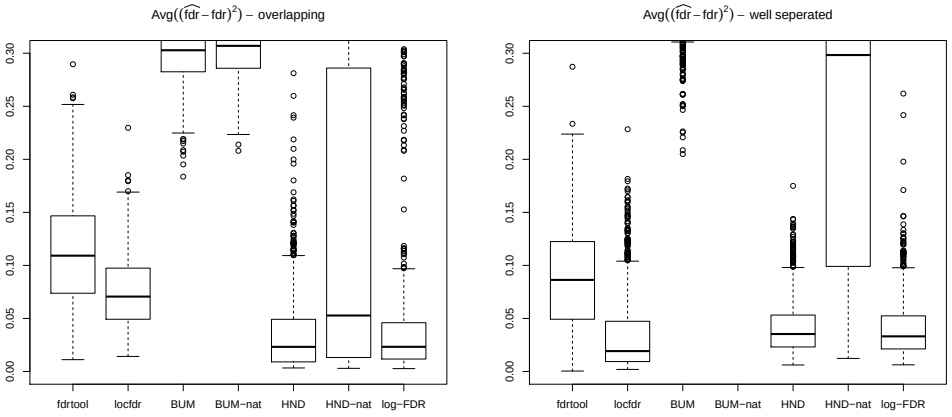


Figure 3.8: Comparison of the accuracy of fdr estimates for the simulated data for the subset of large $|z|$ with $|z| \geq 2$. The left figure shows the results for the overlapping scenario while the right one gives the results for the well separated case.

be clearly seen. If the null and alternative densities are well separated, the HND–native model is not appropriate, but in the more difficult case of overlapping mixture components, HND–native performs rather well. This conclusion is further strengthened by looking at the fdr estimates on the subset of large statistics, corresponding to potentially "significant" cases. Fig. **3.8** shows the same plots as the right column of Fig. **3.6** for all statistics z with $|z| \geq 2$. It can be noticed that in both scenarios, HND–native has large variation. Again, log–FDR performs intriguingly well, especially in the overlapping case, while in the well separated scenario locfdr is a bit ahead. In summary, the HND model works well for both Fdr and fdr estimation if the correct parameters for the null model are being supplied. HND–native estimation of the empirical null requires that model and data are not misspecified. In contrast, the BUM model is only suited for Fdr estimation and empirical null estimation failed for both investigated scenarios. Fdrtool and locfdr yield good results in general, however, log–FDR usually has at least a small lead over the competing algorithms, its sole downside being the more variable null model estimation process.

3.6.3 Empirical Null Analysis of Real Data

Tab. **3.2** shows the estimated null model for four experimental data sets concerning prostate cancer, education (mathematics competency), breast cancer and HIV already analyzed by Rice and Spiegelhalter [2008]. Additionally, the number of "significant"

Table 3.2: Empirical null parameter estimates obtained for four real data sets.

	Prostate	Education	BRCA	HIV
$\hat{\eta}_0$:				
fdrtool	0.9855	0.9671	1	0.9587
locdr	0.9981	0.9436	1.0388	0.9342
BUM–native	1	1	1	0.9984
HND–native	0.9829	0.9536	1	0.9370
MixFdr	0.9504	0.9171	0.9876	0.8285
log–FDR	0.9966	0.9657	1	0.9053
$\hat{\sigma}$:				
fdrtool	1.0649	1.7204	1.5730	0.7999
locdr	1.0871	1.6549	1.5752	0.9342
BUM–native	1.1350	1.9911	1.4313	0.9220
HND–native	1.0588	1.6810	1.4311	0.7652
MixFdr	1.0699	1.6903	1.4298	0.7001
log–FDR	1.0843	1.7152	1.7678	0.7297
No. of statistics with fdr ≤ 0.2 **:**				
fdrtool	49	62	0	99
locdr	19	74	0	160
BUM-native	0	0	0	0
HND-native	12	63	0	155
MixFdr	19	67	0	117
log–FDR	0	49	0	208

statistics given an fdr cutoff of 0.2 is displayed for each model. In this section, a short description of the data sets are given. In general, these data correspond to the overlapping case common in real high dimensional data, i.e. null and alternative model are close to each other.

First, I investigated the prostate cancer data set of Singh et al. [2002]. This consists of gene expression measurements of $d = 6033$ genes for $n = 102$ patients, of which 52 are cancer patients and 50 are healthy. The z–scores are based on two sample t–statistics comparing the two categories.

The education data set consists of 3748 California high schools. The test statistics are based on a binomial test of proportion of advantaged vs. disadvantaged students passing mathematics competency tests.

The HIV data are taken from van 't Wout et al. [2003] and consists of 7680 z–scores stemming from a microarray study comparing four HIV–positive patients with four HIV–negative controls. The goal is to detect genes which are differentially expressed

between the two groups.

In agreement with the simulations in section 3.6.2, Tab. **3.2** shows that BUM–native performs rather poorly, while HND–native and especially MixFdr underestimate η_0 relative to fdrtool, locfdr and log–FDR. However, in these data examples all methods except for BUM–native are more or less in broad agreement, which implies that the implicit alternative density of the HND model is appropriate here. The analysis furthermore proves that empirical truncation point finding by smoothing, used as a null model estimation method by log–FDR, also works satisfactorily on real data sets. The number of statistics considered significant is also quite similar among the methods. Notably, however, the η_0 estimates of the HIV data are in disagreement, with estimates ranging from 82% (MixFdr) to 95% (fdrtool). This also results in a large span of genes considered as differentially expressed between the two groups, ranging from 99 up to 208. In conclusion, the real data analysis conducted shows the applicability of fdrtool, locfdr and log–FDR in real data analysis. Since with all data sets null and alternative can be considered as overlapping, HND–native performs satisfactorily as well.

3.7 Accuracy and Variability of fdr Estimates for Various Algorithms

In this section, test statistics distributed according to the HND model will be simulated and several fdr estimation techniques will be compared with respect to their bias and variability. The motivation for this study Muralidharan's [2010] observation that constrained maximum likelihood estimation of the marginal density can lead to highly variable fdr estimates, while algorithms such as locfdr (cf. section 3.5.1) and the author's own MixFdr (cf. section 3.5.2) algorithm are supposed to show less bias and variability than constrained maximum likelihood estimation. The attention is limited to local FDR (fdr) here, since it is based on densities and thus much harder to estimate than the tail area based Fdr. Algorithms that estimate fdr well will thus also be able to estimate the Fdr easily.

MixFdr (cf. section 3.5.2) uses a normal mixture model to estimate Eq. **2.3** and Eq. **2.4**. The number of mixture components has to be defined beforehand in MixFdr, with one of them being the null model. Then the whole model is fit to the data at hand using a penalized EM algorithm. I will use three components and an penalization parameter of $P = 50$ throughout. As Muralidharan [2010] shows in his simulation studies, the choice of the penalization parameter is of minor importance, so fixing it is justified. Note that MixFdr was not included in previous comparisons because it does not work well for moderate sample sizes (< 1000). The claim about high variability of nonparametric estimation methods made in Muralidharan [2010] is based on simulations using a normal mixture. Here the HND model is used instead. It does not correspond directly to the assumptions of any of the algorithms considered except for HND–native. Thus HND–

native should perform best and all other approaches can be judged according to their behavior relative to HND–native. HND–native can therefore be viewed as the "gold standard" in this section. Again, $B = 1000$ runs are performed but this time 1000 instead of 200 test statistics are generated at each run.

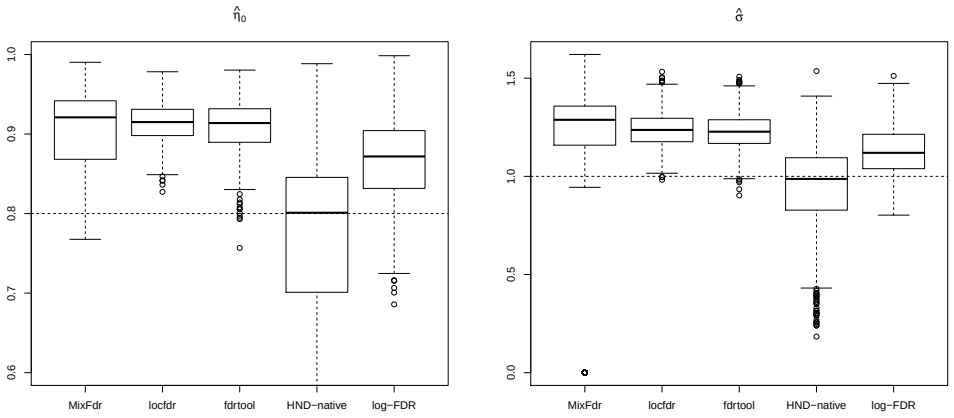


Figure 3.9: Comparison of the of σ and η_0 estimates for 1000 statistics sampled from the HND model.

Fig. **3.9** shows that HND–native gives unbiased null model parameter estimates — as one would expect. In Fig. **3.10**, we see that the same is true for the fdr estimates of HND–native. Log–FDR and HND share the second position with regard to the bias. Intriguingly, log–FDR is able to capture the *shape* of the true fdr curve perfectly yielding almost the same result as the *parametric* HND model. MixFdr performs worst — yet another indication that EM type algorithms do not work reliably in FDR estimation, while fdrtool and locfdr give satisfying results. Fig. **3.11** shows a graphical comparison of the fdr estimation results, which is similar to the plots in section 3.6.2. Lastly, the standard deviation of the fdr estimates is studied in Fig. **3.12**. Interestingly, MixFdr has the highest standard deviation of all algorithms, while all the other ones except for log–FDR, show a distinctively lower variability. log–FDR has a variability that is similar to HND for small test statistics. For larger test statistics, its variability is similar to that of MixFdr. In summary, the constrained log–concave estimation of log–FDR shows considerable variability, especially for large $|z|$–values. However, MixFdr is even more variable. Therefore, it can be concluded that the claims made in Muralidharan [2010] cannot be confirmed. Constrained maximum likelihood estimation does not automatically lead to a high bias and variance of the fdr estimates. Surprisingly, MixFdr even shows the highest variance of all methods considered. Additionally, it exhibits a considerable bias when applied to data generated from the HND model.

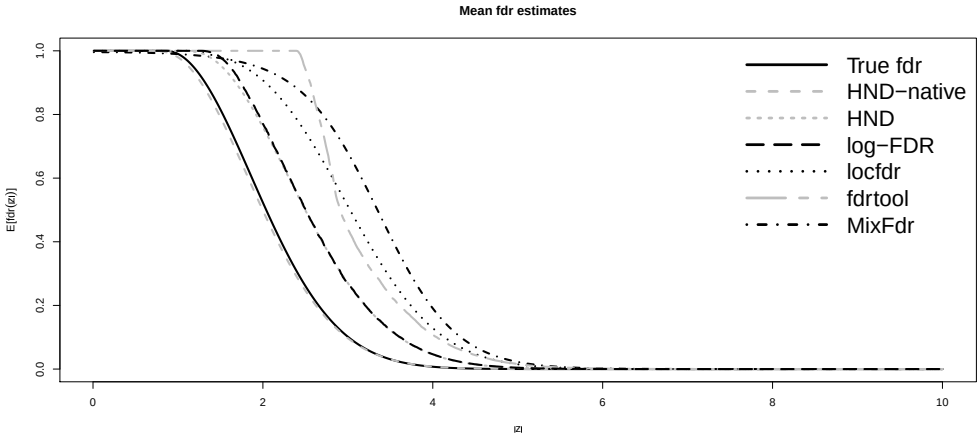


Figure 3.10: Comparison of the of fdr estimates for 1000 statistics sampled from the HND model.

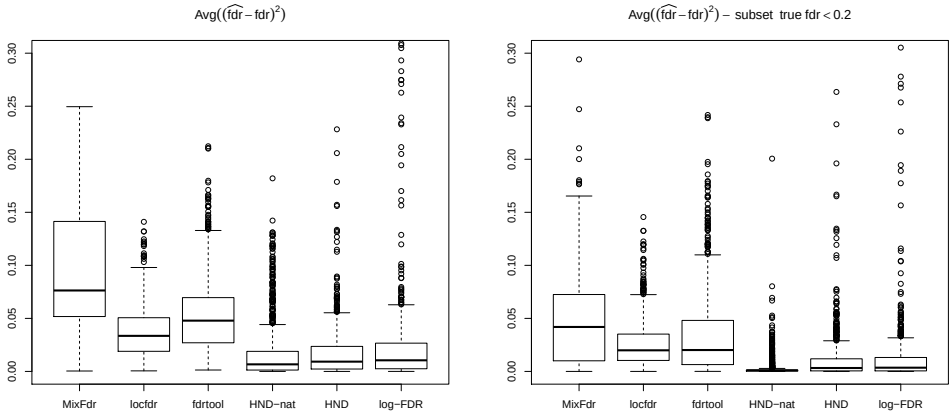


Figure 3.11: Comparison of the standard deviation of the fdr estimates for 1000 statistics sampled from the HND model using boxplots as in section 3.6.2.

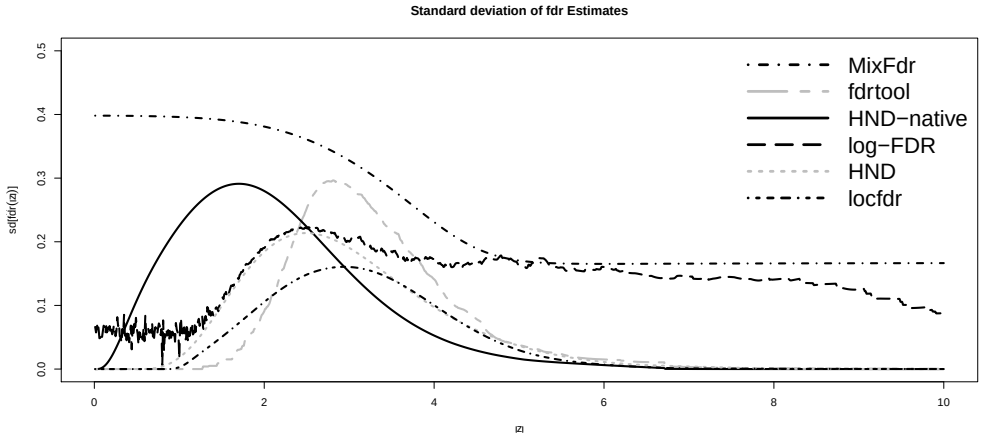


Figure 3.12: Comparison of the standard deviation of the fdr estimates for 1000 statistics sampled from the HND model.

4 Effect Size Estimation and Variable Selection in Linear Discriminant Analysis

After the introduction of false discovery rates, I now want to focus on their applications. Specifically, I will look at variable ranking and selection in the context of linear discriminant analysis (LDA), which is a simple, yet very effective, approach to linear classification [Hand, 2006].

In this chapter, several fundamental notions related to LDA will be established and the estimation of effect sizes will be discussed. A competitor to FDR methods called "Higher Criticism" (HC) will be introduced and discussed in great detail in chapter 5. In this chapter, I will focus on variable selection by thresholding using false discovery rates, HC and the misclassification rate.

A particularly interesting area of application for variable selection is modern medical research, which has been revolutionized by the possibility of characterizing diseases at a molecular level using microarrays. This classification of biological samples based on their gene expression continues to be a field of active research, cf. Pang et al. [2009], Cao et al. [2011], Xiaosheng and Simon [2011] and Shao et al. [2011]. Current reviews of the subject can be found in Schwender et al. [2008], Slawski et al. [2008] as well as in Kim and Simon [2011].

In order to develop classifiers which are potentially useful for molecular diagnostics, it is important to construct them based on a selection of genes (variables) strongly associated with the respective class labels (e.g. cancer and healthy tissue). These genes have a large effect size which is generally measured by standardized differences.

Three distinct, but closely related, objectives need to be achieved to identify a group of genes with high effect sizes [Ahdesmäki and Strimmer, 2010, Matsui and Noma, 2011]:

(i) to establish a reliable variable ranking,

(ii) to provide a reasonable estimate of the effect size for each gene, and

(iii) to find a suitable cutoff point that allows to disregard (the usually large) number of noise-features.

Problems (ii) and (iii) are the main concerns of the current chapter. For the ranking problem (obj. (i)), I will rely on correlation adjusted t–scores (a.k.a. "cat" – scores) introduced by Zuber and Strimmer [2009]. The cat–score is a t–type statistic which takes correlation into account and has been shown to induce a reliable variable ranking even in the presence of correlation among the variables. I therefore am going to use cat–scores to obtain effect size estimates (obj. (ii)). Based on these estimates, a nominal prediction error is computed. It is dependent on the number of variables included. Variable selection is then performed (ob. (iii)) by determining the number of variables necessary to achieve a certain nominal error level.

The approach presented here is similar to that of Efron [2009] and Dabney and Storey [2007]. However, in contrast to Efron [2009], my method applies to any number of classes and allows empirical null modeling. In contrast to Dabney and Storey [2007], it does not need a computationally expensive greedy algorithm to select variables due to the variable ranking performed beforehand.

The chapter is organized as follows: I will present basic theory on LDA in chapter 4.1, then I obtain effect size estimates based on cat–scores and compare them to other effect size estimation approaches in section 4.2. Notably, the methods of Efron [2009] and Matsui and Noma [2011] are presented in a unifying way using cat–scores, which sheds new light on their similarities. Section 4.3 shows how to perform variable ranking and selection using different methods based on a variable ranking. Results of variable selection methods on simulated and real data are then presented in chapter 4.5.

4.1 Linear Discriminant Analysis (LDA) and its Misclassification Rate

4.1.1 LDA and Effect Sizes

LDA forms the basis of most classification algorithms currently employed, e.g. Nearest Shrunken Centroids commonly abbreviated as NSC, and also known as PAM [Tibshirani et al., 2003], Shrinkage Discriminant Analysis — SDA [Ahdesmäki and Strimmer, 2010] — and many more. It starts by assuming a mixture model for the d-dimensional data $\boldsymbol{x}$

$$f(\boldsymbol{x}) = \sum_{k=1}^{K} \pi_k f(\boldsymbol{x}|k),$$

where each class k is represented by a multivariate normal density

$$f(\boldsymbol{x}|k) = (2\pi)^{-d/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_k)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_k)\},$$

with group–specific centroids $\boldsymbol{\mu}_k$ and a common covariance matrix $\boldsymbol{\Sigma}$. A sample $\boldsymbol{x}$ is assigned to the class yielding the highest LDA discriminant score defined as the log posterior probability $d_k^{\text{LDA}}(\boldsymbol{x}) = \log\{\text{prob}(k|\boldsymbol{x})\}$. This score can be written as

$$d_k^{\text{LDA}}(\boldsymbol{x}) = \boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{1}{2}\boldsymbol{\mu}_k^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \log(\pi_k). \tag{4.1}$$

The standard form of the LDA predictor function shown in Eq. **4.1** can be transformed into a scalar product which is given by

$$\Delta_k^{\text{LDA}}(\boldsymbol{x}) = \left(\boldsymbol{\omega}^{(k,\text{pool})}\right)^T \boldsymbol{\delta}_k(\boldsymbol{x}) + \log(\pi_k). \tag{4.2}$$

See Ahdesmäki and Strimmer [2010] for details. In Eq. **4.2** we have an inner product of Mahalanobis transformed variables (commonly called features) $\boldsymbol{\delta}(\boldsymbol{x})$ and a corresponding feature weight vector $\boldsymbol{\omega}^{(k,\text{pool})}$ given by

$$\boldsymbol{\delta}_k(\boldsymbol{x}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_{\text{pool}}}{2}\right) \tag{4.3}$$

and

$$\boldsymbol{\omega}^{(k,\text{pool})} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_{\text{pool}}), \tag{4.4}$$

respectively. In this equation, the pooled mean is calculated as $\boldsymbol{\mu}_{\text{pool}} = \sum_{k=1}^{K}\frac{n_k}{n}\boldsymbol{\mu}_k$ and the covariance matrix $\boldsymbol{\Sigma}$ is decomposed as: $\boldsymbol{\Sigma} = \boldsymbol{V}^{1/2}\boldsymbol{P}\boldsymbol{V}^{1/2}$, with a diagonal matrix containing the variances $\boldsymbol{V} = \text{diag}\{\sigma_1^2,\ldots,\sigma_d^2\}$ and the correlation matrix $\boldsymbol{P} = (\rho_{ij})$. Remarkably, both $\boldsymbol{\omega}^{(k,\text{pool})}$ and $\boldsymbol{\delta}_k(\boldsymbol{x})$ are vectors and not matrices.

The decomposition in Eq. **4.2** shows that $\boldsymbol{\omega}^{(k,\text{pool})}$ gives the influence of the *transformed* variables $\boldsymbol{\delta}(\boldsymbol{x})$ in prediction. Zuber and Strimmer [2009] have shown that this Mahalanobis–transformation leads to an improved ranking of the *original* variables since it removes the effect of correlation. Thus, as in Ahdesmäki and Strimmer [2010], the feature weights $\boldsymbol{\omega}$ will serve as a measure of variable importance and the terms variables and features will be used interchangeably in the following sections.

Additionally, from Eq. **4.4** it can be seen that the components of $\boldsymbol{\omega}^{(k,\text{pool})}$ are decorrelated and standardized differences (i.e. effect sizes) between the class k and the "pooled class" [Matsui and Noma, 2011]. This is readily generalized. The effect size vector $\boldsymbol{\omega}^{(k,l)}$ between any two classes k and l is defined as the difference between the two respective

feature weight vectors $\boldsymbol{\omega}^{(k,\text{pool})}$ and $\boldsymbol{\omega}^{(l,\text{pool})}$

$$\boldsymbol{\omega}^{(k,l)} := \boldsymbol{\omega}^{(k,\text{pool})} - \boldsymbol{\omega}^{(l,\text{pool})} = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{\mu}_k - \boldsymbol{\mu}_l)\,. \tag{4.5}$$

Note that $\boldsymbol{\omega}^{(k,l)}$ is up to the scale factor $(1/n_k + 1/n_l)^{-1/2}$ equivalent to the cat–score vector between the classes k and l on the population level, i.e. assuming known model parameters [Zuber and Strimmer, 2009]. Hence there is a close relationship between test statistics and effect sizes: The effect size is simply a sample size independent version of the test statistic. The statistic is denoted by a "cat" subscript in this article, i.e.

$$\boldsymbol{\omega}^{(k,l)}_{\text{cat}} = (1/n_k + 1/n_l)^{-1/2}\boldsymbol{\omega}^{(k,l)}\,.$$

4.1.2 The Misclassification Rate of Linear Discriminant Analysis

In this section, I am going to look at an unconditional (i.e. not depending on the data) misclassification error of LDA on the population level. This quantity is called (unconditional) misclassification rate in the literature [Dabney and Storey, 2007, Shao et al., 2011].

Let $\boldsymbol{x}^{(k)}$ be a sample vector drawn from the multivariate normal distribution $N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ associated with class k. In the LDA algorithm, it is assigned to the class yielding the highest score (Eq. **4.1**). Using the scalar product of Eq. **4.2** a misclassification (on the population level) of $\boldsymbol{x}^{(k)}$ occurs if $[\boldsymbol{\omega}^{(k,\text{pool})}]^T\boldsymbol{\delta}_k(\boldsymbol{x}^{(k)}) + \log(\pi_k) < \max_l [\boldsymbol{\omega}^{(l,\text{pool})}]^T\boldsymbol{\delta}_l(\boldsymbol{x}^{(k)}) + \log(\pi_l)$. It is easily verified that this is equivalent to the condition

$$\min_{l\neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x}^{(k)} - \frac{\boldsymbol{\mu}_k+\boldsymbol{\mu}_l}{2}\right)] + \log\left(\frac{\pi_k}{\pi_l}\right)}{\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}]}} < 0\,.$$

Since $\boldsymbol{x}^{(k)} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ holds for all $k \in \{1,\ldots,K\}$, the *unconditional* (i.e. expected) probability of misclassifying a sample from class k into a wrong class $j \neq k$ can be deduced from the above formula as:

$$\text{prob}(j \neq k|k) = \Phi\left(-\min_{l\neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}] + 2\log\left(\frac{\pi_k}{\pi_l}\right)}{2\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T[\boldsymbol{\omega}^{(k,l)}]}}\right).$$

This results in a misclassification rate (total error probability) of

$$\begin{aligned}\text{prob}(\text{error}) &= \sum_{k=1}^{K} \text{prob}(j \neq k|k) \times \text{prob}(k) \\ &= \sum_{k=1}^{K} \Phi\left(-\min_{l \neq k} \frac{[\boldsymbol{\omega}^{(k,l)}]^T [\boldsymbol{\omega}^{(k,l)}] + 2\log\left(\frac{\pi_k}{\pi_l}\right)}{2\sqrt{[\boldsymbol{\omega}^{(k,l)}]^T [\boldsymbol{\omega}^{(k,l)}]}} \right) \times \pi_k \,. \end{aligned} \tag{4.6}$$

Observe that Eq. **4.6** is the result of applying an expectation operator *twice*, once with regard to the model parameters $\boldsymbol{\omega}^{(k,l)}$ and once with regard to the transformed data $\boldsymbol{\delta}_k(\boldsymbol{x}^{(k)}) - \boldsymbol{\delta}_l(\boldsymbol{x}^{(k)}) = \boldsymbol{P}^{-1/2}\boldsymbol{V}^{-1/2}\left(\boldsymbol{x}^{(k)} - \frac{\boldsymbol{\mu}_k + \boldsymbol{\mu}_l}{2}\right)$. The first application leads to the *population version* of the statistical model, with $\hat{\boldsymbol{\omega}}^{(k,l)}$ replaced by $\boldsymbol{\omega}^{(k,l)}$, the second results in an unconditional (not dependent on the data) error rate.

4.2 Effect Size Estimation in LDA

For two given classes k and l, a feature i with a large corresponding effect size $\omega_i^{(k,l)}$ is most influential in differentiating between class k and l. However, a "naive" estimation of $\omega_i^{(k,l)}$ (e.g. estimation by plug-in estimates) suffers from the so–called "selection bias": Estimates of $\omega_i^{(k,l)}$ are biased upwards in general. For example, an estimated effect size of 1.5 based on t–scores might correspond to a true effect size of 0.7, see Fig. **4.1**. Therefore, reliable estimates of $\omega_i^{(k,l)}$ are needed in order to furnish a good estimate of Eq. **4.6**.

4.2.1 Three Empirical Bayes Approaches

Bayesian approaches are "immune" to selection effects [Dawid, 1994, Senn, 2008]. Thus, Efron [2009] as well as Matsui and Noma [2011] employ empirical Bayes estimates to tackle the estimation of effect sizes.

I am going to present their ideas in a unified way using cat–scores. This will show similarities between the two methods that are not readily apparent from studying the two original papers. Therefore, both methods are presented in considerable detail in order to clearly demonstrate the conceptual overlap between them. This will also help to indicate their respective weaknesses.

Furthermore, the current section can be read as a concise and yet comprehensive review of both methods, which can be of great help to the interested reader. The empirical Bayes estimator presented in section 4.2.1 is an attempt to combine the strengths of both approaches while adressing their shortcomings.

Let k and l be any two classes. For the sake of simplicity, the feature index i ($i \in \{1,\ldots,d\}$) will be dropped in the upcoming subsections.

Efron's Method

Efron [Efron, 2009] begins by transforming the statistics $\omega_{\text{cat}}^{(k,l)}$ into z–scores via a t–distribution with $n_l + n_k - 2$ degrees of freedom:

$$z = \Phi^{-1}\left(F_{n_l+n_k-2}(\omega_{\text{cat}}^{(k,l)})\right),$$

where $F_{n_l+n_k-2}$ denotes the distribution function of a t–distribution with $n_l + n_k - 2$ degrees of freedom. He then assumes a prior density g on $\omega_{\text{cat}}^{(k,l)}$ given by the mixture

$$g(\omega_{\text{cat}}^{(k,l)}) = \eta_0 I_0(\omega_{\text{cat}}^{(k,l)}) + (1-\eta_0) g_A(\omega_{\text{cat}}^{(k,l)}), \tag{4.7}$$

where I_0 is a delta-function at 0 and η_0 the proportion of genes having a true effect size of zero. The alternative group, i.e. the nonzero effect sizes are represented by g_A. In the following, I will in general abbreviate conditioning on the alternative group with an "A" subscript. The statistic z is assumed to be distributed as

$$z|\omega_{\text{cat}}^{(k,l)} \sim N(\omega_{\text{cat}}^{(k,l)},1).$$

Together with Eq. **4.7**, this results in the following mixture model for z

$$f(z) = \eta_0 \varphi(z) + (1-\eta_0) f_A(z), \tag{4.8}$$

where $\varphi(z)$ is the normal distribution density and f_A is a mixture of the densities $\varphi(z - \omega_{\text{cat}}^{(k,l)})$:

$$f_A(z) = \int_{-\infty}^{\infty} \varphi(z - \omega_{\text{cat}}^{(k,l)}) g_A(\omega_{\text{cat}}^{(k,l)})\, d\omega_{\text{cat}}^{(k,l)}.$$

Eq. **4.8** is a typical case of two–groups mixture model as introduced in section 2.2.1. It consists of a theoretical (i.e. no additional parameters) "null" model $f_0 = \varphi$ and an alternative component f_A from which the "interesting" cases are assumed to be drawn [Efron, 2008]. Note, however, that in contrast to section 2.2.1 the original signed z–values are used. In order to present the ideas of both Matsui and Noma [2011] and Efron [2009] in a unified fashion, I will start with computing the posterior density conditioned on the alternative i.e. $f(\omega_{\text{cat}}^{(k,l)}|z, z \in \text{"alternative"}) = f(\omega_{\text{cat}}^{(k,l)}|z, \omega_{\text{cat}}^{(k,l)} \neq 0)$. As introduced above, the "A" subscript indicates conditioning on the alternative so that $f_A(\omega_{\text{cat}}^{(k,l)}|z) = f(\omega_{\text{cat}}^{(k,l)}|z, z \in \text{"alternative"})$. Finally, using Bayes' rule this density can be computed as

$$
\begin{aligned}
f_A(\omega_{\text{cat}}^{(k,l)}|z) &= \frac{f_A(z|\omega_{\text{cat}}^{(k,l)}) \cdot g_A(\omega_{\text{cat}}^{(k,l)})}{f_A(z)} \\
&= \exp(\omega_{\text{cat}}^{(k,l)} z - \log\{f_A(z)/\varphi(z)\})[\exp\{-(\omega_{\text{cat}}^{(k,l)})^2/2)\}] g_A(\omega_{\text{cat}}^{(k,l)}) .
\end{aligned}
$$

It has the form of a natural exponential family with natural parameter $\omega_{\text{cat}}^{(k,l)}$, sufficient statistic z and cumulant generating function $\log\{f_A(z)/\varphi(z)\} = \log\{[(1-\text{fdr}(z))/\text{fdr}(z)]\} \cdot \eta_0(1-\eta_0)\}$, where

$$
\text{fdr}(z) = \text{prob}(\text{"null"}|z) = \eta_0 \frac{\varphi(z)}{f(z)} = \eta_0 \frac{f_0(z)}{f(z)} \tag{4.9}
$$

is the local false discovery rate [Efron, 2008, cf. Eq. **2.5**]. Conditional on the alternative component, this leads to an effect size estimate of the simple form

$$
E_A\left(\omega^{(k,l)}|z\right) = -(1/n_l + 1/n_k)^{1/2} \frac{d}{dz} \log\left(\frac{1-\text{fdr}(z)}{\text{fdr}(z)} \frac{\eta_0}{1-\eta_0}\right) . \tag{4.10}
$$

Since by Eq. **4.9** the relationship $\text{prob}(\text{"alternative"}|z) = 1 - \text{prob}(\text{"null"}|z) = 1 - \text{fdr}(z)$ holds, the unconditional effect size estimate is:

$$
\begin{aligned}
E\left(\omega^{(k,l)}|z\right) &= E_A\{\omega^{(k,l)}|z\}\{1-\text{fdr}(z)\} \\
&= -(1/n_l + 1/n_k)^{1/2} \frac{d}{dz} \log\{\frac{1-\text{fdr}(z)}{\text{fdr}(z)} \frac{\eta_0}{1-\eta_0}\} \{1-\text{fdr}(z)\} ,
\end{aligned} \tag{4.11}
$$

which after some further calculations becomes

$$
E\left(\omega^{(k,l)}|z\right) = -(1/n_l + 1/n_k)^{1/2} \frac{d}{dz} \log\{\text{fdr}(z)\} . \tag{4.12}
$$

Note that if one used an empirical null $N(0,\sigma^2)$ with estimated σ as null density f_0, the connection to the natural exponential family would be lost. Then both the natural parameter and the sufficient statistic would depend on σ.

Unfortunately, in this case the elegant formula (4.12) no longer holds. This basically is the only downside of Efron's approach: It is conceptually simple and computationally efficient but it is not possible to include an additional variance parameter in the null model without "destroying" Eq. **4.12**.

Matsui and Noma's Method

Matsui and Noma [2011] introduce empirical null modeling into the approach of Efron [2009] via an empirical Bayes method. They start with a similar z–score transform. However, as a starting point absolute values are used:

$$z = \Phi^{-1}\left[1 - 2 \cdot \{1 - F_{n_l+n_k-2}\left(|\omega_{\text{cat}}^{(k,l)}|\right)\}\right].$$

Additionally, only a prior on the absolute non-null effect sizes $g_A\left(|\omega_{\text{cat}}^{(k,l)}|\right)$ is assumed. The non–null z have the conditional density

$$f_A\left(z|\,|\omega_{\text{cat}}^{(k,l)}|\right) = \varphi\left(\frac{|\omega_{\text{cat}}^{(k,l)}| - z}{V\left(|\omega_{\text{cat}}^{(k,l)}|\right)}\right).$$

The variance function V and the prior g_A are estimated from the data. As in Efron [2009], they also assume a two-group mixture model for the z–scores:

$$f(z) = \eta_0 \varphi\left(\frac{z - \mu_0}{\sigma_0}\right) + (1 - \eta_0) f_A(z).$$

The null density is (in contrast to Efron) an empirical null, i.e. mean and variance are estimated from the data: $f_0(z) = \varphi\left((z - \mu_0)/\sigma_0\right)$. The alternative density f_A is computed as:

$$\begin{aligned} f_A(z) &= \int_0^\infty f_A\left(z|\,|\omega_{\text{cat}}^{(k,l)}|\right) g_A\left(|\omega_{\text{cat}}^{(k,l)}|\right) d\,|\omega_{\text{cat}}^{(k,l)}| \\ &= \int_0^\infty \varphi\left(\frac{|\omega_{\text{cat}}^{(k,l)}| - z}{\sqrt{V\left(|\omega_{\text{cat}}^{(k,l)}|\right)}}\right) g_A\left(|\omega_{\text{cat}}^{(k,l)}|\right) d\,|\omega_{\text{cat}}^{(k,l)}|\,. \end{aligned}$$

The application of Bayes' rule gives a posterior expectation of $|\omega_{\text{cat}}^{(k,l)}|$ which is unfortunately not as simple as Eq. **4.10**:

$$E_A\left(|\omega_{\text{cat}}^{(k,l)}|\,|z\right) = \int_0^\infty |\omega_{\text{cat}}^{(k,l)}| \frac{f_A\left(z|\,|\omega_{\text{cat}}^{(k,l)}|\right) g_A\left(|\omega_{\text{cat}}^{(k,l)}|\right)}{f_A(z)}\, d\,|\omega_{\text{cat}}^{(k,l)}|$$
$$= \int_0^\infty |\omega_{\text{cat}}^{(k,l)}| \frac{\varphi\left(\frac{|\omega_{\text{cat}}^{(k,l)}|-z}{\sqrt{V\left(|\omega_{\text{cat}}^{(k,l)}|\right)}}\right) g_A\left(|\omega_{\text{cat}}^{(k,l)}|\right)}{f_A(z)}\, d\,|\omega_{\text{cat}}^{(k,l)}|\,.$$

The statistic $|\omega_{\text{cat}}^{(k,l)}|$ is then transformed back into an absolute value effect size:

$$E_A\left(|\omega^{(k,l)}|\,|z\right) = (1/n_l + 1/n_k)^{1/2} F^{-1}_{n_l+n_k-2}\left(1 - \frac{1}{2}\left[1 - \Phi\left\{E_A\left(|\omega_{\text{cat}}^{(k,l)}|\,|z\right)\right\}\right]\right).$$

As in Eq. **4.12**, the final effect size estimate is:

$$E\left(|\omega^{(k,l)}|\,|z\right) = E_A\left(|\omega^{(k,l)}|\,|z\right)(1 - \text{fdr}(z))\,. \tag{4.13}$$

In contrast to Efron's method, the approach of Matsui and Noma [2011] allows empirical null modeling and thus leads to better effect size estimates in general, as Matsui and Noma [2011] convincingly show in their article.

However, this increased accuracy comes at a price. The estimation of variance function V can take up to two hours. Furthermore, it has to be estimated for every number of class samples n_k and n_l separately. This makes cross-validation based assessment of predictive accuracy extremely time consuming. Additionally, even if V has been computed for fixed n_k and n_l , the estimation of the final effect size will take up to several minutes.

In summary, while Matsui and Noma [2011] provide a method that is superior to Efron's method in terms of bias, it is at the same time computationally very demanding.

A Simple Empirical–Bayes Approach

In this section I will derive another more heuristic approach to the reliable estimation of effect sizes that tries to combine the advantages of Matsui and Noma's [2011] as well as Efron's [2009] methods. Empirical null modeling will be included, it will be computationally tractable and provide sufficient accuracy.

Observe that in non–empirical Bayes frameworks, reliable estimation of effect sizes is generally achieved by shrinking initial estimates of statistics playing the same role as

$\omega_{\text{cat}}^{(k,l)}$. For example, in the popular PAM algorithm [Tibshirani et al., 2003], the estimated t–scores are shrunk using a parameter λ estimated by cross validation.

Therefore, an appropriate adaptive shrinkage of the original should provide us with reasonable effect size estimates. As it turns out, this adaptive shrinkage can easily be achieved by employing false discovery rates.

The first step in my heuristic approach to achieve a shrinkage of $\omega^{(k,l)}$ is the assumption of a two–component mixture model on the effect sizes:

$$f(\omega_{\text{cat}}^{(k,l)}) = \eta_0 f_0(\omega_{\text{cat}}^{(k,l)}) + (1-\eta_0) f_A(\omega_{\text{cat}}^{(k,l)}), \tag{4.14}$$

leading to corresponding fdr estimates of Eq. **4.9**. Assuming a centered null distribution, we can now make use of the "naive" estimates $E_A\left(\omega^{(k,l)}\right) = \omega^{(k,l)}$ and correspondingly $E_0\left(\omega^{(k,l)}\right) = 0$ (since f_0 is centered). The 0 subscript indicates a conditioning on the null distribution, $E_0\left(\omega^{(k,l)}\right) = E\left(\omega^{(k,l)} \,|\omega^{(k,l)} \in \text{"null"}\right)$. It now holds by the law of total probability and Eq. **4.9** that the effect size is given by

$$\begin{aligned}
E\left(\omega^{(k,l)}\right) &= (1/n_l + 1/n_k)^{1/2}\{E_0\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot \text{prob}\left(\omega_{\text{cat}}^{(k,l)} \in \text{"null"}|\omega_{\text{cat}}^{(k,l)}\right) \\
&\quad + E_A\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot \text{prob}\left(\omega_{\text{cat}}^{(k,l)} \in \text{"alternative"}|\omega_{\text{cat}}^{(k,l)}\right)\} \\
&= (1/n_l + 1/n_k)^{1/2} E_A\left(\omega_{\text{cat}}^{(k,l)}\right) \cdot \text{prob}\left(\omega_{\text{cat}}^{(k,l)} \in \text{"alternative"}|\omega_{\text{cat}}^{(k,l)}\right) \\
&= E_A\left(\omega^{(k,l)}\right) \cdot \left(1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})\right) \\
&= \omega^{(k,l)}\left(1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})\right).
\end{aligned} \tag{4.15}$$

Eq. **4.15** is very similar to Eq. **4.13** and Eq. **4.11**, however, no full Bayesian posterior is computed. Instead, simple non–Bayesian estimates for the expectations in the two–groups model Eq. **4.14** are employed. This makes the implementation of Eq. **4.15** computationally efficient.

There is an obvious downside though: Large (with respect to their absolute value) statistics usually have a high fdr value close to 0. Therefore, they are hardly shrunk at all although their effect size is usually grossly overestimated. Thus, it is necessary to impose a minimum shrinkage. From the results of the real data analysis in table 1 of Matsui and Noma [2011], it can easily be seen that the empirical Bayes method that these authors apply imposes a shrinkage of at least 50% on the top 5 test statistics. I therefore also set the minimum shrinkage to 50% leading to the formula

$$\boldsymbol{\omega}_{\text{fdr}}^{(k,l)} = \boldsymbol{\omega}^{(k,l)} \cdot \min\{0.5; [1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})]\}. \tag{4.16}$$

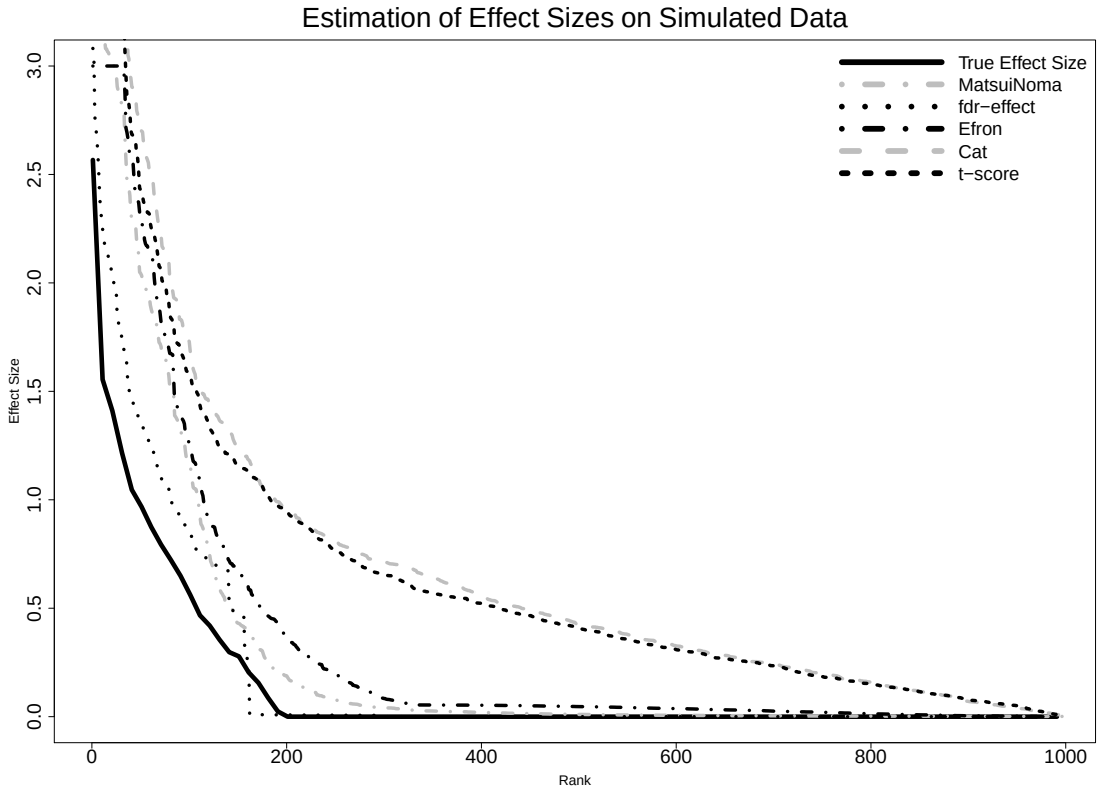


Figure 4.1: Comparison of effect size estimates on simulated data following the Smyth [2004] model.

I call this fdr–effect size estimation (fdr–effect) and abbreviate $\omega^{(k,l)}\left(1 - \text{fdr}(\omega_{\text{cat}}^{(k,l)})\right)$ by $\omega_{\text{fdr}}^{(k,l)}$. Note that a fdr cutoff of 50% is conceptually very close to Higher Criticism Thresholding, see chapter 5 and Klaus and Strimmer [2012].

Perhaps surprisingly, in next section it will be shown that it is competitive with regard to the attained accuracy, even though no sophisticated posterior estimates are used. The adaptive shrinkage performed in Eq. **4.16** can be interpreted as being in between the full empirical Bayes approaches of Efron [2009] or Matsui and Noma [2011] and soft thresholding using a single shrinkage parameter for all statistics as in Tibshirani et al. [2003].

4.2.2 Evaluation of Effect Size Estimation Methods on Real and Simulated Data

A comparison of effect size estimation methods using simulated data is shown in Fig. **4.1**. Specifically, I will compare the effect size estimation using "naive" approaches

(simple cat and t–scores) and the more sophisticated ones described in the previous section abbreviated as MatsuiNoma, Efron and fdr–effect, respectively. For the methods MatsuiNoma and Efron, I use the implementations offered by the authors, for fdr–effect, I perform cat–score and fdr estimation using the `R`–packages [R Development Core Team, 2012] `st` and `fdrtool` [Strimmer, 2008a]. In the real data analysis displayed in Fig. **4.2**, the package `locfdr` [Efron, 2004, 2007b, 2008, section 3.5.1] is applied since this allows a straightforward use of an empirical null as it has been suggested in Matsui and Noma [2011] and Efron [2004] for this data set.

I am going to follow closely the setup used in Smyth [2004], Opgen-Rhein and Strimmer [2007] and Zuber and Strimmer [2009] to simulate gene expression data. The parameters are chosen in such a way that effect sizes between 1 and 3 are obtained, which roughly corresponds to the range considered in the simulation studies of Matsui and Noma [2011].

The number of statistics was fixed at $d = 1000$ with 200 statistics designated to be differentially expressed. The variances across genes were drawn from a scale–inverse–chi–square distribution Scale–inv–$\chi^2(d_0, s_0^2)$ with $s_0^2 = 1$ and $d_0 = 1$, i.e. the variances vary moderately from gene to gene. Furthermore, the difference of means for the differentially expressed genes (1–200) were drawn from a normal distribution with mean zero and the gene-specific variance multiplied with a scale factor set to 0.3. For the non–differentially expressed genes (201–1000), the difference was set to zero. The data were generated by drawing from group–specific multivariate normal distributions with the given variances and means employing a block diagonal correlation structure intended to mimic gene expression data. This structure was generated as in Guo et al. [2007] with block size 100 and block entries equal to $0.9^{|i-j|}$. Furthermore, the sample sizes n_1 and n_2 are equal with $n_1 = n_2 = 8$.

The effect size estimates are plotted in Fig. **4.1** according to their rank. It is important to note that this does not tell us whether the respective ranking is correct. Thus, even though the effect size estimates of the cat–score and an ordinary t–score are very similar, this does not mean that their induced ranking is comparable. Efron's and Matsui and Noma's method will also change the ranking of the supplied cat–scores at least slightly.

It can be seen that fdr–effect and MastsuiNoma yield good results, while Efron's method has a higher bias for effect sizes up to 1, a phenomenon already observed by Matsui and Noma [2011]. The "naive" approaches (cat–scores and t–scores) are far off for effect sizes up to 1.5. However, all methods overestimate large effect sizes. It follows that variable selection methods relying on effect size estimates will generally have a tendency of choosing only a relatively small number of variables in data sets with large effects.

This is in fact a phenomenon already observed by Ahdesmäki and Strimmer [2010] for the Efron algorithm applied to the Singh [Singh et al., 2002] prostate cancer gene

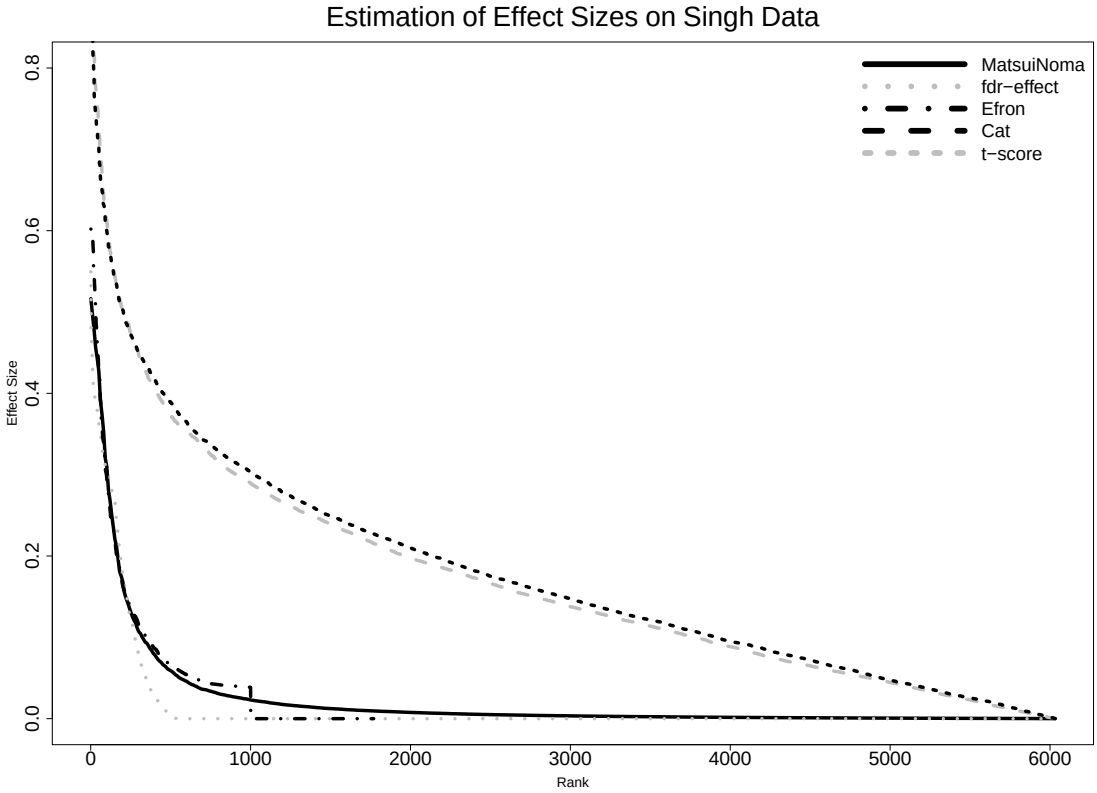


Figure 4.2: Comparison of effect size estimates for the Singh et al. [2002] data.

expression data. This data consists of gene expression measurements of $d = 6033$ genes for $n = 102$ patients, of which 52 are cancer patients and 50 are healthy. It has already been analyzed in Efron [2009] and Matsui and Noma [2011]. Fig. **4.2** shows the analysis results. As in the simulated data, the "naive" approaches are far off, while Efron and MatsuiNoma are quite similar. Note, however, that MatsuiNoma gives significantly lower estimates of large effect sizes than Efron. This is a phenomenon already noted in Matsui and Noma [2011]. The fdr–effect method yields similar results to MatsuiNoma for large effect sizes but reaches zero estimates much faster than MatsuiNoma and Efron. In conclusion, all empirical Bayes methods considered seem to give sound results here, while the empirical methods are probably grossly overestimating the effect sizes.

4.3 Variable Ranking and Selection

4.3.1 Variable Ranking

Before being able to select variables, a variable ranking needs to be established (obj. (i)). In the two class case, this is straightforward since the feature weight vector for class one

$\boldsymbol{\omega}^{(1,\text{pool})}$ is up to a scale factor of n_2/n equal to the effect size vector $\boldsymbol{\omega}^{(1,2)}$ ($\boldsymbol{\omega}^{(1,\text{pool})} = (n_2/n)\boldsymbol{\omega}^{(1,2)}$). Correspondingly, the feature weight vector for class two $\boldsymbol{\omega}^{(2,\text{pool})}$ is equal to the effect size vector $-\boldsymbol{\omega}^{(1,2)}$ up to a scale factor of n_1/n ($\boldsymbol{\omega}^{(2,\text{pool})} = (-n_1/n)\boldsymbol{\omega}^{(1,2)}$). Thus, variables can be ranked according to the absolute value of $\boldsymbol{\omega}^{(1,2)}$. In the the case of multiple classes, the situation is more complicated. The feature weight vectors of the different classes need to be summarized in a certain way to obtain the importance of each feature i in class prediction. Here, I am going to use the summary statistic S_i proposed by Ahdesmäki and Strimmer [2010] and given by

$$S_i = \sum_{k=1}^{K} \left(\omega_{\text{cat},i}^{(k,\text{pool})}\right)^2, \tag{4.17}$$

where $\omega_{\text{cat},i}^{(k,\text{pool})} = (1/n_k - 1/n)^{-1/2}\omega_i^{(k,\text{pool})}$. Since false discovery rates are generally assumed to be monotone, Eq. **4.15** shows that using fdr–effect effect size estimates $\boldsymbol{\omega}_{\text{fdr}}^{(k,\text{pool})}$ would produce the same ranking as the cat–scores if they were used instead of $\boldsymbol{\omega}_{\text{cat}}^{(k,\text{pool})}$ to compute S_i in Eq. **4.17**. Observe further that the statistics S_i fullfill the properties of a generalized test statistic as introduced in section 2.2.1.

4.3.2 Misclassification Rate Based Variable Selection

Having obtained estimates $\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}$ of $\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}$ and $\widehat{\pi_k}$ of π_k, we can now compute an estimate of the misclassification rate using Eq. **4.6**. Let $\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)$ be the vector of the t top–ranked variables according to the ranking induced by the vector $\boldsymbol{S}$ of all statistics S_i given by Eq. **4.17**. This gives an estimate of the misclassification rate, which depends on t:

$$\widehat{\text{prob}}(\text{error})(t) = \sum_{k=1}^{K} \Phi\left(-\min_{l\neq k} \frac{[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]^T[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)] + 2\log\left(\frac{\widehat{\pi_k}}{\widehat{\pi_l}}\right)}{2\sqrt{[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]^T[\widehat{\boldsymbol{\omega}_{\text{fdr}}^{(k,l)}}(t)]}}\right) \times \widehat{\pi_k}. \tag{4.18}$$

Efron performs feature selection by choosing a level $\alpha = 0.05$ as a target misclassification rate for the estimate in Eq. **4.18**. Although one could view α as a tuning parameter, I follow his suggestion in this regard. Experiments with lower α led to very large feature sets showing only a negligible improvement of the classification performance.

After the target error α has been set, a feature threshold t^* is obtained by including as many features as necessary to reach it, i.e. $\widehat{\text{prob}}(\text{error})(t^*) = \alpha$. Since usually a lot of features are shrunken to zero, it is possible that the target error can not be reached. Then, all the features will be included. This, however, is extremely unlikely to happen in real high dimensional data analysis. Finally, all features fulfilling $S_i \geq S_t^*$ are included in the classifier. I call the approach presented in this section misclassification rate (MR) based

variable thresholding (MRT).

4.3.3 Variable Selection via fndr Thresholding

An optimal threshold for the statistic S_i of Eq. **4.17** can also be found by false discovery rates. FDR has its "classic" application area in genomics, especially in the detection of differentially expressed genes.

A standard approach to obtain an FDR threshold to identify differentially expressed genes is to refer to the rule of Benjamini and Hochberg [1995] with a cutoff such as $\text{Fdr} \leq 0.05$. Alternatively, an fdr cutoff of 0.2 has been suggested in Efron et al. [2001].

However, it is important to observe that in the problem of constructing classifiers the FDR approach *cannot* be applied in the same fashion as in differential expression. In the latter case, the aim is to compile a set of genes one has confidence in to be differentially expressed. This is controlled by a tight FDR criterion, e.g. fdr< 0.2 as in Tab. **3.2**.

In contrast, when constructing classifiers one aims to identify the set of null features that are not informative with regard to group separation in order to eliminate them from the classifier. This is best done by controlling the fndr. Ahdesmäki and Strimmer [2010] suggest to include all genes with $\text{fndr}(S_i) < 0.2$ in the classifier. The local false discovery and local false non discovery rates add up to one, $\text{fndr}(S_i) = 1 - \text{fdr}(S_i)$. Hence, the set of features to be retained in the classifier has local false discovery rates smaller than 0.8 — instead of 0.2. This way, the features included in the predictor form a superset of the differentially expressed variables.

4.3.4 Variable Selection by HC–Thresholding

Donoho and Jin [2008] proposed the Higher Criticism–HC approach to variable thresholding. Here the test statistics $S_1, \ldots, S_d$ are first transformed into p–values via a (possibly empirical) null model using Eq. **2.9**.

HC Thresholding (HCT) then works as follows. First, the p–values are arranged from smallest to largest: $p_{(1)}, \ldots, p_{(d)}$. Then, each p–value is centered and standardized using the expected mean (i/d) and standard deviation of the corresponding order statistic $(\sqrt{i/d(1-i/d)/d})$. This results in the HC statistic

$$\text{HC}(p_{(i)}) = \frac{p_{(i)} - i/d}{\sqrt{i/d(1-i/d)/d}}.$$

Finally, the maximizing argument p_{HC}

$$p_{\text{HC}} = \arg\max_{i \in 1,\ldots,d} \text{HC}(p_{(i)})$$

is taken as the HC decision threshold for variable selection. Then, all variables with $p_i < p_{\text{HC}}$ are included in the classifier. HCT is a conceptually simple procedure that nonetheless works remarkably well. It has a close relationship to fdr thresholding and misclassification rate based variable selection. A more detailed exposition of these connections will be given in section 4.4 and chapter 5.

4.3.5 Estimation of the Prediction Rule and FDR

For the estimation of the prediction rule (Eq. **4.2**), I will mostly employ James-Stein-type estimators as in shrinkage discriminant analysis — SDA, Ahdesmäki and Strimmer [2010]. The group centroids $\boldsymbol{\mu}_k$ are estimated by the empirical means, for the correlations $\boldsymbol{P}$ the ridge-type estimator from Schäfer and Strimmer [2005] is used and the variances $\boldsymbol{V}$ are estimated by the shrinkage estimator from Opgen-Rhein and Strimmer [2007]. Finally, the proportions π_k are obtained by using the frequency estimator from Hausser and Strimmer [2009]. For SDA, I am going to employ the implementation provided by the `R`-package [R Development Core Team, 2012] `sda`. The local false discovery rates used in the fdr–effect approach are learned by using the Grenander density estimator and truncated maximum likelihood for the empirical null as presented in sections 3.1.1 and 3.2.2. As in section 4.2, the implementation offered by the `R`– package [R Development Core Team, 2012] `fdrtool` is employed.

4.4 The Relationship Between MR–Based Variable Selection and HCT

In this section, it will be shown that the HCT approach to variable selection can also be interpreted as a variant of MR–based variable selection (section 2.2.1) in the case of $K = 2$ classes.

The case $K = 2$ entails that thresholding the feature weight vector $\boldsymbol{\omega}_1$ is equivalent to thresholding the vector $\boldsymbol{\omega}^{(1,2)} = -\boldsymbol{\omega}^{(2,1)}$ (cf. section 4.3.1). Let $\boldsymbol{\omega} := \boldsymbol{\omega}^{(1,2)}$ and $\boldsymbol{\omega}(t)$ be the vector of effect sizes under the influence of feature selection as in section 4.3.2. Remember that then features with $|\omega| \geq \omega_t$ are included in the classifier, while features with $|\omega| \leq \omega_t$ are excluded and hence *set to zero* in $\boldsymbol{\omega}(t)$.

Supposing a two–groups model (Eq. **2.3** and Eq. **2.4**) on $\boldsymbol{\omega}$ the (expected) fractions of true positives (TP), true negatives (TN), false negatives (FN) and false positives (FP)

depending on the cutoff ω_t are given by the formulae

$$\begin{aligned}&\mathrm{TP}(\omega_t) = (1-\eta_0)(1-F_A(\omega_t)),\ \mathrm{TN}(\omega_t) = \eta_0 F_0(\omega_t),\\ &\mathrm{FN}(\omega_t) = (1-\eta_0)F_A(\omega_t) \text{ and } \mathrm{FP}(\omega_t) = \eta_0(1-F_0(\omega_t)).\end{aligned}$$

These quantities will be used to connect HCT with MRT. This will show that the HC criterion can be viewed as an approximation of the misclassification rate under the influence of feature selection.

To demonstrate this connection, I first go back on the expectation performed with respect to the data $\boldsymbol{x}$ in order to obtain Eq. **4.6**. This gives the conditional (i.e. depending on the data $\boldsymbol{x}$) population level error rate:

$$\begin{aligned}\Pr(\text{error} \mid \text{x}) = \Phi\left(-\frac{[\boldsymbol{\omega}^{(1,2)}(t)]^T[\boldsymbol{\delta}_1(\boldsymbol{x}) - \boldsymbol{\delta}_2(\boldsymbol{x})] + 2\log\left(\frac{\pi_1}{\pi_2}\right)}{2\sqrt{[\boldsymbol{\omega}^{(1,2)}(t)]^T[\boldsymbol{\omega}^{(1,2)}(t)]}}\right) \times \pi_1 + \\ \Phi\left(-\frac{[\boldsymbol{\omega}^{(2,1)}(t)]^T[\boldsymbol{\delta}_2(\boldsymbol{x}) - \boldsymbol{\delta}_1(\boldsymbol{x})] + 2\log\left(\frac{\pi_2}{\pi_1}\right)}{2\sqrt{[\boldsymbol{\omega}^{(2,1)}(t)]^T[\boldsymbol{\omega}^{(2,1)}(t)]}}\right) \times \pi_2\,.\end{aligned}$$

If equal class probabilities are assumed, the class frequencies can be dropped from the model and the abbreviation $\boldsymbol{\omega} = \boldsymbol{\omega}^{(1,2)}$ introduced above can be used. This leads to the formula:

$$\Pr(\text{error} \mid \text{x}) \propto \Phi\left(-\frac{[\boldsymbol{\omega}(t)]^T[\boldsymbol{\delta}_1(\boldsymbol{x}) - \boldsymbol{\delta}_2(\boldsymbol{x})]}{2\sqrt{[\boldsymbol{\omega}(t)]^T[\boldsymbol{\omega}(t)]}}\right).$$

Furthermore, it is assumed that all features posses the same strength, i.e. $\omega_i = c$ for all $i \in 1,\ldots,d$. Additionally, assume that the true null variables are *known*. Applying the expectation with respect to the x the entries of the vector $E[\boldsymbol{\delta}_1(\boldsymbol{x}) - \boldsymbol{\delta}_2(\boldsymbol{x})]$ will then equal $\frac{\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2}{2}$ if they belong to the alternative (are true positives) and 0 otherwise. Therefore,

$$E[\boldsymbol{\delta}_1(\boldsymbol{x}) - \boldsymbol{\delta}_2(\boldsymbol{x})] = 0.5c \cdot \mathrm{TP}(\boldsymbol{\omega})\,,$$

and correspondingly

$$[\boldsymbol{\omega}(t)]^T E[\boldsymbol{\delta}_1(\boldsymbol{x}) - \boldsymbol{\delta}_2(\boldsymbol{x})] = 0.5c[\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2]d\mathrm{TP}(\boldsymbol{\omega}(t))\,.$$

This gives

$$\begin{aligned}\Pr(\text{error }) &\propto \Phi\left(-\frac{c0.5[\boldsymbol{\mu}_1-\boldsymbol{\mu}_2]d\mathrm{TP}(\boldsymbol{\omega}(t))}{\sqrt{d\cdot c^2(\mathrm{TP}(\boldsymbol{\omega}(t))+\mathrm{FP}(\boldsymbol{\omega}(t)))}}\right)\\ &= \Phi\left(-0.5\sqrt{d}[\boldsymbol{\mu}_1-\boldsymbol{\mu}_2]\frac{\mathrm{TP}(\boldsymbol{\omega}(t))}{\sqrt{\mathrm{TP}(\boldsymbol{\omega}(t))+\mathrm{FP}(\boldsymbol{\omega}(t))}}\right).\end{aligned}$$

Donoho and Jin [2009] show (their Eq. **5.1**) that the HC objective function approximates the fraction

$$\frac{\mathrm{TP}(\boldsymbol{\omega}(t))}{\sqrt{\mathrm{TP}(\boldsymbol{\omega}(t))+\mathrm{FP}(\boldsymbol{\omega}(t))}}.$$

Hence, Higher Criticism thresholding can be viewed as a *very special* case of MR–based thresholding under relatively strong assumptions such as equal effect sizes for all non–null features. However, HCT nonetheless has competitive operating characteristics, as demonstrated in chapter 5.

4.5 Analysis of Real and Simulated Data

4.5.1 Simulations

In this section, I will compare variable selection based on the misclassification rate (MR) with several other state of the art thresholding variable selection approaches, namely false-non discovery rate (FNDR) thresholding (section 4.3.3 and Ahdesmäki and Strimmer [2010]), HCT (section 4.3.4 and Donoho and Jin [2008]) and the PAM/NSC algorithm [Tibshirani et al., 2003]. All methods are performed using empirical null modeling. As a base line classifier, I also include the results of classification with all features, i.e. performing no variable selection.

The simulations closely follow the setup of Witten and Tibshirani [2011]. A training set of size 100 and a test set of 1000 samples are created with a dimension of $d = 500$ variables. In total, 25 runs of each simulation setup are performed.

Simulation Setup 1

In this setup, there are four classes with equal probability (0.25) no correlation and unit variance. In each class 25 features are differentially expressed with an effect size of 0.7, yielding a total number of 100 differentially expressed features. Since there is no correlation, I perform Diagonal Discriminant Analysis (DDA), i.e. LDA with identity covariance $\boldsymbol{\Sigma} = I_d$. The results are displayed in Tab. **4.1**.

It can be seen that thresholding the summary statistic S (Eq. **4.17**) by false-non dis-

Table 4.1: Prediction errors and number of selected features for simulation setup 1, the number in the round brackets is the estimated standard error over 25 runs. The true number of differentially expressed features is 100.

Method	Prediction Error	Features
DDA–MR	0.1077 (0.0177)	156.48 (64.70)
DDA–FNDR	0.2482 (0.1272)	39.24 (23.72)
DDA–HC	0.1880 (0.0626)	152.32 (193.48)
PAM	0.0923 (0.0163)	253.6 (116.26)
DDA–ALL	0.1555 (0.0180)	500

covery rates or Higher Criticism yields hardly any significant features in most runs. Consequently, the estimated prediction errors are quite high.

Misclassification rate based feature selection as well as PAM, however, identify features useful for classification. This indicates that "analytical" thresholding methods, which do not rely on the optimization of a tuning parameter, may not work reliably when the effect sizes are small.

Simulation Setup 2

In this simulation, I am going to use a Guo et al. [2007] type block correlation with 5 blocks of size 100×100. As in section 4.2, each block entry is given by $0.9^{|i-j|}$, thus we have some highly correlated variables within blocks but variables in different blocks are independent.

Note that Witten and Tibshirani [2011] report using an entry size of 0.6. This is probably a misprint since my results obtained for PAM are quite similar to the ones reported in their article, while for 0.6 the error of PAM is only about 5%.

There are two classes with equal probability (0.5) and 200 features are differentially expressed with effect size 0.6, all of them are attributed to class 2. Since there is correlation present in this setting, I will perform LDA.

It can be seen in Tab. **4.2** that all feature selection methods except for PAM, which does not take correlation into account, perform quite well here.

4.5.2 Gene Expression Data

In Ahdesmäki and Strimmer [2010], the relative effectiveness of the FNDR and HC thresholds to select relevant genes in shrinkage discriminant analysis applied to gene

Table 4.2: Prediction errors and number of selected features for simulation setup 2, the number in the round brackets is the estimated standard error over 25 runs. The true number of differentially expressed features is 200.

Method	Prediction Error	Features
LDA–MR	0.000 (0.000)	63.16 (7.215)
LDA–FNDR	0.000 (0.000)	60.96 (6.567)
LDA–HC	0.000 (0.000)	85.04 (8.677)
PAM	0.088 (0.018)	294.0 (69.43)
LDA–ALL	0.093 (0.014)	500

expression data has already been compared. I am going follow their setup here and will analyze four clinical gene expression data sets related to prostate cancer [Singh et al., 2002], B-cell lymphoma [Alizadeh et al., 2000], colon cancer [Alon et al., 1999] and brain cancer [Pomeroy et al., 2002].

Specifically, balanced 10–fold cross–validation with 20 repetitions was performed to obtain error estimates and their standard deviations. The number of selected features is inferred by a single run of the respective variable selection method on the whole data set. Only for PAM this was repeated several times since the number of selected variables selected by this algorithm varies considerably between several runs in a row on the same data set.

In Tab. **4.3**, it can bee seen that the MRT approach has a performance similar to the other approaches. Interestingly, the MRT approach shows a more "adaptive" feature selection, leading to appropriate feature sets for each problem. In the brain data set, a very compact set of features is selected yielding a prediction error which is nonetheless in the range of the other approaches. The same is true for the Lymphoma and Colon data sets. This demonstrates that a variable selection method based on effect sizes leads to compact and yet effective molecular signatures. Furthermore, fndr and HC thresholding yield very similar results.

Table 4.3: Analysis of four cancer gene expression data sets with shrinkage discriminant analysis. The number of selected features are determined by a single feature selection run on the whole data set.

Data / Method	Prediction Error	Selected Variables
Prostate ($d = 6033, n = 102, K = 2$)		
LDA–MR	0.0630 (0.0050)	134
LDA–FNDR	0.0550 (0.0048)	131
LDA–HC	0.0497 (0.0045)	116
PAM	0.0850 (0.0061)	172–377
Lymphoma ($d = 4026, n = 62, K = 3$)		
LDA–MR	0.0211 (0.0039)	34
LDA–FNDR	0.0036 (0.0018)	392
LDA–HC	0.0000 (0.0000)	345
PAM	0.0234 (0.0041)	2796–2383
Colon ($d = 2000, n = 62, K = 2$)		
LDA–MR	0.1291 (0.0093)	28
LDA–FNDR	0.1278 (0.0088)	168
LDA–HC	0.1233 (0.0087)	122
PAM	0.1160 (0.0921)	13–23
Brain ($d = 5597, n = 42, K = 5$)		
LDA–MR	0.1628 (0.0126)	56
LDA–FNDR	0.1525 (0.0120)	102
LDA–HC	0.1417 (0.0108)	131
PAM	0.2023 (0.0118)	42–5587

5 Signal Identification for Rare and Weak Features: Higher Criticism or False Discovery Rates?

In this chapter, I will look in more detail at variable selection methods using thresholding. The chapter will mainly be concerned with an in–depth analysis of HCT and its relationship to f(n)dr thresholding (see sections 4.3.3 and 4.3.4). It will be shown that thresholding with an fdr of 0.5 is conceptually very similar to HCT.

Variable selection by thresholding can be understood as a special case of *signal identification*. Especially identification of sparse and weak signals in complex high–dimensional data is a challenging statistical problem that has many important applications in fields as diverse as astronomy, finance, genetics, medicine and proteomics. A typical biomedical task is the search for biomarkers using data from genome-wide association studies [Xie et al., 2011]. Signal *identification* is much more difficult than the closely related problem of signal *detection*. Whereas in detection, we are concerned purely with the presence or absence of a signal, in identification, we additionally seek to locate the signal.

In a series of recent publications, the method of Higher Criticism (HC) was powerfully advocated in settings with rare and weak features as an efficient means for signal detection [Donoho and Jin, 2004] as well as signal identification [Donoho and Jin, 2008, 2009, section 4.3.4]. Originally, HC was introduced by Tukey [1976] as an approach to multiple significance testing using a second–level test statistic computed from p-values. Importantly, in Donoho and Jin [2004] it was shown that HC provides a procedure which is optimal for signal detection in the sense that it achieves the best possible theoretical detection limit discovered earlier by Ingster [1999]. Subsequently, HC was also employed in a thresholding procedure to determine relevant features for prediction. Again, it was demonstrated that the HC approach to signal identification outperforms other commonly employed selection strategies, in particular those based on false discovery rates [Donoho and Jin, 2008, 2009].

In the previous chapter, the utility of HC for variable selection in classification was confirmed, but at the same time it was also empirically shown that in the signal identification problem controlling the false *non*-discovery rate is equivalent to the HC procedure.

Furthermore, it was discovered by Jager and Wellner [2007] that HC is not unique in achieving the detection limit. Given the success of HC, this raises questions about the fundamental principles that may underlie this approach.

Here, I will explore signal identification using the HC and false (non)-discovery rate approaches, with the aim to provide a better understanding of HC as well as offering a simple explanation for the favorable performance of HC. Specifically, I am going to argue that the decision threshold provided by HC may also be viewed as an approximation to a natural class boundary (CB) in classification, which, in turn, is easy to understand from a false discovery rate perspective. In particular, in the rare-weak setting in the region of the phase space where identification is actually possible we show that the HC and CB thresholds are nearly indistinguishable.

The remainder of the chapter is structured as follows: First, a non-technical introduction to HC both on the sample and the population level is provided. Second, I will derive the ideal thresholds corresponding to HC and false discovery rate approaches, and explore their mutual relationships. Next, I am going to investigate these thresholds in the rare-weak model and establish the near identity of HC and a natural CB threshold in the rare-weak identification setting. Finally, the validity of the theoretical considerations will be demonstrated by simulation and by analyzing data from four gene expression experiments (3 of which have already been analyzed in chapter 4).

5.1 Higher Criticism

In the following, the section 4.3.4 is expanded, describing the HC approach to signal identification in greater detail. Various properties of the HC threshold, both from a sample and population point of view, will be discussed. In this section and throughout the whole chapter, "x" will always denote a variable on the p–value scale, in contrast to the previous chapter, where it denoted multivariate normal data.

5.1.1 Empirical HC Threshold Based on p-Values

The already familiar situation with d observed generalized test statistics $y_1, \ldots, y_d$ (section 2.2.1) is assumed here. For each statistic, a corresponding p-value $p_1, \ldots, p_d$ via Eq. **2.9** is computed. The dimension d is potentially very large, as in many current applications in genomics or proteomics.

The HC approach to signal identification then proceeds as follows:

- First, by arranging the p-values from smallest to largest $p_{(1)}, \ldots, p_{(d)}$, the empirical

distribution function of the p-values is obtained

$$\hat{F}(x) = i/d \text{ for } p_{(i)} \leq x < p_{(i+1)}$$

with $x \in [0;1]$, $p_{(0)} = 0$, and $p_{(d+1)} = 1$.

- Second, the empirical HC objective function

$$\widehat{HC}(x) = \frac{|\hat{F}(x) - x|}{\sqrt{\hat{F}(x)(1 - \hat{F}(x))/d}} \tag{5.1}$$

is computed [Donoho and Jin, 2008, 2009].

- Third, the HC statistic $\widehat{HC}^{\star}$ is obtained as the maximum of the empirical HC objective

$$\widehat{HC}^{\star} = \max_i \widehat{HC}(p_{(i)}) = \widehat{HC}(x^{\text{HC}}).$$

- Finally, the maximizing argument x^{HC} is taken as the HC decision threshold for signal identification. As shown in Fig. **5.1a**, all $p_i < x^{\text{HC}}$ are considered "significant" and likely correspond to non-null cases.

Informally, the empirical HC objective function $\widehat{HC}(x)$ may be interpreted as z–scores constructed from p-values — recall that $\text{Var}(\hat{F}(x)) = F(x)(1 - F(x))/d$. Indeed, it is this second–level assessment of the p-values that was the original motivation for the HC approach [Tukey, 1976] and that gave rise to its name "Higher Criticism".

5.1.2 Population HC Objective Function and Goodness–of–Fit Statistics

By definition, p-values have a uniform $U(0,1)$ null distribution with $F_0(x) = x$. Moreover, the marginal distribution of the p-values may be viewed as a two-component mixture

$$F(x) = \eta_0 F_0(x) + (1 - \eta_0) F_A(x)$$

of the null model $F_0(x)$ and an alternative model $F_A(x)$ where $\eta_0 \in [0;1]$ is the proportion of the null model (cf. Eq. **2.11**). With this in mind, the squared empirical HC objective function can be written as

$$\widehat{HC}(x)^2 \propto \frac{(\hat{F}_A(x) - F_0(x))^2}{\hat{F}(x)(1 - \hat{F}(x))}.$$

The proportionality factor $d(1 - \eta_0)^2$ has been left out as it does not depend on x and hence is irrelevant for determining the decision threshold x^{HC}. Thus, for maximization

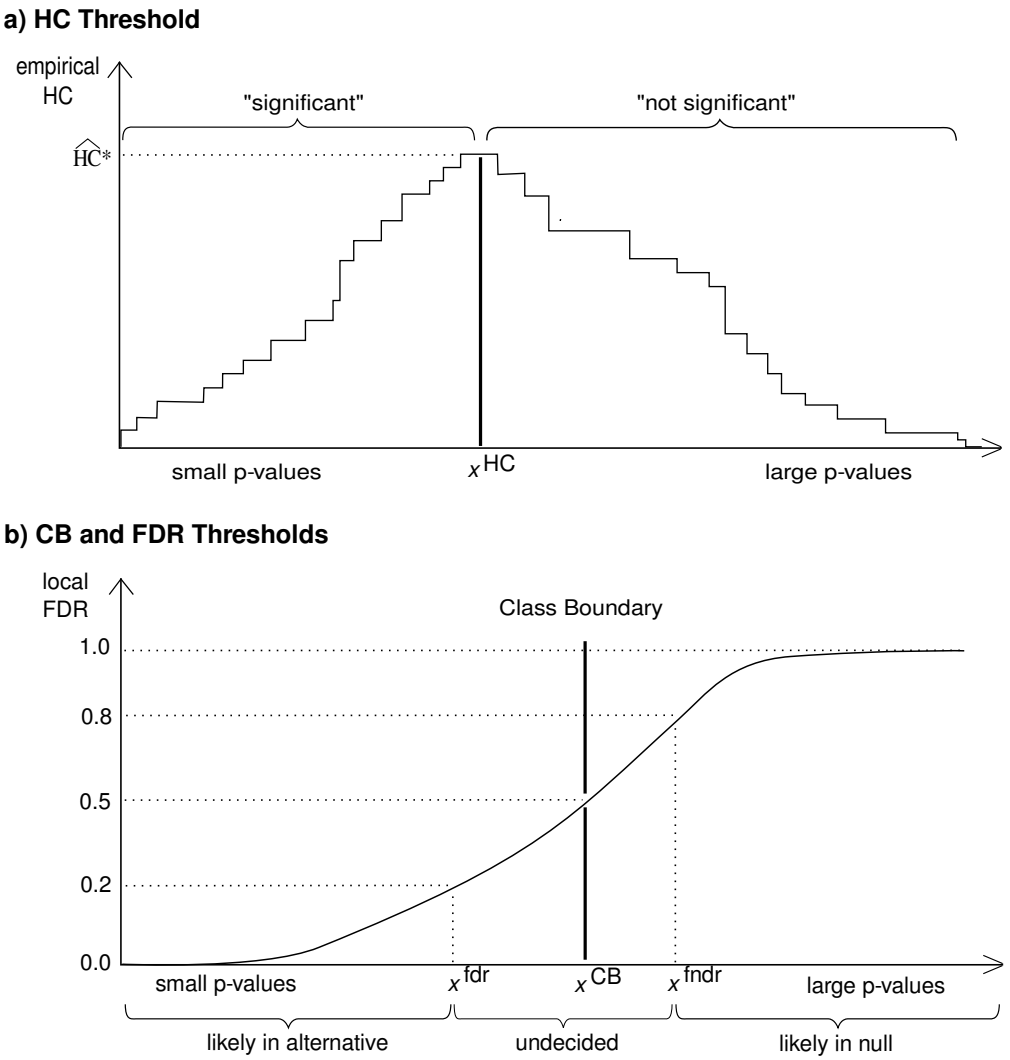


Figure 5.1: (a) Empirical HC decision threshold x^{HC} obtained by maximizing the empirical HC objective, and (b) Class boundary threshold x^{CB} given by fdr = 1/2 and its relationship to the neighboring fdr and fndr thresholds.

Table 5.1: Relationship of HC statistic with other goodness-of-fit statistics.

	Supremum	Expectation
Not standardized	Kolmogorov-Smirnov: $\sup_x \lvert F_A(x) - F_0(x)\rvert$	Cramér-von Mises: $\mathrm{E}_F\{(F_A(X) - F_0(X))^2\}$
Standardized	Higher Criticism: $\sup_x \{\frac{\lvert F_A(x)-F_0(x)\rvert}{\sqrt{F(x)(1-F(x))}}\}$	Anderson-Darling: $\mathrm{E}_F\{\frac{(F_A(X)-F_0(X))^2}{F(X)(1-F(X))}\}$

we can use the above formula rather than Eq. **5.1**. Furthermore, it has the advantage of immediately generalizing to *population level* (i.e. to $d \to \infty$)

$$HC(x)^2 \propto \frac{(F_A(x) - F_0(x))^2}{F(x)(1 - F(x))}, \tag{5.2}$$

which greatly facilitates the conceptual understanding of the HC approach.

The function Eq. **5.2** is well known from the goodness-of-fit statistic of Anderson and Darling [1954], which is proportional to the expectation $\mathrm{E}_F(HC(X)^2)$. Hence, the HC statistic bears the same relationship to the Anderson-Darling statistic as does the Kolmogorov-Smirnov statistic to the Cramér-von Mises statistic [Darling, 1957]. Moreover, as can be seen in Tab. **5.1** the HC statistic is the standardized Kolmogorov-Smirnov (KS) statistic. In fact, the KS statistic may be used in the same fashion as HC to derive a decision threshold x^{KS}.

In the mixture model for p-values it is commonly assumed (see also the next section on false discovery rates) that $F_A(x) \geq F_0(x)$ for all x, i.e. that the alternative component is stochastically smaller than or equal to the null component (cf. section 2.2.2). Thus, on the population level (though not on the sample level) we may leave out the absolute value signs in the first column of Tab. **5.1**.

5.1.3 Invariance of HC Objective Function

By the inspection of Eq. **5.2**, we derive a number of interesting properties of the HC objective function.

First, it is completely symmetric with regard to the two components in the underlying mixture model for the p-values. The alternative model F_A and the null model F_0 play the same role in Eq. **5.2**.

Second, for computing the HC objective it is not necessary to explicitly specify the null

proportion η_0.

Third, Eq. **5.2** is invariant against transformation of the underlying test statistic. This can be seen as follows: Under a change of variables from x to $y = y(x)$ the distribution function changes according to

$$F^Y(y) = \begin{cases} F\Big(x(y)\Big) & \text{for increasing } x(y), \text{ and} \\ 1 - F\Big(x(y)\Big) & \text{for a decreasing transformation.} \end{cases}$$

Applied to Eq. **5.2**, this leads to

$$HC(y)^2 = HC(y(x))^2 \propto \frac{(F_A^Y(y) - F_0^Y(y))^2}{F^Y(y)(1 - F^Y(y))}.$$

Remarkably, the HC objective function Eq. **5.2** retains its functional form under a change of variables. Thus, Eq. **5.2** is *not* constrained to p-values only and may instead be applied to any test statistic y without the need of prior conversion to the p-value scale. The HC decision threshold as the location of the maximum of Eq. **5.2** transforms accordingly from x^{HC} to $y^{\text{HC}} = y(x^{\text{HC}})$.

5.2 Signal Identification with FDR and FNDR

A standard approach to obtain a decision threshold with the FDR is to refer to the rule of Benjamini and Hochberg [1995] with a cutoff such as $\widehat{\text{Fdr}}(p_{(i)}) \leq 0.05$. Alternatively, a threshold may be found by controlling the local FDR, for instance by requiring $\widehat{\text{fdr}}(p_{(i)}) \leq 0.2$ — cf. section 4.3.3.

This ensures that the identified features are mostly from the alternative with only a little contamination by unwanted null features. Conversely, if the interest is to identify true null features, then similar thresholds may be imposed on the FNDR rather than the FDR.

This is illustrated for the fdr and the fndr in Fig. **5.1b** where the signal space is divided by the decision thresholds x^{fdr} and x^{fndr} into three distinct zones corresponding to areas where one is very sure about membership to the null (local FNDR < 0.2 or local FDR > 0.8) or to the alternative (local FDR < 0.2) and one additional intermediate region.

From a classification perspective, there exists another threshold — the class boundary (CB) threshold x^{CB} — that provides a natural separation between null and non-null components. At x^{CB}, the probabilities of membership to the alternative component and

to the null component both equal $1/2$. Hence, in terms of the fdr we have

$$\mathrm{fdr}(x^{\mathrm{CB}}) = \mathrm{fndr}(x^{\mathrm{CB}}) = \frac{1}{2}.$$

As can be seen in Fig. **5.1b**, by construction x^{CB} is located in between x^{fndr} and x^{fdr}. From the definition $\mathrm{fdr}(x^{\mathrm{CB}}) = 1/2$ and Eq. **2.5** we obtain the condition

$$\eta_0 f_0(x^{\mathrm{CB}}) = (1-\eta_0) f_A(x^{\mathrm{CB}}) \tag{5.3}$$

for the CB threshold.

5.3 Comparison of CB and HC Decision Thresholds

It is now instructive to study the mutual connections among the various decision thresholds, in particular x^{HC}, x^{KS} and x^{CB}.

5.3.1 Kolmogorov-Smirnov (KS) Decision Threshold

The location x^{KS}, where the Kolmogorov-Smirnov objective function $|F_A(x) - F_0(x)|$ is maximized, is given by

$$f_0(x^{\mathrm{KS}}) = f_A(x^{\mathrm{KS}}). \tag{5.4}$$

Thus, the KS decision threshold coincides with the class boundary x^{CB} if $\eta_0 = 1/2$. Therefore, the KS threshold implicitly assumes that null and non-null components have the same prior probability.

5.3.2 HC Decision Threshold

Using Eq. **5.2**, we may determine the population decision threshold that one tries to estimate by maximizing the empirical HC objective $\widehat{HC}(x)$. This leads to the general condition

$$\begin{aligned} f_0 \left\{2F(1-F) + (F_A - F_0)(1-2F)\,\eta_0\right\} = \\ f_A \left\{2F(1-F) - (F_A - F_0)(1-2F)\,(1-\eta_0)\right\} \end{aligned} \tag{5.5}$$

that must be satisfied by the HC decision threshold x^{HC} (note that in Eq. **5.5** the arguments have been left out for the sake of clarity).

There are two cases where the HC threshold condition simplifies substantially. First, if the null and alternative components are well separated, then $F_A(x^{\mathrm{HC}}) = 1$ and $F_0(x^{\mathrm{HC}}) =$

0 and consequently $F(x^{\mathrm{HC}}) = 1 - \eta_0$ so that Eq. **5.5** reduces to

$$\eta_0 f_0(x^{\mathrm{HC}}) = (1 - \eta_0) f_A(x^{\mathrm{HC}}).$$

Thus, for well-separated null and alternative, the HC threshold is identical to the CB threshold.

Second, if null and alternative components are very close, then $F_A(x^{\mathrm{HC}}) \approx F_0(x^{\mathrm{HC}})$ and Eq. **5.5** becomes

$$f_0(x^{\mathrm{HC}}) = f_A(x^{\mathrm{HC}}),$$

i.e. the HC threshold becomes identical to the KS threshold.

Hence, the HC threshold may be viewed as a compromise between the CB threshold and the KS threshold. This is directly observed in the study of the "rare-weak" model (cf. Tab. **5.2a**).

5.4 The Rare Weak Model

The use of Higher Criticism is particularly advocated in settings where the signal is sparse and weak. This situation is described by the so-called rare weak (RW) model that has been used to study the performance of HC. In the following, I am going introduce the RW model and compare corresponding decision thresholds.

5.4.1 Setup of the RW Model

The RW model is a sparse normal mean mixture model with

$$Z \sim (1 - \epsilon)N(0,1) + \epsilon N(\tau,1). \tag{5.6}$$

Its two parameters $\tau \in [0;\infty]$ and $\epsilon \in [0;1]$ describe intensity and sparsity of the signal, respectively. If ϵ is small, then the non–null features are rare, and likewise if τ is small, then the effect size is weak (hence the name of the model). From this mixture, we observe z–scores $z_1, \ldots, z_d$, which provide the data from which decision thresholds are inferred.

Despite its simplicity, this model is sufficiently rich to study the behavior of signal detection and signal identification methods [Ingster, 1999, Donoho and Jin, 2004, 2008, 2009, Xie et al., 2011]. A generalized RW model with an additional variance parameter in the alternative is discussed in Cai et al. [2011].

A typical scenario where the RW model naturally arises is in classification. For example, consider a two–class setting with means $\boldsymbol{\mu}_1 = \boldsymbol{\mu}$ and $\boldsymbol{\mu}_2 = -\boldsymbol{\mu}$ where $\boldsymbol{\mu} =$

$(\ldots,\mu_0,\ldots,0,\ldots)^T$ is a d-dimensional vector containing either 0 or μ_0 as components, with ϵ describing the proportion of non-zero entries. Further, assume an identity covariance $\boldsymbol{I}_d$ and equal number of observations $n_1 = n_2 = n/2$ from the two classes. Then the corresponding cat-score vector $(1/n_1 + 1/n_2)^{-1/2}(\hat{\boldsymbol{\mu}}_1 - \hat{\boldsymbol{\mu}}_2)$ used for variable selection (cf. sections 4.1 and 4.3.1) simplifies to $\boldsymbol{Z} = \sqrt{n}\hat{\boldsymbol{\mu}}$. The d components of $\boldsymbol{Z}$ follow the RW model of Eq. **5.6** with $\tau = \sqrt{n}\mu_0$. Note the confounding of n and μ_0, so a small number of observations n and large μ_0 gives rise to the same RW model as large sample size n and a small μ_0.

Instead of ϵ and τ, it is sometimes convenient to use the alternative parameterization

$$\beta_\epsilon = -\log(\epsilon)/\log(d)$$

and

$$r_\tau = \left(\frac{\tau^2}{2}\right)/\log(d)$$

with corresponding backtransformations $\epsilon_\beta = d^{-\beta}$ and $\tau_r = \sqrt{2r\log(d)}$.

The motivation to use β instead of ϵ to measure sparsity is that for d observations the smallest possible fraction of the alternative is $1/d$. The change of variables maps $\epsilon \in [\frac{1}{d};1]$ to $\beta \in [0;1]$. A sparse setting in the RW model is characterized by $\beta \in [\frac{1}{2},1]$ or equivalently $\epsilon < d^{-1/2}$. Similarly, the alternative intensity parameter is a map of $\tau \in [0;\sqrt{2\log(d)}]$ to $r \in [0;1]$. As for d observed z–scores, their maximum is bounded in expectation by $\sqrt{2\log(d)}$, an RW model with $r > 1$ contains comparatively well-separated null and alternative components ,whereas in a model with $r < 1$ the signal is weak.

5.4.2 Decision Boundaries for the RW Model

The RW model is simple enough to allow analytical calculations of some decision boundaries.

Using the null and alternative densities $f_0(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$ and $f_A(z) = \frac{1}{\sqrt{2\pi}}e^{-(z-\tau)^2/2}$, respectively, and distribution functions $F_0(z) = \Phi(z)$ and $F_A(z) = \Phi(z-\tau)$, the KS decision threshold (Eq. **5.4**) for the RW model is

$$z^{\mathrm{KS}} = \frac{\tau}{2}.$$

Similarly, the classification class boundary (Eq. **5.3**) simplifies for the RW model to

$$z^{\mathrm{CB}} = \frac{\tau}{2} + \frac{1}{\tau}\log\left(\frac{1-\epsilon}{\epsilon}\right).$$

For $\epsilon = 1/2$, the CB threshold reduces to the KS threshold and, for $\epsilon \leq 1/2$ we have $z^{\mathrm{CB}} \geq z^{KS}$. For fixed ϵ and large enough effect size τ, the second term above also vanishes and hence also leads to the KS threshold. As the proportion of non-null features becomes smaller ($\epsilon \to 0$), the decision threshold moves to infinity ($z^{\mathrm{CB}} \to \infty$). Thus, if $\epsilon = 0$, no feature will be classified as non-null.

For the HC decision threshold, unfortunately, no analytic expression for z^{HC} is available. From the general considerations above (cf. Section 5.3.2), we know that for larger τ the HC threshold approximates the CB threshold, and that both reduce to the KS threshold for $\epsilon = 1/2$. Furthermore, Donoho and Jin [2009, Appendix Eq. 1.1] show that, for the RW model, $\mathrm{fdr}(z^{\mathrm{HC}}) \geq 1/2$. This, together with the monotonicity of the local FDR in the RW model, implies that

$$z^{\mathrm{HC}} \leq z^{\mathrm{CB}}.$$

Thus, in general using the HC decision threshold causes the inclusion of more features than using the CB threshold.

Of particular interest is the behavior of the HC threshold for small values of ϵ. Specifically, if $\epsilon = 0$ and τ is finite, then the HC threshold is also finite. For example, $\epsilon = 0$ and $\tau = 2$ leads to $z^{\mathrm{HC}} \approx 3.35$, which is distinctly different from the class boundary threshold $z^{\mathrm{CB}} \to \infty$. Thus, by construction the HC criterion (and also the KS threshold) encourages false positives in signal identification.

A comparison of the KS, HC, and CB thresholds for some settings of ϵ and τ is given in Tab. **5.2a**. As expected, with increasing τ the HC and CB thresholds become very similar and for $\epsilon = 1/2$ both HC and the CB thresholds reduce to the KS threshold. Thus, the pattern confirms the general relationships of these decision thresholds discussed above.

In addition, in the RW model, there exists a further close link between the HC and CB thresholds. This results from the special structure of the parameter space of the RW model discussed next.

5.4.3 Phase Space of the RW Model

Within the RW model, the behavior of signal detection and identification procedures have been studied extensively. This has led to the remarkable insight that there exist several fundamental boundaries in its phase space that give rise to four distinct regions, as illustrated in Fig. **5.2a**.

Table 5.2: a) Comparison of the KS, HC and CB decision thresholds in the RW model, and b) Analysis of four cancer gene expression data sets with shrinkage discriminant analysis.

a) Comparison of Thresholds

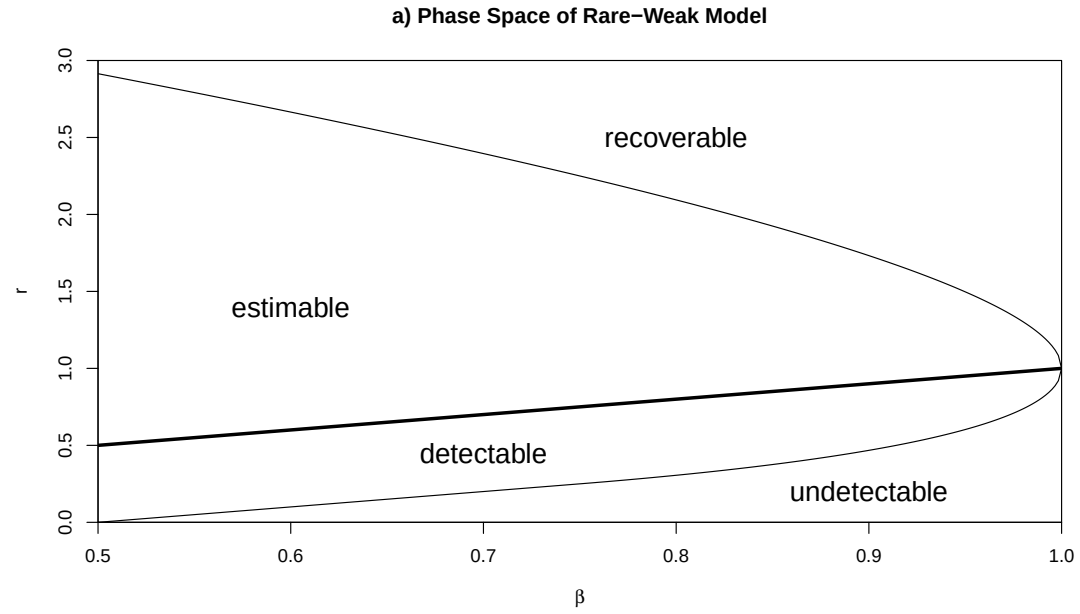
Setting	z^{KS}	z^{HC}	z^{CB}
$\tau = 2$			
$\epsilon = 0$	1	3.3514	∞
$\epsilon = 0.001$	1	3.0707	4.4534
$\epsilon = 0.01$	1	2.5203	3.2976
$\epsilon = 0.1$	1	1.7574	2.0986
$\epsilon = 0.5^*$	1	1.0000	1
$\tau = 4$			
$\epsilon = 0$	2	3.3514	∞
$\epsilon = 0.001^*$	2	3.6377	3.7267
$\epsilon = 0.01^*$	2	3.0965	3.1488
$\epsilon = 0.1^*$	2	2.5268	2.5493
$\epsilon = 0.5^*$	2	2.0000	2
$\tau = 6$			
$\epsilon = 0$	3	8.1607	∞
$\epsilon = 0.001^*$	3	4.1454	4.1511
$\epsilon = 0.01^*$	3	3.7631	3.7659
$\epsilon = 0.1^*$	3	3.3652	3.3662
$\epsilon = 0.5^*$	3	3.0000	3

* Signal identification is possible as $\epsilon \geq \exp(-\tau^2/2)$, see section 5.4.3.

b) Cancer Gene Expression Data

Data / Method	Prediction Error		Selected Variables
Prostate ($d = 6033, n = 102, K = 2$)			
CB	0.0637	(0.0053)	115
HC	0.0497	(0.0045)	116
FNDR	0.0550	(0.0048)	131
Lymphoma ($d = 4026, n = 62, K = 3$)			
CB	0.0211	(0.0042)	178
HC	0.0000	(0.0000)	345
FNDR	0.0036	(0.0018)	392
SRBCT ($d = 2308, n = 63, K = 4$)			
CB	0.0000	(0.0000)	88
HC	0.0007	(0.0007)	174
FNDR	0.0000	(0.0000)	89
Brain ($d = 5597, n = 42, K = 5$)			
CB	0.1633	(0.0120)	78
HC	0.1417	(0.0108)	131
FNDR	0.1525	(0.0120)	102

K: Number of classes in the response variable.



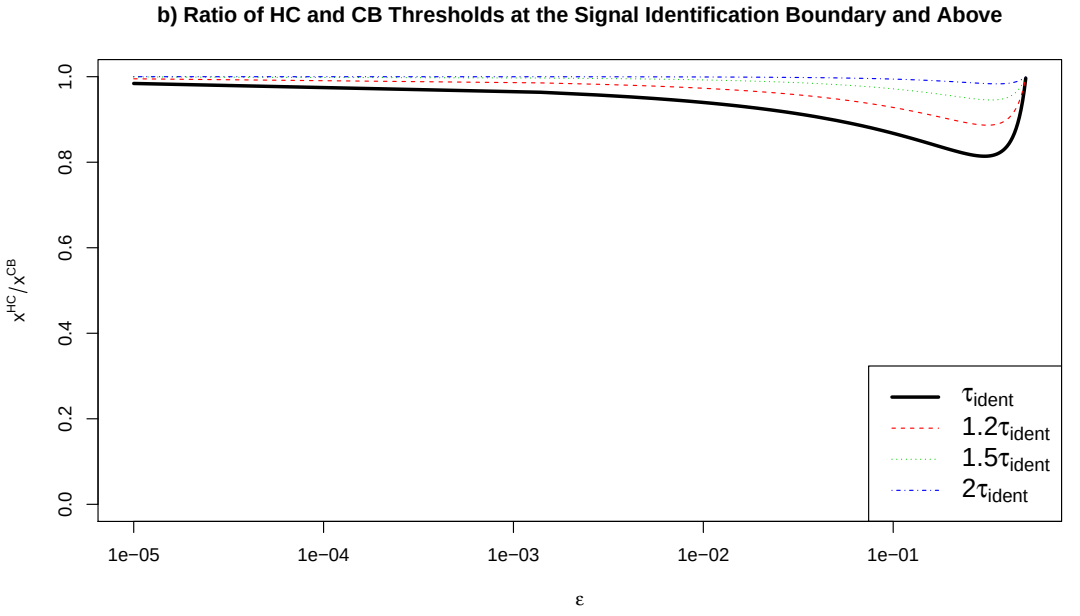


Figure 5.2: a) Phase space of the RW model following Xie et al. [2011]. The bold line shows the signal identification boundary $r_{\text{ident}}(\beta) = \beta$ above which signal identification is possible. For details on the four regions see the description in Section 5.4.3. b) Ratio of x^{HC} and x^{CB} thresholds at the signal identification boundary (solid line) and above (dotted lines). Note that $\tau_{\text{ident}}(\epsilon) = \sqrt{-2\log(\epsilon)}$.

Ingster [1999] discovered the *detection boundary*

$$r_{\text{detect}}(\beta) = \begin{cases} \beta - \frac{1}{2} & \beta \in [\frac{1}{2}; \frac{3}{4}] \\ (1 - \sqrt{1-\beta})^2 & \beta \in [\frac{3}{4}; 1] \end{cases}.$$

Below this boundary lies the "undetectable" region in which even signal detection is impossible, i.e. no method is able to decide whether $\epsilon \neq 0$. Conversely, above the detection boundary it is possible to consistently estimate ϵ [Cai et al., 2007].

Donoho and Jin [2004] report the *identification boundary*

$$r_{\text{ident}}(\beta) = \beta.$$

It is only above this boundary in the "estimable" and "recoverable" regions that signal identification by thresholding is actually possible. In terms of original parameters, this corresponds to the conditions $\tau \geq \sqrt{-2\log(\epsilon)}$ or $\epsilon \geq \exp(-\tau^2/2)$. Directly below this boundary lies the "detectable" region where detection of a signal is possible but not identification. This shows that signal identification is more difficult than signal detection.

Finally, Xie et al. [2011] demonstrated the existence of the *recovery boundary*

$$r_{\text{recov}}(\beta) = (1 + \sqrt{1-\beta})^2$$

above which in the "recoverable" region almost all signal can be completely identified.

5.4.4 HC Threshold as Approximation of the Natural Class Boundary

When comparing the KS, HC, and CB decision thresholds in Tab. **5.2a**, a striking phenomenon can be observed: Whenever signal identification is possible, i.e. if $\epsilon \geq \exp(-\tau^2/2)$, then z^{CB} and z^{HC} are very similar.

To investigate this further, I computed the ratio of the HC and CB threshold directly at the signal identification boundary and above (Fig. **5.2b**). Already at the boundary this ratio is close to 1, especially for small values of ϵ. Moving further into the "estimable" and "recoverable" regions, the differences between the two thresholds become negligible.

Hence, in the RW model, in the area where signal identification is possible, z^{HC} and z^{CB} are in the worst case very similar and mostly indistinguishable for practical purposes.

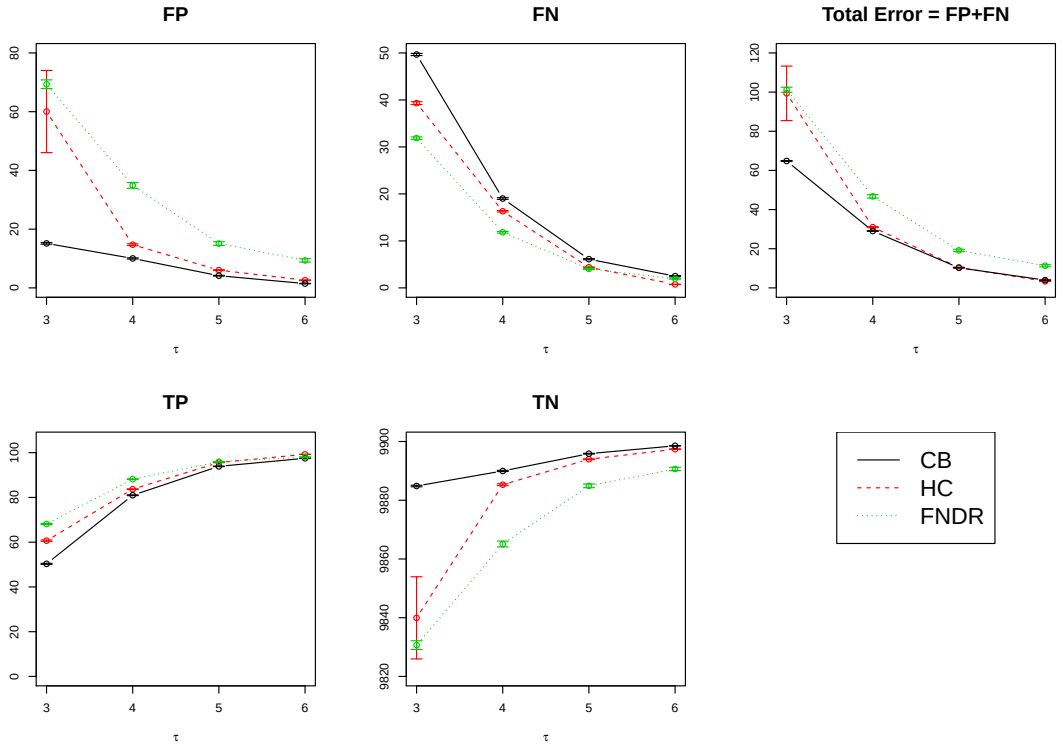


Figure 5.3: Comparison of errors when using the HC, CB and FNDR decision thresholds on data simulated from the RW model located directly at the detection boundary ($\epsilon = 0.01$ and $\tau = 3$) and above ($\tau > 3$).

5.5 Data Examples

To further study the relationship among the HC, CB, and FNDR decision thresholds, I will analyze both simulated as well as experimental data next.

5.5.1 Synthetic Data

The data is simulated from the RW model at the signal identification boundary and above as follows:

1. I sample $d = 10,000$ z–scores from the mixture model Eq. **5.6** with $\epsilon = 0.01$ and $\tau \in \{3,4,5,6\}$. For $\tau = 3$, this is is a sparse and weak scenario located directly at the signal identification boundary ($\epsilon \approx \exp(-\tau^2/2)$).

2. From the test statistics $z_1,\ldots,z_d$ the p–values according to $p_i = 1 - F_0(z_i)$ are computed.

3. Subsequently, the empirical HC threshold is obtained by maximization of Eq. **5.1**.

4. In addition, the fdr was estimated using the `fdrtool` algorithm [Strimmer, 2008a,b] and correspondingly the CB (local FDR = 0.5) and FNDR (local FDR = 0.8) decision thresholds are identified.

5. For each of the three investigated thresholds (HC, CB, FNDR), the number of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN) are determined.

6. The simulations are repeated $B = 1000$ times to estimate the mean errors and their standard deviations.

The results are visualized in Fig. **5.3**. As expected, the HC and CB thresholds yield similar results with growing τ. However, if the signal is weak (small τ), signal identification with HC leads to many more false positives and in addition the variability of the error rates for HC is very large. Conversely, in this situation the CB threshold is more cautious and thus results in more false negatives. For all settings, the error rates of HC are found in between those of CB and FNDR. Interestingly, the total error (FP+FN) is smallest when using the CB threshold.

I also repeated this study with other sparsity settings $\epsilon > 0.01$. The resulting error plots all showed exactly the same pattern of convergence of the CB and and HC methods as Fig. **5.3**.

5.5.2 Gene Expression Data

Next, I will also analyze four clinical gene expression data sets related to prostate cancer [Singh et al., 2002], lymphoma [Alizadeh et al., 2000], small round blue cell tumors (SRBCT) [Khan et al., 2001] and brain cancer [Pomeroy et al., 2002]. In chapter 4, the relative effectiveness of the fndr and HC thresholds to select relevant genes in shrinkage discriminant analysis using cat–scores has already been shown [cf. section 4.5.2].

In Tab. **5.2b**, I show in addition the estimated prediction error and the number of selected variables for the CB threshold. Generally, using the CB decision threshold leads to the smallest predictor sets. Except for the prostate data, the number of selected genes is roughly half compared to using the HC threshold as criterion. As the predictor error is only slightly increased, I conclude that most of the additionally included predictors by HC are false positives.

For practical analysis of gene expression data, this implies that using x^{CB} yields — in comparison with x^{HC} — smaller and hence more interpretable predictor gene sets without compromising prediction error.

6 Summary and Outlook

False discovery rate analysis is a major recent statistical innovation that has found widespread application in the study of high-dimensional data. The FDR estimation methodology introduced in chapter 3 helps to separate signal from noise. Both of them are very often overlapping, making decisions very difficult. However, once a mixture model, composed of a "null" component for the noise, and an "alternative" component that represents the signal, have been fit to the data, false discovery rates allow intuitive and simple signal identification.

The FDR analysis is best understood from an estimation perspective. Then, the incredible number of test statistics encountered in today's multiple testing problems is actually a blessing and not a curse, since it is this high dimension which allows to estimate the FDR from data. Truncated maximum likelihood estimation has been shown to be a powerful approach for estimating the null component, yielding reliable null model parameter estimates [Efron, 2004, 2007b, 2008, Strimmer, 2008b]. By data analysis and simulations (sections 3.6 and 3.7), I have demonstrated that my new approach to truncated maximum likelihood estimation (section 3.1.2) yields accurate null model parameter estimates. This null model estimation is complemented by constrained maximum likelihood estimation for the alternative density using log–concave density estimation. Log–concave density estimation provides a non–parametric, tuning–parameter free and yet very "smooth" estimator for the alternative density. Thus, it has many advantages over the conceptually similar non–parametric Grenander estimator [Strimmer, 2008b].

Since new, even higher dimensional techniques for genome analysis, such as next generation sequencing [Metzker, 2010, NGS], are now routinely used, false discovery rate methods will continue to be of high importance. With next generation sequencing technologies (RNA–Seq), the abundance of these sequences in a sample can be measured directly requiring no prior gene definitions. The results of these measurements are count data, FDR methods developed for continuous data have to be adapted in order to work with this kind of data. With discrete data, the histogram of the p–values

under the null distribution is no longer uniform, making adaptions necessary. One way to do this is the calculation of "randomized p–values" to obtain a continuous and uniform p–value distribution [Muralidharan et al., 2012]. In principle, the classical approaches can then be used again. Unfortunately, this is not the end of the story: Due to various different sources of variation in RNA–Seq, the null model fitting process is much more complicated with RNA–Seq data than with microarrays. Most approaches to the identification of differential expression in so far thus rely on permutation–based methods to infer a null distribution [Anders and Huber, 2010]. However, even if a permutation null is valid for each gene *indidividually*, correlation between genes can still make it unreliable for the ensemble [Efron, 2007b]. Thus, novel empirical null modeling approaches are needed for next generation sequencing.

As indicated in 2.3, mixture models are difficult to fit. Although the truncated maximum likelihood approach combined with a non–parametric alternative works quite well, it is rather ad–hoc from a theoretical point of view. An interesting research direction would be the analysis of mixture models consisting of a parametric null and a non–parametric alternative. Bordes et al. [2006] propose an estimation method for these mixtures. However, no monotonicity constraints are imposed on the alternative density in their paper.

FDR methods are not only useful for multiple testing situations in the strict sense, they are also applicable to related problems. In chapter 4, I looked at several kinds of applications of FDR in linear classification. I presented and extended statistical techniques related to effect size estimation using false discovery rates and showed how to use these for variable selection. The fdr–effect method proposed for effect size estimation has been shown to work as well as competing approaches, while being conceptually simple and computationally inexpensive. Variable selection by minimizing the misclassification rate has been somewhat neglected in the literature but I showed in accordance with Dabney and Storey [2007], Efron [2009] and Matsui and Noma [2011] that it is indeed very well suited for real world problems. In addition, it is also much more intuitive than selecting a non-interpretable regularization parameter, as for example in the PAM algorithm, and leads to compact and interpretable feature sets.

High expectations are associated with the promise of a personalized medicine delivering tailored treatments based on genetic and other information of the patient. In order to develop molecular diagnostics guiding these treatments, statistical approaches for effective and interpretable classification are indispensable. Thus, classification of individuals based on individual information will continue to be an important field of research. The methodology presented in chapter 4 delivers interpretable gene signatures and therefore provides applicability for biological study and medical use. Reliable effect size estimates allow one to identify genes having discriminative power, while variable selection based on these effect size estimates allows the selection of the most important genes for the construction of classification algorithms.

In chapter 5, I took a look at a prominent competitor of FDR, Higher Criticism (HC). Recently, HC was shown to be an effective means for determining appropriate signal identification decision thresholds [Donoho and Jin, 2008, 2009]. Thus, the investigation of the relationship of the HC and FDR methods in chapter 5 started with the aim to better understand HC as a method for signal identification. In the context of variable selection for classification, it had been demonstrated empirically earlier [Ahdesmäki and Strimmer, 2010, cf. section 4.5] that using Higher Criticism thresholding is similar to competing procedures, in particular to those using a threshold based on FDR. In chapter 5, I studied this further and argued that the HC decision threshold may also be viewed as an approximation of the natural class boundary (CB) between the null and alternative groups in the rare–weak (RW) mixture model. This CB threshold can be directly expressed in terms of local FDR and local FNDR. Importantly, in the RW model, in the region of the phase space where signal identification is possible, both thresholds are either very similar or practically indistinguishable.

If the two thresholds are notably different, then using the HC threshold leads to the inclusion of more false positives, and conversely the CB threshold yields a more compact feature set but with slightly increased prediction error. In short, the CB threshold is more cautious than the HC threshold (and the FNDR threshold). Hence, the study in chapter 5 provides further support to the excellent performance of HC for signal identification. However, my conclusions and recommendations are different from those of Donoho and Jin [2008, 2009]. I showed that false discovery rates, properly applied, are indeed perfectly useful for signal identification, which had been disputed earlier. Rather than considering HC as a fundamental criterion, I recommend using the CB threshold for signal identification and suggest employing the HC threshold only in situations where having many false positives is harmless.

In general, estimation of the CB threshold is a challenging problem as this requires the fit of a mixture model and estimation of the mixing density. In contrast, the empirical HC threshold can readily be determined using p-values computed from F_0 alone. Thus, for signal identification the HC approach provides a simple yet effective means to approximate the CB threshold. Due to the difficulties encountered when estimating mixture models, such as lack of identifiably, methods such as HC that allow to infer sensible decision thresholds without full mixture modeling will continue to be of importance for practical data analysis in the future.

A Available Software

Program files to be used with the statistical software R [R Development Core Team, 2012], published under the GNU General Public License 3.0, are available from my homepage:

http://b-klaus.de.

These programs require some additional packages available from *the comprehensive R archive network (CRAN)*, (http://cran.r-project.org) or the Bioconductor platform (http://www.bioconductor.org) which can be easily installed within an R session.

Estimation of FDR: The log–FDR appraoch

The log–FDR estimation approach introduced in sections 3.1.2 and 3.2.2 is implemented in the function log.fdr.R. An example analysis of simulated z–scores can be found in the file log-fdr-example.R.

Variable Selection in Classification

The file CMA-Ana-Singh.R performs cross validation based prediction error estimation for the Singh et al. [2002] prostate cancer gene expression data. In this sample script misclassification rate based variable selection is used. Other variable selection schemes are implemented in the file predfun-CMA.R. The CV based prediction error estimation itself is implemented in the file predfun-CMA.R. This function uses a data split procedure implemented in the Bioconductor package CMA [Slawski et al., 2008].

Higher Criticism

The Higher Criticism statistic can be computed with the function hc.score of the R package fdrtool available from the CRAN archive (Version 1.2.10 or later).

B Article Abstracts

Thresholding methods for feature selection in genomics: higher criticism versus false non-discovery rates
Klaus and Strimmer [2010]

In high-dimensional genomic analysis it is often necessary to conduct feature selection, in order to improve prediction accuracy and to obtain interpretable classifiers. Traditionally, feature selection relies on computer-intensive procedures such as cross-validation. However, recently two approaches have been advocated that both are computationally more efficient: False Non-Discovery Rates (FNDR) and Higher Criticism (HC). Here, we describe the rationale behind the two approaches, conduct an empirical comparison based on synthetic and real data, and discuss the respective merits of HC-based and FNDR-based feature selection.

Learning false discovery rates by fitting sigmoidal threshold functions
Klaus and Strimmer [2011]

False discovery rates (FDR) are typically estimated from a mixture of a null and an alternative distribution. Here, we study a complementary approach proposed by Rice and Spiegelhalter (2008) that uses as primary quantities the null model and a parametric family for the local false discovery rate. Specifically, we consider the half-normal decay and the beta-uniform mixture models as FDR threshold functions. Using simulations and analysis of real data we compare the performance of the Rice-Spiegelhalter approach with that of competing FDR estimation procedures. If the alternative model is misspecified and an empirical null distribution is employed the accuracy of FDR estimation degrades substantially. Hence, while being a very elegant formalism, the FDR threshold approach requires special care in actual application.

Signal identification for rare and weak features: higher criticism or false discovery rates?

Klaus and Strimmer [2012]

Signal identification in large-dimensional settings is a challenging problem in biostatistics. Recently, the method of higher criticism (HC) was shown to be an effective means for determining appropriate decision thresholds. Here, we study HC from a false discovery rate (FDR) perspective. We show that the HC threshold may be viewed as an approximation to a natural class boundary (CB) in two-class discriminant analysis which in turn is expressible as FDR threshold. We demonstrate that in a rare-weak setting in the region of the phase space where signal identification is possible both thresholds are practicably indistinguishable, and thus HC thresholding is identical to using a simple local FDR cutoff. The relationship of the HC and CB thresholds and their properties are investigated both analytically and by simulations, and are further compared by application to four cancer gene expression data sets.

Effect Size Estimation And Misclassification Rate Based Variable Selection In Linear Discriminant Analysis

Klaus [2012]

Supervised classifying of biological samples based on genetic information, (e.g. gene expression profiles) is an important problem in biostatistics. In order to find both accurate and interpretable classification rules variable selection is indispensable. This article explores how an assessment of the individual importance of variables (effect size estimation) can be used to perform variable selection. I review recent effect size estimation approaches in the context of linear discriminant analysis (LDA) and propose a new conceptually simple effect size estimation method which is at the same time computationally efficient. I then show how to use effect sizes to perform variable selection based on the misclassification rate which is the data independent expectation of the prediction error. Simulation studies and real data analyses illustrate that the proposed effect size estimation and variable selection methods are competitive. Particularly, they lead to both compact and interpretable feature sets.

C Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialen oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, _____
Ort, Datum

_____
Bernd Klaus

Bibliography

M. Ahdesmäki and K. Strimmer. Feature selection in omics prediction problems using cat scores and false non-discovery rate control. *Ann. Appl. Statist.*, 4:503–519, 2010.

A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown, and L. M. Staudt. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000. URL http://dx.doi.org/10.1038/35000501.

U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, 96:6745–6750, 1999.

S. Anders and W. Huber. Differential expression for sequence count data. *Genome Biology*, 11:R106, 2010.

T. W. Anderson and D. A. Darling. A test of goodness of fit. *J. Amer. Statist. Assoc.*, 49: 765–769, 1954.

Y. Benjamini. Comment: Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23:23–27, 2008.

Y. Benjamini. Simultaneous and selective inference: current successes and future challenges. *Biom. J.*, 52:708–721, 2010a.

Y. Benjamini. Discovering the false discovery rate. *J. R. Statist. Soc. B*, 72:405–416, 2010b.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, 57:289–300, 1995.

C. E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60, Rome, 1935.

L Bordes, C Delimas, and P Vandekerkhove. Semiparametric estimation of a two-component mixture model where one component is known. *Scandinavian Journal of Statistics*, 33(4):733–752, 2006.

P. Broberg. A comparative review of estimates of the proportion unchanged genes and the false discovery rate. *BMC Bioinformatics*, 6:199, 2005.

T. T. Cai, J. Jin, and M. G. Low. Estimation and confidence sets for spare normal mixtures. *Ann. Statist.*, 35:2421–2449, 2007.

T. T. Cai, X. J. Jeng, and J. Jin. Optimal detection of heterogeneous and heteroscedastic mixtures. *J. R. Statist. Soc. B*, 73:629–662, 2011.

K-A Lê Cao, S Boitart, and P Besse. Sparse pls discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics*, 12:253, 2011.

A. R. Dabney and J. D. Storey. Optimality driven nearest centroid classification from genomic data. *PLoS ONE*, 2:e1002, 2007.

D. A. Darling. The Kolmogorov-Smirnov, Cramér-von Mises tests. *Ann. Math. Stat.*, 28: 823–838, 1957.

A. P. Dawid. Selection paradoxes of bayesian inference. *Multivariate analysis and its applications (Hong Kong, 1992), volume 24 of IMS Lecture Notes Monogr. Ser. Hayward, CA: Inst. Math. Statist.*, pages 211–220, 1994.

T. Dickhaus. *False Discovery Rate and Asymptotics*. PhD thesis, Heinrich Heine Universität Düsseldorf, 2008.

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

D. Donoho and J. Jin. Higher criticism thresholding: optimal feature selection when useful features are rare and weak. *Proc. Natl. Acad. Sci. USA*, 105:14790–15795, 2008.

D. Donoho and J. Jin. Feature selection by higher criticism thresholding achieves the optimal phase diagram. *Phil. Trans. R. Soc. A*, 367:4449–4470, 2009.

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

L. Dümbgen and K. Rufibach. logcondens: computations related to univariate log-concave density estimation. *Journal of Statistical Software*, 39:1–28, 2011.

L. Dümbgen and K. Rufibach. Maximum likelihood estimation of a log-concave density and its distribution function: basic properties and uniform consistency. *Bernoulli*, 15: 40–68, 2009.

B. Efron. Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *J. Amer. Statist. Assoc.*, 99:96–104, 2004.

B. Efron. Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.*, 102:93–103, 2007a.

B. Efron. Size, power and false discovery rates. *Ann. Applied Statistics*, 35:1351–1377, 2007b.

B. Efron. Microarrays, empirical Bayes, and the two-groups model. *Statist. Sci.*, 23:1–22, 2008.

B. Efron. Empirical Bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.*, 104:1015–1028, 2009.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher. Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.*, 96:1151–1160, 2001.

H. Finner, T. Dickhaus, and M. Roters. On the false discovery rate and an asymptotically optimal rejection curve. *Annals of Statistics*, 37:596–618, 2009.

C. Genovese and L. Wassermann. Operating characteristics and extensions of the false discovery rate procedure. *J. R. Statist. Soc. B*, 64:499–517, 2002.

U. Grenander. On the theory of mortality measurement, part ii. *Skan. Aktuarietidskr.*, 39: 125—153, 1956.

Y. Guo, T. Hastie, and T. Tibshirani. Regularized discriminant analysis and its application in microarrays. *Biostatistics*, 8:86–100, 2007.

D. J. Hand. Classifier technology and the illusion of progress. *Statist. Sci.*, 21:1–14, 2006.

J. Hausser and K. Strimmer. Entropy inference and the James-Stein estimator, with application to nonlinear gene association networks. *J. Mach. Learn. Res.*, 10:1469–1484, 2009.

F. Ibanez. Sur une nouvelle application de la théorie de l'information à la description des séries chronologiques planctoniques. *J.Plankton res.*, 4:619–632, 1982.

Y. I. Ingster. Minimax detection of a signal for l_n^p balls. *Math. Methods. Statist.*, 7:401–428, 1999.

L. Jager and J. A. Wellner. Goodness-of-fit tests via phi-divergences. *Ann. Statist.*, 35: 2018–2053, 2007.

H. K. Jankowski. Maximum likelihood estimation under shape constraints. Teaching manuscript, 2009. URL http://www.math.yorku.ca/~hkj/Teaching/Bristol/notes.pdf.

J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Med.*, 7:673–679, 2001.

I K Kim and R Simon. Probabilistic classifiers with high-dimensional data. *Biostatistics*, 12:399 – 412, 2011.

B. Klaus. Effect size estimation and misclassification rate based variable selection in linear discriminant analysis. *ArXiv Preprint*, 2012. URL http://arxiv.org/abs/1205.6653.

B Klaus and K Strimmer. Thresholding methods for feature selection in genomics: higher criticism versus false non-discovery rates. In *Proceedings of the 7th International Workshop on Computational Systems Biology, WCSB 2010*, pages 59–62, 2010.

B Klaus and K Strimmer. Learning false discovery rates by fitting sigmoidal threshold functions. *Journal de la Société Française de Statistique*, 152, No. 2:39–50, 2011.

B Klaus and K Strimmer. Signal identification for rare and weak features: higher criticism or false discovery rates? *Biostatistics*, in press, 2012. doi: 10.1093/biostatistics/kxs030.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to DNA microarray data. *J. R. Statist. Soc. B*, 67:565–572, 2005.

Lucien Le Cam. Likelihood: An introduction. *International Statistical Review*, 58:153–177, 1990.

S Matsui and H Noma. Estimation and selection in high-dimensional genomic studies for developing molecular diagnostics. *Biostatistics*, 12:223–233, 2011.

G. J. McLachlan, R. W. Bean, and L. B.-T. Jones. A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics*, 22:1608–1615, 2006.

M L Metzker. Sequencing technologies – the next generation. *Nature Rev. Genet.*, 11: 31–46, 2010.

O. Muralidharan. An empirical Bayes mixture model for effect size and false discovery rate estimation. *Ann. Applied Statistics*, 4:422–438, 2010.

O. Muralidharan, G. Natsoulis, J. Bell, H. Ji, and N.R. Zhang. Detecting mutations in mixed sample sequencing data using empirical Bayes. *Ann. Appl. Statist.*, in press, 2012.

R. Opgen-Rhein and K. Strimmer. Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Statist. Appl. Genet. Mol. Biol.*, 6:9, 2007.

H Pang, T Tong, and H Zhao. Shrinkage-based diagonal discriminant analysis and its applications in high-dimensional data. *Biometrics*, 65:1021–1029, 2009.

N. G. Polson and J. G. Scott. Good, great, or lucky? screening for firms with sustained superior performance using heavy-tailed priors. *Ann. Appl. Statist.*, 6(1):161–185, 2012.

S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, L. M. Sturla, M. Angelo, M. E. McLaughlin, J. Y. H. Kim, L. C. Goumnerova, P. M. Black, C. Lau, J. C. Allen, D. Zagzag, J. M. Olson, T. Curran, C. Wetmore, J. A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D. N. Louis, J. P. Mesirov, E. S. Lander, and T. R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–442, 2002. URL http://dx.doi.org/10.1038/415436a.

S. Pounds and S. W. Morris. Estimating the occurrence of false positives and false negatives in microarray studies by approximating and partitioning the empirical distribution of *p*-values. *Bioinformatics*, 19:1236–1242, 2003.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2012. URL http://www.R-project.org. ISBN 3-900051-07-0.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis. 2nd Edition.* Springer Verlag, New York, 2005.

K. Rice and D. Spiegelhalter. Comment: Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.*, 23:41–44, 2008.

B. Rüger. *Test- und Schätztheorie: Band 2 Statistische Tests*. Oldenbourg Wissenschaftsverlag, 2002.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.*, 4: 32, 2005.

T. Schweder and E. Spjøtvoll. Plots of *p*-values to evaluate many tests simultaneously. *Biometrika*, 69:493–502, 1982.

H. Schwender, K. Ickstadt, and J. Rahnenführer. Classification with high-dimensional genetic data: assigning patients and genetic features to known classes. *Biometr. J.*, 50: 911–926, 2008.

S. Senn. A note concerning a selection "paradox" of Dawid's. *The American Statistician*, 62:206–210, 2008.

J. Shaffer. Multiple hypothesis testing: a review. *Ann. Rev. Psychol.*, 46:561–584, 1995.

Jun Shao, Yazhen Wang, Xinwei Deng, and Sijian Wang. Sparse linear discriminant analysis by thresholdig for high dimensional data. *Annals of Statistics*, 39:1241–1265, 2011.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1:203–209, 2002.

M. Slawski, M. Daumer, and A.-L. Boulesteix. CMA - a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC Bionformatics*, 9:439, 2008.

G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Statist. Appl. Genet. Mol. Biol.*, 3:3, 2004.

J. D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Ann. Statist.*, 31:2013–2035, 2003.

J. D. Storey and R. Tibshirani. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA*, 100:9440–9445, 2003.

K. Strimmer. fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, 24:1461–1462, 2008a.

K. Strimmer. A unified approach to false discovery rate estimation. *BMC Bioinformatics*, 9:303, 2008b.

R. Tibshirani, T. Hastie, B. Narsimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Statist. Sci.*, 18:104–117, 2003.

J. W. Tukey. T13 N: the higher criticism. Course Notes, Statistics 411, Princeton Univ., 1976.

B. B. Turnbull. Optimal estimation of false discovery rates. Technical report, Stanford University, 2007. URL http://www.stanford.edu/~bkatzen/optimal-FDR.pdf.

Angelique B van 't Wout, Ginger K Lehrman, Svetlana A Mikheeva, Gemma C O'Keeffe, Michael G Katze, Roger E Bumgarner, Gary K Geiss, and James I Mullins. Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4(+)-T-cell lines. *J Virol*, 77(2):1392–1402, 2003.

G. Walther. Inference and modeling with log-concave distributions. *Statist. Sci.*, 24: 319—-327, 2009.

D. M. Witten and R. Tibshirani. Penalized classification using Fisher's linear discriminant. *J. R. Statist. Soc. B*, 73:753–772, 2011.

W Xiaosheng and R Simon. Microarray-based cancer prediction using single genes. *BMC Bioinformatics*, 12:391, 2011.

J. Xie, T. T. Cai, and H. Li. Sample size and power analysis for sparse signal recovery in genome-wide association studies. *Biometrika*, 98:273–290, 2011.

V. Zuber and K. Strimmer. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25:2700–2707, 2009.