Multiple Testing Procedures under Dependence, with Applications

Alessio Farcomeni

November 2004

Dottorato di ricerca in Statistica Metodologica
Dipartimento di Statistica, Probabilità e Statistiche Applicate
Università di Roma "La Sapienza"

Commissione Esaminatrice:
Prof. Domenico Marinucci (Univ. Tor Vergata), Enzo Orsingher (Univ. "La Sapienza"), Silvia Terzi (Univ. Roma3)

To Orbelina
and
Nicko McBrain

Contents

Foreword

Multiple hypothesis testing is concerned with maintaining low the number of false positives when testing several hypotheses simultaneously, while achieving a number of false negatives as small as possible. It will become clear that a multiple testing situation presents many substantial differences with the single hypothesis setting.

As noted in Bayarri and Berger (2004), there are quite a few distinct and competing methodologies to deal with multiple tests. We will not attempt to unify them here, but rather argue that this diversity is a tool to the researcher, who should know the properties and behavior of the procedures in the different situations that go under the wide multiple hypotheses framework. It is intuitive that the problems posed by simultaneously testing a hundred thousand hypotheses under independence are different than the problems posed by testing only ten hypotheses under dependence.

In general, anyway, it is critical that procedures for testing many hypotheses simultaneously be distribution free and robust with respect to known or possibly unknown dependency among the test statistics.

The procedures available in the literature can work very well with an unknown distribution of the data, but the problem of robustness with respect to dependence has not been completely solved up to now, apart for the classical error measures. Moreover it is not yet clear which error measure to control in certain cases (Bickel (2004)). Finally, the multiple testing procedures devised to control the most recent Type I error rate, called $tFDP(c)$, are not efficient in certain cases (in particular when the number of tests is very large), as we will see in Chapter 1. $tFDP(c)$ is defined as the probability that the proportion of false positives on the number of rejections exceeds a fixed threshold c. Among other open problems, which will not be addressed in this dissertation, there is the derivation of a framework for power analysis; and a closely related problem, that is, a method to choose the sample size for each test.

The present dissertation mainly deals with dependence in multiple testing. The primary goal is to provide sufficient conditions on the dependency between the test statistics in order to use the procedures under dependence *without any correction*. As additional results, and when the primary goal will not be met, extension to dependence will be given with suitable modifications.

Main contributions of this dissertation are reported below.

- A new procedure to control the $tFDP(c)$ is provided. Innovations with respect to existing methods are shown.

- Dependency among the test statistics is considered. We show what kind of considerations are to be made on the dependence, which procedure to use in light of the considerations, and what modifications may be needed.

  We will mainly impose "weak dependence" conditions (like mixing conditions) or conditions on the "direction" of the dependence, i.e., that it is either all positive or all negative (association conditions). More general conditions are also discussed.

- A family of estimators of the number of the false null hypotheses is proposed. These estimators are shown to be robust with respect to dependence. Power of the procedures is increased via such estimators.

- Application to DNA microarrays, a common motivating example, is discussed.

- An application of multiple testing procedures to wavelet thresholding is discussed, with suggestions of when they prove more efficient than classical thresholding methods.

- Some inequalities of interest, like Hoeffding and Vapnik-Cervonenkis inequality, are derived under dependence.

The dissertation is organized as follows: In Chapter 1 we will review the existing literature on multiple testing, together with a critical comparison of the results. Subsequent chapters will describe our contributions: in Chapter 2 we will propose a new multiple testing procedure and show how it solves part of the open problems related to $tFDP(c)$ control. Asymptotic results for multiple testing under dependence will be given in Chapter 3, together with extensive simulations. Some results for multiple testing under dependence for fixed number of tests will be given in Chapter 5. In Chapter 4 we will introduce a family of estimators of the number of false nulls among the hypotheses, and illustrate how to use these estimators to increase the power of multiple testing procedures. Effects of dependency on such a family is discussed. Finally, Chapter 6 will show some applications.

There are many possibilities of further work, which will be usually pointed out through the exposition.

Chapter 1

Introduction

1.1 Notation

The following summarizes the most recurring notation.

Symbol	Description
$\lvert \cdot \rvert$	cardinality of a set
$1_{\{A\}}$	indicator function, 1 if condition A is true, 0 otherwise
p_j	j-th p-value
$p_{(j)}$	j-th ordered p-value
H_j	indicator of the j-th null hypothesis to be false
m	number of tests
M_0	number of true null hypotheses
M_1	number of false null hypotheses
a	M_1/m
S_0	set of indexes of true null hypotheses
T_{Pr}	threshold (reject if $p_j < T_{Pr}$) under procedure Pr or controlling error rate Pr
$Er(T_{Pr})$	value of the error rate using threshold T_{Pr}
R_{Pr}	number of rejected hypotheses under procedure Pr or controlling error rate Er
F	distribution of $p_j \lvert H_j = 1$, $j = 1, \ldots, m$
G	marginal distribution of p_j, $j = 1, \ldots, m$
U	uniform CDF
Γ	number of false rejections divided by R_{Pr}, FDP process

1.2 Motivation and Outline

In many new areas of statistics, in particular in bioinformatics, conclusions are drawn by testing hundreds, often thousands, of hypotheses simultaneously. This can be the case of identifying the spots of the brain where there is neuronal activity after a stimulus (Worsley *et al.* (1996), Ellis *et al.* (2000), Merriam *et al.* (2003)) or the identification of differentially expressed genes in DNA microarray experiments (Drigalenko and El-

ston (1997), Weller *et al.* (1998), Heyen *et al.* (1999), Bovenhuis and Spelman (2000), Mosig *et al.* (2001))). For a review of multiple testing methods in the context of microarray data analysis see Dudoit *et al.* (2003a). Among the other possible applications, there are medicine (Khatri *et al.* (2001)), pharmacology (Schlaeppi *et al.* (1996)), epidemiology (Ottenbacher (1998)), marketing (Schaffer and Green (1998)), psychometrics (Vedantham *et al.* (2001)).

Moreover, multiple tests are often used as a key part of another statistical procedure, like variable selection (George (2000), George and Foster (2000)), item-response modeling (Ip (2001)), structural equation modeling (Green and Babyak (1997)), decision trees (Benjamini and Yekutieli (2002)), etc.

The procedures we review here, and the new ones we will propose, are useful and valid for any of these applications, and some of them we will consider throughout. Among them, a particularly motivanting example is given by the kind of problems faced in bioinformatics, in particular DNA microarray experiments. Bioinformatics and genetics applications are usually characterized by a number of variables (or occasions) much larger than the sample size, with a complex correlation pattern and possibly unknown (joint) distributions. Most recent DNA microarray experiments measure the expression levels of around 30000 genes, with a sample size (number of individuals) that is almost always lower than a hundred. One usually tests one hypothesis per gene, to verify if it is differentially (over or under) expressed between two groups (case-control studies).

Multiple hypothesis methods are concerned with the problem of simultaneously testing all those hypotheses, controlling a suitably defined Type I error rate; and maximizing the number of correct rejections at the same time. The best methods are the ones that take advantage of the large number of hypotheses by efficiently using the information given by one test to make the others. This will be usually done by estimating the marginal distribution of the test statistics. It is well known that the procedures must be distribution free, since the distribution of the test statistics under the alternative hypothesis is usually unknown (and possibly different among tests). Fortunately, they usually are. It is all the same desired that the procedures be robust under known or unknown dependence structure. Variables, and hence test statistics, are often dependent; and usually there is no design of the experiment that can yield independent test statistics. For instance in neurology, different neurons are dependent by their intrinsic nature. For this reason, it is crucial to have also procedures that are robust under flexible (or even arbitrary) dependence structures. In this dissertation, we will sometimes give sufficient conditions on the dependency for a certain procedure to work; other times we will introduce procedures that work under arbitrary dependence. The researcher should choose the procedure that works best under the conditions he/she can prove to be true. It is obvious that, in general, more general hypotheses lead to more conservative procedures.

Many attempts have been made to tackle the problem of robustness of multiple tests under dependence, or to extend the available ones. Among the others see Benjamini and Yekutieli (2001), van der Laan *et al.* (2003b), Storey and Tibshirani (2001), Pollard and

	H_0 **not rejected**	H_0 **rejected**	**Total**
H_0 **True**	$N_{0\|0}$	$N_{1\|0}$	M_0
H_0 **False**	$N_{0\|1}$	$N_{1\|1}$	M_1
Total	$m - R$	R	m

Table 1.1: Categorization of the outcome

van der Laan (2003b), Yekutieli and Benjamini (1999), Sarkar (2002). We will briefly review these works below.

This dissertation, in particular, is dedicated to a recent and, at the moment, most used error measure: the false discovery proportion (FDP), loosely defined as the proportion of false positives with respect to the number of rejections. Next section introduces a framework for multiple hypothesis testing, with a review of the methods under dependence and independence assumptions, and few new results. Chapter 2 presents a procedure to control the tail of the FDP under independence which proves better in terms of power, and more flexible, than others available in the literature. Chapter 3 shows asymptotic results on the control under dependence of the expected value of the FDP, the FDR. Chapter 4 proposes a generalized iterative estimation method for the true number of false nulls, which can improve significantly the power of the procedures. Chapter 5 shows results on the control under dependence of the quantiles of the FDP and of the FDR for any finite number of tests. Finally, Chapter 6 presents applications of the procedures.

1.3 The Multiple Hypotheses Framework

Consider a multiple testing situation in which m tests are being performed. Suppose M_0 of the m hypotheses are true, and M_1 are false. Table 1.1 shows a categorization of the outcome of the tests. R is the number of rejections. $N_{i|j}$, with $i, j \in \{0, 1\}$, is the number of H_i accepted when H_j is true. Note that $N_{0|1}$ and $N_{1|0}$ denote the exact (unknown) number of errors made after testing. All quantities in capital letters are random, or at least not observable. To observe the realization of all these quantities, apart from R, one would need the knowledge of which hypotheses are actually true and which are actually false.

In the usual (single) test setting, one controls the probability of false rejection (Type I error) while looking for a procedure that possibly minimizes the probability of observing a false negative (Type II error). In the multiple case, there are a variety of possible generalizations for the Type I error rate, all involving the counts of false positives $N_{1|0}$. In the same way, one can generalize the concept of Type II error. From a frequentist point of view, one still wants to get a sufficiently small $N_{1|0}$ while rejecting the maximum number of hypotheses, i.e. minimizing the number of false negatives. All classical multiple Type I error measures are based only on the distribution of $N_{1|0}$, i.e., on what

happens for the tests corresponding to the true null hypotheses :

- *Family-wise error rate* (FWER), the probability of a least one Type I error:

$$FWER = \Pr(N_{1|0} \geq 1) \tag{1.1}$$

- *Generalized family-wise error rate* (FWER), the probability of a least $k+1$ Type I errors[1], $k = 0, \ldots, m-1$:

$$gFWER(k) = \Pr(N_{1|0} \geq k+1) \tag{1.2}$$

Up to few years ago, FWER was by far the most used error measure. Define now the False Discovery Proportion (FDP) to be the proportion of erroneously rejected hypotheses:

$$FDP = \begin{cases} \frac{N_{1|0}}{R} & \text{if } R > 0 \\ 0 & \text{if } R == 0 \end{cases} \tag{1.3}$$

Benjamini and Hochberg (1995) propose a new error measure, the FDR, based on the FDP, that hence depends also on the distribution of R, i.e., on what happens for the hypotheses for which H_0 is false. Perone Pacifico *et al.* (2003) and van der Laan *et al.* (2003b) propose another similar error measure:

- *FDR*, expected proportion of Type I errors:

$$FDR = E(FDP) \tag{1.4}$$

- *Exact control of the FDP*, tail probability of the FDP:

$$tFDP(c) = \Pr(FDP > c). \tag{1.5}$$

This is a slight contamination of concepts: failing to reject an hypothesis increases the Type I error measure. The justification of this stems from classification theory and a simple reasoning: the researcher is prepared to accept an higher number of Type I errors when many rejections are made; since the false positives will likely not bias the conclusion of the analysis.

It is easy to see (Benjamini and Hochberg (1995), Genovese and Wasserman (2002)) that FDR control is also a loose control on the FWER, and it is FWER control when $R = 1$.

[1]Note that $gFWER(k) = gFWER(k')$ for $k \geq m_0 - 1$ and $k' \geq m_0 - 1$.

Relationship between FDR and tFDP control

As noted by Genovese and Wasserman (2004a) any $tFDP$ controlling procedure can be easily modified to control the FDR: in fact, if $c \in (0, \alpha)$ and $tFDP(c) \leq \frac{\alpha - c}{1 - c}$, then $FDR \leq \alpha$. It is easily seen that, in general, if $tFDP(c) \leq \alpha$,

$$FDR < c + (1 - c)\alpha \tag{1.6}$$

A partial converse is given by an application of Markov inequality, which yields: $FDR < \alpha \Rightarrow tFDP(c) < \alpha/c$. Another partial converse follows from next Lemma, which is a simple extension of the mean value theorem:

Lemma 1.3.1. *Let $f(\cdot)$ be a function $f : \mathcal{R} \to \mathcal{R}$, continuous, integrable, and monotone on an interval $[a, b]$. If $f(\cdot)$ is concave in $[a, b]$, then there exists $\xi \in [a, (a + b)/2]$ such that $f(\xi)(b-a) = \int_a^b f(x)\, dx$. If $f(\cdot)$ is convex in $[a, b]$, there there exists $\xi \in [(a+b)/2, b]$ such that $f(\xi)(b - a) = \int_a^b f(x)\, dx$.*

Proof. The well known mean value theorem of calculus states that if $f(\cdot)$ is continuous and integrable in the interval $[a, b]$, then there exists $\xi \in [a, b]$ such that $f(\xi)(b - a) = \int_a^b f(x)\, dx$. First, it is straightforward to prove, by contradiction, that if $f(\cdot)$ is strictly monotone then ξ is unique. If $f(\cdot)$ is only monotone, there there exists at most an interval $[\xi_1, \xi_2]$ such that the equality holds for all points in the interval. If moreover $f(\cdot)$ is concave, we have an easy application of Jensen inequality for integrals to its (convex) inverse:

$$\begin{aligned} \xi &= f^{-1}\left(\frac{1}{b-a}\int_a^b f(x)\, dx\right) \\ &\leq \frac{1}{b-a}\int_a^b f^{-1} \cdot f(x)\, dx \\ &= \frac{1}{b-a}\int_a^b x\, dx = \frac{b+a}{2}, \end{aligned}$$

i.e., the point(s) ξ are all in the first half of the interval. If $f(\cdot)$ is convex, it is straightforward to see the reverse inequality. □

Now we are able to prove that, for some c, $tFDP(c)$ is controlled at the same level as the FDR:

Theorem 1.3.2. *If $\Pr(FDP < c)$ is convex[2] as a function of c, then if $FDR \leq \alpha$ we have that $tFDP(c) \leq \alpha$ for any $c > 0.5$.*

Proof. If $\Pr(FDP < c)$ is convex, then $\Pr(FDP > c) = tFDP(c)$ will be concave. Since $E[FDP] = \int_0^1 tFDP(c)\, dc$, we have that $\exists\ \xi$ such that $tFDP(\xi) = E[FDP]$. By Lemma 1.3.1, hence, if $FDR \leq \alpha$ then $tFDP(0.5) \leq \alpha$. It is obvious that $tFDP(0.5) \geq tFDP(c)$ for any $c > 0.5$. □

[2]This is a safe assumption if one takes the stochastic dominance assumption defined at page 10.

As final remarks, note that $FDR = E[FDP] = \int_0^1 tFDP(c)\, dc$; which implies that FDR control is a control on the expected $tFDP(c)$ with respect to Lebesgue measure. Moreover, it is straightforward to see that $tFDP(0) = FWER$.

1.3.1 The Statistical Model

Let $X_1, \ldots, X_n$ be n random vectors in $\mathcal{X} \subseteq R^m$: $X_i = (X_{ij} : j = 1, \ldots, m)$. Suppose $X_i \sim \mathcal{P}$, where the (multivariate) data generating distribution $\mathcal{P}$ is an element of a family of distributions $\mathcal{M}$. $\mathcal{M}$ depends on a finite or infinite vector of unknown parameters. Thus, we refer to a statistical model $\{\mathcal{X}^n, \sigma_X, \mathcal{P}\}, \mathcal{P} \in \mathcal{M}$; where σ_X is an opportune σ-field on the subsets of the sample space $\mathcal{X}$.

In usual applications, the sample size n is typically much smaller than m. For instance, one can measure the gene expression of thousands of genes for less than one hundred patients, together with biological covariates and risk factors. This obviously complicates the shape and definition of $\mathcal{P}$.

We define a parameter to be a function of the unknown distribution $\mathcal{P}$. Parameters of interest typically include means, differences in means, variances, ratios of variances, regression coefficients, etc.

We define m null hypotheses as the indicator functions of m possible submodels: $1 - H_j = 1_{\{\mathcal{P} \in M_j\}}$, with $M_j \subseteq \mathcal{M}$. The m alternatives will just be defined as the complements: $H_j = 1_{\{\mathcal{P} \notin M_j\}}$, so that $H_j = 0$ if the null hypothesis is true and $H_j = 1$ otherwise. Let S_0 be the set of all true null hypotheses: $S_0 = \{j : H_j = 0\}$. The goal of a multiple testing procedure is the accurate estimation of S_0^c, while probabilistically controlling a pre-specified function of $N_{1|0} = \{S_0 \cap \widehat{S}_0^c\}$, i.e., a Type I error rate.

When $m = 1$, we have a classical statistical test, a procedure which partitions the sample space in two: a subset $\mathcal{X}_1 \subseteq \mathcal{X}^n$ such that observing a realization of $X_1, \ldots, X_n$ in $\mathcal{X}_1$ leads to rejection of the single null hypothesis; and the complementary subset, leading to failure of rejection. This idea is easily generalized to the $m > 1$ case: a multiple test is a procedure which partitions the sample space in 2^m subsets, some of which can be empty. Each subset corresponds to one of the possible 2^m estimates of S_0^c, i.e., to a subset of the set of indexes $\{1, \ldots, m\}$. For obvious reasons, in practice one never directly refers to these 2^m subsets of $\mathcal{X}^n$.

The partitioning is made through an m-vector of *test statistics*, $T_n = (T_n(j) : j = 1, \ldots, m)$, that are functions of the data $X_1, \ldots, X_n$. We define the j-th p-value to be

$$p_j = Pr(|T_n(j)| > |t_n(j)||H_j = 0)$$

where $t_n(j)$ is the observed value of the test statistic $T_n(j)$. Throughout the dissertation, we adopt the notation $p_{(j)}$ to denote the j-th ordered statistic of the vector of p_js, with $p_{(0)} = 0$ and $p_{(m+1)} = 1$. The p-value p_j is again a random variable. It is well known that, if the null hypothesis is simple, it is uniformly distributed on $[0, 1]$ under the null; which we will assume throughout. Until now, the literature on FDR and multiple testing in general does not seem to be interested in extensions to composite null hypotheses. As

an aside, we refer the reader to Bayarri and Berger (2000), where an objective Bayesian approach is used to derive significance levels (i.e., alternative p-values) that are always uniformly distributed under the null. If one makes use of their *partial posterior predictive p-value*, *U-conditional predictive p-value*, or similar alternative p-values; the procedures are immediately and directly extended to the case of composite null hypotheses. See references therein for the problem of calibration of p-values, which is also partly solved with the use of alternative significance levels.

Let $p_j|(H_j = 1) \sim F_j$, where F_j is any distribution on $[0, 1]$. It is easy to see that, in the single test setting, rejection of the hypothesis when $p_j < \alpha$ controls automatically the probability of Type I error at a pre-specified level α.

We will denote the true proportion of false nulls for fixed m, M_1/m, by a. When m changes, namely for asymptotic results, we will always assume $M_1/m \xrightarrow{P} a$. Here M_1 is considered random since it is not known (nor even observable). We will moreover say that $\Pr(H_j = 1) = a$, since if the j-th hypothesis is declared at random to be false[3], the chance that it is actually false is M_1/m.

Then, the marginal distribution of p_j is the mixture $G_j(t) = \Pr(H_j = 0)\Pr(p_j < t|H_j = 0) + \Pr(H_j = 1)\Pr(p_j < t|H_j = 1) = (1 - a)t + aF_j(t)$. We often will write F instead of F_j, and $G(t)$ instead of $G_j(t)$, to simplify the notation. Unless stated otherwise the procedures will not require the distribution of the p_j under the alternative to be equal (in general, they never are). For a discussion of this and more details of the definition of the "marginal model" $G_j(t)$, see for instance Genovese and Wasserman (2002). The empirical distribution of the p-values will be denoted by $\widehat{G}(t)$ and is defined as $\widehat{G}(t) = \frac{1}{m}\sum 1_{\{p_i<t\}}$.

In this dissertation, we will always use multiple testing procedures based on the vector of p-values.

1.3.2 Multiple Testing Procedures

As stated before, a multiple testing procedure (MTP) produces a set S_{MTP} of rejected hypotheses, which is an estimation of the set S_0^c of false null hypotheses. For a fixed MTP, the set S_{MTP} depends on:

- The data, often only through the vector of p-values $p_1, \ldots, p_m$
- A level α, that is, an upper bound for an appropriate Type I error rate,

and it will usually be defined as $S_{MTP} = \{j : p_j \leq T\}$ for the (random) cut-off T. The only problem, then, is to specify a way to determine a (data dependent) cut-off T.

Some authors, e.g. Dudoit *et al.* (2003b) and Westfall and Young (1993), prefer to leave the cut-off T fixed to α and introduce the concept of *adjusted p-value*, i.e., a level of significance related to the entire MTP (instead of a relationship with the single j-th

[3]By "declared at random" we mean that an hypothesis is sampled with a simple random sampling scheme and it is then declared to be false.

test). In practice, this involves defining a new p-value $\tilde{p}_j$, which will be a function (often, a scale transformation) of the old p_j. Then, these authors define $S'_m = \{j : \tilde{p}_j < \alpha\}$. It can be shown that, for each MTP procedure, it is perfectly equivalent to consider adjustment of threshold or of p-values.

We prefer, like other authors (e.g., Genovese and Wasserman (2004b)), not to refer to the concept of adjusted p-values for two reasons: first of all, working on a data dependent cut-off seems more intuitive and direct; secondly, we will make extensive use of the marginal distribution of the p-values, which is not easily tractable in the case of adjustment of the vector of p-values.

MTPs are usually categorized as:

- **One-Step**: In one-step procedures, the p-values are compared to a predetermined cut-off that is only a function of α and m, with no dependence on the data.

- **Step-down**: In step-down procedures, each p-value is compared with a cut-off dependent on its rank. The p-values are examined in order, from smallest to largest. Once one is found to be greater than its cut-off, the threshold T is set equal to the previous p-value. Hence, all hypotheses corresponding to the first p-value greater than its cut-off and to larger ones are not rejected.

- **Step-up**: Step-up procedures are similar to step-down procedures. p-values are examined from the largest to the smallest. Once one is found to be smaller than its cut-off, the hypotheses corresponding to that one and to smaller p-values are rejected, and the threshold T is set equal to that p-value.

Hochberg and Tamhane (1987) note that step-up procedures are less powerful than step-down ones, if controlling the same error rate at the same level.

Let $ER_m(P)$ be one of the error rates defined at page 4 for a fixed m and under distribution P for the data. We introduce now two definitions of Type I error rate control:

A multiple testing procedure S_{MTP} provides *finite sample control* of a pre-specified Type I error rate at a level $\alpha \in (0, 1)$ if $ER_m(P) \leq \alpha$ for any m. A multiple testing procedure S_{MTP} provides *asymptotic control* of a pre-specified Type I error rate at level α if $\limsup_m ER_m(P) \leq \alpha$.

Note that the actual error rate is the one determined under the true distribution of the data P, and the asymptotics is defined as the number of tests increases, no matter the actual sample size. In Chapter 3 we propose ways to provide asymptotic control of the FDR under dependence, while in Chapter 5 we propose procedures to get finite sample control of the FDR and of the quantiles of the FDP under various hypotheses on the dependence.

It is the case to comment on the concept of asymptotic control. While it is natural in the statistical literature to think of a growing number of subjects, it is far less common to think about the number of variables as possibly growing to infinity. Historically,

statisticians work very well with cases in which the number of subjects is larger than the number of variables. Many recent applications have a number of variables much higher than the number of observations. For this reason, it may not be sensible to retain the common notion of asymptotics, in which the number of subjects is increasing. In multiple testing it is natural to think about asymptotics in m. See for instance Finner and Roters (2002), and references therein. There are two interpretations to that: in the first case, m can in principle grow to infinity. It is the case of many applications in epidemiology, medicine, environmental statistics, psychometrics, non parametric estimation through wavelets. The second interpretation of asymptopia with m is as an approximation for "big" m. This may be the case of DNA microarrays, neurology, etc. There are cases, moreover, in which the number of possible tests is actually infinite, for instance when testing all possible linear combinations of parameters in an ANOVA model (cfr. Westfall and Young (1993), pag. 199).

The control of the error rate can be categorized as:

- **Weak control**: there is weak control of the Type I error rate $ER_m(P)$ if there is finite sample control of $ER_m(P_0)$, where P_0 is the distribution that would have generated the data if all the null hypotheses were true (the so called "complete null").

- **Strong control**: there is strong control of the Type I error rate $ER_m(P)$ if there is finite sample control of $\max_{I \subseteq \{1,\ldots,m\}} ER_m(P_I)$, where P_I is the distribution that would have generated the data if the null hypotheses indexed in I were true and the other false.

- **Exact control**: The direct control of $ER_m(P)$, i.e., having $ER_m(P) \leq \alpha$ is defined as exact control.

Obviously, strong control implies weak control, and the ultimate goal is exact control. Weak and strong control are introduced since it may not be doable to work with $ER_m(P)$, with unknown P. If attaining weak control, one hopes that $ER_m(P) \leq ER_m(P_0)$. This is often true. The same idea applies to strong control, since it is intuitive that it should happen that $ER_m(P) \leq \max_{I \subseteq \{1,\ldots,m\}} ER_m(P_I)$. Nevertheless, Pollard and van der Laan (2003b) note that in certain cases strong control may not imply exact control of the Type I error rate, and propose a way to overcome this problem.

We introduce also the idea of *subset pivotality*, as defined in Condition 2.1 of Westfall and Young (1993). The vector $p = \{p_j, j = 1, \ldots, m\}$ has the subset pivotality property if, for any $k \in 1, \ldots, m$, it happens that $\{p_{j_1}, \ldots, p_{j_k}\} \stackrel{d}{=} \{p_{j'_1}, \ldots, p_{j'_k}\}$. If subset pivotality holds, strong control is implied by weak control of the Type I error rate.
Another crucial hypothesis, in this setting, is what is usually called Close World Assumption (Reiter (2001)): the tests we do are all and the only one we need to do. If this was not true, we would be estimating the "wrong" distribution of the test statistics. This is far from being a mere philosophical issue: it is in fact not difficult to cheat.

Many MTPs can be easily tricked to reject the hypotheses we are interested in rejecting by artificially adding enough tests for which H_0 is false, or to fail to reject by artificially adding enough tests for which H_0 is clearly true. For the same reason, if in an experiment there are hypotheses that are known to be true or false *by design*, these should be left out of the MTP.

The role of $F(\cdot)$

As we will see, all MTPs are fully distribution free, i.e., they control the chosen Type I error rate under any distribution for $p_i|H_i = 1$, which we denote by $F(\cdot)$. It is important to keep in mind that, even if the error control is independent of $F(\cdot)$, many other features of the MTPs will not. Power, performance of estimators of M_1, and how much the MTP is conservative will always depend on $F(\cdot)$. Moreover, many results get trivial if we let $F(\cdot)$ vary among all the possible distributions on $[0, 1]$. Hence, finest approaches will take into account an estimator of $F(\cdot)$ and/or include some assumptions on $F(\cdot)$. Among the most common assumptions, there are:

- **Identifiability**: $F \neq U[0, 1]$
- **Stochastic Dominance**: $F(t) \geq t$ for any $t \in [0, 1]$, and $\exists\, t_0$ such that $F(t_0) > t_0$.
- **Parametric Model**: $F(t) = F_v(t) \geq 1 - e^{-vc(t)}$, with $c(t) > 0$ and $v > 0$.
- **Adaptive Parametric Model**: $F(t) = F_{v(m)}(t) \geq 1 - e^{-v(m)c(t)}$, with $c(t) > 0$ and $v(m)$ such that $\frac{v(m)}{\log(m)} \overset{m}{\to} +\infty$.

Note that the last assumption implies that the p-values, under the alternative, get more and more concentrated towards zero as the number of tests increases. As noted also by Genovese and Wasserman (2004b), this is a particularly good situation, in the sense that asymptotically no false negative will be made for any threshold t; and it happens for instance if the test statistics satisfy the standard large deviation principle (Van der Vaart (1998), pag. 209) and there is a common sample size n and/or a common sampling distribution; or if each test is based on measurements from a counting process, where v represents exposure time.

1.3.3 Bayesian Multiple Testing

As noted by Berry and Hochberg (1999), in the case of multiple testing

> "In the simplest Bayesian view, there is no need for adjustments and the Bayesian perspective is similar to that of the frequentist who makes inferences on a per-comparison basis."

Scott and Berger (2003) and Bayarri and Berger (2004) strengthen this view, by claiming that "a correct adjustment is automatic within the Bayesian paradigm". Bayesian testing of many hypotheses does not pose problems different than testing a single hypothesis, no adjustment is needed.

Nevertheless, Berry and Hochberg (1999) make a review of available Bayesian procedures to control frequentist error measures, and propose a hierarchical model based on a Dirichlet process prior distribution to allow for exchangeability of the tests. If independent priors are used, they formally conclude that from a Bayesian point of view no modification is needed to the standard single setting procedure.

Scott and Berger (2003) propose a way to choose suitable prior distributions on the quantities of interest, and develop efficient sampling methods to deal with multiplicity, i.e., they provide a way to speed up computations via importance sampling (see Robert and Casella (1999)).

Efron *et al.* (2001) put FDR controlling procedures under an empirical Bayes framework, and Storey (2003), when proposing the positive false discovery rate, discusses an interesting Bayesian interpretation.

1.3.4 General ideas behind a Multiple Testing Procedure

Before reviewing the procedures to control each Type I error rate, we state some general ideas:

- We want to fix a cut-off T such that the error rate is at most equal to a pre-specified $\alpha \in [0, 1]$.

- We want this cut-off to be as high as possible, provided the specified error rate is controlled. The higher T, more tests are rejected and the more powerful the procedure.

- If two procedures control the same error rate, we prefer the one which is better in terms of power, i.e., achieves a smaller Type II error rate we fix. We will use two Type II error rates: one is the FNR, which we define in (1.8), and the other is the average count of Type II errors $N_{0|1}$, which is usually called *per family Type II error rate* in literature.

Multiple testing procedures aim to balance between false positives (given by larger Ts) and false negatives (given by smaller Ts). We follow here the principle that, as long as the Type I error rate is controlled at the desired level, we prefer to make more false positives in order to have less false negatives. We defer to Section 1.4 a discussion of why it is not advisable to perform *uncorrected* testing, i.e., to simply reject all p-values smaller than $T = \alpha$. With a simplification of notation, we will say "reject $p_{(j)}$ such that..." to indicate "reject the hypotheses for which $p_{(j)}$ is such that".

1.3.5 Procedures Controlling the FWER

In this subsection we briefly review procedures to control the Family-Wise Error Rate, as defined in (1.1). More details can be found in Hochberg and Tamhane (1987) or Westfall and Young (1993).

Bonferroni Correction

The most famous and used procedure to control the FWER is the Bonferroni correction, which is a one-step procedure fixing $T = \alpha/m$. Hence, one would reject only the hypotheses for which $p_j \leq \alpha/m$. It is easily seen that this controls the FWER under arbitrary dependence:

$$\begin{aligned} P(N_{1|0} \geq 1) &= P(min_{j \in S_0} p_j < \alpha/m) \\ &\leq P(p_{(1)} < \alpha/m) \\ &\leq m * P(p_j < \alpha/m) = \alpha \end{aligned}$$

Step-down Holm

Holm (1979) proposed a step-down procedure to control the FWER, which consists in rejecting $p_{(j)} \leq \alpha/(m-j+1)$. This is a direct improvement in terms of power on the Bonferroni correction.

Step-down minP

Let $qbeta_\alpha(a,b)$ indicate the α percentile of a $beta(a,b)$. The "Step-down minP" procedure is as follows:

1. If $p_{(1)} > qbeta_\alpha(1,m)$, don't reject any hypothesis (i.e., set $T_{minP} := 0$).

2. While $p_{(i)} \leq qbeta_\alpha(1, m-i+1)$, set $i := i+1$.

3. As soon as $p_{(i)} > qbeta_\alpha(1, m-i+1)$, set $T_{minP} := p_{(i-1)}$.

4. If $p_{(m)} \leq qbeta_\alpha(1,1)$, reject all hypotheses.

Quantiles are computed from the distribution of the minima of the last k p-values, which under the complete null hypothesis is well known to be a $beta(1, m-k+1)$. It can be proved, hence, that the "minP" procedure provides *weak* control of the FWER. See also van der Laan *et al.* (2003a) for comments on this procedure, which proves in many cases better than the other ones controlling FWER, as the number of tests increases.

Other Procedures Controlling the FWER

Among the many other procedures that provide control of the FWER, there are:

- **One-step Sidak**: The one-step Sidak procedure consists in controlling each test at a level $1 - \sqrt[m]{1-\alpha}$.
- **Step-down Sidak**: A step-down version of Sidak correction consists in rejecting $p_{(j)} \leq 1 - \sqrt[m-j+1]{1-\alpha}$.
- **Step-up Hochberg**: Proposed in Hochberg and Benjamini (1990), it consists in a step-up version of step-down Holm.

For the Sidak procedures, see Sidak (1967) and Sidak (1971).

1.3.6 Procedures Controlling the FDR

The FDR, as defined in (1.4), was introduced by Benjamini and Hochberg (1995) to fulfill the need of an error measure that would provide less strict control on the number of false rejections, in particular with large m. It is easy to see that in many cases FWER controlling procedures are such that the number of rejections is $o_P(1)$, and in general all have very low power (see Table 1.2).

We will describe FDR control referring mainly to articles by Genovese and Wasserman (Genovese and Wasserman (2002), Genovese and Wasserman (2004b)), who put the problem under a stochastic process framework; rather than to Benjamini and Hochberg (1995), who introduced the FDR.

First, note that the FDP can be seen as a stochastic process indexed by the threshold t:

$$\Gamma(t) = \frac{\sum (1 - H_i) 1_{\{p_i < t\}}}{\sum 1_{\{p_i < t\}} + \prod (1 - 1_{\{p_i < t\}})},$$

note in fact that $(1 - H_i)1_{\{p_i<t\}}$ is one if and only if $H_i = 0$ *and* $p_i < t$, i.e., if the i-th hypothesis has been rejected while being true. $\Gamma(t)$ is a stochastic process since, for each $t \in [0, 1]$, $\Gamma(t)$ is a random variable. If t is a fixed cut-off (i.e., we actually reject $p_j < t$), then $\Gamma(t)$ is the realized FDP, the proportion of false rejections.

As also Genovese and Wasserman (2004b) note, the cut-offs T are usually random, which implies the non trivial problem of evaluating a stochastic process at a random point:

> "One of the essential difficulties in studying a procedure T is that $\Gamma(T)$ is the evaluation of the stochastic process $\Gamma(\cdot)$ at a random variable T. Both depend on the observed data, and in general they are correlated. In particular, if $\widehat{Q}(t)$ estimates $FDR(t)$ for each fixed t, it does not follow that $\widehat{Q}(T)$ estimates well $FDR(T)$ at a random T. The stochastic process point of view provides a suitable framework for addressing this problem."

There are two procedures to control the FDR:

BH Reject $p_j < T_{BH}$, where T_{BH} as $\sup\{t : \widehat{G}(t) = \frac{t}{\alpha}\}$, where α is the desired upper bound for the FDR and $\widehat{G}(t)$ is the empirical distribution of the p-values.

plug-in Reject $p_j < T_{BH}$, where T_{BH} as $\sup\{t : \widehat{G}(t) = \frac{(1-\widehat{a})t}{\alpha}\}$, where $\widehat{a}$ is any estimator of a.

The BH procedure was originally proposed in Simes (1986), but it didn't receive much attention at that time since it didn't control the FWER in the strong sense (while it did in the weak sense). It can be seen that it controls the FDR: $FDR(T_{BH}) \leq (1-a)\alpha \leq \alpha$ (see Benjamini and Hochberg (1995) or Storey *et al.* (2004)).

The *plug-in* procedure was first proposed in Genovese and Wasserman (2002). It exploits information given by the sequence of p-values through a suitable estimator of a. The plug-in procedure controls the FDR and is more powerful than the BH procedure. Uncertainty brought about by the estimation is not usually incorporated. We will see in Chapter 4 how to do this.

The most common estimator used is Storey's estimator, proposed in Storey (2002), and defined as:

$$\widehat{a} = \frac{\widehat{G}(t_0) - t_0}{1 - t_0} \tag{1.7}$$

for some $t_0 \in (0, 1)$. Among the other possibilities, there are estimators proposed in a completely different context: see Swanepoel (1999), or Woodroofe and Sun (1999). All these estimators are seen to break down under dependence (see Chapter 3). For this reason, we will propose in Chapter 4 a class of estimators robust with respect to dependence.

Genovese and Wasserman (2002) also introduce the FNR, the dual of the FDR, a measure of power defined as

$$E\left[\frac{N_{0|1}}{m - R + 1_{(m-R)=0}}\right]. \tag{1.8}$$

We will use this Type II error measure when comparing procedures in terms of power (the other possibility will be to count the average Type II errors).

Storey (2003) introduced the positive False Discovery Rate, which proves even more powerful than the original BH method, and provided interesting extensions and insights of the methods. He also provides a way to estimate the FDR for fixed number of rejections.

Storey *et al.* (2004) propose a unified estimation approach for the FDR, showing methods to estimate the FDR fixing the threshold or the rejection region, or asymptotically over all rejection regions simultaneously. They suggest a way to control the FDR through their estimates. In particular, they present several theorems that all require almost sure pointwise convergence of the empirical distributions of the subsequence of p-values for which the null is true and the subsequence of p-values for which the alternative hypothesis is true.

Genovese and Wasserman (2004b) introduce estimators for a and $F(\cdot)$, suggest ways to build confidence thresholds for the FDP and prove asymptotic results, providing some limiting distributions. In Chapter 3 we provide some extensions of their results under dependence.

1.3.7 Exact control of the FDP

$tFDP(c)$, as defined in (1.5), is a much more recent error rate, first proposed by Perone Pacifico *et al.* (2003) and then by van der Laan *et al.* (2003b). Basically, control of the tail of the distribution of the FDP is performed, while the FDR controls the mean.

An intuition of why and when this should be preferred is shortly given.

Quantile control is more protective against extremes[4]: interest is taken in the tails of the distribution of the FDP rather than in its central part. This is useful in cases in which the expected value in not a suitable representative of the random variable, and extreme realizations should be avoided, controlled or forecasted.

The case of FDP is an excellent example of such a problem: it can obviously happen that $FDR = E[FDP] < \alpha$ but the realized FDP is very close to 1. $tFDP(c)$ requests that the tail of the distribution of the FDP is light enough, i.e. that large FDP is realized with small probability. In some sense, $tFDP(c)$ is a more finer error measure than FDR. In general, anyway, $tFDP(c)$ and FDR control respond differently to the distribution of the p-values under the alternative, $F(\cdot)$; and $tFDP(c)$ control may lead to more or less rejections than FDR control on a case by case basis.

We will moreover note that, in general, $Var[FDP]$ is increased by dependence among the test statistics. As the variance increases, the FDP is less and less concentrated around its expected value, so it becomes less and less meaningful to control the FDR. A similar remark is made in Bickel (2004) and Owen (2004).

A simple example of problems posed by dependence is provided below:

Example 1.3.1. *Let $p_1 \sim Unif(0,1)$. Let $p_2 = \cdots, p_m = p_1$. Suppose the $FDR = \alpha$. The number of false rejections is m with probability α. Here, $\Gamma(T)$ has a distribution with mass $1-\alpha$ at 0 and α at 1.*

This suggests that aiming for confidence thresholds on the FDP, i.e., quantile control, is even more desirable in the presence of dependence.

We will give some results under dependence and propose some improvements to the available procedures in Chapters 4 and 5. We will propose a procedure to control $tFDP(c)$ in Chapter 2, that proves more powerful than the other available ones; and also extend it to dependence in Chapter 5.

[4]Note that the idea is similar to the one of "quantile regression" (Koenker and Bassett (1978)), where attention is shifted from the (conditional) mean to the entire distribution of the response variable.

Augmentation

van der Laan *et al.* (2003b) propose a very clever way of controlling at level α both the $tFDP$ and the $gFWER(k)$ as defined in (1.2). They start from the idea that any procedure requiring something less stringent than FWER control will result in the rejection of at least the same hypotheses. For this reason, they start by controlling the FWER (any procedure will be fine). Then, they augment by rejecting the previously selected hypotheses and an opportune additional number. I.e., they propose a universal method to identify additional rejections among the null hypotheses which were not rejected with a procedure controlling the FWER (asymptotically or exactly). This idea is easily understood by examining their augmentation method to control the $gFWER(k)$. It is straightforward to see that the set $S_{gFWER(k)} = S_{FWER} + \{j_1, \ldots, j_k\}$ for any choice of $j_1, \ldots, j_k \notin S_{FWER}$ will be suitable. For power considerations, one obviously adds the k most significant non rejected p-values. Here, S_{FWER} stands for the set of rejected hypotheses using a FWER controlling procedure.

The *augmentation procedure* for control of $tFDP(c)$ is as follows:

1. Control the FWER with any procedure, and reject $|S_{FWER}|$ hypotheses.

2. If $|S_{FWER}| > 0$, let

$$k_n(c, \alpha) = \max\{j \in \{0, \ldots m - |S_{FWER}|\} : \frac{j}{j + |S_{FWER}|} \leq c\}$$

   and

$$c^* = \frac{k_n(c, \alpha)}{k_n(c, \alpha) + |S_{FWER}|} \leq c.$$

3. Any choice of $k_n(c, \alpha)$ additional hypotheses will control $tFDP(c^*)$ at the desired level. Again, for obvious power considerations, the $k_n(c, \alpha)$ most significant p-values not previously rejected will be selected.

Understanding of this augmentation procedure is less straightforward. The idea is that the target of the FWER procedure is to avoid false rejections. This happens with probability at least $1 - \alpha$. Hence, after augmentation, with probability $1 - \alpha$ the FDP is at most c^* (which happens if all the $k_n(c, \alpha)$ added null hypotheses are true).

The great advantage of this augmentation procedure is that it is valid under arbitrary joint distributions of the test statistics, i.e., under any form of dependence; if the FWER is controlled under dependence.

As we will show later via simulations, the drawback is that the power may be unacceptably low for large number of tests. In fact, the number of rejections in $o_P(1)$ in many cases. The reason is easily understood: a first insight is the fact that augmentation is strongly linked to FWER procedures. The smaller $|S_{FWER}|$, the smaller $k_n(c, \alpha)$. Just by looking at the definition of $k_n(c, \alpha)$, one can see that when no test is rejected at the first stage, none will be at the second stage for any $c < 1$.

Moreover, in the choice of the number of additional hypotheses to be rejected the observed p-values are not considered. In the proof, the number of errors produced by augmentation is roughly approximated by $k_n(c, \alpha)$; while it will be much smaller when the most significant p-values are selected.

Inversion

The inversion method was first proposed in Genovese and Wasserman (2004b) and then more extensively examined in Genovese and Wasserman (2004a).

This method involves inverting a set of uniformity tests. The steps are as follows:

1. For every $I \subseteq \{1, \ldots, m\}$, test at level α the hypothesis that the random variables $\{p_j : j \in I\}$ are identically distributed like a $U(0, 1)$. I.e., let $\{\psi_I : I \in \{1, \ldots, m\}\}$ be a set of non-randomized level α tests of uniformity.

2. Let $U = \{I : \psi_I(p_I) = 0\}$, the collection of all subsets not rejected in the previous step.

3. For any $C \neq \varnothing$ let $\bar{\Gamma}(C) = \max_{B \in U} \frac{|C \cap B|}{|C|}$. Let R be the biggest set such that $\bar{\Gamma}(R) \leq c$. R is a rejection set that yields $tFDP(c) \leq \alpha$.

Note that $\bar{\Gamma}(C)$ can always be rewritten in terms of threshold T, so we loosely indicate this mapping with $\bar{\Gamma}(t)$. This is a $1 - \alpha$ confidence upper envelope for the $\Gamma(t)$ process, in the sense that $\Pr(\bar{\Gamma}(t) > \Gamma(t)) > 1 - \alpha$. The rejection set is determined through inversion of the confidence envelope, i.e., $T = \sup_t\{t : \bar{\Gamma}(t) \leq c\}$. Under independence, it is easily seen that the 2^m tests at Step 1 reduce to m tests. We will prove that this is true also under arbitrary dependence in Chapter 5. For more insights on the inversion procedure, refer to Genovese and Wasserman (2004a). They note that one can choose any uniformity test at Step 1, and suggest a few possibilities; among which the $minP$ test of van der Laan *et al.* (2003a). Genovese and Wasserman (2004a) prove that, with this choice, the augmentation and inversion procedures lead to the same rejection regions. For this reason, we call the two methods the $p_{(1)}$-approach, which we will extend to the dependent case in Chapter 5. As stated, the $p_{(1)}$-approach can lead to very small power for $tFDP(c)$ and FDR control (see also Table 1.2).

DKW-approach

A very close idea to the inversion is the DKW-approach of Perone Pacifico *et al.* (2003):

1. Let $\varepsilon_m = \sqrt{\frac{1}{2m} \log(2/\alpha)}$.

2. Define now

$$R(t) = \begin{cases} \frac{t(1-\widehat{a})}{\widehat{G}(t) - \varepsilon_m} & \text{if } \widehat{G}(t) > t(1 - \widehat{a}) + \varepsilon_m \\ 1 & \text{otherwise} \end{cases} \tag{1.9}$$

3. Let $T_{DKW} = \sup\{t : R(t) \leq c\}$ and reject $p_j \leq T_{DKW}$.

This approach is based on the DKW-inequality (Dvoretzky *et al.* (1956), Massart (1990)):

$$\Pr(\sqrt{n}\sup_x |\widehat{F}(x) - F(x)| > \varepsilon) \leq 2e^{-2\varepsilon^2}, \tag{1.10}$$

where $F(\cdot)$ is any CDF and $\widehat{F}(\cdot)$ is the corresponding empirical distribution function based on an i.i.d. sample of size n from F. Let then $\varepsilon_m = \sqrt{\frac{1}{2m}\log(2/\alpha)}$. By DKW-inequality,

$$\Pr(||G - \widehat{G}||_\infty > \varepsilon_m) \leq \alpha,$$

where $G(\cdot)$ is the marginal distribution of the p-values and $\widehat{G}(\cdot)$ is the empirical distribution. A lower confidence bound can then be obtained for the empirical distribution, which implies that if $\widehat{a} \leq a$, by DKW inequality, $\Gamma(t) \leq R(t)$ with probability at least $1 - \alpha$. Hence $T_{DKW} = \sup\{t : R(t) \leq c\}$ controls $tFDP(c)$ at level α. Unfortunately, this procedure can have very low power for small m and weak signal ($M_0 \cong m$ and/or $F \cong U[0,1]$), resulting in no rejections as will be seen in simulations. It is a good procedure, anyway, for big m or if the signal is strong (it will prove very good in applications of wavelet thresholding, as discussed to Chapter 6).

In Chapter 5 we generalize this approach under dependence, proving a form of DKW-inequality under dependence.

1.4 Simulation of the procedures

In this section we will briefly compare the procedures on the basis of the counts of errors $N_{1|0}$ and $N_{0|1}$. This provides an immediate comparison in power on what is usually done in applications. Table 1.2 shows the average counts for a thousand of simulated normal data sets, under independence, with $m = 5000$ and $M_0 = 4500$. The alternative expected values were sampled from a uniform random variable in $(0, 5)$. The word *Storey* indicates that the estimator defined in (1.7) was used, to estimate the proportion of false nulls a.

Among the procedures controlling the tail of the FDP, with $c = 0.1$, at level α; $p_{(1)}$-approach and augmentation with Bonferroni correction at first step proved to be unsatisfactory, as we expected.

Remark 1.4.1. *A good framework for understanding MTPs is that, in our setting, all are nothing but a "uniformity" check on each single p-value, possibly using information from the other tests.*

1.5 Type I Error Rates Control Under Dependence

FWER Control

Control of FWER under dependence has never been an issue. It is easily seen that the Bonferroni correction is valid under arbitrary dependence. The extension of "Step-down

Method	$\mathbf{E}[\mathbf{N_{1\|0}}]$	$\mathbf{E}[\mathbf{N_{0\|1}}]$
Control of Single Type I Error		
Uncorrected	225.57	166.63
Control of FWER		
Bonferroni	0.040	412.58
Step-Down Holm	0.055	411.84
One-Step Sidak	0.050	412.71
Step-Down Sidak	0.057	411.43
Step-Up Hochberg	0.051	412.42
Control of FDR		
BH	10.27	282.85
Plug-in (Storey)	11.81	276.99
Control of tFDP(0.1)		
DKW (Storey)	10.588	284.52
$p_{(1)}$-approach	0.091	401.774
Augmentation with Bonferroni at first step	0.082	403.488

Table 1.2: Average error counts for $m = 5000$ tests, $M_0 = 4500$ for different methods controlling different error measures at level $\alpha = 0.05$

minP" procedure to arbitrary dependence has been recently devised in van der Laan *et al.* (2003a), who propose a way to estimate the distribution of the minima of the last k p-values under dependence, and to substitute the quantiles of the *beta* distribution with the opportune quantiles in the algorithm at page 12. Genovese and Wasserman (2004a) note that estimation of this distribution in general is not a good path to follow, usually being very unstable, since the number of observations n is usually much smaller than the number of p-values m. In the setting of DNA microarrays, van der Laan and Bryan (2000) argue that one needs $\frac{n}{\log(m)} \to \infty$ as $n, m \to \infty$ for consistent estimates of the correlation matrix of the test statistics. In Chapter 6, anyway, we will provide a real data example with $\frac{n}{\log(m)} \cong 8$, $n = 62$, where we will argue that the variance/covariance matrix is *not* efficiently estimated. We will propose in Chapter 5 two procedures, based on minima of p-values, that do not need the estimation part and thus work very well under different dependence situations.

FDR and $tFDP(c)$ Control

When they introduced the FDR, Benjamini and Hochberg (1995) proved that the Simes (1986) procedure controlled the FDR under independence of the M_0 test statistics corresponding to the true nulls. Providing results under dependence of the whole sequence of p-values has been an open problem since then.

The best results in our opinion are achieved in Benjamini and Yekutieli (2001), who prove with that the BH procedure can never control the FDR at level higher than $\alpha \sum_{i=1}^{m} 1/i$. Hence, taking into account a factor of $\sum_{i=1}^{m} 1/i$ will allow to control the FDR under general dependence. We will call this the BY approach throughout. Note that this is unacceptably conservative. They also prove that, under conditions of Positive Regression Dependency on S_0, the BH procedure is still valid, controlling the FDR at level α. The condition of PRDS introduced in Benjamini and Yekutieli (2001) is as follows: for any increasing set D and for each $i \in S_0$, let $\Pr(X \in D | X_i = x)$ be non decreasing in x. This is a relaxed version of Positive Regression Dependency, a slightly more general form of association (see Esary *et al.* (1967), and Chapter 5). Recall that a set is said to be increasing if for any $x \in D$ and $y \geq x$, $y \in D$. They note, together with Sarkar (2002), that distributions satisfying this property include multivariate normal distributions with positive correlations and few other cases. Sarkar (2002) also extend the results of Benjamini and Yekutieli (2001) by generalizing their results to a whole class of step-up/step-down procedures to control the FDR. As we will see in the next chapter, simulations show that strong positive dependence, including PRDS, is likely to make the procedure more conservative. For this reason in Chapter 3 we will introduce completely different conditions on the dependence which will lead the procedure under dependence to asymptotically behave like in the independent case.

Storey *et al.* (2004), as said, provide several theorems that all require almost sure pointwise convergence of the empirical distributions of the null p-values and alternative p-values. They argue that this may be true also under dependence; and in fact Bickel (2004) shows a process with long-range correlations that satisfies the conditions of Storey *et al.* (2004).

Storey and Tibshirani (2001) show how to estimate the $pFDR$, positive false discovery rate of Storey (2003), under general dependence between the test statistics and apply the methodology to estimate the FDR under dependence for pre-fixed rejection region.

Yekutieli and Benjamini (1999), Pollard and van der Laan (2003b) propose resampling based procedures to control the FDR when the test statistics are correlated.

Chapter 2

Generalized Augmentation Procedure

In the previous chapter we provided a review of many multiple testing procedures. We argued that the procedures controlling $tFDP(c)$ as defined in (1.5) have some drawbacks, in the sense that *augmentation* (and $p_{(1)}$-approach) can lead to no rejections as the number of tests increases, and likewise for the DKW approach when the signal is weak ($M_0 \cong m$, $F \cong U[0,1]$) and/or for small m. We propose here a simple generalization, under independence, of the augmentation approach of van der Laan *et al.* (2003b) (as described at page 16). For this reason, we call this "Generalized augmentation procedure". In the next section we propose the algorithm and prove it controls $tFDP(c)$ at the desired level. Section 2.2 will provide some comments and simulations. Section 2.3 will provide some insights on how to choose the parameter c when using the proposed procedure in order to control FDR instead of $tFDP(c)$. This insights are generalized to use of any $tFDP(c)$ controlling procedure for FDR control.

2.1 The Procedure

As we already noted, the *augmentation* approach of van der Laan *et al.* (2003b) can lead to no rejections as the number of tests increases. It is not difficult to realize that this happens because this approach is strongly linked to the FWER controlling procedure used at the first step. When the FWER controlling procedure does not lead to rejection of any hypothesis (a common situation for big number of tests) then also the augmentation approach will not lead to rejections. Generalized augmentation procedure is based on the idea that FWER control at the first step can be replaced by uncorrected testing. This will lead to (almost always) reject a certain number of hypotheses at the first step.

Then, if uncorrected testing at a certain level q rejects fewer hypotheses than it is needed to get control $tFDP(c)$, a suitable augmentation can be applied. If too many hypotheses are rejected, so that control of the error measure is not achieved, a *negative*

augmentation is to be applied.

The algorithm is as follows:

1. Reject the $S = |S_q|$ p-values smaller than a certain $q \in (0, 1)$.

2. Let i^* be
$$\min\{i : \sum_{k=i}^{S} \binom{m}{k} q^k (1-q)^{m-k} \leq \alpha\}. \tag{2.1}$$
Note that i^* is easily found for fixed m and α, consisting in the evaluation of the binomial distribution with parameters m and q.

3. If $\frac{(i^*-1)}{|S_q|} \leq c$, let $k_n(c, \alpha) = \max\{j \in \{0, \ldots, m - |S_q|\} : \frac{j+i^*-1}{j+|S_q|} \leq c\}$.

If $k_n(c, \alpha)$ exists and is positive, any choice of that number of additional hypotheses will control $tFDP(c^*)$ at level α. If $\frac{(i^*-1)}{|S_q|} > c$ or i^* does not exist, then at the first step we rejected too many hypotheses. One can pick any of this three choices:

1. Choose a smaller q (for instance, divide by 2 the previous one), and repeat the procedure.

2. Note that the tail of the FDP will be controlled at level $\alpha + \sum_{k=i^{**}}^{i^*-1} \binom{m}{k} q^k (1-q)^{m-k} > \alpha$, where $i^{**} = min\{i : i/|S_q| > c\}$. This, depending on q and m, may be way too far from α or acceptably close to the desired level α.

3. Do a negative augmentation in this way: Let
$$\begin{aligned} k'_n(c, \alpha) \quad = \quad & \min\{k \in \{0, \ldots, |S_q|\} : \\ & 1_{\{|S_q|-k>0\}} \left(\sum_{i=0}^{|S_q|-k} \sum_{j=0}^{\min(k,i)} \right. \\ & 1_{\{(i-j)/(|S_q|-k)>c\}} \binom{m}{i} q^i (1-q)^{m-i} \\ & \left. \frac{\binom{i}{j}\binom{|S_q|-i}{k-j}}{\binom{|S_q|}{k}} \right) < \alpha\}. \end{aligned} \tag{2.2}$$
Then reject only the $|S_q| - k'_n(c, \alpha)$ most significant p-values.

We will now provide a proof of the generalization, and a simulation showing how much this actually improves the approach of van der Laan *et al.* (2003b). A formal proof of the fact that choosing a fixed q instead of one going to zero with m brings about better results in terms of power is straightforward and thus omitted. Further work may provide a general statement on how to choose an optimal q (in terms of power), and better negative augmentation approaches. We will propose one based on estimation of M_0 in Chapter 4. Next theorem proves that this method controls the $tFDP(c)$ at the desired level:

Theorem 2.1.1. *The Generalized augmentation procedure controls $tFDP(c)$ at level α.*

Proof. From an immediate extension of the results of Finner and Roters (2002) we know that $N_{1|0}$ (the number of false positives) in a single-step method is $Binomial(M_0, q)$, where M_0 is as in Table 1.1 and q is such that we reject $p_j < q$. Then[1]:

$$\begin{aligned}
\Pr(FDP > c) &\leq \sum_{i=0}^{S} \Pr\left(\frac{k_n(c,\alpha)+i}{|S_q|+k_n(c,\alpha)} > c\right)\binom{M_0}{i} q^i (1-q)^{M_0-i} \\
&= \sum_{i=0}^{S} 1_{\{\frac{k_n(c,\alpha)+i}{|S_q|+k_n(c,\alpha)}>c\}}\binom{M_0}{i} q^i (1-q)^{M_0-i} \\
&= \sum_{i=i^*}^{S} \binom{M_0}{i} q^i (1-q)^{M_0-i} \\
&\leq \sum_{i=i^*}^{S} \binom{m}{i} q^i (1-q)^{m-i} \leq \alpha,
\end{aligned}$$

which proves the positive augmentation step.

Let V_n be the number of false rejections at Step 1 and W_n be the number of hypotheses not rejected after negative augmentation that were in fact true nulls. The effect of negative augmentation is seen by:

$$\begin{aligned}
\Pr(FDP > c) &= \sum_i \sum_j \Pr(FDP > c | V_n = i, W_n = j) \\
&\qquad \Pr(W_n = j | V_n = i) \Pr(V_n = i) \\
&\leq 1_{\{|S_q|-k'_n(c,\alpha)>0\}} \Bigg(\sum_{i=0}^{S-k'_n(c,\alpha)} \sum_{j=0}^{\min(i,k'_n(c,\alpha))} 1_{\{(i-j)/(|S_q|-k'_n(c,\alpha))>c\}} \\
&\qquad \binom{m}{i} q^i (1-q)^{m-i} \frac{\binom{i}{j}\binom{S-i}{k'_n(c,\alpha)-j}}{\binom{S}{k'_n(c,\alpha)}} \Bigg) \\
&\leq \alpha,
\end{aligned}$$

since it is obvious that $W_n|V_n$ follows an Hypergeometric distribution with parameters (S, i, k), and where the majorization for the tail of the distribution of V_n used in the previous proof was applied again. □

2.2 Comments and Simulations

Note that the proposed procedure is an actual generalization of the *augmentation* of van der Laan *et al.* (2003b), in the sense that if $q = \alpha/m$ then $i^* = 1$, which gives back the van der Laan *et al.* (2003b) procedure with Bonferroni at the first step. The purpose

[1]Note that $\Pr(FDP > c)$ is actually $\Pr(FDP > c|S)$.

	$q = \alpha/m$ (van der Laan *et al.* (2003b))	$q = \alpha$, then $q := q/2$	$q = \alpha$, neg. aug.
$m = 10$	0.0517 (0.043)	0.0553 (0.029)	0.0801 (0.016)
$m = 100$	0.0674 (0.036)	0.0686 (0.033)	0.0829 (0.019)
$m = 200$	0.0710 (0.030)	0.0677 (0.036)	0.0776 (0.039)
$m = 500$	0.0737 (0.003)	0.0635 (0.032)	0.0707 (0.030)
$m = 700$	0.0753 (0.002)	0.0632 (0.025)	0.0673 (0.021)
$m = 800$	0.0756 (0.000)	0.0620 (0.026)	0.0669 (0.031)
$m = 1000$	0.0764 (0.000)	0.0622 (0.012)	0.0653 (0.028)
$m = 5000$	0.0823 (0.000)	0.0574 (0.006)	0.0598 (0.001)

Table 2.1: FNR and $tFDP(c)$ (in parentheses) for Generalized Augmentation Procedure at level $\alpha = 0.05$, $c = 0.1$

behind the generalization is to avoid FWER control at the first step. This avoids all the problems connected with FWER control when m is large (namely, $R \xrightarrow{P} 0$).

An open problem is the choice of q. Any $q \in (0, 1)$ brings about $tFDP(c)$ control, the optimal is obviously the one that yields the highest cut-off. This, anyway, implies repeating the procedure many times. We will always set $q = \alpha$ (i.e., do uncorrected testing at the first step) in this dissertation.

Table 2.1 compares, in simulation, the augmentation procedure with the generalized augmentation with $q = \alpha$, i.e., augmentation of uncorrected multiple testing. The comparison is done in terms of power (expressed as average observed FNR as defined in (1.8)). The average observed $tFDP(c)$ is reported in parentheses. It is apparent that our methods are better in terms of power than van der Laan *et al.* (2003b) augmentation method (recall that the lower the FNR, the higher the power), especially for large values of m. It is intuitive that any choice of q that is not infinitesimal with m in general will not result in $R \xrightarrow{P} 0$. Note that, for small m, a choice of q lower than α may lead to more rejections.

Finally, we simulated the procedure and counted the average number of errors for $m = 5000$ tests, with $M_0 = 4500$, as in Table 1.2. The average number of false positives ($E[N_{1|0}]$) for the generalized augmentation procedure, setting $q := q/2$ when negative augmentation is needed, is 14.65, while the average number of false negatives ($E[N_{0|1}]$) is 271.96. If negative augmentation is actually performed, then $E[N_{1|0}] = 9.94$ and

$E[N_{0|1}] = 285.94$.

2.3 FDR control via $tFDP(c)$ control: Choice of c

Recall that any $tFDP(c)$ procedure controls the FDR at level $c + (1 - c)\alpha$. Hence, if $tFDP(c) < \frac{\alpha - c}{1 - c}$, $FDR \leq \alpha$. If one wants to use the generalized augmentation procedure to control the FDR at level α, then, a choice of $c \in (0, \alpha)$ must be done. Similarly if the $p_{(1)}$-approach, or any other method to control $tFDP(c)$, is used. The purpose of this subsection is to give insights on how to choose a value for c.

It is obvious that, since any $c \in (0, \alpha)$ is fine, the optimal is the one that yields the highest cut-off. This, anyway, implies running the procedure many times since the cut-off is data dependent, and no general statement can be made.

We will now get a sense, via simulation, of what happens for different values of c, and see that the optimal c is usually very close to zero. We will declare the optimal c as the one which yields the highest FDR, or the lowest FNR (as defined in (1.8)).

At each iteration a different data set was generated, $B = 1000$ times for each value of c. As usual, normal random variables were taken, and the expected value under the alternative was generated as a random uniform in $(0, 5)$. Note that we are not looking for the highest cut-off for a single given data set.

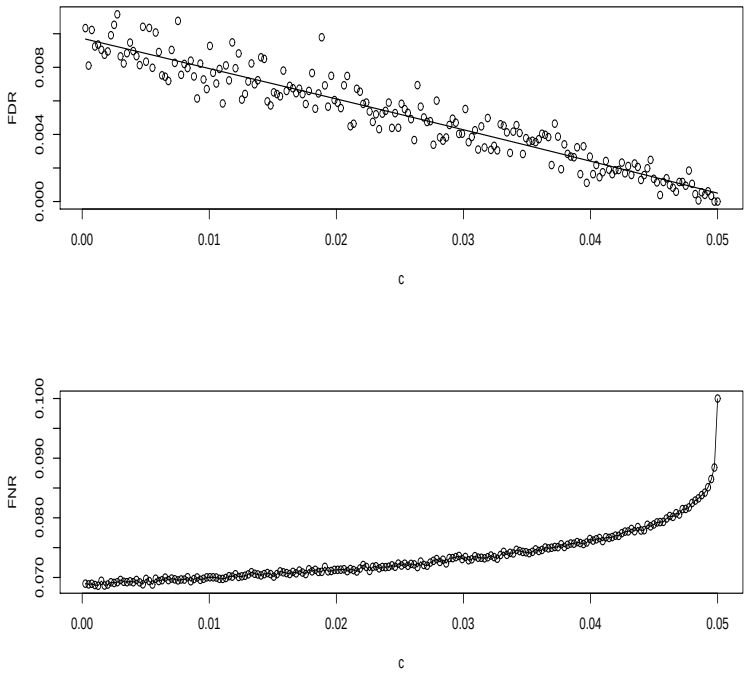

Figure 2.1: Generalized Augmentation Approach, $m = 100, mo = 90$

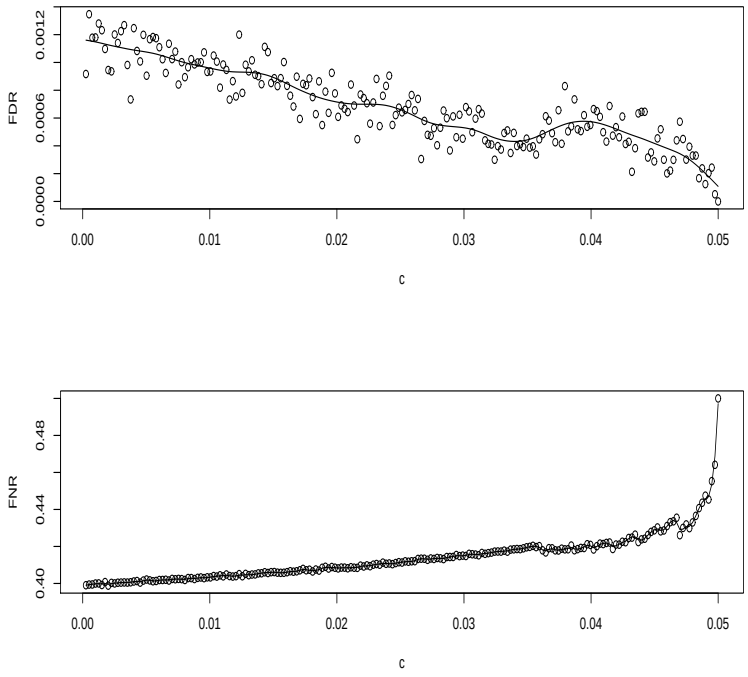

Figure 2.2: Generalized Augmentation Approach, $m = 100, mo = 50$

Figure 2.1 shows the results of the simulations for $m = 100$ and $m_0 = 90$. Similar results are observed in Figure 2.2, where simulations are done for $m_0 = 50$ and Figure 2.3, where $m_0 = 10$. The dots are the simulated FDR and FNR, while the line is a fitted cubic smoothing spline, with amount of smoothness estimated by cross validation (see for instance Green and Silverman (1994) for these non-parametric methods). We can see that, unless the number of true nulls is very small, the optimal c is always very close to zero. Moreover, the procedure is always conservative. This is the price we pay for using a $tFDP(c)$ controlling procedure to control the FDR.

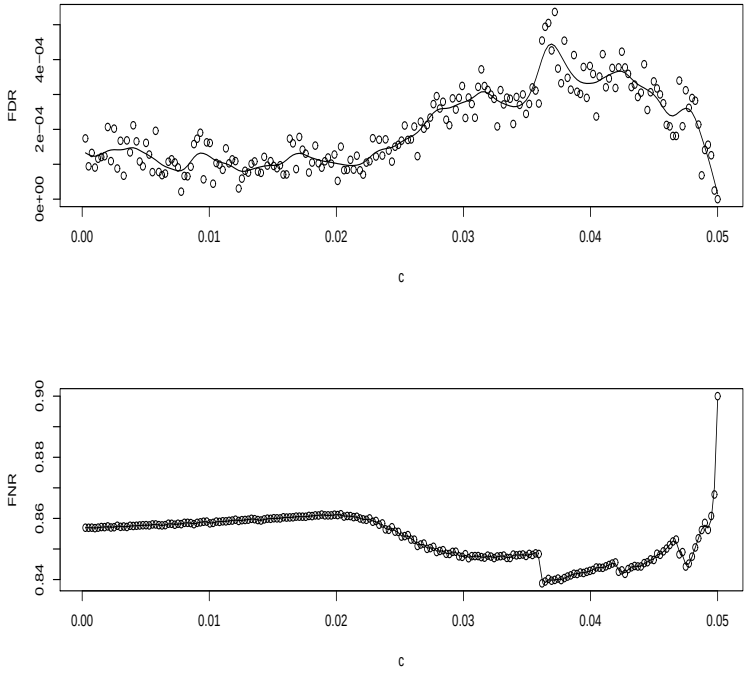

Figure 2.3: Generalized Augmentation Approach, $m = 100, mo = 10$

Chapter 3

Asymptotic Control of the FDR Under Dependence

In the previous chapters we made a brief review of the problem of multiple testing in a frequentist setting, with particular attention to methods of control of the mean, or quantiles, of FDP. We also proposed a new method, the "generalized augmentation procedure", to control the $tFDP(c)$ under independence of the test statistics, which proved more powerful than currently available procedures. A general comparison was made through simulations. In this chapter we will focus on the case of dependent tests (i.e., dependence among the p-values). We will prove that, under very broad hypotheses on the dependence among the test statistics, the BH and Plug-in procedure, with an appropriate estimator of a, asymptotically control the FDR at the desired level; with no need for any correction. We will provide moreover asymptotic distributional results under dependence for the threshold T_{PI} and the entire FDP stochastic process. In the light of this results, one can make use of these two procedures without even knowing if the test statistics are independent, but just assuming weak dependence among them. The dependence will be measured by mixing coefficients. For an extensive discussion on mixing, see for instance Doukan (1994). For the main concepts on asymptotics used in this chapter, refer for instance to Van der Vaart (1998), Shorack and Wellner (1986) and van der Vaart and Wellner (1996).

For a comment on the sense of asymptotic results in multiple testing, refer to pag. 8. Applications we will mainly have in mind in this chapter are wavelet thresholding, follow-up studies, case control studies, etc. We furthermore provide extensive simulations under dependence, and argue that even wider hypotheses should be sufficient; in the sense that when our main result does not hold, then the procedures become conservative, thus still controlling the FDR under a pre-fixed threshold.

The setting is as follows: Section 3.1 will present our main theoretical results, Section 3.2 will describe the outcomes of simulations of the BH and plug-in methods applied to correlated data. Section 3.3 will show how the common estimator for a breaks down under dependence, and will propose a much more robust estimator for this quantity.

This new estimator will have a good performance for all m, not only asymptotically. For an extensive discussion on how to estimate a, refer to Chapter 4 where many iterative procedures are proposed.

3.1 Theoretical Considerations

3.1.1 Asymptotic Validity of the Plug-in Method Under Dependence

We show that under conditions of weak dependence for the p-value and H_i sequences, the plug-in procedure (with a good estimator of a) is able to control the FDR at the desired level α. This is a generalization of the results from Genovese and Wasserman (2004b). We consider here H_i as a random variable, with a commonly used mixture model. The results obviously remain valid if H_i is considered as fixed, just by assuming that $M_1/m \xrightarrow{P} a$.

We will need the following two definitions:

Definition 3.1.1 (α-Mixing)**.** *The k-th α-mixing coefficient is defined as:*

$$\alpha(k) = \sup_j \{|P(E_1)P(E_2) - P(E_1 \cap E_2)| : E_1 \in \mathcal{M}_1^j, E_2 \in \mathcal{M}_{j+k}^{+\infty}\}; \tag{3.1}$$

where $\mathcal{M}_i^j$ is the $\sigma-$algebra generated by $\{p_i, \ldots, p_j\}$.

Definition 3.1.2 (Association)**.** *A vector of random variables $X_1, \ldots, X_n$ is associated if, for all monotonically coordinate-wise non-decreasing functions g_1 and g_2, $Cov[g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)] \geq 0$, when it exists.*

For more details on association and examples of associated random variables refer to Appendix C.

We will prove our main results under any of the following conditions:

1. If the p-values are independent (Genovese and Wasserman (2004b)).

2. If $\alpha(k)$ are the mixing coefficients of the p-values, there exists $\delta > 0$ such that

$$\alpha(k) \leq Ck^{-3-\delta}$$

   for some constant C, and the vector $(p_j)_{j\geq 1}$ is stationary.

3. If $(p_j)_{j\geq 1}$ is stationary, associated and

$$\sum_k k^{13/2+\delta} Cov(p_1, p_k) < +\infty$$

   for some $\delta > 0$.

4. If $(p_j)_{j\geq 1}$ is stationary and
$$\sum_k \alpha(k) < +\infty$$

5. If $(p_j)_{j\geq 1}$ is stationary, associated and
$$\sum_{k=2}^{+\infty}[P(X_1 \leq s, X_k \leq t) - G(s)G(t)] < +\infty.$$

If $m_0 = m$, Yu (1993) proves in a different setting that this is equivalent to
$$\sum Cov^{1/3}(p_1, p_k) < +\infty$$

Analogous conditions will be assumed on $(H_j)_{j\geq 1}$. As a (not negligible) aside, note that in the first three cases the convergence of the processes will be in $D[0,1]$, otherwise in $L^2[0,1]$. Note also that Condition 4 is more general than Condition 2.

We will make use of the following three lemmas, which are proved in the Appendix A. Let $\Lambda_0(t) = \sum(1-H_i)1_{\{p_i<t\}}$ and $\Lambda_1(t) = \sum H_i 1_{\{p_i<t\}}$; let H_i be a Bernoulli such that $\Pr(H_i = 1) = a$.

Lemma 3.1.3. *Let $\widehat{G}(t) = \frac{1}{m}\sum 1_{\{p_i<t\}}$ be asymptotically equicontinuous. Then $\Lambda_j(t),\ j = 0,1$ will be too.*

Lemma 3.1.4. *Let $\Lambda_j(t),\ j = 0,1$ be asymptotically equicontinouos. Then the vector $\Lambda = (\Lambda_0(t), \Lambda_1(t))$ will be too.*

Lemma 3.1.5. *Assume any of the proposed conditions holds. The vector $(W_0(t), W_1(t))$ will be convergent in distribution for any t; where $W_0(t) = \sqrt{m}(\Lambda_0(t) - (1-a)t)$ and $W_1(t) = \sqrt{m}(\Lambda_1(t) - aF(t))$.*

Note that the condition on the mixing coefficients only requires that the dependence between a fixed p-value in the sequence and the following others decreases fast enough. This will be implied by other (more strict) conditions, like m-dependence, Gaussian processes with covariance tending to 0, strictly stationary ARIMA models, block dependence with bounded block dimension, etc. In practice, one usually proves m-dependence, which is more intuitive than mixing (see Chapter 6 and Appendix C for other comments and references on mixing). Mixing conditions will be true in many applications where tests are on points in the space or time getting further and further. Many times series and environmental/spatial statistics applications will fall under the conditions on the mixing coefficients. They can be thought to be true also in many applications of thresholding of wavelet coefficients, in the case of testing of points in an image (like fMRI or else) when the tests start from a spot and then are moved away from the starting point.

The following theorem states our main result: if the dependence decreases fast enough, then there is no need to modify the BH and plug-in procedures used under independence, because they remain valid.

Theorem 3.1.6. *Let $\{p_i\}_{i\in\mathcal{N}}$ be a random sequence of p-values from tests. Let H_i be the indicator of the i-th hypothesis to be false. Let $\Pr(p_i < t|H_i = 1) \sim F(t)$, where $F \neq U$ and $\Pr(H_i = 1) = a$. Assume any of the specified conditions (1-5) holds. Assume the quantity a is known.*

Then, $E[\Gamma(T_{PI})] = \alpha + o(1)$, where $\Gamma(t)$ is the FDP for threshold t and T_{PI} is the plug-in threshold.

For a proof and distributional results, refer to the Appendix A.

Corollary 3.1.7. *Let $\widehat{a}$ be a consistent estimator of the quantity $a_0 \leq a$. Let the hypotheses from Theorem 3.1.6 hold. Then the plug-in method will asymptotically control the FDR at α level.*

Proof. The thesis will follow from Theorem 3.1.6 and same reasoning as Theorem 5.2 in Genovese and Wasserman (2004b). □

Remark 3.1.8. *As stated in Benjamini and Yekutieli (2002), PRDS condition is very similar but not completely overlapping with association. For a review of these general conditions on dependence, see Lehmann (1966).*

General Weak Dependence Assumptions

Doukhan and Louhichi (1999) and then Nze *et al.* (2002) define a more general framework for weak dependence, which includes processes that satisfy mixing and association conditions, together with cases in which these two properties fail to hold, like Bernoulli shifts driven by discrete innovations. They define the set $\mathcal{L}_1 = \{h : h\ is\ Lipschitz,\ ||h||_\infty \leq 1\}$, and they define a weak dependent sequence $\{X_n\}_{n\in\mathcal{N}}$ to satisfy

$$|Cov(h(X_{i_1}, \ldots, X_{i_u}), k(X_{j_1}, \ldots, X_{j_v})| \leq \theta_r \psi_i(h, k, u, v),\ i = 1, 2.$$

where k and h are in $\mathcal{L}_1$, θ_r is a sequence of numbers decreasing to zero, $r = j_1 - i_u$, and $\psi_1(h, k, u, v) = \max(Lip(h), Lip(k))(u + v)$, $\psi_2(h, k, u, v) = Lip(h)Lip(k)\min(u, v)$; where $Lip(h)$ is the Lipschitz dimension of h. It is apparent that this is a definition similar to the one of mixing processes.

Conditions on weak dependence can be given for the same results of Theorem 3.1.6 to hold: if ψ_1 function is used and $\theta_r = O(r^{-5-v})$ or ψ_2 function is used and $\theta_r = O(r^{-15/2-v})$, then it is possible to prove weak convergence of the processes in $D[0, 1]$ to a centered Gaussian process indexed by $[0, 1]$, and the key lemmas 3.1.3, 3.1.4, 3.1.5. The covariance kernel of $\sqrt{n}(\widehat{G}(t) - G(t))$ is $2\sum_{k=0}^{+\infty} Cov(1_{p_1\leq s}, 1_{p_k\leq t})$, which reduces to the covariance kernels defined in Appendix A in case of mixing or association. Proofs and generalization of the covariance kernels is straightforward and omitted for brevity.

3.1.2 Mean and Variance of the FDP

We argue in this section that the expected value of the FDP is unchanged by dependence between the test statistics; while its variance will be. Suppose that the variables on the

field are not independent, so that the p values will not be independent. Let $I^m = (I_1, \ldots, I_m)$, where $I_j = 1_{\{p_j \leq t\}}$. Let moreover $Q(t) = (1-a)t/G(t)$.

Genovese and Wasserman (2002) prove that $E\Gamma(t) = Q(t)(1-(1-G(t))^m)$. As long as I_i is independent of H_j given I_j, it's easy to see that[1] $E(\Gamma(t)|I^m) = Q(t)1_{\{some\ p_i \leq t\}}$. Taking expectations, we have $E(\Gamma(t)) = Q(t)(1-(1-G(t))^m)$, which is the same expression as above.

Lemma 3.1.9. *The variance of the FDP process $V(\Gamma(t))$ is equal under dependence to* $Q(t)^2(1-(1-G(t))^m)(1-G(t))^m + E[\frac{\sum_{i\neq j} I_i I_j[\Pr(H_i=0,H_j=0|I^m)-Q(t)^2]}{(\sum I_i + \prod(1-I_i))^2}]$.

Proof. To compute the variance, we will just apply the well known formula:

$$V(\Gamma(t)) = V(E(\Gamma(t)|I^m)) + E(V(\Gamma(t)|I^m)).$$

Note that $\Pr(H_i = 0|I_i = 1) = Q(t)$.

Then, independently of the correlation structure,

$$E(\Gamma(t)|I^m) = Q(t)1_{\{some\ p_i \leq t\}}$$

and[2]

$$V(Q(t)1_{\{some\ p_i \leq t\}}) = Q^2(t)(1-(1-G(t))^m)(1-G(t))^m.$$

On the other hand,

$$V(\Gamma(t)|I^m) = E(\Gamma^2(t)|I^m) - Q^2(t)1_{\{some\ p_i \leq t\}},$$

and

$$\begin{aligned} E(\Gamma^2(t)|I^m) &= E(\frac{\sum_{ij} I_i I_j (1-H_i)(1-H_j)}{(\sum I_i + \prod(1-I_i))^2}|I^m) \\ &= \frac{\sum_i I_i \Pr(H_i = 0|I_i)^2 + \sum_{i\neq j} I_i I_j \Pr(H_i = 0, H_j = 0|I^m)}{(\sum I_i + \prod(1-I_i))^2} \\ &= Q(t)^2 1_{\{some\ p_i < t\}} + \frac{\sum_{i\neq j} I_i I_j [\Pr(H_i = 0, H_j = 0|I^m) - Q(t)^2]}{(\sum I_i + \prod(1-I_i))^2}. \end{aligned}$$

In the end,

$$V(\Gamma(t)) = Q(t)^2(1-(1-G(t))^m)(1-G(t))^m + E[\frac{\sum_{i\neq j} I_i I_j [\Pr(H_i = 0, H_j = 0|I^m) - Q(t)^2]}{(\sum I_i + \prod(1-I_i))^2}].$$

□

[1] If the p-values are dependent, it is not reasonable to assume independence between H_j and p_i, $i \neq j$, tout court. It is usually reasonable, by the way, to assume conditional independence between H_j and p_i, $i \neq j$.

[2] Note that $I_i \sim Bernoulli < G(t) >$.

It is apparent that, if all the p values are independent; then the second term in the expression of $V(\Gamma(t))$ as derived in Lemma 3.1.9 will be zero. This reasoning suggests that in general it is not sensible to control the FDR under strong dependence: since the variance of the FDP may be increased by dependence, control of the quantiles of the distribution of the FDP is in general more desirable. Refer to the following chapter for methods to control the $tFDP$ under dependence for any value of m. A similar argument is made in Bickel (2004).

3.2 The Simulations

3.2.1 Gaussian Data

We applied the BH and the plug-in method to correlated spatial data; to illustrate the effects of dependence on the outcomes of standard i.i.d. methods for controlling the FDR. The case of stationary and isotropic spatial data is particularly interesting; since data are usually correlated when close to each other, but as distance increases the correlations fade to zero. Good references on spatial statistics are Smith (2001) and Banerjee *et al.* (2004).

We will simulate data on a regular quadratic grid of r by r pixels, with $m = r^2$. We want to test the mean on each pixel, to discover if it is different than zero. Note that this can be a very common setting, for instance, in neuroimaging; where some variables in m spots of the brain are measured to see if there is neuronal activity, or in environmental statistics where the presence of a certain pollutant is verified in different points of a city. We will randomly assign M_1 pixels to a non-zero mean (uniform in $(0, 5)$); and the variance/covariance matrix will remain the same throughout the iterations. For each set of parameters, we will do 1000 iterations.

The Covariance Structures

We will use two covariance structures. The first is a simplified version of an exponential covariance structure commonly used in spatial statistics: the covariance between two different pixels will always be non-negative, and determined by $e^{-\frac{1}{\tau}\ d(x,y)}$, where x and y are the coordinates on the plane of the two pixels, $d(\cdot, \cdot)$ is the euclidean distance function and τ is just a tuning parameter. The higher τ, the more slowly decaying the correlation. The second covariance structure will allow for both positive and negative covariances, and a suitable structure will be given by the "damped cosine" function:

$$e^{-\frac{1}{\tau}\ d(x,y)} \cos(\frac{1}{\tau}\ d(x, y)).$$

Of course, we will need to make sure that the variance/covariance matrix is positive definite. Abrahmsen (1997) proves that the lower bound for the correlation value, in order to maintain positive definiteness, is -0.4 in two dimensions. The "damped cosine"

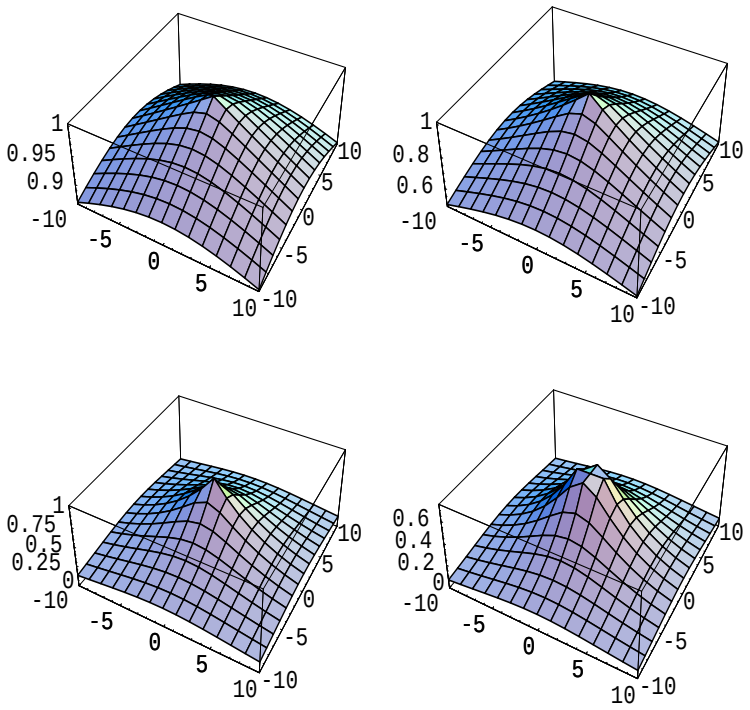

Figure 3.1: Covariance structure, "simplified exponential" covariance function

structure will allow us to have correlations as low as -.39; thus being almost as extreme as possible.

Figure 3.1 shows a plot of the values of the covariance computed as $e^{-\frac{1}{\tau}\ d(x_0,x)}$ for a point x_0 in the middle of the grid and different values of τ. Figure 3.2 shows shapes of the "damped cosine" structure for a point x_0 in the middle of the grid and other values of τ.

The following analyses are done for $r = 10$ by 10, 40x40 and 100x100 grids. Note that with this two functions, $cov(x_0, x_0) = 1$, so that the covariances will also be correlation values.

Results, All Positive Correlations

Tables 3.1, 3.2, 3.3 show the range of the covariances on each grid for the tuning parameters chosen, with the "simplified exponential" covariance function[3].

Figure 3.3 shows the results for the FDR using the standard BH method. Figure 3.4 shows the results for the FDR using the plug-in method.

It is evident that, as long as the relationship between the variables fades to independence fast enough, the methods are still working; and the plug-in is sensibly less conservative than the BH. When the correlation becomes too strong, the BH is valid

[3]All the simulations were programmed in C language.

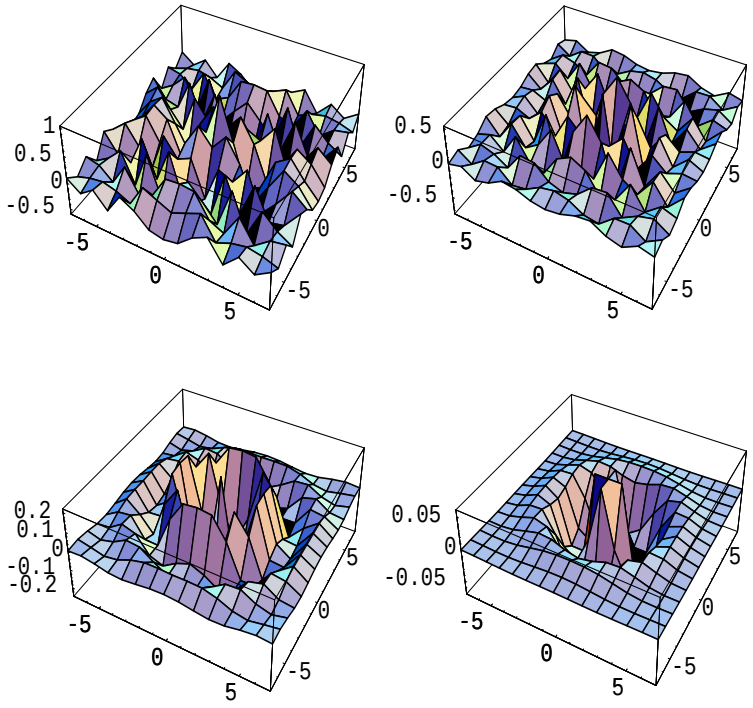

Figure 3.2: Covariance structure, "damped cosine" covariance function

$1/\tau$	.02	.05	.1	.3	.5	.6	.7	.8	1	1.2
min	0.77	0.53	0.28	.02	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	.90	.74	0.61	.54	.50	.45	.37	.30

Table 3.1: Correlation Ranges, 10 by 10 grid, exponential model

τ	50	20	10	3.33	2	1.66	1.43	1.25	1	0.833
min	0.33	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	0.90	0.74	0.61	0.54	.50	0.45	0.37	0.30

Table 3.2: Correlation Ranges, 40 by 40 grid, exponential model

τ	50	20	10	3.33	2	1.66	1.43	1.25	1	0.833
min	0.06	0.00	0.00	.00	0.00	0.00	0.00	0.00	0.00	0.00
max	0.98	0.95	.90	.74	0.61	.54	.50	.45	.37	.30

Table 3.3: Correlation Ranges, 100 by 100 grid, exponential model

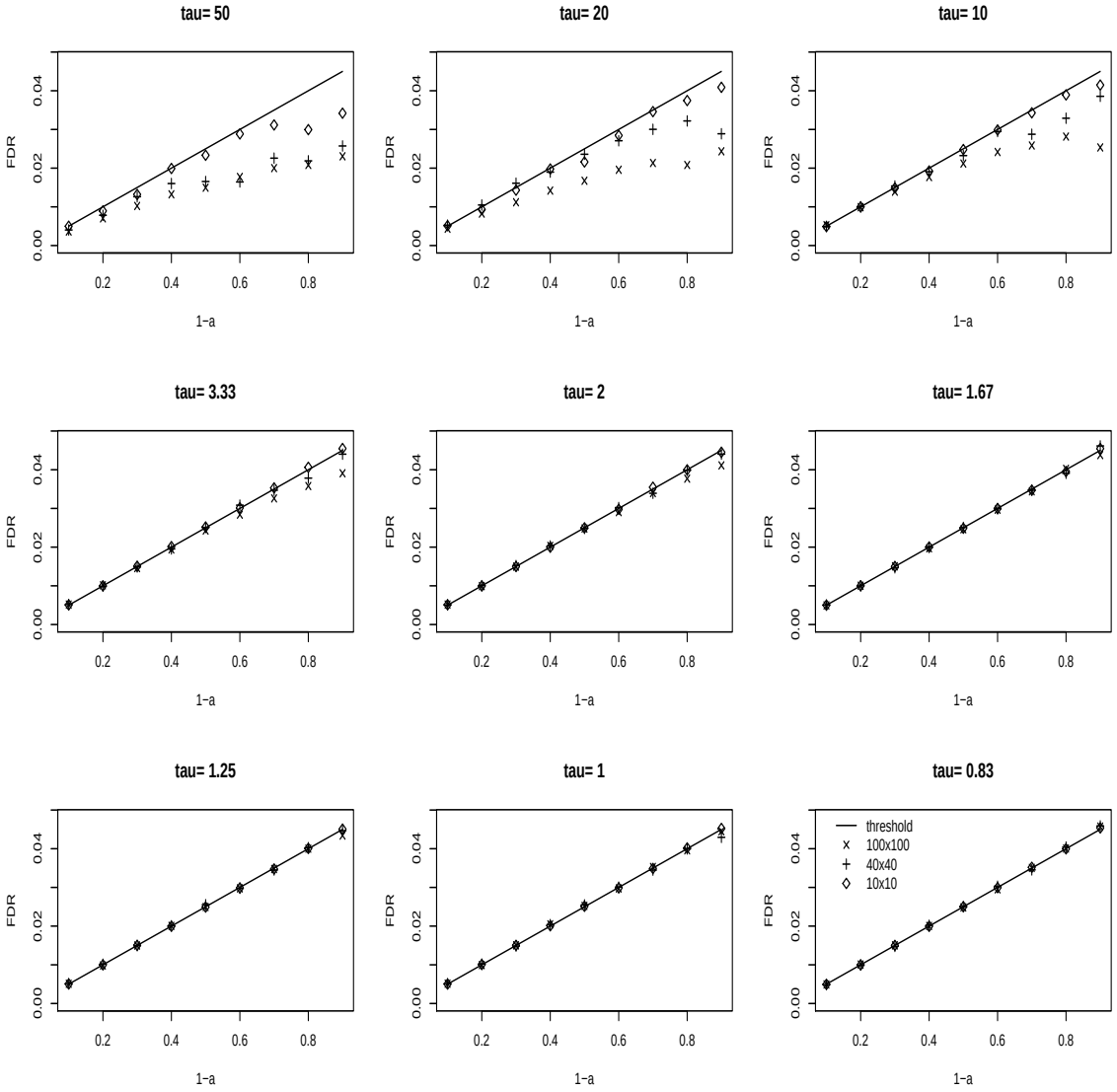

Figure 3.3: FDR, BH, positive case, Normal Data

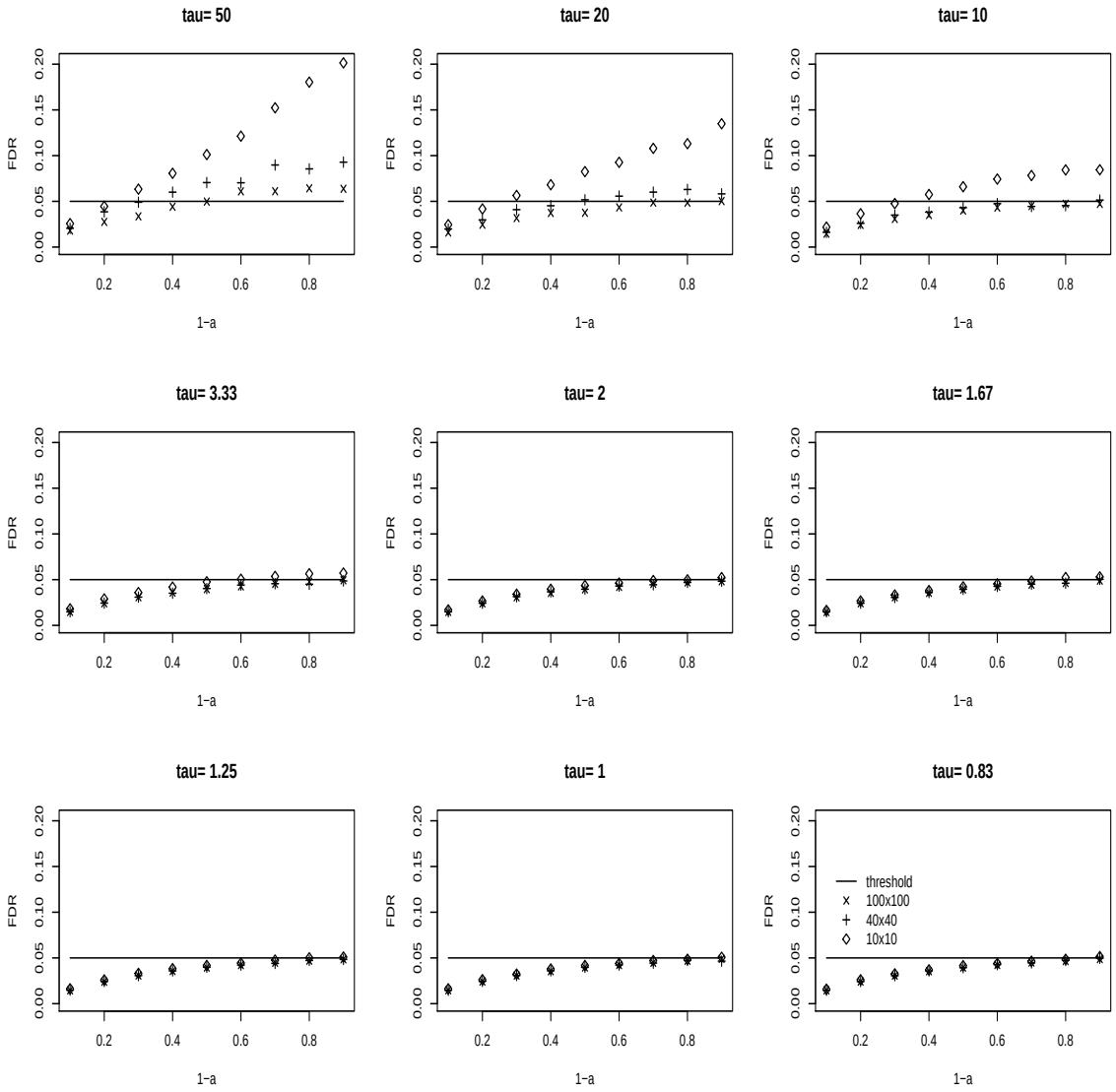

Figure 3.4: FDR, Plug-in, positive case, Normal Data

τ	0.2	0.33	0.73	0.77	0.83	0.98	14.29	16.67	20
min	.00	-.05	-.10	-.16	-.24	-.36	-.39	-.34	-.22
max	.00	.01	.14	.14	.11	.13	.91	.92	.94

Table 3.4: Correlation Ranges, "damped cosine"

but becomes even more conservative, while the plug-in violates the threshold and gets bigger than $\alpha = .05$. We will show later that this problem is determined by the estimator of a.
A comment should be given in relationship with the results of Benjamini and Yekutieli (2001): in the examples in this subsection, their assumption of PRDS is always satisfied. Nevertheless, the BH procedure becomes overly conservative, in some cases, under their assumptions. Note that under the assumptions of our Theorem 3.1.6 the BH procedures behave just like under independence, starting from reasonably small values of m.

Results, Negative Correlations

Table 3.4 shows the range of the covariances for the "damped cosine" covariance function, for all the grids (it is so because the highest and smallest covariance values are attained between close pixels, and then the covariance gets closer and closer to zero as the distance increases). During the simulation of the bigger grids, certain cases were dropped because the variance covariance matrix lost the positive definiteness property due to machine error in approximation.

Figure 3.5 shows the results for the FDR using the standard BH method. Figure 3.6 shows the results for the FDR using the plug-in method.

The negative case gives almost same, when not better, results as the positive one. Note that there were problems when the correlation became too high (bigger than 0.9), and nothing wrong was observed in the cases in which the correlation was low.

3.2.2 Pearson Type VII Data

The same random fields were simulated with Pearson Type VII random variables. See Johnson (1987) for a review of multivariate statistical simulation when data are not normal. The degrees of freedom were chosen to be 3, because these random variables are the most dissimilar from normality in the Pearson Type VII family, but still with finite mean and variance. Using three degrees of freedom makes also straightforward the comparison with the normal case, since the covariance matrix, using the same parameters, will be unchanged.

Figure 3.7 shows the results for the FDR using the standard BH method. Highest correlations made the procedure be conservative also in this case. There is also more instability in the observed FDRs (and even a couple of FDRs just over the threshold), most likely due to higher simulation variability. Figure 3.8 shows the results for the

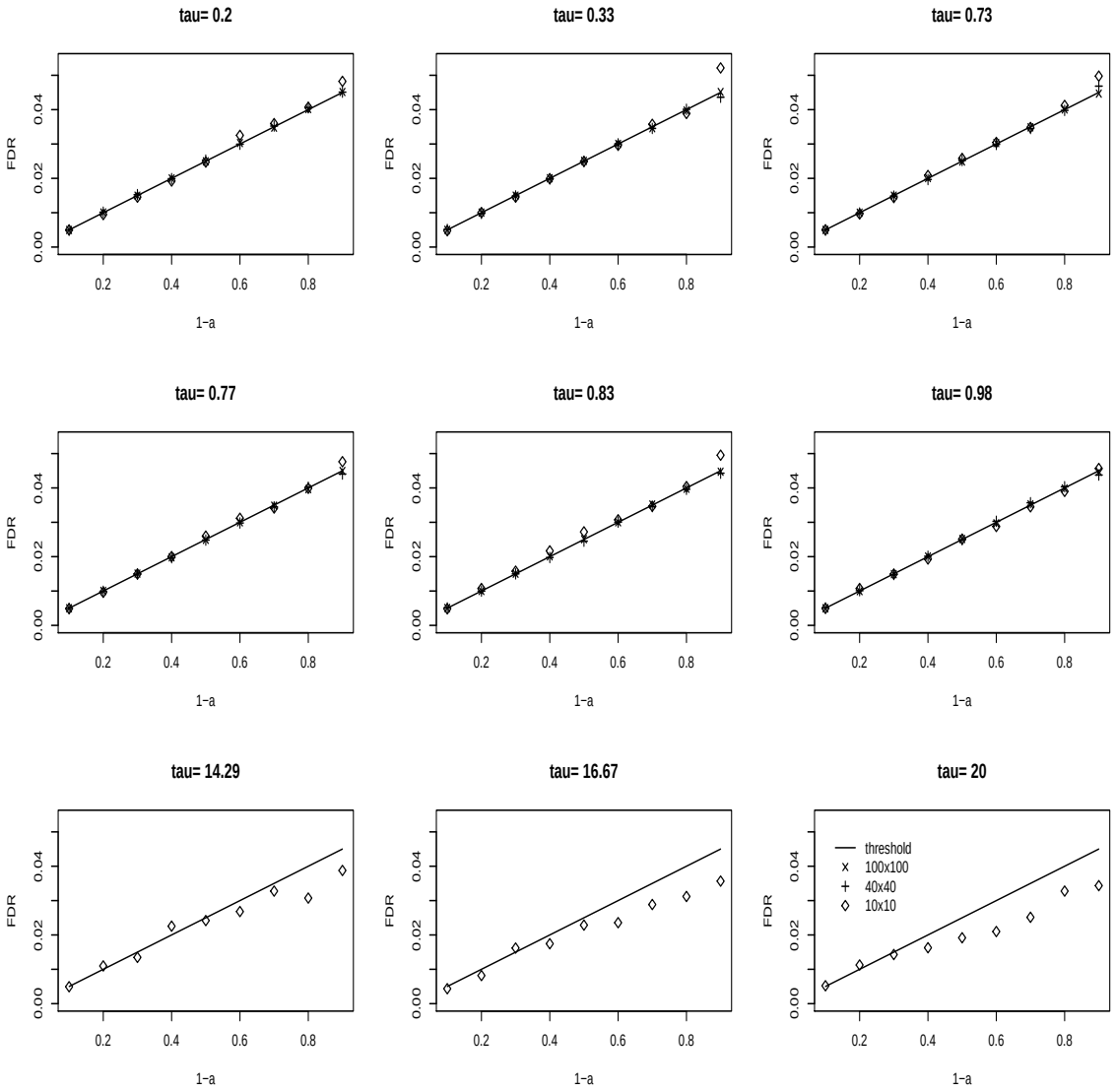

Figure 3.5: FDR, BH, negative case, Normal Data

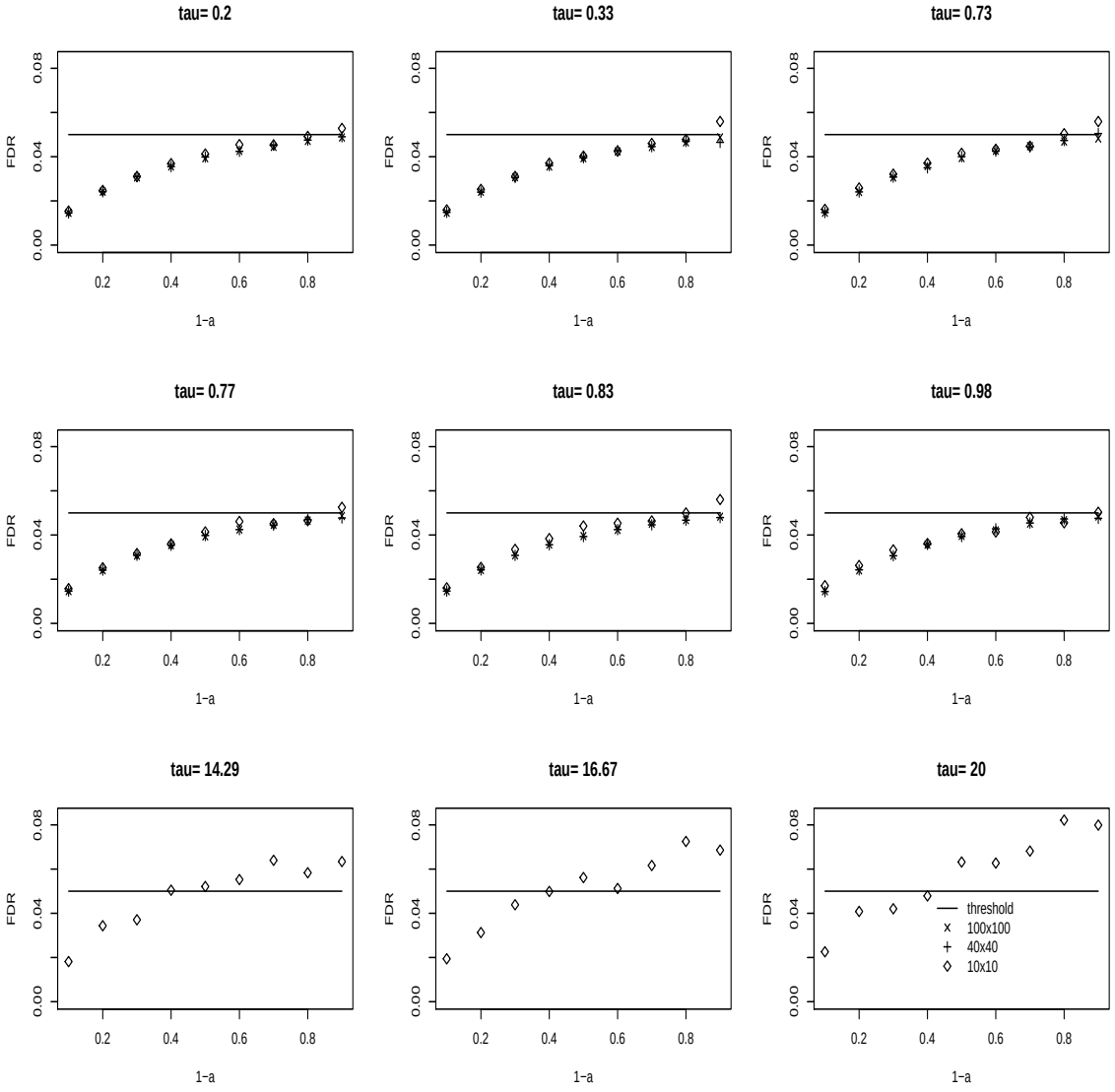

Figure 3.6: FDR, Plug-in, negative case, Normal Data

FDR using the plug-in method. The break down for high correlation is slightly worse than the normal case. Note that previously $\tau = 3.33$ didn't show any violation of the threshold.

Remark 3.2.1. *It seems like strong positive correlations can lead the standard BH method to be more conservative than it already is; but it will not lead it violation of the threshold for the FDR. On the other hand, the simulations show that the plug-in method is not robust under dependence; since many violations of the threshold are seen. Note moreover that the simulations show that the hypotheses of Theorem 3.1.6 are only sufficient.*

In the previous section we proved that under wide hypotheses on the dependence of the test statistics the BH method remains valid. This section gives an illustration of this behavior through extensive simulations. On the other hand, the same results were proved for the Plug-in procedure with a suitable estimator for a. The simulations show that the common estimator used for a is not robust under dependence. An explicit proposal for a suitable estimator is given in the next section.

3.3 Estimating a with Dependent Data

A key thing in the plug-in method is the choice of the estimator for a. It is obvious that, as long as $0 \leq \widehat{a} \leq a$ the power is increased with respect to the BH procedure; while the FDR is still below the desired threshold. So, if anything, a statistic that underestimates a is desirable because it improves on BH while being at least conservative.

3.3.1 Oracle Simulations

In an attempt to understand why the plug-in method failed when the correlation between close variables was very high, we implemented a simulation in which the proportion of true nulls, $1 - a$, was considered known and used in the procedure. Figure 3.9 shows the results: the estimator used for a was the thing that broke down when using strong correlations, while now the plug-in method works just as it should. The case of strong correlations, when parameter is close to 0, brings about just an increase in the variance (less stability)[4].

It is easy to see how Storey's estimator breaks down by strongly overestimating a. Figure 3.10 shows $\frac{\widehat{a}}{a}$, where $\widehat{a}$ is Storey's estimator, in the usual $10x10$ Gaussian random fields. We observed values as much as 3.5 times the real a.

[4]Note that in real cases it is reasonable to assume sparseness, i.e., $1 - a > .5$ or even much more. Then, as we can see, it is reasonable not to expect problems even from Storey's estimator.

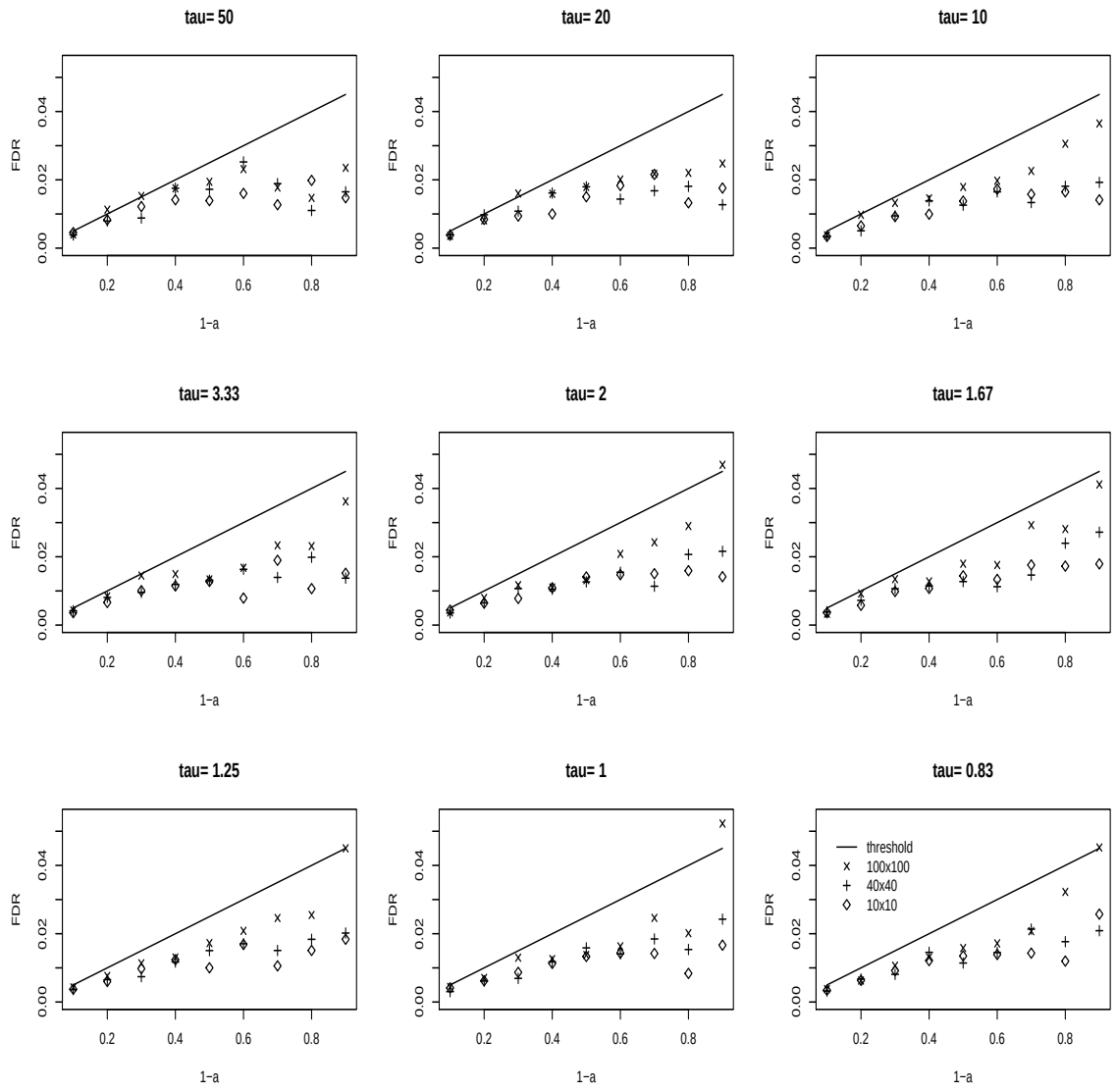

Figure 3.7: FDR, BH, positive case, T_3 Data

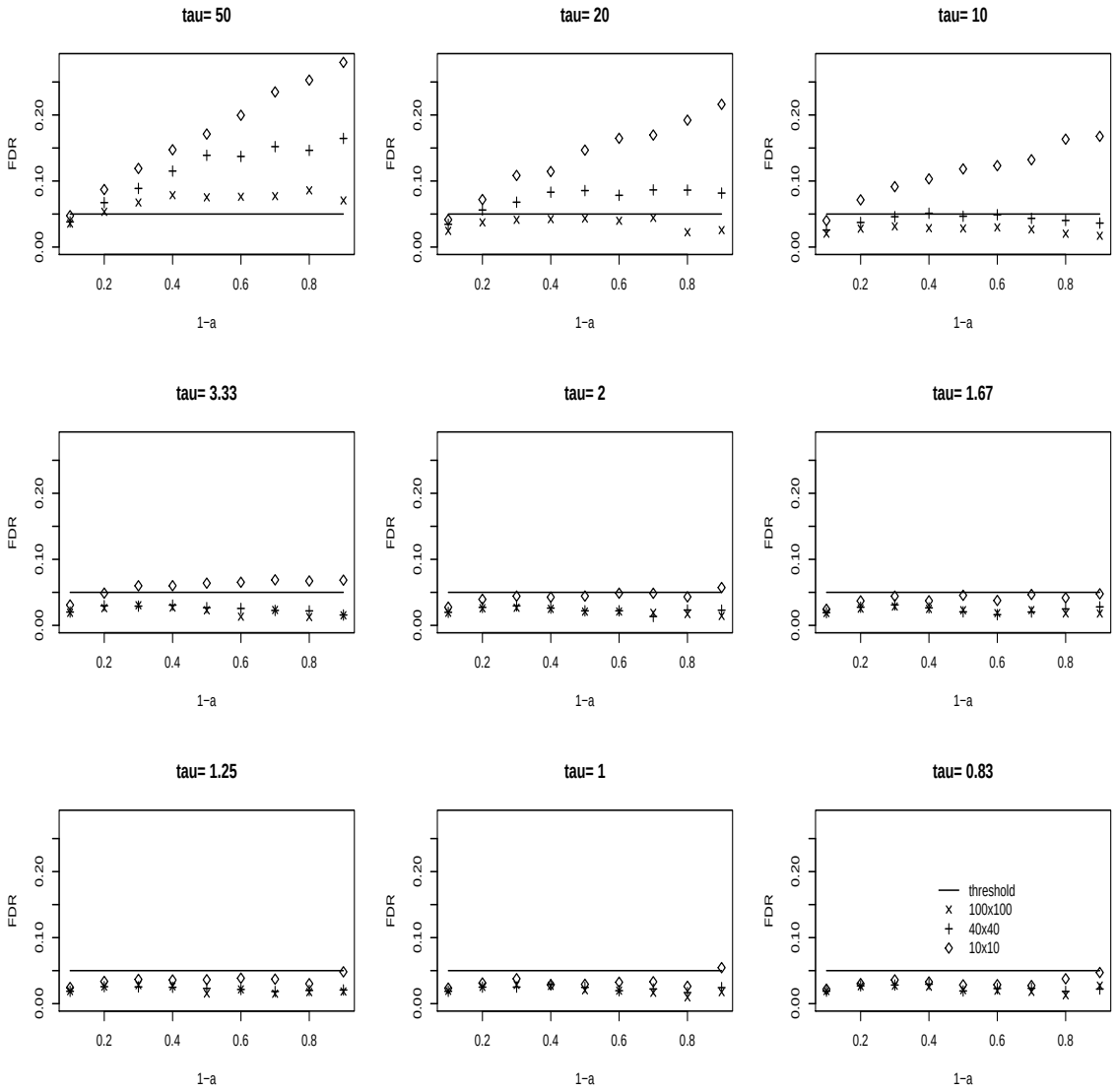

Figure 3.8: FDR, Plug-in, positive case, T_3 data

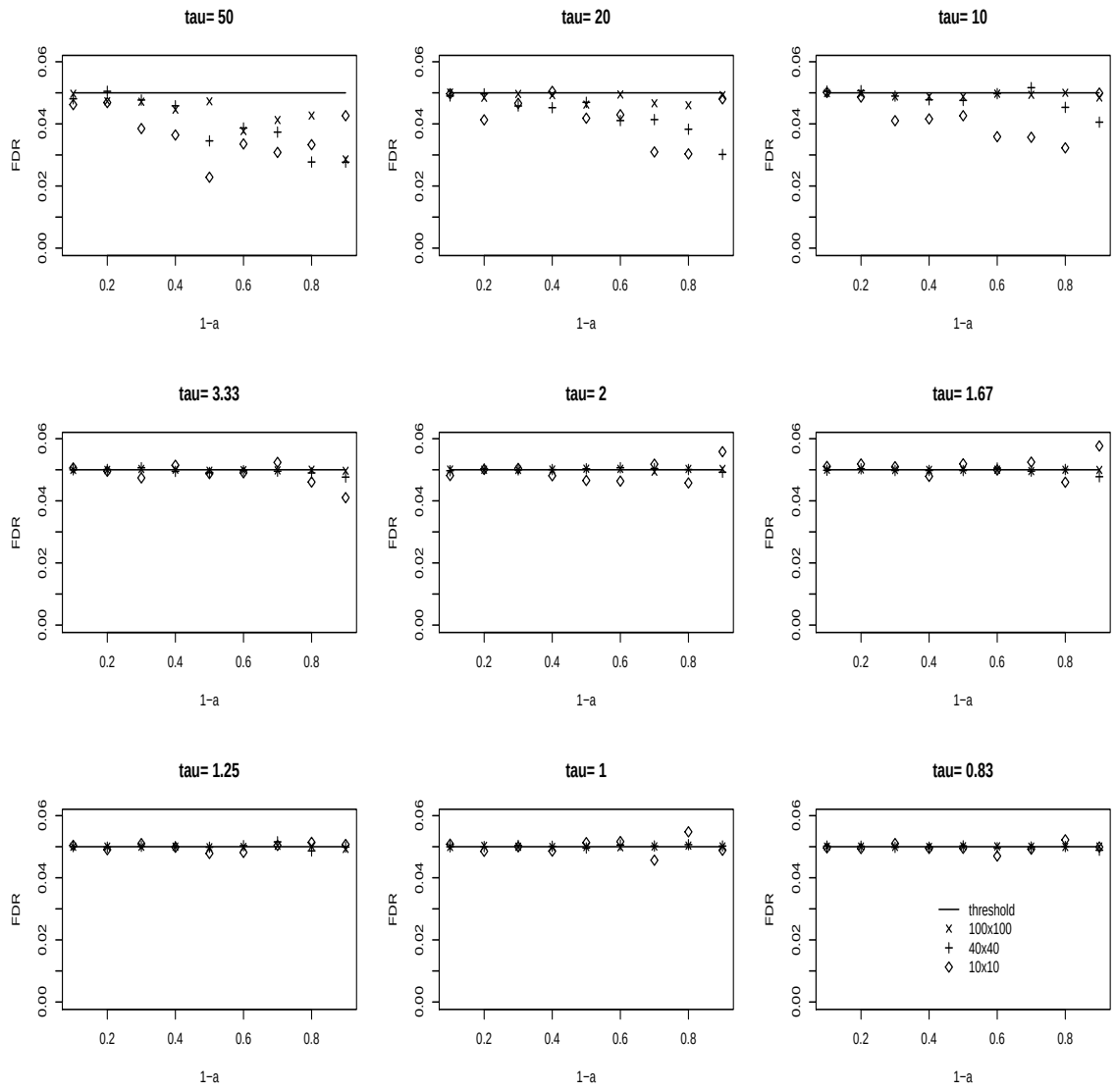

Figure 3.9: FDR, oracle simulation, 1000 iterations

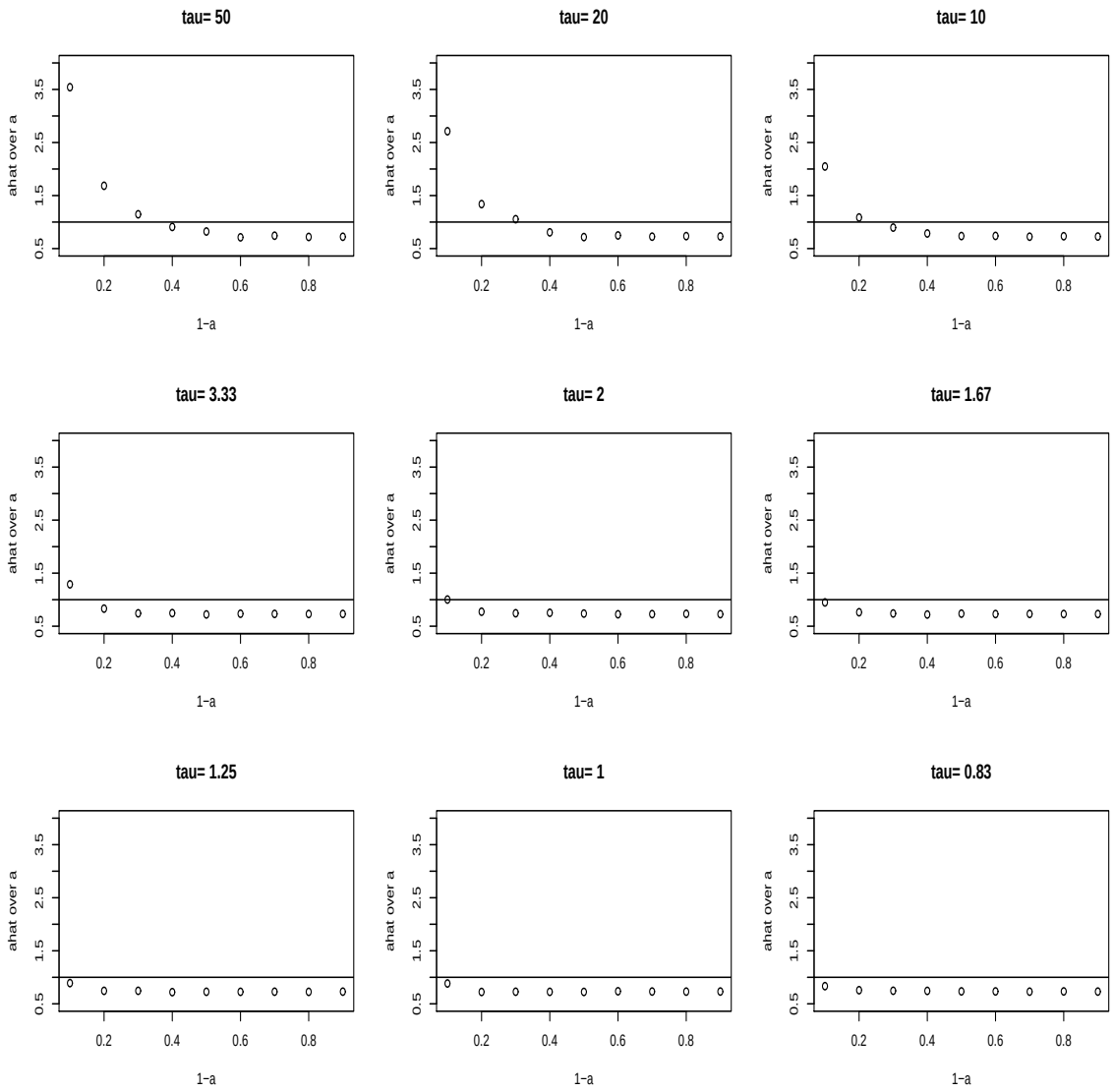

Figure 3.10: Storey Estimator relative to a, 1000 iterations

3.3.2 Iterative Simulation of a

The "iterative plug-in method", which we will describe in this section, proves much more robust with respect to dependence than Storey's estimator. The other classical estimators, like the one in Swanepoel (1999) or the one in Woodroofe and Sun (1999), are also seen to break down under dependence[5]. Since we need to estimate M_1, the number of false nulls, we thought the most natural estimator was the number of rejected hypotheses. The proportion a is then estimated iteratively as the proportion of rejected nulls in the previous plug-in step; till a does not change in two subsequent iterations, with a BH method at the first iteration (i.e., the first estimator is always set to 0). A similar single-stage estimator (i.e., always doing a single iteration) was independently derived in Benjamini *et al.* (2004). They prove, under independence, that the single-stage estimator is in fact conservative. They also suggest an iterative estimator similar to our proposal, and note that, as in our case, this kind of estimators possess an interesting internal coherence property: the final number of rejected hypotheses is used as an estimator of the number of false nulls.

For the usual Gaussian simulations with positive correlation structure, Figure 3.11 shows the results of the plug-in procedure with the iterative estimator, with $a_0 = 0$. The average number of steps was always between 2 and 7. Note that this procedure manages to control the FDR at level 5% when the correlation is very strong (robustness), while it behaves just like the old one-step procedure when the correlation is weak (it just seems to be a little bit more conservative).

On the estimator level, Figure 3.12 shows the ratio between the iterative estimator and the real a. In all cases we succeeded in being conservative.

It is straightforward to prove that the number of iterations is finite: there are only $m+1$ possible values for $\widehat{a}$. Then, the random variable given by the difference between the current and the previous estimate is discrete and puts a non null probability mass at 0, which is our stopping rule.

3.3.3 Iteration in theory

We will first investigate the iterative estimator at a population level. Let $R(t) = t/G(t)$. Let a_n be the estimate for a at the n-th iteration, and t_n the corresponding deciding point. We have: $a_n = G(t_{n-1})$ and $t_n = R^{-1}(\frac{\alpha}{1-a_n})$, with $a_0 = 0$.

It is straightforward to see that $G(R^{-1}(s)) = \frac{R^{-1}(s)}{s}$. With this, one can prove that:

$$\begin{cases} a_n = \frac{1}{\alpha^n} \sum_{i=0}^{n-1} (\prod_{k=n-i}^{n-1} t_k) \alpha^{n-i-1} (-1)^{i-1} \\ t_n = R^{-1}(\frac{\alpha}{1-a_n}). \end{cases}$$

One could easily obtain an expression for a_n independent of t_k, just by substituting t_k

[5]Simulations of the other estimators not shown for reasons of space.

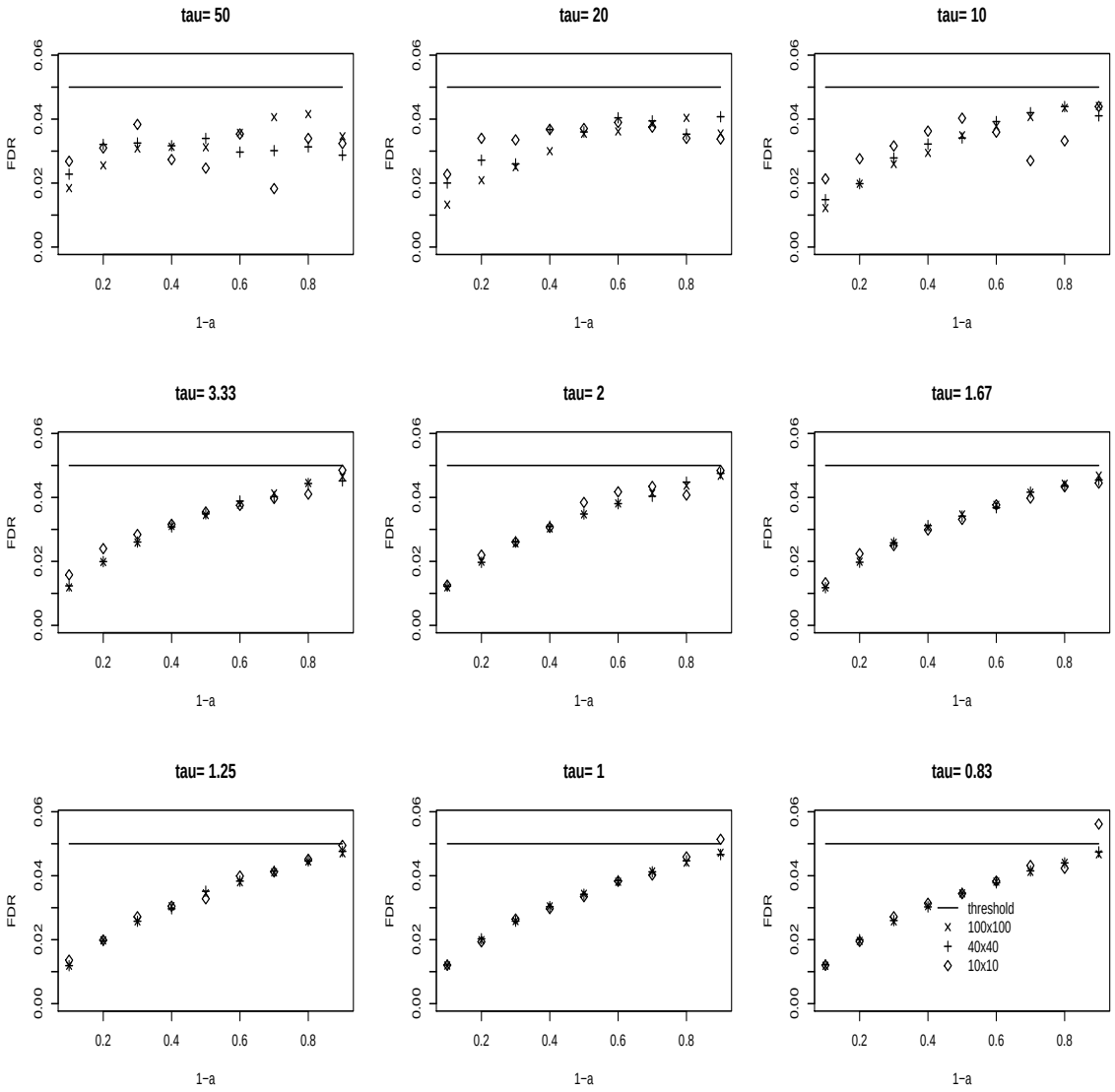

Figure 3.11: FDR, iterative simulation, 1000 iterations

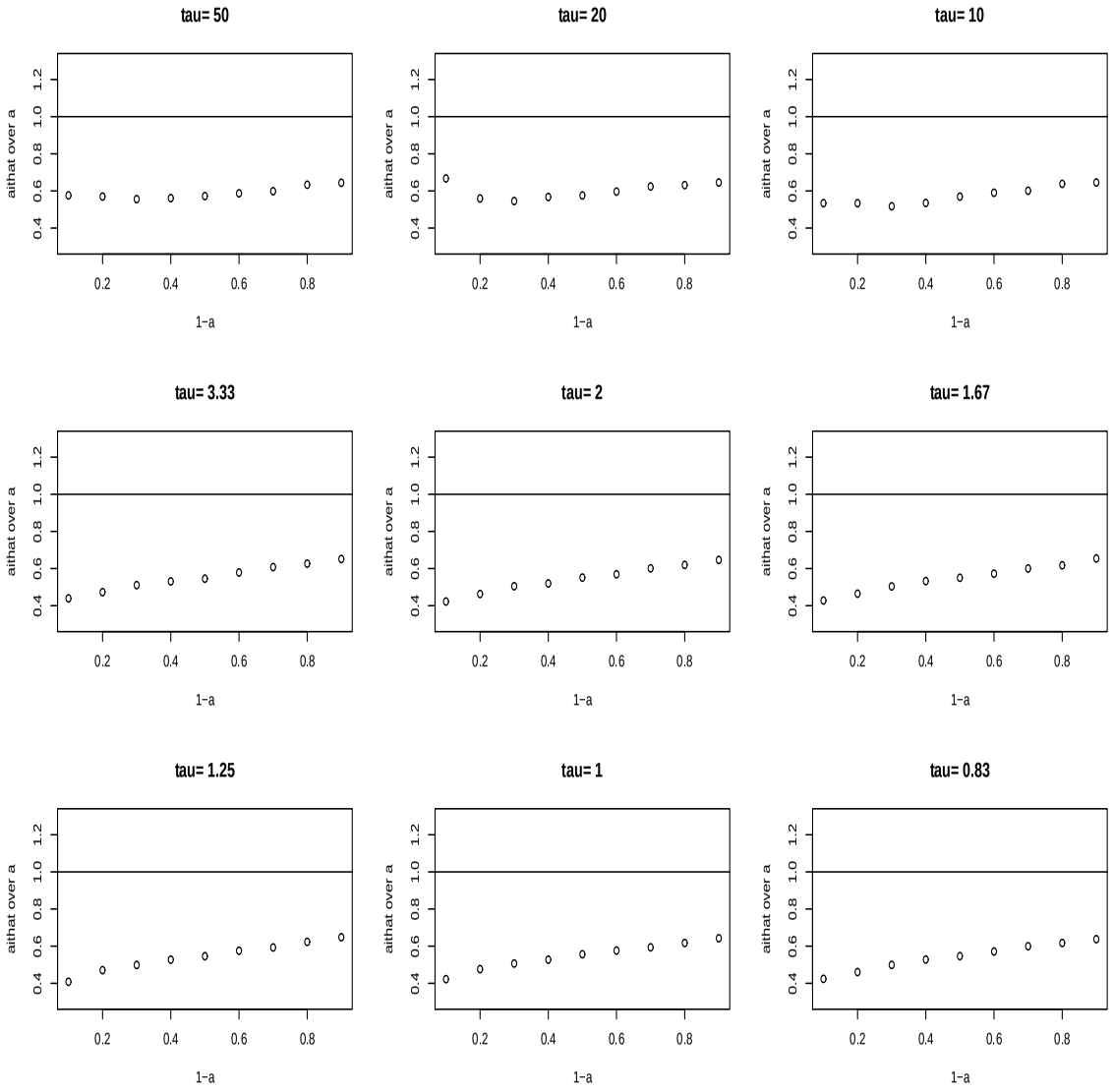

Figure 3.12: Iterative Estimator relative to a, 1000 iterations

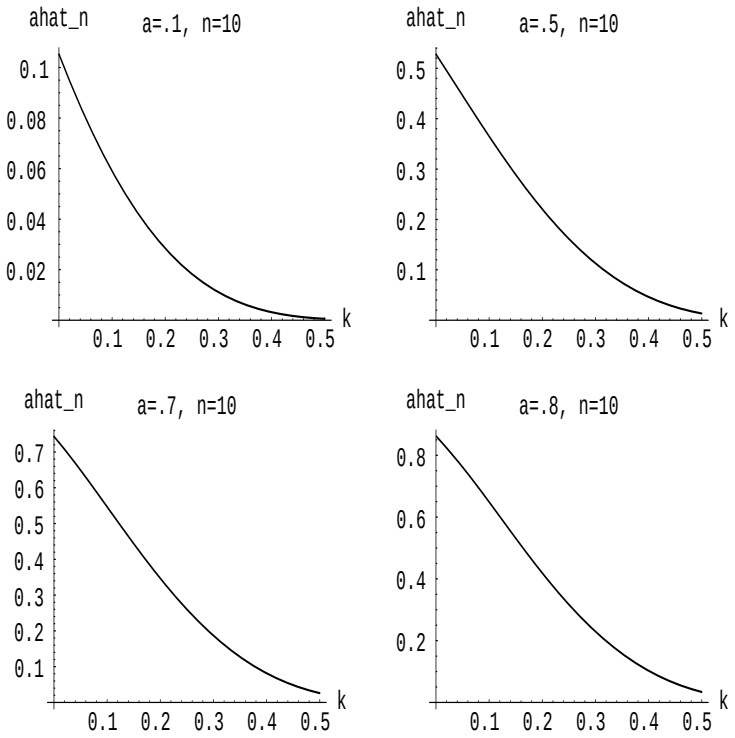

Figure 3.13: Iterative $\widehat{a}$ in the Beta Example

with $R^{-1}(\frac{\alpha}{1-a_k})$. The iterative estimator at sample level is just the one obtained by substituting $G(t)$ with the empirical distribution.

An example

In a completely different context, Sellke *et al.* (2001) propose an example in which F is a $beta(\xi, 1)$ with $\xi \in (0, 1)$. In that case we get a closed form expression for $G(t) = (1-a)t + at^{\xi}$ and $R^{-1}(y) = \left(\frac{1-y-ay}{ay}\right)^{\frac{1}{\xi-1}}$.

It is easy to see that:

$$\begin{cases} a_0 = 0 \\ a_n = \left(\frac{1-a_{n-1}-\alpha+a\alpha}{a\alpha}\right)^{\frac{1}{\xi-1}} \frac{1-a_{n-1}}{\alpha}. \end{cases}$$

Figure 3.13 shows the estimator as a function of ξ, after 10 iterations, for various values of the true a. The estimator converges to a certain value after very few iterations, and 10 were enough in all cases.

So as we get closer to the unidentifiable case, $\widehat{a}$ gets closer and closer to 0. In general, one could expect this strong dependence on the alternative distribution F. Moreover, it is possible, even if hard, to overestimate a, thus being anti-conservative. As we see, this happens only for ξ close to 0. In this sense, one could think about the iterative

estimator for a as a conservative estimator which outperforms taking $\widehat{a} = 0$ like in the BH procedure; and that it is robust with respect to the dependence structure. In Chapter 4, we will generalize the iterative approach and provide different estimators that are conservative under dependence.

Chapter 4

Estimating the number of False Null Hypotheses

In this Chapter we will briefly give some results and explicit proposals for estimating the number of false null hypotheses. Such estimates, as seen, can greatly improve the power of many MTPs, for instance in passing from the BH to the plug-in method. Many other procedures can benefit from a good estimator of M_1.

Not surprisingly, there is not a $tFDP(c)$ controlling procedure which can make use of an estimator of M_1 available yet. We propose one at the end of the chapter, based on the generalized augmentation procedure proposed in Chapter 2.

The usual estimators are seen to break down under dependence (see Chapter 3). We will propose here a class of estimators robust with respect to dependence.

Many of the ideas that will be proposed are grounds for further work.

First, recall that a suitable estimator for M_1 is a conservative one. I.e., we want $\widehat{m}_1 \leq M_1$, but as close as possible to the upper bound. This will increase the power without violating the condition on the Type I error rate. This is equivalent to looking for confidence intervals for M_1, which will be in the form $[\widehat{m}_1, m]$.

The basic idea is as follows: first of all, note that, in the notation of Table 1.1 of page 3,

$$M_1 = R - N_{1|0} + N_{0|1}.$$

We usually don't know much directly on M_1, while the random variables on the right hand side are dealt with in MTPs.

In the previous chapter we proposed a multi-step procedure that took $\widehat{m}_1 = R$, where R is the number of rejections at the previous stage. We will give here some insights of when this is a conservative choice, and propose a family of procedures to estimate M_1 in this fashion. We will follow two tracks: we will conservatively approximate $N_{0|1}$ to 0 and require that with high probability $\widehat{m}_1 < R - N_{1|0}$, on which we have bounds. A more complex path is to include $N_{0|1}$ in the considerations.

Note that whenever $N_{0|1} \geq N_{1|0}$, R is a good conservative estimator of M_1. All depends on the controlled error measure, α, m and F. In general, anyway, experience

and simulations suggest that this is often true, especially for big m.

In what follows, we will always take into account the uncertainty brought about by the estimation of M_1 when controlling the desired error measure. This will be done by controlling the error measure at a certain level $\alpha_2 \leq \alpha$, which will be exactly determined. It is common in literature *not* to incorporate this uncertainty. Obviously, in what follows, this corresponds to using $\alpha_2 = \alpha$. We will make some comparisons at the end of the chapter.

4.1 Two-step Procedures

We define a k-step procedure as a procedure that estimates M_1 through $k-1$ MTP steps, and then controls a pre-specified error rate using the estimate found in the previous steps. Note that the iterative procedure in the previous chapter is a k-step procedure with unknown, random, k. A particular case is given by two step procedures. In our calculations, we will always condition on R, since we will know it from the previous step.

4.1.1 Two-step procedures based on FWER control

Many $FWER$ controlling procedures work under arbitrary dependence (like Bonferroni). Let R_{Bonf}, in the usual notation, denote the number of rejected hypotheses controlling the FWER at level α_1.

It is easily seen than, under arbitrary dependence, $\widehat{m}_1 = R_{Bonf}$ is conservative with high probability: $\Pr(R_{Bonf} < R_{Bonf} - N_{1|0} + N_{0|1}) \geq \Pr(N_{1|0} = 0) > 1 - \alpha_1$. Note that the second inequality in practice is always strict, and that typically $\Pr(N_{0|1} > 0) >> 0$, so that the bound is far from being sharp.

Two-step control of FWER under arbitrary dependence

At the second step, one can reject all the p-values smaller than $\alpha_2/(m - R_{Bonf})$ and control $FWER$ at level $\alpha = 1-(1-\alpha_1)(1-\alpha_2)$ in this way, under arbitrary dependence. Note in fact that if m_1 is known, rejecting if $p_j < \alpha/(m - m_1)$ controls FWER at level α. Hence,

$$\begin{aligned}
\Pr(FWER = 0) &= \Pr(FWER = 0 | R_{Bonf} < M_1) \Pr(R_{Bonf} < M_1) \\
&\quad + \Pr(FWER = 0 | R_{Bonf} > M_1) \Pr(R_{Bonf} > M_1) \\
&\geq (1 - \alpha_2)(1 - \alpha_1) + \Pr(FWER = 0 | R_{Bonf} > M_1) \Pr(R_{Bonf} > M_1) \\
&\geq (1 - \alpha_2)(1 - \alpha_1)
\end{aligned}$$

Two-step control of FDR under PRDS

Now we can improve on the results of Benjamini and Yekutieli (2001). If at the second step a Plug-in procedure at level α_2 is done, FDR is controlled at level $\alpha = \alpha_2(1 +$

$(R_{Bonf}/(m - R_{Bonf}))\alpha_1)$ for any m under PRDS assumptions. In fact:

$$\begin{aligned} E[FDP] &= E[FDP|R_{Bonf} < M_1]\Pr(R_{Bonf} < M_1) \\ &\quad + E[FDP|R_{Bonf} > M_1]\Pr(R_{Bonf} > M_1) \\ &\leq \alpha_2 \Pr(R_{Bonf} < M_1) + \frac{(m - M_1)\alpha_2}{m - R_{Bonf}}\alpha_1 \\ &\leq \alpha_2(1 - \alpha_1) + \frac{m}{m - R_{Bonf}}\alpha_2\alpha_1. \end{aligned}$$

Note that the we can explicitly use R_{Bonf} in the bound since it is known, i.e., the randomness of the event $R_{Bonf} \lessgtr M_1$ is given only by M_1. So, taking any $\alpha_1 \in (0, 1)$ and $\alpha_2 = \alpha/(1 + (R_{Bonf}/(m - R_{Bonf}))\alpha_1)$ controls the FDR at the desired level α for any finite m under PRDS assumptions. This is a slight generalization of the results of Benjamini and Yekutieli (2001): they proved the result only for $\alpha_1 = 0$. Moreover, it is now possible to achieve better power using a sort of plug-in procedure under dependence. Note moreover that taking no correction and using $\alpha_2 = \alpha$ is sensible since in many cases FDR will be controlled with high probability at the desired level (see below). It is so because $(R_{Bonf}/(m - R_{Bonf}))\alpha_1$ is always very close to zero, and ignoring it is just a weak counter part of the conservativeness of the entire procedure.

Table 4.1 compares the choice of $\alpha_1 = 0$ (BH procedure) with the choice of $\alpha_1 = \alpha$. Simulation shows that there is improvement in choosing $\alpha_1 = \alpha$. Recall that the higher the FDR (still below α), the better the procedure in terms of power. This improvement is more and more evident as M_1 increases, for smaller m. Note that a very small increase in the FDR, especially for big m, can result in a much higher number of rejections, and hence in a much higher power.

4.1.2 Two-step procedures based on $tFDP(c)$ control

It is easily seen that, if at the first step a $tFDP(c)$ controlling procedure is used then a good estimator is $\lfloor R_{tFDP}(1 - c)\rfloor$:

$$\Pr(R_{tFDP}(1 - c) > R_{tFDP} - N_{1|0}) = \Pr(N_{1|0}/R_{tFDP} > 1 - (1 - c)) < \alpha_1.$$

Recall that, among the $tFDP(c)$ controlling procedures, there are many working under arbitrary dependence (augmentation, for instance). Refer to next chapter for other $tFDP(c)$ controlling procedures working under dependence for any finite m.

The second steps are the same as the ones used in the previous subsection.

Table 4.2 compare the BH procedure ($\alpha_1 = 0$) with $\alpha_1 = \alpha$, with and without correction. The $tFDP(c)$ controlling procedure chosen was the generalized augmentation procedure, proposed in Chapter 2, with level q divided by two in case of need of negative augmentation. The parameter c was taken to be 0.1. It is seen that the Bonferroni at the first step is better for small m, while using a good $tFDP(c)$ controlling procedure brings about an improvement for big m. This can be appreciated in Table 4.3, where estimators of M_1 are compared.

	$\alpha_1 = 0$ (BH)	$\alpha_1 = 0.05$, no correction	$\alpha_1 = 0.05$, corrected
	$M_1 = 0.1 * m$		
m=10	0.0418	0.0471	0.0468
m=30	0.0448	0.0467	0.0465
m=100	0.0408	0.0424	0.0423
m=500	0.0445	0.0459	0.0459
m=1000	0.0449	0.0465	0.0464
m=5000	0.0451	0.0465	0.0464
	$M_1 = 0.5 * m$		
m=10	0.0266	0.0349	0.0343
m=30	0.0241	0.0311	0.0305
m=100	0.0254	0.0310	0.0307
m=500	0.0249	0.0287	0.0285
m=1000	0.0251	0.0285	0.0283
m=5000	0.0248	0.0287	0.0286
	$M_1 = 0.9 * m$		
m=10	0.0053	0.0099	0.0094
m=30	0.0047	0.0081	0.0078
m=100	0.0050	0.0072	0.0070
m=500	0.0050	0.0067	0.0065
m=1000	0.0048	0.0064	0.0064
m=5000	0.0048	0.0061	0.0060

Table 4.1: Observed FDR for two-step control, using Bonferroni at first step

	$\alpha_1 = 0$ (BH)	$\alpha_1 = 0.05$, no correction	$\alpha_1 = 0.05$, corrected
	$M_1 = 0.1 * m$		
m=10	0.0418	0.0481	0.0480
m=30	0.0448	0.0467	0.0465
m=100	0.0408	0.0419	0.0419
m=500	0.0445	0.0463	0.0463
m=1000	0.0449	0.0472	0.0472
m=5000	0.0451	0.0476	0.0475
	$M_1 = 0.5 * m$		
m=10	0.0266	0.0327	0.0324
m=30	0.0241	0.0308	0.0304
m=100	0.0254	0.0337	0.0332
m=500	0.0249	0.0338	0.0332
m=1000	0.0251	0.0343	0.0337
m=5000	0.0248	0.0349	0.0342
	$M_1 = 0.9 * m$		
m=10	0.0053	0.0088	0.0085
m=30	0.0047	0.0087	0.0084
m=100	0.0050	0.0098	0.0093
m=500	0.0050	0.0103	0.0098
m=1000	0.0048	0.0102	0.0097
m=5000	0.0048	0.0103	0.0099

Table 4.2: Observed FDR for two-step control, using the Generalized Augmentation Procedure at the first step

Two-step procedure based on DKW approach

We propose here a two-step procedure based on a completely different reasoning, which is a simple refinement of Storey's estimator. Let $\varepsilon_m = \sqrt{\frac{1}{2m}\log(2/\alpha_1)}$. Define

$$\overline{(1-a)} = \min(\inf_s \frac{1-\widehat{G}(s)+\varepsilon_m}{1-s}, 1); \tag{4.1}$$

and let $\widehat{M_1} = m(1-\overline{(1-a)})$.

This estimator, that cannot be generalized to more than two steps, is based on the fact that

$$a \geq \frac{G(s)-s}{1-s} \tag{4.2}$$

for any $s \in (0,1)$, as easily seen by looking at the definition of $G(\cdot)$. By DKW inequality as defined in (1.10), $\Pr(1-a \leq \overline{(1-a)}) \geq 1-\alpha_1$. This directly provides a $1-\alpha_1$ confidence interval for $1-a$: $[0, \overline{(1-a)}]$, and another for M_1: $Pr(M_1 > (1-\overline{(1-a)})m) \geq 1-\alpha_1$.

It is easily seen that this is a refinement of Storey's estimator, which is also based on (4.2). Storey's estimator estimates the empirical distribution, since $a \geq \frac{\widehat{G}(s)-s}{1-s}$ for any $s \in (0,1)$ as m gets bigger. Here we take a lower confidence bound for the empirical distribution, substituting $G(s)$ with $\widehat{G}(s)-\varepsilon_m$ in (4.2), and take the infimum on $[0,1]$ instead of fixing an arbitrary s. This leads to a conservative estimator, with high probability, also for small m. Moreover, in Chapter 5 we will generalize DKW inequality under dependence, which will make possible to use this estimator also under certain hypotheses on the dependence among the test statistics.

4.1.3 Two-step procedures based on FDR control

Suppose BH procedure is used at the first step. It is easy to extend our approach to estimation of M_1 by taking an estimator that is good "on average". It is straightforward to see, in fact, that $\lfloor R(1-\alpha_1) \rfloor$ is on average smaller than M_1 if the FDR is controlled at level α_1 at the first step.

Alternatively, recall that any FDR controlling procedure is a $tFDP(c)$ controlling procedure at level α_1/c, so that $R_{BH}(1-c)$ is conservative with probability α_1/c. If a correction for this uncertainty is used (like the ones proposed in the previous sections), the chosen error measure will be controlled at the desired level even if α_1/c is not close to zero. Otherwise, if convexity assumption on the CDF of the FDP is taken, the FDR controlling procedure is $tFDP(0.5)$ controlling procedure at the same level α_1, as proved at page 5.

4.2 Multi-step procedures

A generalization of the iterative estimator proposed in the previous chapter is as follows:

1. Pick any procedure to estimate M_1.
2. Update the estimator of M_1 by repeating Step 1 with the previous estimate $\widehat{m}_1$.
3. Iterate k-times or till Step 1 and Step 2 give the same estimator.
4. Control the desired error measure making use of the more recent value of $\widehat{m}_1$.

It is intuitive that multi-step procedures are less conservative than two-step procedures, and that iterating till the estimator does not change in two subsequent steps is the least conservative method of all. In practice, the change in the estimate $\widehat{m}_1$ will be smaller and smaller as the number of iterations increase. An appreciation of the improvement in passing from one-step to multi-step estimation of M_1 is given in a comparison of the first two columns of Table 4.3.

To fix the ideas, we describe the algorithm for a particular choice of Step 1 and Step 2:

1. Let $R_B := 0$.
2. Let $R_B := |\{j \,:\, p_j < \alpha_1/(m - R_B)\}|$.
3. Iterate k-times or till Step 1 and Step 2 give the same estimator.
4. Let $\widehat{m}_1$ be the number of rejected hypotheses at the previous step. Do a plug-in method to control the FDR, taking $\widehat{a} = \widehat{m}_1/m$.

Table 4.4 compares the BH procedure ($\alpha_1 = 0$) with the multi step just described, with and without correction (note that, as said, no correction is usually taken on the level α_2).

4.3 Generalized Augmentation Procedure estimating M_1

To our knowledge there is not a $tFDP(c)$ controlling procedure that can make use of a suitable estimator of M_1. We propose here one. It is easily seen that, in a single step method at threshold q, $N_{1|0}$ is a binomial with parameters M_0 and q. Hence, if $\widehat{m}_0 = m - \widehat{m}_1$ is an anti-conservative estimator of M_0, one can substitute $\widehat{m}_0$ to m in (2.1) and (2.2), and the generalized augmentation procedure is still valid. The improvement in power is obvious, since for any $k \in \mathcal{N}$, if $X_1 \sim Bin < m_1, q >$ and $X_2 \sim Bin < m_2, q >$ with $m_1 \leq m_2$, it happens that $\Pr(X_1 \geq k) \leq \Pr(X_2 \geq k)$.

m	M_1	$E[\widehat{m}_1]$ Bonferroni	$E[\widehat{m}_1]$ multistep Bonferroni	$E[\widehat{m}_1]$ Gen. Aug.
5	1	0.581	0.589	0.477
10	1	0.526	0.529	0.430
30	3	1.309	1.316	1.165
100	10	3.48	3.506	3.036
200	20	6.184	6.211	6.311
500	50	13.47	13.53	18.29
1000	100	23.58	23.67	37.68
5000	500	86.95	87.25	218.172
5	3	1.621	1.694	1.364
10	5	2.438	2.531	2.034
30	15	6.272	6.495	5.918
100	50	17.317	17.855	23.903
200	100	31.010	31.866	50.400
500	250	66.261	67.941	129.236
1000	500	118.257	120.957	264.429
5000	2500	483.219	486.191	1629.731
5	4	2.133	2.306	1.802
10	9	4.394	4.772	3.696
30	27	11.187	12.007	12.471
100	95	32.930	35.160	52.000
200	190	58.760	62.333	106.034
500	475	126.606	133.376	272.398
1000	950	222.750	233.520	568.560
5000	4750	826.700	857.420	2903.570

Table 4.3: Comparison of estimators of M_1

	$\alpha_1 = 0$ (BH)	$\alpha_1 = 0.05$, no correction	$\alpha_1 = 0.05$, corrected
	$M_1 = 0.1 * m$		
m=10	0.0418	0.0471	0.0468
m=30	0.0448	0.0467	0.0465
m=100	0.0408	0.0424	0.0423
m=500	0.0445	0.0459	0.0459
m=1000	0.0449	0.0465	0.0464
m=5000	0.0451	0.0465	0.0464
	$M_1 = 0.5 * m$		
m=10	0.0266	0.0354	0.0347
m=30	0.0241	0.0315	0.0309
m=100	0.0254	0.0312	0.0309
m=500	0.0249	0.0288	0.0286
m=1000	0.0251	0.0286	0.0284
m=5000	0.0248	0.0289	0.0288
	$M_1 = 0.9 * m$		
m=10	0.0053	0.0112	0.0104
m=30	0.0047	0.0084	0.0078
m=100	0.0050	0.0074	0.0073
m=500	0.0050	0.0067	0.0063
m=1000	0.0048	0.0064	0.0064
m=5000	0.0048	0.0061	0.0060

Table 4.4: Observed FDR for multi-step control, using Bonferroni at estimation steps

Chapter 5

Finite Sample Control of FDR and tFDP(c) Under Dependence

In Chapter 3 we provided broad conditions on the dependence of a sequence of p-values for the usual methods for FDR control to work asymptotically. We also argued that, if the conditions do not hold or the number of tests m is too small, dependence can increase unacceptably $Var[FDP]$, so that FDR control may be no longer advisable. In this chapter we will provide many results on FDR and $tFDP(c)$ control under dependence for finite m. The chapter is organized as follows: Section 5.1 will generalize the $p_{(1)}$-approach. Section 5.2 will generalize the DKW approach, Section 5.3 will extend the generalized augmentation approach to the dependent case. For a description of $p_{(1)}$ and DKW approach see Chapter 1, for the generalized augmentation approach see Chapter 2. Section 5.4 will focus on a particular case of dependence structure, known in literature as *clumpy*, or *block*, dependence (for istance in Storey *et al.* (2004)). Discussion and simulations will be given in Section 5.5. Many of the results in this chapter are open to further developments, which we will also point out through the exposition.

First of all note that the results in Benjamini and Yekutieli (2001) can be immediately extended to the plug-in procedure, though noone has still noticed this fact. Benjamini and Yekutieli (2001) prove that, under PRDS assumptions on the vector of p-values (as discussed the end of Chapter 1), the BH procedure controls the FDR at level $\frac{M_0}{m}\alpha$. It is straightforward to see that, if $\widehat{a}$ is a conservative estimator of the quantity a, under any dependence assumption that includes PRDS (even arbitrary dependence), the Plug-in method with that estimator is valid for any finite m under PRDS assumptions. In the previous chapter we devised a whole family of such suitable estimators.

5.1 $p_{(1)}$-approach under dependence

Recall that the $p_{(1)}$-approach as described by Genovese and Wasserman (2004a) (i.e., inversion) consists in testing uniformity of all the possible subsets of $(p_1, \ldots, p_m)$, and taking the union of indexes of all the accepted tests as an estimate of S_0. Then, inversion

can be applied to control $tFDP(c)$. The test statistic more often considered is the minimal p-value in the selected subset, from which stems the name of the approach. Under independence, this test statistic follows a Beta distribution. Under dependence, van der Laan *et al.* (2003b) propose to estimate this distribution. This is doable in certain cases, but in many other cases the estimates are going to be inefficient (see van der Laan and Bryan (2000),Genovese and Wasserman (2004a) and Chapter 6). We thus need to avoid estimation. We will provide here conditions on the dependence (namely, association) which let us avoid the estimation of the joint distribution of the test statistics, and proceed to apply the $p_{(1)}$-approach without any modification. We will moreover prove that, if the test statistics are normally distributed, the $p_{(1)}$-approach is valid under *arbitrary* dependence among the tests from $m > 3$ tests, together with further results. This is an important result, since normal or approximately normal test statistics are often used in practice.

$\mathcal{A}$ is an increasing set

Call $\mathcal{A}$ the union of indexes of all subsets of $(p_1, \ldots, p_m)$ not rejected after testing for uniformity. First of all, we will prove that under arbitrary dependence, $\mathcal{A}$ is an increasing set. This is a key result to avoid estimation of the distribution of the minimum of a given set of p-values. Note that for different uniformity tests it can happen that $\mathcal{A}$ is not increasing; i.e., $p_{(k)}$ is not rejected, while $p_{(k+1)}$ is, for some k. This can never happen under independence (Genovese and Wasserman (2004a)).

Lemma 5.1.1. *Suppose A_1 and A_2 are subsets of the set of indexes $(1, \ldots, m)$, that $A_1 \subseteq A_2$ and that $pmin = \min\{p_i | i \in A_1\} = \min\{p_i | i \in A_2\}$. If $A_1 \in \mathcal{A}$, then $A_2 \in \mathcal{A}$.*

Proof. We have that $A_1 \in \mathcal{A}$ iff $\Pr(\min\{p_i | i \in A_1\} > pmin) > \alpha$. Since $A_1 \subseteq A_2$, $\min\{p_i | i \in A_1\} \geq \min\{p_i | i \in A_2\}$. Hence, $\Pr(\min\{p_i | i \in A_2\} > t) \geq \Pr(\{p_i | i \in A_1\} > t)$ for any $t \in [0, 1]$. This implies $\Pr(\min\{p_i | i \in A_2\} > pmin) > \alpha$, hence $A_2 \in \mathcal{A}$. □

The previous Lemma implies that, no matter the dependence structure, $\mathcal{A}$ will always be of the kind $[J, \ldots, m]$ for a certain J:

Corollary 5.1.2. *The set $\mathcal{A}$ is increasing, i.e., in the form $[J, J+1, \ldots, m]$ for a certain $J \subseteq (1, \ldots, m)$, or the empty set.*

Proof. Suppose by contradiction that $\mathcal{A}$ is not in this form. Call $J = min\{i \in \mathcal{A}\}$. Consider the set $[J, J+1, \ldots, m]$. We clearly have that $\mathcal{A} \subset [J, J+1, \ldots, m]$, and they have the same minimum, that is, $p_{(J)}$. Hence, by Lemma 5.1.1, the set $[J, J+1, \ldots, m] \subseteq \mathcal{A}$, leading to contradiction. □

Known Joint Distribution of the p-values

Call $F_{k,A}$ the CDF of the minimum of the set $A \subseteq \{1, \ldots, m\}$ of $k = |A|$ p-values. In general, it will not be a $beta < 1, k+1 >$ as in the independent case. The subscript A

indicates the specific dependence structure of the chosen set of k variables. I.e., different sets of the same number of p-values will have different CDFs, and this is indicated by A. Suppose we can determine $F_{k,A}$ for any k and A. This will be easy in case of normal random variables (see below, where we examine the case of $b \geq 1$ blocks of dependent normal random variables). The next theorem shows that in this case the reduction of the number of uniformity tests is, like in the independent case, from 2^m to m, i.e., we get to a step-down method; and shows how to do the $p_{(1)}$-approach.

Theorem 5.1.3. *Let $p = (p_1, \ldots, p_m)$ be a set of dependent p-values. Suppose we can determine $F_{k,A}(\cdot)$, the CDF of the minimum of a set A of k p-values. Call $p_{(j)}$ the j-th ordered p-value. Let $J = \min\{j : F_{m-j,A_{m-j}}(p_{(j)}) \geq \alpha\}$. Here A_{m-j} stands for the set of biggest $m-j$ p-values. Then*

1. *the set $A_J^* = [J, J+1, \ldots, m] \in \mathcal{A}$*
2. *for any $A \in \mathcal{A}$, $A \subseteq A_J^*$.*

Proof. The first assertion is true by construction. To prove the second, proceed by contradiction. Suppose $A \in \mathcal{A}$ but $V = min\{j : j \in A\} < J$. Consider the set $A_V^* = [V, V+1, \ldots, m]$. We have $A \subseteq A_V^*$. So, by Lemma 5.1.1, $A_V^* \in \mathcal{A}$, which contradicts the definition of A_J^*. □

This is a very close approach to van der Laan *et al.* (2003a) *minP* procedure, the only practical difference being in the fact that $F_{k,A}$ is considered known and not estimated, and no normality assumption is made. This is also a more general test: *minP* procedure is a FWER controlling procedure, while we proved here it also can be considered as a uniformity test on each and every subset of $(p_1, \ldots, p_m)$.

Unknown Joint Distribution of the p-values

If the CDF of the minimum of a set of $k \leq m$ p-values cannot be determined, then we need to make assumptions on the dependence. Recall that the random variables $X_1, \ldots, X_n$ are said to be associated if $Cov[g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)] \geq 0$, when it exists, for all monotonically coordinate-wise non-decreasing functions g_1 and g_2. Refer to Esary *et al.* (1967) and Tong (1980) for further details and the properties of association. In Appendix C we will give some examples of vectors of associated random variables.

We will now prove that if the test statistics are associated, then the $p_{(1)}$-approach is still valid:

Theorem 5.1.4. *If the test statistics are associated random variables, then the $p_{(1)}$-approach as defined in Section 1.3.7 controls the $tFDP(c)$ at the desired level α.*

Proof. If $X_1, \ldots, X_n$ are associated random variables, then

$$P(\bigcap_{i=1}^{n}\{X_i \leq z_i\}) \geq \prod P(X_i \leq z_i) \tag{5.1}$$

(and similarly

$$P(\bigcap_{i=1}^{n}\{X_i > z_i\}) \geq \prod P(X_i > z_i),) \tag{5.2}$$

for $z_i \in R$, $i = 1, \ldots, n$. Moreover, non decreasing functions of associated random variables are still associated. I.e., if the test statistics are associated then also the p-values are associated. In our case, this means that $Beta < 1, k+1 >$ stochastically dominates $F_{k,A}$ for any A, if the vector of p-values is associated[1]:

$$\begin{aligned} F_{k,A}(t) &= \Pr(\min_{j\in A} p_j \leq t) \\ &= 1 - \Pr(p_j \leq t,\ \forall j \in A) \\ &\leq 1 - \prod_{j\in A} P(p_j \leq t) \\ &= \Pr(Beta < 1, |A| + 1 > \leq t), \end{aligned} \tag{5.3}$$

where we used (5.1) at the third step. It is easy to prove that there exists a monotonicity also among dependency structures, which will not be reported here for shortness.

Call now $J_1 = \min\{j : 1 - (1 - p_{(j)})^{m-j} \geq \alpha\}$. This will be a "worst case" scenario: in the case of association of the random variables, $A^*_{J_1} = [J_1, J_1 + 1, \ldots, m] \notin \mathcal{A}$; but for any $A \in \mathcal{A}$, $A \subseteq A^*_{J_1}$. I.e., by Corollary 5.1.2 and (5.3), there exists $J \geq J_1$ such that $A^*_J \in \mathcal{A}$ and for any $A \in \mathcal{A}$, $A \subseteq A^*_J$. Hence, under association, one can reject $p_{(1)}, \ldots, p_{(J)}$ and then apply augmentation; or do inversion on the set $\{1, \ldots, J\}$, i.e., use $U = \{1, \ldots, J\}$ in Step 2 at pag. 17. This extends the $p_{(1)}$-approach to the case of association. □

There are cases in which also an upper bound for J can be obtained:

Example 5.1.1. *Suppose the p-values from tests are dependent in blocks, but the number and size of the blocks are unknown. Suppose it is possible to determine $F_{k,A}$ if all the p-values come from the same block. This is the strongest dependence possible for a set of k p-values. As a realistic example, which will be developed below, consider normal random variables dependent in blocks; with unknown blocks. It obviously is easy to compute all the possible CDFs but not to determine the CDF of the minimum of a given set $(p_{i_1}, \ldots, p_{i_k})$. Let $J_2 = \min\{j : F_{m-j,A_{m-j}} \geq \alpha\}$. Here A_{m-j} indicates conditioning on the event that all the $m - j$ p-values come from the same block. Under any dependence assumption, $J_2 \geq J$. Under association, it can be computed also J_1 such that $J_2 \geq J \geq J_1$, as in Theorem 5.1.4. Moreover, if $J_2 = J_1$, then $J = J_1$ and the set $A^*_{J_1}$ is exactly the union of indexes of all subsets of $(p_1, \ldots, p_m)$ not rejected after testing for uniformity at level α.*

[1]Recall also that any subset of a set of associated random variables is associated.

Positive Dependence by mixture representation

Tong (1980) proves the same main properties of association for a particular class of mixtures of distributions. If a vector of random variables $(X_1, \ldots, X_n)$ has CDF $F(X) = \int \prod_{i=1}^{n} G_u^i(x_i)\ dH(u)$, and the family $\mathcal{G}_i = \{G_u^i(t(x_i)) : u \in \mathcal{U}\}$ is stochastically increasing for any i; then (5.1) and (5.2) hold. Note that this is equivalent to asking that X_i is positively regression dependent on an opportune latent variable U for each i. Hence, $p_{(1)}$-approach is still valid in case of mixture representation via a stochastically increasing $\mathcal{G}$.

5.1.1 The case of Normal Random Variables

If the original test statistics are normal random variables, it is possible to give conditions for the $p_{(1)}$-approach to remain valid even if association does *not* hold. We will actually prove that the $p_{(1)}$-approach can be applied without changes under arbitrary dependence, when $m > 3$, and otherwise we will make use of the following definition:

Definition 5.1.5 (Structure l)**.** *A k by k positive definite covariance matrix $V = (v_{ij})$ is said to have the structure l if there exist real numbers $\lambda_1, \ldots, \lambda_k$ in $(-1, 1)$ and $\sigma_1, \ldots, \sigma_k$ in $\mathcal{R}^+$ such that $v_{ii} = \sigma_i^2$ for all i and $v_{ij} = \sigma_i \sigma_j \lambda_i \lambda_j$ for all $i \neq j$.*

A multivariate normal random variable is said to have structure l when its covariance matrix has the structure l.

Theorem 5.1.6. *If the test statistics are normal random variables and $m > 3$, then the $p_{(1)}$-approach as defined in Section 1.3.7 controls the $tFDP(c)$ at the desired level α. If $m \leq 3$, then this is true only if the test statistics are structure l normal random variables.*

Proof. We are interested in inequalities:

$$\Pr(\bigcap |X_i| \leq a_i) \geq \prod \Pr(|X_i| \leq a_i) \tag{5.4}$$

and

$$\Pr(\bigcap |X_i| \geq a_i) \geq \prod \Pr(|X_i| \geq a_i) \tag{5.5}$$

Inequality (5.4) is proved in Dunn (1958) for a structure l multivariate normal with $k \leq 3$, and in Khatri (1967) for $k > 3$ and arbitrary positive definite correlation matrix. Inequality (5.5) is proved again by Khatri (1967) for a structure l multivariate normal, for any[2] k.

If $a_i = \Phi^{-1}(\alpha)$ is taken, inequalities of the form of (5.1) and (5.2) for the p-values follow from (5.4) and (5.5); when the alternative hypothesis is two-sided.

[2]A conjecture, still not proving but without any counterexample, is made that (5.5) holds for arbitrary correlation matrix.

In particular, using (5.4), it is possible to prove (5.3). Since Corollary 5.1.2 is valid for arbitrary dependence, this proves that, when the test statistics are normal, the $p_{(1)}$-approach is valid for *arbitrary* dependence when $m > 3$ and for structure l multivariate normals when $m \leq 3$. □

Since usually $m >> 3$, Theorem 5.1.6 is a particularly strong result, essentially proving that what is true under independence, in this case, extends to arbitrary dependence.

5.2 DKW approach under dependence

We will here extend to the dependent case the DKW procedure described at pag. 17. We use the property of negative association (see Kumar and Proschan (1983), Block *et al.* (1982) and Appendix C for further details and examples of negatively associated random variables), which is a sort of dual of association:

Definition 5.2.1 (Negative association)**.** *A vector of random variables $X_1, \ldots, X_n$ is negatively associated if, for all monotonically coordinate-wise non-decreasing functions g_1 and g_2, $Cov[g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)] \leq 0$, when it exists.*

Some examples of negatively associated random variables are given in Appendix C.

We will now propose procedures, making use of two different DKW-Type inequalities which are proved in Appendix B, together with Hoeffding inequality (which we don't use in this dissertation) and other technical results. We will use the assumption of negative association between the test statistics. In this case it hasn't been possible to extend the approach without any modification.

5.2.1 Type I DKW approach under dependence

For our first extension, we will make use of this extension of DKW inequality:

Lemma 5.2.2 (DKW-Type Inequality)**.** *Let $X_1, \ldots, X_n$ be a sequence of identically distributed negatively associated random variables. Let $F(z)$ be the CDF of X_1, and $\widehat{F}(z)$ the empirical distribution of the sequence $X_1, \ldots, X_n$. Then,*

$$\Pr\{\sup_{z \in \mathcal{R}} |F(z) - \widehat{F}(z)| > \varepsilon\} \leq 4(2n+1)e^{-n\varepsilon^2/8}.$$

Lemma 5.2.2 is Lemma B.2.4 in Appendix B, at page 111. Proof is given in the Appendix.

The DKW approach can now be used just at the price of taking a bigger confidence set for $G(t)$, under the assumptions of Lemma 5.2.2:

1. Let

$$\varepsilon_m = \sqrt{\frac{8}{m} \log\left(\frac{4(2m+1)}{\alpha}\right)}.$$

2. Plug-in ε_m and a suitable estimator of a, $\widehat{a}$, in $R(t)$ as defined in (1.9).

3. Fix $T_{DKW} = \sup\{t : \ R(t) \leq c\}$ and reject $p_j < T_{DKW}$.

It is easily seen that this controls $tFDP(c)$ at level α.

5.2.2 Type II DKW approach under dependence

Another possibility is to use a different DKW-Type Inequality

Lemma 5.2.3 (DKW-Type Inequality 2)**.** *Let $X_1, \ldots, X_n$ be a sequence of identically distributed negatively associated random variables. Let $F(z)$ be the CDF of X_1, and $\widehat{F}(z)$ the empirical distribution of the sequence $X_1, \ldots, X_n$. Then,*

$$\Pr\{\sup_{z \in \mathcal{R}} |F(z) - \widehat{F}(z)| > \varepsilon + \frac{24\sqrt{2\pi}}{\sqrt{n}}\} \leq e^{-2n\varepsilon^2}.$$

Lemma 5.2.3 is Lemma B.3.4 in Appendix B, at page 114. Proof is given in the Appendix.

Type II extension of the DKW approach can now be performed via a different definition of the function $R(t)$, which we call $\tilde{R}(t)$. Under the assumptions of Lemma 5.2.2:

1. Let
$$\varepsilon_m = \sqrt{\frac{1}{2m} \log\left(\frac{1}{\alpha}\right)}.$$

2. Define now
$$\tilde{R}(t) = \begin{cases} \frac{t(1-\widehat{a})}{\widehat{G}(t) - \varepsilon_m - \frac{24\sqrt{2\pi}}{\sqrt{m}}} & \text{if } \widehat{G}(t) > t(1-\widehat{a}) + \varepsilon_m + \frac{24\sqrt{2\pi}}{\sqrt{m}} \\ 1 & \text{otherwise} \end{cases} \tag{5.6}$$
for a suitable estimator of a, $\widehat{a}$.

3. Fix $T_{DKW} = \sup\{t : \ \tilde{R}(t) \leq c\}$ and reject $p_j < T_{DKW}$.

It is easily seen that this controls $tFDP(c)$ at level α.

The two extensions are based on two different exponential tail inequalities, but as will be seen in simulations they usually give very close results. Usual problems linked with the DKW approach (very small number of rejections, or no rejection at all, when m is small and/or the signal is weak) are carried over from the independent to the dependent case. We already noted that the DKW approach may work well when the number of tests is very big and should be avoided in the other cases.

5.3 Generalized Augmentation Procedure under dependence

Extension to dependence of the Generalized Augmentation Procedure for any $m \in \mathcal{N}$ is not straightforward; since evaluation of the distribution of $N_{1|0}$ for the one step method at level q is needed, which needs the knowledge of the joint distribution of $(p_j, j \in S_0)$, and an evaluation of 2^{M_0} slices of it.

Now, in a given one-step method at level q, $N_{1|0} = \sum_{j \in S_0} 1_{p_j < q}$. Each indicator function in the sum is distributed like a Bernoulli, with mean q.

A sufficient condition can be given: in the usual notation, if it holds that $\Pr(\sum 1_{\{p_i<q\}}(1-H_i) \leq k) \geq \sum_{j=0}^{k} \binom{M_0}{k} q^j (1-q)^{M_0-j}$ for $k > 0$, then it is straightforward to see that the generalized augmentation approach is valid without any correction. We are requesting that a sum of *dependent* Bernoullies has a lighter tail than a sum of independent Bernoullies (i.e., a Binomial), each with the same parameter q. Unfortunately, at the moment no more general condition on the dependence is known that can imply this sufficient condition. Note that (negative) association causes the condition to fail for some k. Note moreover that the condition on the sum of the Bernoullies is the only technical condition needed. The Hypergeometric distribution in the negative augmentation part is the same under arbitrary dependence, since it arises from pure combinatorial reasoning.

We need to turn here to the wide literature of binomial and Poisson approximation, a problem which has received attention starting from the 50s, in order to give a survey of possible modifications of the procedure under different dependence situations. The researcher should then choose the best one for his/her application (usually, it will just be a matter of how big is m). A milestone in Poisson approximation is the well known result of Le Cam (1960), stating that the total variation distance between a $Binomial < m, q >$ and a $Poisson < mq >$ is bounded above by mq^2. Chen (1975) prove that, if the mixing coefficients of the sequence of random indicators $\{X_i\}_{i \in \mathcal{N}}$ are $O(e^{-\eta k})$ for some $\eta > 0$ then the total variation between the distribution of $\sum_{i=1}^{m} X_i$ and a $Poisson < \lambda = \sum_i \Pr(X_i = 1) >$ is bounded above by

$$C(\eta) \min(\lambda^{-1/2}, 1)[Var(\sum X_i) - \lambda + (\lambda + 1)^2 m^{-1} \log m],$$

for a certain $C(\eta)$. If the statistics are m-dependent of order h, then the bound is

$$6 \min(\lambda^{-1/2}, 1)[|\sum_{i \neq j} Cov(X_i, X_j)| + 4(h+1)(\lambda^2)/m].$$

This results can be used to extend the generalized augmentation procedure to mixing sequences *for any finite number of tests*. One needs only substitute the binomial distributions in (2.1) and (2.2) with the distribution of a Poisson with parameter $m * q$. Since q is typically small, the approximation is good when the covariance between the p_is is small in absolute value.

Binomial approximation of sums of dependent random variables can also be used. Boutsikas and Koutras (2000) prove that, if the random indicators are either associated

or negatively associated[3], then

$$\sup_k |P(\sum_i X_i \leq k) - P(Bin < m, q >\leq k)| \leq |\sum_{i<j} Cov(X_i, X_j)|,$$

where $q = \Pr(X_i = 1)$. Note that, obviously, $|\sum_{i<j} Cov(X_i, X_j)| = |\sum_{i<j}(\Pr(X_i = 1, X_j = 1) - q^2)|$. This is a useful result: one can estimate either $Cov(X_i, X_j)$ or $\Pr(X_i = 1, X_j = 1)$ and add $|\sum_{i<j} \widehat{Cov(X_i, X_j)}|$ to the binomial distribution expressions in (2.1) and (2.2). This is particularly easy in the case of block dependence, in which $Cov(X_i, X_j) = 0$ for many combinations of i and j. If the constant covariance is assumed between the test statistics (which is reasonable in many real applications, see below), then $\Pr(X_i = 1, X_j = 1)$ is easily and efficiently estimated.

Shao (2000) proves that, if the vector of indicators is negatively associated, then $E[\sum_{i=1}^m X_i] \leq mq$. By an immediate application of Markov inequality, it is seen that one can bound $\Pr(\sum_i X_i \geq k) \leq \frac{mq}{k}$. Once again, substitute the bound in (2.1) and (2.2). Note that this is a very clean substitution, since no estimation is needed; though it may prove conservative for certain m and q.

Soon (1996) proposes a different approach, using the Chen-Stein approximation method (Stein (1971, 1986),Chen (1974, 1975)), to the same problem. He proposes approximating the distribution of a sum of m dependent random indicators with a binomial with parameters $m' = \left\lceil \frac{2m^2q^2}{2mq^2+1} \right\rceil$ and $q' = q + 1/2np$. He proves that the total variation distance between the distribution of the sum of indicators and the binomial so defined is bounded above by $C_{mq}(0.5 + |\sum_{i \neq j} Cov(X_i, X_j)|)$ in case of positive or negative association. The constant C_{mq} is equal to $\frac{1-q'^{m'+1}-(1-q')^{m'+1}}{(m'+1)q'(1-q')}$, which is seen to be small for small q. Since the Kolmogorov distance is obviously bounded by the total variation distance, this result can be used to approximate the distributions in (2.1) and (2.2) in the usual fashion. Note that this proves also *asymptotic* validity of the generalized augmentation method *without any modification* since $q' \xrightarrow{m} q$ and $\frac{m'}{m} \xrightarrow{m} 1$.

Gabriel (1959) derives the distribution of a sum of M_0 dependent indicators in case of Markovian dependence (which requires also ordering of the tests). Let $q_0 = \Pr(X_k = 1|X_{k-1} = 0)$ and $q_1 = \Pr(X_k = 1|X_{k-1} = 0)$. Call q the probability of success on the initial trial, then $\Pr(\sum X_i = k) = q \Pr(\sum X_i = k|X_0 = 1) + (1-q) \Pr \sum X_i = k|X_0 = 0)$. Call $C_1 = M_0 + 0.5 - |2k - 0.5 + M_0|$, $C_0 = M_0 + 0.5 - |2k - 0.5 - M_0|$ and let a_i and b_i be the least integers not smaller than $1/2C_i - 1$ and $1/2C_i$ respectively. Then,

$$\Pr(\sum X_i = k|X_0 = 1) = q_1^k(1-q_0)^{M_0-k} \sum_{C=1}^{C_1} \binom{k}{a_1} \binom{M_0 - k - 1}{b_1 - 1} \left(\frac{1-q_1}{1-q_0}\right)^{b_1} \left(\frac{q_0}{q_1}\right)^{a_1}$$

and

$$\Pr(\sum X_i = k|X_0 = 0) = q_1^k(1-q_0)^{M_0-k} \sum_{C=1}^{C_0} \binom{k-1}{b_0 - 1} \binom{M_0 - k}{a_0} \left(\frac{1-q_1}{1-q_0}\right)^{a_0} \left(\frac{q_0}{q_1}\right)^{b_0}.$$

[3]Actually, they need only $(X_i, \sum_{j<i} X_j)$ to be positively (negatively) quadrant dependent, as defined in Lehmann (1966). This is implied by (negative) association, and usually defined as positive (negative) cumulant dependence.

So if the tests form a Markov chain, the exact distribution of the number of errors by rejecting all p-values smaller than a certain q can be exactly derived, and again substituted in (2.1) and (2.2). Ladd (1975) describes a fast algorithm for computing these probabilities.

As a last useful result for applying the binomial/Poisson approximation approach, note that (Doukan (1994)):

$$|Cov(X_i, X_{i+k})| \leq 8\alpha(k)^{1/r}(E|X_i|^p)^{1/p}(E|X_{i+k}|^q)^{1/q},$$

for any $p, q, r \geq 1$ and $1/p + 1/q + 1/r = 1$, where $\alpha(k)$ are the alpha mixing coefficients defined in (3.1).

Finally, asymptotic results can easily be given by applying any form of CLT under dependence. Many sufficient conditions on dependence for CLT results are given in Chapter 3, at page 30. More general conditions for the standard CLT are established in Ibragimov (1962, 1975), on the α-mixing coefficients: if $\alpha(k)$ is infinitesimal and $E|X_1|^{2+\delta} < \infty$ for some $\delta > 0$, then CLT holds for identically distributed random variables. Refer also to Billingsley (1999) for conditions on different mixing coefficients.

If any of the conditions hold, $N_{1|0}$ is asymptotically distributed like a normal random variable. One can now substitute the CDF of the opportune normal in (2.1) and (2.2), or simply use the well known asymptotic binomial approximation to the normal, and keep the binomial PMF in the same formulas.

This essentially proves same results of Chapter 3 for $tFDP(c)$ control via Generalized Augmentation Procedure.

5.4 The case of block dependence with known blocks

We provide in this section specific results on a common case of dependent random variables, the so called block/clumpy dependence. This is a kind of dependence arising in many real applications, such as DNA microarrays, environmental surveys, multi-center studies. This kind of dependence arises also in case of spatially dependent random variables for which $X_{ij} \perp X_{i'j'} = 0$ whenever $d(X_{ij} - X_{i'j'}) > d_{min}$, for some d_{min}.

Formally, let $p = \{p_{i,b}\}$, $i = 1, \ldots, r_b$; $b = 1, \ldots, k$ be a sequence of p-values such that $p_{i,b}$ is independent of p_{j,b_1} for $b \neq b_1$ and for any i and j; with $m = \sum_b r_b$. This is the case of p values dependent in blocks and independent otherwise. Unless stated otherwise, we will assume $r_b \geq 1$, $k \geq 1$, and that the dependence between $p_{i,b}$ and $p_{j,b}$ (within blocks) is arbitrary.

Obviously, a rough first approach to block dependence is to ignore it, and consider the experiment as a whole dependent sequence of tests, essentially applying results prove so far. Note that if the size of each block is not allowed to grow arbitrarily (while their number is), we automatically fall under the m-dependence assumption, which implies mixing at any rate, i.e., one of the hypotheses used in Chapter 3 to prove Theorem 3.1.6. See Appendix C for a discussion of this.

Another good approach is to apply an intuitive generalization of the so called BY technique. BY (Benjamini and Yekutieli (2001)) consists in applying a correction factor of $\sum_{i=1}^{m} 1/i$ to the level α. This is a worst case scenario if there is a single block of m dependent test statistics. It is reasonable to expect that, if the random variables are dependent in blocks, the FDR in each block will be bounded by $\alpha * \sum_{i=1}^{r_b} 1/i$. Hence, a correction factor of $\sum_{i=1}^{max_b r_b} 1/i$ should be used. Note that this reduces to $\sum_{i=1}^{m} 1/i$ in case of a single block, and to 1 in case of independent random variables (i.e., m blocks).Note that only the size of the biggest block, and not the composition, is to be known. This is a great advantage. For instance, genes in microarray experiments are commonly believed to be dependent in blocks of at most 50 genes, but the blocks are unknown. A correction factor of $\sum_{i=1}^{50} 1/i = 4.499$ is to be applied, instead of a correction factor of around $\sum_{i=1}^{51000} 1/i = 11.417$ for a full human genome scan. Increase in power is substantial.

We will show now other possibilities to control the FDR with block dependent tests, assuming blocks are known. Extension to unknown blocks will be partly discussed at the end of the section.

Note that, when $b = 1$, as stated, we are working with a whole vector of dependent random variables. In this case, the methods we are going to propose are still valid (though, like the one in the next subsection, may become trivial).

5.4.1 Aggregating after FDR control in each block

We begin by considering the possibility of controlling the FDR in each block, and then aggregating the indexes of the rejected tests to estimate S_0^c. Storey (2003) uses a nice approximation for the FDR. He assumes $FDR(t) \cong \frac{E[\sum I_i(1-H_i)]}{E[\sum I_i + \prod(1-I_i)]}$, where $I_i = 1_{\{p_i<T\}}$ and T is the cut-off. Using this approximation, it is easy to prove a general result, which anyway we don't use further:

Theorem 5.4.1. *Consider the case of block dependent test statistics. If the FDR is controlled in each block, then it is controlled on the whole sequence.*

Proof. Call $FDR_i(T_i)$ the FDR in the i-th block, where T_i is the threshold in the i-th block. Let p_{ij} be the j-th p-value of the i-th block. Let $I_{ij} = 1_{\{p_ij<T_i\}}$. Using the approximation, we have

$$\begin{aligned} FDR_i(T_i) &\cong \frac{E[\sum I_{ij}(1-H_{ij})]}{E[\sum I_{ij} + \prod(1-I_{ij})]} \\ &\leq \alpha, \end{aligned}$$

which implies $E[\sum I_{ij}(1-H_{ij})] \leq \alpha * E[\sum I_{ij} + \prod(1-I_{ij})]$. Then:

$$\begin{aligned} FDR &\cong \frac{E[\sum\sum I_{ij}(1-H_{ij})]}{E[\sum\sum I_{ij} + \prod\prod(1-I_{ij})]} \\ &= \frac{\sum_i E[\sum_j I_{ij}(1-H_{ij})]}{E[\sum\sum I_{ij} + \prod\prod(1-I_{ij})]} \\ &\leq \alpha \sum_i \frac{E[\sum_j I_{ij}(1-H_{ij})]}{E[\sum\sum I_{ij} + \prod\prod(1-I_{ij})]} \end{aligned}$$

Note that $E[\sum\sum I_{ij} + \prod\prod(1-I_{ij})] \geq \sum_i E[\sum I_{ij} + \prod(1-I_{ij})]$. Hence, it is straightforward to see that the last expression will be smaller than or equal to α. □

Remark 5.4.2. *This means that different experiments can be put together without having to compute a common threshold, and without getting an FDR over the desired level. Aggregation of experiments can be thought of being aggregation of blocks.*

Remark 5.4.3. *The theorem will work no matter how each T_i is derived. It is obvious that applying BY to each block will greatly increase the power than doing BY on the whole sequence.*

5.4.2 Sampling independent vectors from each block

The basic idea of this subsection is that any vector, of size k, of p-values coming from each and every block, is a vector of independent p-values. If the blocks are known, we can derive such vectors, apply methods which are known to work under independence, and then aggregate. For simplicity, assume we are in the balanced case, i.e., $r_b = r_{b_1} = r$ for all $b, b_1 \in (1, \ldots, k)$.

Theorem 5.4.4. *For the p-value sequence given, let $\tilde{p} = \{\tilde{p}_1, \ldots, \tilde{p}_k\}$ be a vector in which $\tilde{p}_i$ is sampled from the i-th block. $\tilde{p}$ is then a vector of independent p-values. If $\tilde{T}$ is the BH threshold for this vector, we have that the FDR for the sequence $\tilde{p}$ is controlled at desired level α. Let T^* be the minimal cut-off over all r^k possible sampled $\tilde{p}$s. Then, the FDR for the sequence p, using cut-off T^*, will be controlled at level α.*

Proof. The FDP for the original sequence p is:

$$\frac{\sum_{b=1}^{k}\sum_{i=1}^{r} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b})}{\sum_{b=1}^{k}\sum_{i=1}^{r} 1_{\{p_{i,b}<T^*\}} + \prod_{b=1}^{k}\prod_{i=1}^{r}(1-1_{\{p_{i,b}<T^*\}})} =$$

$$\frac{\sum_{i=1}^{r}\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b})}{\sum_{i=1}^{r}\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}} + \prod_{i=1}^{r}\prod_{b=1}^{k}(1-1_{\{p_{i,b}<T^*\}})} \leq$$

$$\frac{\sum_{i=1}^{r}(\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b}))}{\sum_{i=1}^{r}(\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}} + \prod_{b=1}^{k}(1-1_{\{p_{i,b}<T^*\}}))} \leq$$

$$\max_{i=1}^{r} \frac{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b})}{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}} + \prod_{b=1}^{k}(1-1_{\{p_{i,b}<T^*\}})}$$

The last inequality follows from the fact that, if $x_1, \ldots x_n$ and $y_1, \ldots, y_n$ are two sequences of real numbers, if for all i it happens that $y_i \geq x_i \geq 0$; then it is easy to see that $\frac{\sum_i x_i}{\sum_i y_i} \leq \max \frac{x_i}{y_i}$. It is obvious that the expectation of the right hand side of the last inequality is less than or equal of the expectation of the left hand side. We will now prove that $E[\max_{i=1}^{r} \frac{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b})}{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}} + \prod_{b=1}^{k}(1-1_{\{p_{i,b}<T^*\}})}] \leq \alpha$. Thesis will follow from that. Note that, for any p from the joint distribution (for fixed r and k);

$$E[\max_{i=1}^{r} \frac{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}}(1-H_{i,b})}{\sum_{b=1}^{k} 1_{\{p_{i,b}<T^*\}} + \prod_{b=1}^{k}(1-1_{\{p_{i,b}<T^*\}})} | p] \leq \alpha;$$

since T^* is chosen accordingly. Hence, the unconditional expected value is also below α. The thesis follows. □

A possible extension is to include the possibility of not knowing the size and composition of the blocks. The blocks will be known in case of controlled, experimental situations, while in many other cases they may be not known. A good approach should be to estimate the blocks, and then proceed as if they were known; possibly incorporating the uncertainty brought about by estimation.

Note that this approach is very conservative, since the cut-off is defined as the minimum over a set of possibly conservative cut-offs. Moreover if there are very few blocks the "sampled" vectors will be small and cut offs will have a greater variance. The

ideal situation is to have a large number of small blocks, inside which the dependence among the p values is high: this will lead to a fast algorithm (few combinations are possible) and to an efficient procedure (the thresholds will be close to each other, so that the minimal will not be too close to zero).

Remark 5.4.5. *In the same fashion, DKW and $p_{(1)}$-approach can be generalized to block dependence with arbitrary dependence within the blocks, at the price, for instance, of substituting m with k in DKW inequality. This, in practice, extends all the procedures to the case of block dependence with arbitrary dependence within blocks. Nevertheless, one cannot avoid being overly conservative with this approach.*

5.4.3 $p_{(1)}$-approach for normal block dependent random variables

Let us now restrict to a particularly relevant case. We will suppose that the p-values come from normal test statistics, with known variance equal to 1. We will suppose that the test statistics are all independent between blocks, while $Cov(X_i, X_j) = \rho$ whenever X_i and X_j come from the same block; $-1 < \rho < 1$. Of course, to ensure positive definiteness of the matrix, we'll have the condition $\rho > -\frac{1}{\max_b r_b - 1}$, where $\max_b r_b$ is the size of the biggest block. Let's assume there are k blocks. Let Σ_i be the i x i matrix with diagonal elements all equal to 1 and off-diagonal elements equal to ρ. Let $\Phi(\cdot)$ denote the standard normal CDF, and $\Phi_b(\cdot, \mu, \Sigma)$ denote the a multivariate normal CDF in R^b, with mean vector μ and variance-covariance matrix Σ. We will now determine $F_{k,A}$, the CDF of the minimal p-value in a given subset. This will be sufficient to apply the $p_{(1)}$-approach, as described in Section 5.1.

Note that, when $m > 3$, this computation can be avoided since the $p_{(1)}$-approach is valid in its *independent* form for arbitrary dependence (if the alternative is two-sided). Nevertheless, using exact results yields a less conservative and more powerful procedure. It is intuitive that exact results are always to be preferred, when available.

Theorem 5.4.6. *Suppose we pick a subset of r p-values from the m. There will be r_1 from the first block, r_2 from the second, and so on. Of course, $\sum_{j=1}^{b} r_j = r$. We have:*

$$\Pr(\min_i\{p_{i_1}, \ldots, p_{i_r}\} < t) = 1 - (1-t)^{\sum_j 1_{\{r_j=1\}}} \prod_{i=2}^{S} \Phi_i(\Phi^{-1}(1-t), 0, \Sigma_i)^{\sum_j 1_{\{r_j=i\}}}$$

Proof. Let p_i^j denote the i-th p-value of the j-th block, i.e. $p_i^j = \Pr(X_i^j > xobs_i^j)$.

$$\begin{aligned}
\Pr(\min_i\{p_{i_1},\dots,p_{i_r}\} < t) &= 1 - \prod_{i=1}^{b} \Pr(p_1^i > t, \cdots, p_{r_i}^i > t) \\
&= 1 - \prod_{i=1}^{b} \Pr(Xobs_1^i < \Phi^{-1}(1-t), \cdots, Xobs_{r_i}^i < \Phi^{-1}(1-t)) \\
&= 1 - \prod_{i=1}^{b} \Phi_i(\Phi^{-1}(1-t), 0, \Sigma_{r_i}) \\
&= 1 - \prod_{i=1}^{S} \Phi_i(\Phi^{-1}(1-t), 0, \Sigma_i)^{\sum_j 1_{\{r_j=i\}}}
\end{aligned}$$

□

Remark 5.4.7. *Note that the expression is invariant to permutations of the r_i.*

Refer to next section for simulation results.

5.5 Discussion and Simulations

blocks =	2	5	10	20	25	50
ρ =	−0.02000	−0.05200	−0.11100	−0.25000	−0.33300	−0.99900
BH	0.0592	0.0584	0.0583	0.0583	0.0580	0.0587
	(0.0455)	(0.0421)	(0.0440)	(0.0468)	(0.0463)	(0.0418)
Iterative	0.0589	0.0581	0.0579	0.0579	0.0577	0.0586
	(0.0477)	(0.0451)	(0.0467)	(0.0496)	(0.0488)	(0.0447)
BY	0.0695	0.0691	0.0688	0.0687	0.0689	0.0696
	(0.0078)	(0.0068)	(0.0084)	(0.0104)	(0.0102)	(0.0080)
Block	0.0622	0.0638	0.0656	0.0668	0.0673	0.0681
	(0.0322)	(0.0194)	(0.0142)	(0.0141)	(0.0117)	(0.0073)
$P_{(1)}$	0.0607	0.0599	0.0594	0.0600	0.0589	0.0601
	(0.0380)	(0.0391)	(0.0382)	(0.0402)	(0.0384)	(0.0386)
Storey	0.0584	0.0576	0.0576	0.0577	0.0573	0.0581
	(0.0501)	(0.0449)	(0.0483)	(0.0496)	(0.0511)	(0.0466)

Table 5.1: FNR (FDR) for different methods, block dependence

Table 5.1 shows the FNR (as defined in (1.8)) and FDR in parentheses for an array of $m = 100$ normal random variables divided in different numbers of independent blocks, with $m_0 = 90$, for different methods. The correlation inside each block was taken to be

$\tau =$	50	20	10	2
BH	0.0575	0.0563	0.0583	0.0570
	(0.0204)	(0.0239)	(0.0253)	(0.0378)
$Iter.$	0.0572	0.0557	0.0577	0.0567
	(0.0254)	(0.0346)	(0.0324)	(0.0402)
BY	0.0683	0.0667	0.0690	0.0680
	(0.0016)	(0.0062)	(0.0043)	(0.0078)
$P_{(1)}$	0.0662	0.0664	0.0669	0.0664
	(0.0081)	(0.0047)	(0.0050)	(0.0086)
$Stor.$	0.0453	0.0473	0.0524	0.0550
	(0.2825)	(0.2227)	(0.1666)	(0.0706)

Table 5.2: FNR (FDR) for different methods, simplified exponential dependence

constant, equal to a certain parameter ρ. This parameter was taken to be negative, as small as possible. BH stands for the classical method, *Iterative* for the plug-in method applied with the iterative estimator, BY for the classical method in which the level α is divided by $\sum_{i=1}^{m} 1/i$, as suggested in Benjamini and Yekutieli (2001). We proposed here a more fine tuned correction, dividing instead by $\sum_{i=1}^{\max_b r_b} 1/i$, where r_b is the size of b-th block. It is apparent that this method (*Block*) is much less conservative. $P_{(1)}$ stands for the $p_{(1)}$ approach with known blocks, as derived in Section 5.4.3, used to control the FDR, with $c = 0.0002$. Finally, *Storey* stands for the Plug-in method with Storey estimator.

Tables 5.2 and 5.3 show the FNR (FDR) for the same methods, applied to a grid of dependent normals with simplified exponential covariance structure of pag. 34. In this case it is seen that Plug-in with Storey's estimator brings unacceptably high FDRs, and this is the only procedure observed to fail.

Finally, Tables 5.4 and 5.5 show the $tFDP(0.1)$ for different methods, with the usual simplified exponential covariance structure. The Generalized Augmentation procedure, as we said, works in virtue of the normal approximation to the binomial. Augmentation of $minP$ procedure, which we call the $p_{(1)}$-approach, was proved to work under normality of the test statistics. Note instead that the simulated random variables are *not* negatively associated, hence the DKW method, even using the iterative estimator to estimate $1 - a$, may not work. This is an example of how things can go wrong. It is the case for extremely strong dependence (actually, $\tau \geq 300$ is sufficient). Note that our extensions (*Type I and Type II DKW*) under dependence give very close results and behave in the same way. Moreover, as we said, the DKW approach can have very low power if m is not big and the signal is weak. This is the case: in almost independent cases, this approach results in rejection of no hypothesis. Refer to Chapter 6 for cases in which the DKW approach works well.

$\tau =$	1.67	1.25	1	0.83
BH	0.0590	0.0582	0.0584	0.0577
	(0.0432)	(0.0435)	(0.0432)	(0.0438)
$Iter.$	0.0587	0.0579	0.0581	0.0574
	(0.0460)	(0.0483)	(0.0479)	(0.0486)
BY	0.0695	0.0692	0.0691	0.0688
	(0.0074)	(0.0104)	(0.0094)	(0.0101)
$P_{(1)}$	0.0671	0.0671	0.662	0.0675
	(0.0078)	(0.0105)	(0.0135)	(0.0143)
$Stor.$	0.0572	0.0568	0.0574	0.0568
	(0.0678)	(0.0631)	(0.0599)	(0.0496)

Table 5.3: FNR (FDR) for different methods, simplified exponential dependence

As final comments, note that (negative) association of the vector of p-values follows directly from (negative) association of the vector of test statistics; which in general may be easier to work with. It is particularly interesting that mainly we are imposing conditions of "coherence" on the dependence, i.e., that it is "all positive" or "all negative".

$\tau =$	2000	50	20	10	3.33
Gen. Aug., $q := q/2$	0.066	0.067	0.068	0.066	0.069
	(0.049)	(0.034)	(0.033)	(0.034)	(0.036)
$p_{(1)}$-approach	0.066	0.067	0.067	0.066	0.069
	(0.048)	(0.012)	(0.007)	(0.013)	(0.035)
Type I DKW (Iterative)	0.045	0.097	0.097	0.097	0.099
	(0.095)	(0.028)	(0.029)	(0.022)	(0.001)
Type II DKW (Iterative)	0.038	0.097	0.096	0.097	0.099
	(0.096)	(0.026)	(0.034)	(0.024)	(0.004)

Table 5.4: FNR ($tFDP(0.1)$) for different methods, simplified exponential dependence

$\tau =$	2	1.67	1.25	1	0.83
Gen. Aug., $q := q/2$	0.068 (0.042)	0.068 (0.039)	0.069 (0.043)	0.068 (0.042)	0.069 (0.041)
$p_{(1)}$-approach	0.067 (0.041)	0.066 (0.042)	0.067 (0.044)	0.066 (0.045)	0.067 (0.048)
Type I DKW (Iterative)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)
Type II DKW (Iterative)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)	0.1 (0.000)

Table 5.5: FNR ($tFDP(0.1)$) for different methods, simplified exponential dependence

Chapter 6

Applications

In this chapter we will provide some applications and real data examples. Among the many other possible applications, refer to Ip (2001) for an explanation of the use and benefits of MTP procedures for dependent data in testing for local dependency in item response data. Refer to Yekutieli and Benjamini (1999) for the use of MTP procedures in testing for significant correlations in correlation maps.

Example 6.0.1. *First, we will revisit the multiple problem described in Benjamini and Hochberg (1995), with a complete description in Neuhaus* et al. *(1992). Multiple endpoints analysis in clinical trials is one of the most encountered multiplicity problem in medical research. In a randomized multicentre trial in 421 patients with acute myocardial infarction, a new front-loaded administration of rt-PA (thrombolysis with recombinant tissue-type plasminogen activator) has been compared with APSAC (anisoylated plasminogen streptokinase activator). The treatments are both known to reduce mortality in myocardial infarction. In this example, the difference between the treatments is measured on each of 15 endpoints, measuring cardiac and other events (like bleeding complications) after the start of thrombolitic treatment, yielding the ordered p-values: 0.0001, 0.0004, 0.0019, 0.0095, 0.0201, 0.0278, 0.0298, 0.0344, 0.0459, 0.3240, 0.4262, 0.5719, 0.6528, 0.7590, 1.000. The most important hypothesis, which we want to reject, is the one referring to* $p_{(4)} = 0.0095$*, and it refers to reduced in-hospital mortality rate. Note that there is a big jump from 0.0459, the highest p-value smaller than 0.05, and 0.3240. Table 6.1 shows how many hypotheses (corresponding to the smaller p-values) are rejected at level* $\alpha = 0.05$*, for different procedures. Note that many procedures end up rejecting as many hypotheses as the uncorrected testing. Note that the DKW approach, in this case, is very powerful even if m is small; while the other procedures for* $tFDP(0.1)$ *control are not as powerful. In particular, FWER controlling procedures, BY procedure and generalized augmentation procedure with negative augmentation don't lead us to declare significant difference in mortality rate between the treatments.*

Uncorrected	9
Bonferroni	3
Step-down Holm	3
BH	4
BY	3
Plug-in (Storey)	9
Plug-in (DKW)	6
Plug-in Iterative	9
Gen. Aug., $q := q/2$	4
Gen. Aug., neg. aug.	3
$p_{(1)}$-approach	4
DKW (Iterative)	9
DKW (Storey)	9

Table 6.1: Multiple Endpoint Analysis: number of rejections

6.1 DNA Microarrays

6.1.1 The setting of DNA Microarrays

We will not attempt here a complete review of the analysis of DNA Microarrays, a field of biostatistics which is receiving more and more attention. We will point the reader to detailed reviews (Amaratunga and Cabrera (2004), Parmigiani *et al.* (2003), Brown and Botstein (1999), and Duggan *et al.* (1999) also for a survey of the impressive spectrum of biological applications) and only sketch a very simplified explanation of the problem. Refer to Bolsover *et al.* (1997) and Garret and Grisham (2002) for background on biochemistry and genetics. For a survey on microarray literature, refer also to web sites `http://genomicshome.com` and `http://www.nslij-genetics.org/microarray`, and to `http://www.bioconductor.org` for software support.

Advances of the technology have made it possible to obtain the expression levels of tens of thousands of genes from a single biological sample. For a review of such technologies, refer to Schena *et al.* (1995), Velculescu *et al.* (1995) and Lockhart *et al.* (1996)[1], and also to Cabras (2004) for a detailed review of the statistical issues related to the use of different technologies and relationships with multiple testing. The general idea is as follows: more or less every cell contains a copy of its entire genome. Genes are encoded in the DNA of the genome, whose task is to "make" proteins. Proteins are very different from each other, and perform a whole variety of tasks, from regulation to structural or catalytic functions. When a protein is to be generated, DNA is transcribed

[1]By the way, the two most important techniques are cDNA arrays (Schena (2000), `http://www.microarrays.org`) and oligonucleotide arrays (the suggested Lockhart *et al.* (1996) and Affymetrix (1999)).

into RNA (by splitting), which is then translated into a protein. The idea behind the process is similar to the idea behind the generation of an executable from a program in informatics. RNA is the code of the program, the protein is the executable. The essential feature of microarray analysis is to measure mRNA (i.e., the *messenger* RNA) abundance in some sampled cells. These experiments can be performed with many purposes in mind[2]. We list here at least four: first, compare this abundance (the expression of a single gene) with the expression of genes from samples of other individuals in different biological conditions, and identify genes that are less expressed (down-regulated) or over expressed (up-regulated) in the biological condition of interest. For instance, if a particular gene is significantly up-regulated in a sample from a group of ill people, it is reasonable to view it as correlated with the disease, if not even a genetic cause of the disease. The second purpose can be to identify genes that are *not* expressed. It is well known that a great part of our genome is constituted by *junk* DNA, i.e., DNA that never activates. Again, it is particularly interesting to find genes that that are somehow "turned off" or "turned on" by the disease. The third purpose can be identifying *pathways*, i.e., groups of genes that activate in sequence, structured ordering, or interact. Common clustering techniques are applied, like Partitioning Around Medoids (PAM). Clustering is used also to identify groups of active genes without formal testing. Among the other references, see for instance Pollard and van der Laan (2003a) or Tibshirani and Bair (2004) who propose innovative procedures to cluster genes. Rocci and Vichi (2004) propose the "double k-means", a clever procedure to *simultaneously* cluster genes and samples, i.e., rows and columns of the data matrix. A fourth purpose is classification, i.e., prediction of the biological condition through measuring of the genes. If a good predictor can be formed, then the genes in the classifier are related to the disease; and moreover the disease can be diagnosed by measuring the expression levels of some particular genes. For the purpose of our dissertation, the first two tasks are more relevant: usually, a test is done on each gene to determine if it is differentially expressed between the biological conditions, and a test is done under each biological condition to determine if the gene is not expressed in that case. The third and fourth task (clustering and classification) are also relevant, in the sense that significance testing is usually performed *before*, in order to select a subset of relevant genes. Using not relevant genes for class prediction or clustering may lead to inconsistent results due to gene's expression uncertainty. For a review of multiple testing methods in the context of microarray data analysis, among the many possibilities, see also Dudoit *et al.* (2003a).

Description of a Microarray Experiment

A typical (oligonucleotide) microarray experiment is as follows: 3000 to 50000 genes are measured on one slide, and two biological samples are put on the same slide. One of them can be a reference sample, to standardize the results. A green fluorescent dye is

[2]Assuming for the time being no covariates are measured with the genes expression, otherwise we fall into (mainly logistic) modeling.

attached to the mRNA from one sample, and red to the other. Hence, a slide is made of thousands of spots, each with a particular probe of a single mRNA sequence. In general, there are also other kind of spots, like the so called "Negative Checks", i.e., strains of DNA which should not be hybridized by any mRNA sample (and are precious in estimating the distribution of not expressed genes); and "Positive Checks", the dual. The two samples in each spot hybridize to their complementary strand, in a competitive setting. This can be thought of as thousands of experiments going on *simultaneously*. The measure of the green and red signal is a noisy (relative) measurement of the abundance of mRNA for that particular gene. The result of the experiment is thus a data matrix of n rows (n cDNA samples, in general from the same tissue of individuals in two or more biological conditions) with m columns (one for each gene) with the ($\log_2$) of the expression of the gene, measured on the red or green channel. The sample size n typically ranges from 4 to 100 individuals. It is apparent that a data set in which the number of rows is much smaller than the number of columns presents statistical challenges not traditionally dealt with. Among the immediate statistical issues, not linked in general to inference, there is the quantification of the fluorescence signals from each spot (Yang *et al.* (2002a)), filtering of bad spots, normalization within and between slides, forming a test statistic for each gene (usually, a t or F statistic). The filtering phase consists in getting rid of badly measured genes. It is important to keep in mind that the microarray experiment is subject to a lot of experimental artifacts. A tiny grain of dust in a spot can of course invalidate the expression measure for that gene, and so on. Usually, genes with expression too close to the *saturation* level (the maximum expression recordable by the machine) are not considered in the analysis, together with genes corresponding to spots not passing a whole lot of quality tests (signal to noise ratio, spot uniformity, dimension of hybridized area, background noise, etc.). Then, normalization is performed to get rid of part of the experimental variability and systematic bias inside a single slide and between the slides (for a discussion of the problem, see Tseng *et al.* (2001) or Durbin and Rocke (2004)). Again, the conditions under which the experiment is done are crucial to the results: the heat in the room where the experiment is performed, tiny differences in the duration of exposure, etc. introduce bias and increase the variability of the recorded expression levels. Usually, first of all a measure of the background intensity of expression is subtracted to the expression intensity[3]. Then, a normalization method is applied. Among the most used there are: the *global* normalization, in which the expression for the genes is divided by the observed mean in the whole slide, and RI-lowess normalization (Yang *et al.* (2001)). Recently, Wang *et al.* (2001), Yang *et al.* (2002b), and Wang *et al.* (2003) propose the q_{com}, an index of the overall quality of each spot. They merge in this way filtering and normalization, eliminating the genes corresponding to q_{com} too low, and then using the index to perform a lowess based normalization.
Note that microarrays are in general only the first step before a further investigation,

[3]Though some authors, like Yang *et al.* (2002a), suggest that background should be ignored in order not to introduce additional variability in the low intensity range.

the *validation* phase, takes place. The task of the researcher is to restrict the number of suspect genes from the whole genome to some tens[4], which will then be biologically validated (for instance with RT-PCR, RNA blotting, or other techniques: see for instance Zweiger (2001)). Hence, a small proportion of false positives are allowed; while too many false positives would make the validation phase impossible from an economical point of view. As we pointed out in the first chapter, aiming at no false positives, with thousands of tests going on, means no rejections and no candidate genes for validation. Hence, control of FDR or $tFDP(c)$ is naturally desirable in the microarray setting.

Dependence is intuitively present in DNA microarrays, and often ignored in the literature. Genes measured with the same technology in the same laboratory are subject to common sources of noise. While normalization is performed to remove part of the systematic bias, it is reasonable to expect that the random noise is not acting independently on each spot of the slide. There is a form of spatial dependence. Secondly, there is dependence also in the "signal": changes in expression are part of the same biological mechanism, and hence the expression of each gene is not unrelated to the expression of the other genes. In other words, while the individuals can be thought of being independent (i.e., the rows of the data matrix are independent), the genes in a single tissue are dependent. Likely, they present at least a form of block dependence, with blocks identified by the pathways and/or by groups of similar mRNA codes. This is the well known cross-hybridization problem. By cross-hybridization, we mean that a probe in a particular spot can be hybridized not only by its complementary mRNA strand, but also by similar mRNA which correspond to a different gene. The likelihood of cross-hybridization, by the way, varies from technology to technology; and is almost zero in certain cases. Since blocks of dependent genes are reasonably expected to be small (literature investigates pathways of two to five genes, while a maximum of 50 is thought of being possible), block dependence in microarray data can be thought of as m-dependence (i.e., no block can be of arbitrary size); and thus microarray data fall under the hypotheses of Theorem 3.1.6.

6.1.2 Genetic patterns of colon cancer

Alon *et al.* (1999) analyze data on colon cancer. The expression of around 6500 genes is recorded in 40 tumor and 22 normal samples from the colon of 62 patients. After filtering, 2000 genes were normalized using global normalization, and a two-sample t-statistic was computed on each gene to verify if there was a significant difference between the biological conditions. Figure 6.1 shows an histogram of the 2000 t-statistics. p-values are computed from the statistics. Colon cancer is well known to be associated with variations in the expression of many genes, and in fact the histogram itself suggest the presence of a few significant genes.

There are hints of possible dependence in this data. Using an immediate extension of

[4]For this reason, the process is also sometimes called *gene shaving* or *gene screening*.

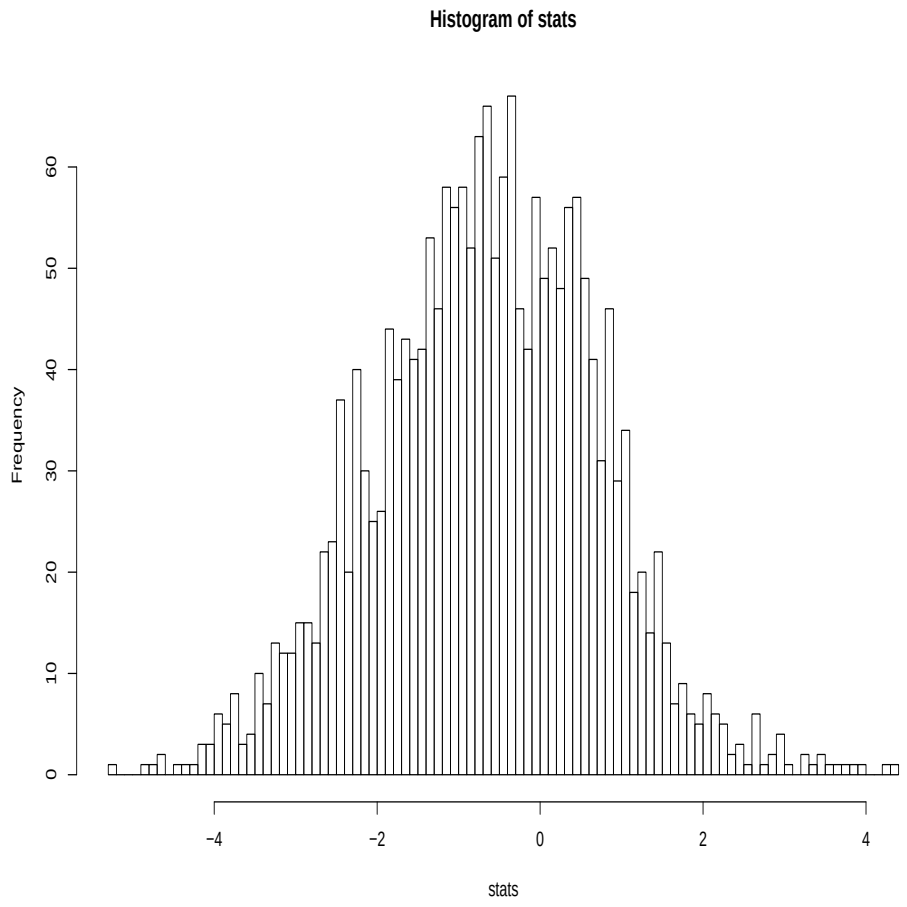

Figure 6.1: Histogram of 2000 two-sample t-statistics for Alon *et al.* (1999) data

the results of van der Laan and Bryan (2000), it is possible to give an idea of how much it is inefficient to estimate the dependence in a microarray experiment. This stems from the fact that the number of subjects/samples n is such that $n << m$, where m is the number of genes. We obtain now confidence intervals for the $\binom{2000}{2}$ correlations between the genes. Let Y_i be the i-th gene expression, which will be assumed to be $N(\mu_i, \sigma_i^2)$. It is straightforward to invert the formulas in Theorem 3.2, pag. 10, of van der Laan and Bryan (2000), to obtain

$$\Pr(\max_{ij} |\widehat{\Sigma}_{i,j} - \Sigma_{i,j}| > \varepsilon) < \delta(\varepsilon), \tag{6.1}$$

where $\delta(\varepsilon) = \min(2 \exp [2 \log m - \frac{3\varepsilon^2 n}{6\sigma_\Sigma^2 + 2W^2 \varepsilon}], 1)$; $\Sigma_{i,j}$ is the covariance ($i \neq j$) between two genes, or the variance of the $i = j$-th gene. σ_Σ^2 is an upper bound for the variance of $Y_i Y_j$, W an upper bound for $Y_i - \mu_i$, m is the number of genes and n the number of samples. We will let $W = 5$ and σ_Σ^2 be twice the estimated maximum on a random subset of 5000 of the $\binom{2000}{2}$ samples, i.e., 36528.93. This leads to $\varepsilon = 139.21$ if one wants to obtain a confidence interval at level $\delta(\varepsilon) = 0.95$. Hence, a 95% confidence interval

for each estimated covariance coefficient is given by $[\Sigma_{ij} - 139.21, \Sigma_{ij} + 139.21]$. This leads all but eight of the confidence intervals to contain the zero, counting also the 2000 confidence intervals for the variances. The sample variance/covariance matrix is not a good estimate of Σ due to high variance of the estimator. Nevertheless, the positive estimates for the correlations outnumber the negative estimates by a factor of 16.2 to 1. This is a (mild) hint of positive dependence (association under normality assumption) in the data. This, and the fact that the dependence *cannot* be efficiently estimated, should lead us to consider and use MTPs valid under dependence in dealing with this data.

Now, p-values are computed from the test statistics. Uncorrected testing leads us to 488 rejections, Bonferroni correction to 11, BH method to 180, Plug-in with Storey estimator to 217, Plug-in with iterative estimator to 197. Generalized augmentation procedure to 198, $p_{(1)}$-approach to 13. Of this methods, we know Bonferroni, BH (by m-dependence), Plug-in with iterative estimator, $p_{(1)}$-approach and generalized augmentation procedure were proved to be valid under the assumed dependence structure on this data.

6.1.3 Classification of Lymphoblastic and Myeloid Leukemia

Acute Lymphoblastic Leukemia (ALL) and Acute Myeloid Leukemia (AML) are two variants of Leukemia, which are treated differently. The goal of this experiment is to build a classifier in order to distinguish between the two variants, through gene expression profiling. Data come from Golub *et al.* (1999). 7129 different genes remained after filtering, with some positive controls and replicates for quality control. This leads to 6817 different genes, normalized using global normalization. We will use 56 of the original 72 samples.

A training set of 22 samples was selected, 11 chosen at random from the ALL samples and 11 chosen at random from the AML samples. A test set of 34 samples, of which 20 ALL and 14 AML, was used to estimate the classification error.

As stated, the first problem is to select a subset of genes which are significantly differentially expressed between the two Leukemia conditions. To do so, a t-test was performed on each gene, obtaining a vector of $m = 6817$ p-values. Uncorrected testing would reject 1552 hypotheses, classical BH 28, Plug-in with iterative estimator 30 and Plug-in with Storey estimator 124. FWER control with Bonferroni correction would reject only 1 p-value. If the goal of this experiment was just to determine a subset of candidate genes for validation, only the set provided by Plug-in with iterative estimator or BH method would be considered.

A further confirmation of the fact that FDR controlling procedures are to be preferred over uncorrected testing and FWER control is given by the estimates of classification error. Classification was performed using k-Nearest Neighbor Classifier (Cover and Hart (1967)), with $k = 3$. Using all the genes would lead to 16 misclassified cases in the test set, 10 would be given using the set of genes selected with uncorrected testing, 8 with

the set of genes selected with Plug-in with Storey estimator, 6 by BH and Plug-in with iterative estimator. Note that, in the light of the simulations in Chapter 3, the fact that Plug-in with Storey estimator is so much different from the other Plug-in method suggests presence of dependence in the data (like in the previous example).

6.2 Wavelet Thresholding

We will now show that our multiple testing procedures can be used as a flexible, automatic and adaptive method for wavelet thresholding. Again, we will not attempt a full review of the complex and large world of non-parametric curve estimation here, but just sketch a general idea. Suppose we observe data y_i, $i = 1, \ldots, n = 2^{J+1}$, where it is known that

$$y_i = g(t_i) + \varepsilon_i, \tag{6.2}$$

with $\varepsilon_i \sim N(0, \sigma^2)$, σ known. We will assume $g : \mathcal{R}^d \to \mathcal{R}$ and that the observation points t_i are equally scattered on a compact cube in $\mathcal{R}^d$. If $d = 1$, this is a very common non parametric regression problem; while when $d = 2$ this is for instance an image reconstruction problem. There exists a variety of methods to estimate $g(\cdot)$, including kernel estimation, spline smoothing, etc. However, these methods are sub-optimal in case the function f is spatially inhomogeneous. An advantageous possibility is to use an orthonormal wavelet basis $\psi_{jk}(t)$, and estimate the wavelet coefficients d_{jk} in the expansion $g(t) = \sum_j \sum_k d_{jk}\psi_{jk}(t)$. $\psi_{jk}(t)$ is a translation and dilation of a single fixed function $\psi(t)$, the so called *mother* wavelet:

$$\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k), \quad j, k \in \mathcal{Z}.$$

For many possible mother wavelets, see Daubechies (1992), who introduced also a general family of smooth wavelets, indexed by the number of vanishing moments of the mother wavelet, N (including the 0^{th} moment). In our examples, we will use $N = 2$. Note that many mother wavelets don't even have a closed form expression. It is well known that the coefficients d_{jk}, with respect to any basis $\psi_{jk}, j, k \in \mathcal{Z}$, are computed as

$$d_{jk} =< g, \psi_{jk} >= \int g(t)\psi_{jk}(t)\ dt.$$

For an essential and clear introduction to wavelets, refer to Ogden (1997), to Härdle *et al.* (1998) for a more detailed discussion, and to Donoho *et al.* (1995) for applications in statistics. A wavelet basis works just like any other possible basis for function spaces, but unlike many other bases (like the sine and cosine basis), it is localized both in frequency and time scale. This is the biggest advantage of wavelet bases, together with the speediness of the computations. For this reason, functions with variable degree of smoothness through the support (i.e., spatially inhomogeneous) are well estimated. Wavelets can efficiently estimate both peaks and constant parts of a function. Wavelets estimate functions in Besov, Sobolev and/or Hölder spaces of smooth functions, and

are adaptive in the sense that they achieve minimax risk "close" to the risk that would achieve the optimal estimator with known function space. Other main properties of wavelet bases are: *multiresolution*, i.e. the fact that the function is analyzed through a nested set of scales; and *compression*, i.e. the fact that wavelet transforms of real world signals tend to be sparse. Wavelet estimators are commonly used for signal denoising, image compression, non-parametric regression, etc. The procedure for wavelet estimation is as follows: through a discrete wavelet transform algorithm (DWT) estimate the wavelet coefficients from the data y_i. For a description of a fast DWT algorithm, which will also be used here, refer to Mallat (1989). There will be $m = n - 1 = 2^{J+1} - 1$ coefficients. White noise obviously contaminates all the estimated $\widehat{d}_{jk}$, $j = 0, \ldots, J$, $k = 0, \ldots, 2^j - 1$. However, due to parsimonious representation by wavelets, it is reasonable to expect that only few $\widehat{d}_{jk}$ contain information about the signal $g(\cdot)$; while all the rest are very close to zero. Hence, the goal is to identify the non zero coefficients and set to zero the other estimates, through a procedure called *thresholding*: set $\widehat{d}_{jk} := \widehat{d}_{jk} 1_{|\widehat{d}_{jk}| \geq \lambda}$, for a certain λ. This is *hard* thresholding. Another possibility is *soft* thresholding, which sets $\widehat{d}_{jk} := sign(\widehat{d}_{jk})(|\widehat{d}_{jk}| - \lambda) 1_{|\widehat{d}_{jk}| \geq \lambda}$. This is done to filter the noise also from the estimates of the non zero coefficients. After the thresholding step, the final estimator of the signal is then $\widehat{g}(t) = \sum_j \sum_k \widehat{d}_{jk} \psi_{jk}(t)$, constructed via a fast algorithm called inverse wavelet transform (IWT). See also Strang (1989) for detailed exposition of possible DWT and IWT algorithms. Among the many possible choices of λ, the most used is the universal threshold proposed in Donoho and Johnstone (1994), which is $\lambda = \sigma\sqrt{2 \log n}$ for $j \geq j_0$ and 0 for $j \leq j_0$, for a certain j_0. j_0 is typically 5 (Donoho and Johnstone (1994)) or 3 (Nason and Silverman (1994)). If σ is unknown, a robust estimator is taken, in general using the coefficients from the highest level, which intuitively are almost completely determined by random noise: $\widehat{\sigma} = \frac{\text{median}(|\widehat{d}_{Jk} - \text{median}(\widehat{d}_{Jk})| : k = 0, \ldots, 2^{J-1} - 1)}{0.6745}$. Among other possible thresholding methods there is SureShrink, proposed in Donoho and Johnstone (1995), which combines an adaptive threshold chosen to minimize the risk, estimated with Stein Unbiased Risk Estimator (SURE), with the universal threshold. Nason (1996) uses a cross validation criterion to minimize the predicted MSE. While asymptotically any choice of j_0 is equivalent in the SureShrink method, for small sample sizes it is crucial, as for the universal method. The same Donoho and Johnstone (1994) and Fan (1994) point out that thresholding is nothing but hypothesis testing on each wavelet coefficient. For this reason, Abramovich and Benjamini (1996) propose to use the BH procedure to test the set of hypotheses $H_i : d_{jk} = 0$, set λ equal to the highest absolute value of the rejected d_{jk}, and then do either hard or soft thresholding. They show that this provides a much more automatic procedure of thresholding, which works very well with the sparseness of wavelet representations. They argue that FDR control can improve the MSE in some cases, and most of all it adapts very well to different smoothness situations and enjoys robustness of MSE-efficiency. They do this by comparing the MSE of thresholding via of testing with BH procedure with the MSE of universal thresholding in simulation.

Here MSE is defined as $n^{-1}||g - \widehat{g}||_2$. Their computations and applications are all in one dimension, for Gaussian independent noise. We will apply in this section more powerful FDR controlling procedures, and procedures to control the $tFDP(c)$, to thresholding in one and two dimensions, with dependence in the noise. We will show that the improvement in using multiple testing thresholding is much more evident in more than one dimension, and impressive when the signal-to-noise ratio is high. Among the possibilities for further work, there are formal optimality results of MTP procedures for thresholding.

Extension of the wavelet model (6.2) to non-Gaussian (Chipman *et al.* (1997), Simoncelli and Adelson (1996)) and/or to dependent situations (Lee *et al.* (1996), Chou and Heck (1994), Crouse *et al.* (1998), Vannucci and Corradi (1999)) is a wide part of literature on wavelets. Usually, a Bayesian framework is adopted. Crouse *et al.* (1998) note that in the non-Gaussian situation, assumption of independence between wavelet coefficients is not realistic, since coefficients are usually *clustered* and *persistent*: they tend to be large/small in groups and the magnitude propagates across scales. In the Gaussian situation, dependence between the coefficients can also come from two sources: correlation in the noise, a typical situation in two dimensions; and correlation in the signal, in the sense that $g(t)$ is a realization of a certain stochastic process $X(t)$. In this last case, Vannucci and Corradi (1999) prove important results on dependence in wavelets: first of all, the recursion

$$cov(d_{jk}, d_{j'k'}) = \sum_h \sum_i (-1)^{i+h-2(k+k')} l_{1-i+2k} l_{1-h+2k'} cov(c_{(j+1)h}, c_{(j'+1)i}), \tag{6.3}$$

where l_k are known *positive* coefficients completely determined by the chosen basis, satisfying $\sum_{k=-\infty}^{+\infty} l_k^2 < \infty$; and c_{jk} are the so called *scaling* coefficients (see Ogden (1997)) satisfying the recursion

$$cov(c_{jk}, c_{j'k'}) = \sum_h \sum_i l_{h-2k} l_{i-2i} cov(c_{j+1,h}, c_{j'+1,i}) \tag{6.4}$$

and with

$$d_{jk} = \sum_h (-1)^{h-2k} l_{1-h+2k} c_{j+1,h}. \tag{6.5}$$

The recursions in (6.3) and (6.4), and the relationship in (6.5) can be used, in a so called *decomposition* algorithm, to make statements on the dependence between the d_{jk}. In many cases, depending on the basis used and considerations on the covariance between the scale coefficients at the coarser levels, it will be possible to prove constant sign of $cov(d_{jk}, d_{j'k'})$. This, together with the normality assumption, will imply (negative) association of the wavelet coefficients.

Another result of Vannucci and Corradi (1999) that can be extended to our setting is this: if the mother wavelet comes from the Daubechies minimum phase family with N vanishing moments, then the coefficients d_{jk} and $d_{j'k'}$ are uncorrelated for integers k and k' such that $k' - 2^l k > 2^l(2N-1)$ or $k' - 2^l k < 1 - 2N$, with $l = |j' - j|$.

I.e., wavelet coefficients are dependent in blocks; and the blocks are known. Note that the size of the blocks is not bounded, hence m-dependence is not proved on the wavelet coefficients. Mixing assumptions, if needed, should be supported by different considerations. Nevertheless, wavelets under dependence show a whole lot of properties which allow us to use multiple testing methods for thresholding.

6.2.1 Non Parametric Regression

We will apply our methods, in this subsection, to a test function commonly used in this setting (Donoho and Johnstone (1994, 1995)), the Doppler function: $g(x) = \sqrt{x(1-x)}\sin\left(\frac{2,1\pi}{x+0.05}\right)$, sampled in $n = 512$ points; and to a public available data set.

Table 6.2 shows the MSE for thresholding with different methods, calculated as the average MSE in $B = 1000$ Doppler signals, with independent noise and different variance. In bold, the worst and best MSE for each signal. It is evident that FDR and $tFDP(0.1)$ control is better than universal thresholding for small noise levels, while it outperforms our methods in case of high noise, when the signal is hard to reconstruct. Figure 6.2 shows the Doppler function, and one Doppler signal for each of the three noise levels. Figure 6.3 shows the estimated functions, with best thresholding method on the first column for the three noise levels, and the worst on the second column of graphs. It is evident that multiple testing methods do a much weaker denoising than universal thresholding, thus not losing information about the signal when the noise is not strong.

$\sigma =$	0.05	0.1	0.25
Hard universal, $j_0 = 3$	0.009174	0.01033	0.0176
Hard universal, $j_0 = 5$	0.001612	0.00525	**0.0128**
Soft universal, $j_0 = 3$	**0.019078**	**0.02027**	0.0283
Soft universal, $j_0 = 5$	0.006469	0.00756	0.0647
BH	0.001610	0.00471	0.0323
Plug-in (Storey)	0.001614	0.00473	**0.0324**
Plug-in (Iterative)	0.001611	0.00472	0.0323
DKW (Storey)	0.0016165	0.00458	0.0318
$p_{(1)}$-approach	0.0016432	**0.00458**	0.0321
Gen. Aug., $q := q/2$	**0.001609**	0.00466	0.0319

Table 6.2: MSE for thresholding Doppler signal with independent noise

Table 6.3 shows the MSE for thresholding with different methods, with dependent noise. The usual simplified exponential dependency structure was used[5], with parameter $\tau = 20$. In bold, the worst and best MSE for each signal. Note that BH and Plug-in with

[5]Note anyway that the mapping between distance of the points and covariance is partly lost, since the function is in one (and not two) dimensions. In next section, this link will not be lost.

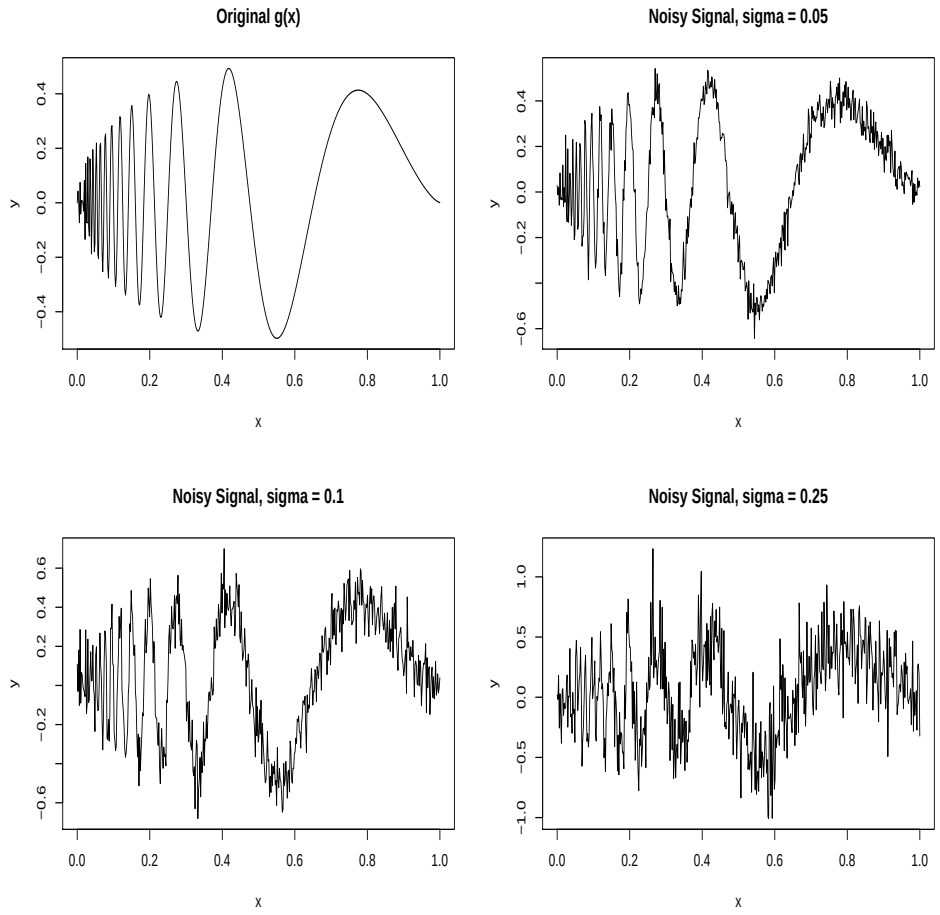

Figure 6.2: Doppler function and signals

iterative estimator approaches, work in light of the results of Chapter 3 (the association hypotheses), while $p_{(1)}$ and gen. aug. approach work in light of the results of Chapter 5. Figure 6.4 shows the Doppler function, and one Doppler signal for each of the three noise levels. Figure 6.5 shows the estimated functions, with best thresholding method on the first column for the three noise levels, and the worst on the second column of graphs.

Same results are observed using the other infamous test functions: Blocks, Bumps, Heavisine, Jumpsine; with Blocks function showing a preference for multiple testing methods till much larger values of σ. This is so because universal thresholding is seen to be preferred when the signal-to-noise ratio is small and the signal is not distinguishable any more. Blocks signal is particularly robust to noise.

We will now apply the methods that proved best to thresholding wavelet coefficients on a real data set. The data were obtained from Silverman (1985) and describe the recorded head acceleration of a motorcycle after an accident, as a function of time in milliseconds since impact. A few tricks are needed to fit this data: we will first dilate

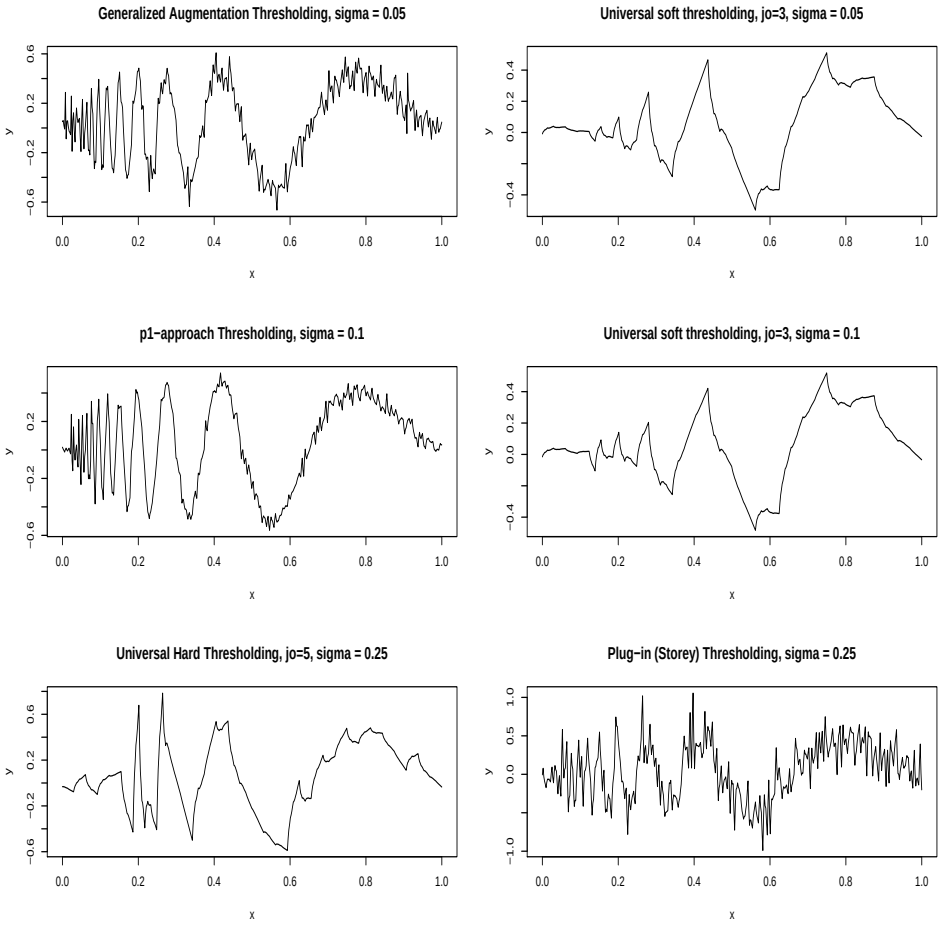

Figure 6.3: Best and worst estimated signals

the data on an equally spaced x-axis, and add zeros at the beginning and end of the data set to make the length of the acceleration vector a power of 2. These zeros will not be counted in the estimate of the standard deviation, calculated with the usual robust estimator, and in thresholding with multiple testing approaches. Then, we will get rid of the initial and final zeros, and dilate back the data and estimated function on the original scale. Figure 6.6 shows the data and the estimates.

6.2.2 Image Reconstruction

We will apply our methods, in this subsection, to a test image downloaded from the Internet, used also by Allen Gersho's lab at the University of California, Santa Barbara, and digitalized to 256*256 pixels, shown in Figure 6.7. Figure 6.8 shows the noisy image, with independent Gaussian noise added, and the denoised image with universal hard thresholding, $j_0 = 5$; Plug-in thresholding with iterative estimator and DKW thresholding. Figure 6.9 shows the same, for dependent noise. The noise was sampled in 64 blocks of 32*32=1024 dependent Gaussian random variables, with usual simplified-

$\sigma =$	0.05	0.1	0.25
Hard universal, $j_0 = 3$	0.009349	0.01074	0.0195
Hard universal, $j_0 = 5$	0.004353	**0.00567**	**0.0164**
Soft universal, $j_0 = 3$	**0.019106**	**0.02036**	0.0289
Soft universal, $j_0 = 5$	0.006570	0.00790	0.0167
BH	0.002035	0.00619	0.0444
Plug-in (Storey)	0.002028	0.00620	**0.0444**
Plug-in (Iterative)	0.002018	0.00618	0.0444
Type I DKW (Iterative)	**0.002248**	0.006192	0.0444
Type II DKW (Iterative)	0.002016	0.006184	0.0444
$p_{(1)}$-approach	0.002413	0.006193	0.0444
Gen. Aug., $q := q/2$	0.002037	0.006195	0.0444

Table 6.3: MSE for thresholding Doppler signal with dependent noise

	Universal, $j_0 = 5$	Plug-in (iterative)	(Type II) DKW (iterative)
Independent Noise	0.218	0.150	0.145
Dependent Noise	0.217	0.161	0.159

Table 6.4: Mean Square Errors for Image Reconstruction Example

exponential covariance structure, with $\tau = 20$. It is apparent that in this case the universal thresholding proves too severe, resulting in bad quality of the image after denoising. Mean square errors are shown in Table 6.4. The lowest MSE in the independent case is achieved by DKW approach with iterative estimator. In the dependent case, the lowest MSE is achieved by Type II DKW approach as derived at page 69. Note that in image compression problems no noise, or very high signal-to-noise ratio, is assumed. In this case, MTP thresholding is strongly advisable for high quality results.

6.3 Multivariate Linear Regression

Multivariate linear regression is a basic tool in data analysis, and a basic textbook topic. The number of purposes and applications in which this kind of tool is used is very large. A linear model is fitted to predict a response variable from many covariates. For each covariate, a test is done on the estimated coefficient in order to determine whether it is significantly different from zero, thus indicating a linear relationship between the corresponding covariate and the response. It is well acknowledged that, in general, the covariates are dependent, which leads often to problems of multi-collinearity. It is all the same well acknowledged that, if the covariates in the model are only "almost" significant, this is not a big problem in the model fitting. It is not uncommon to see linear models in which covariates that have been declared not significant are kept in the

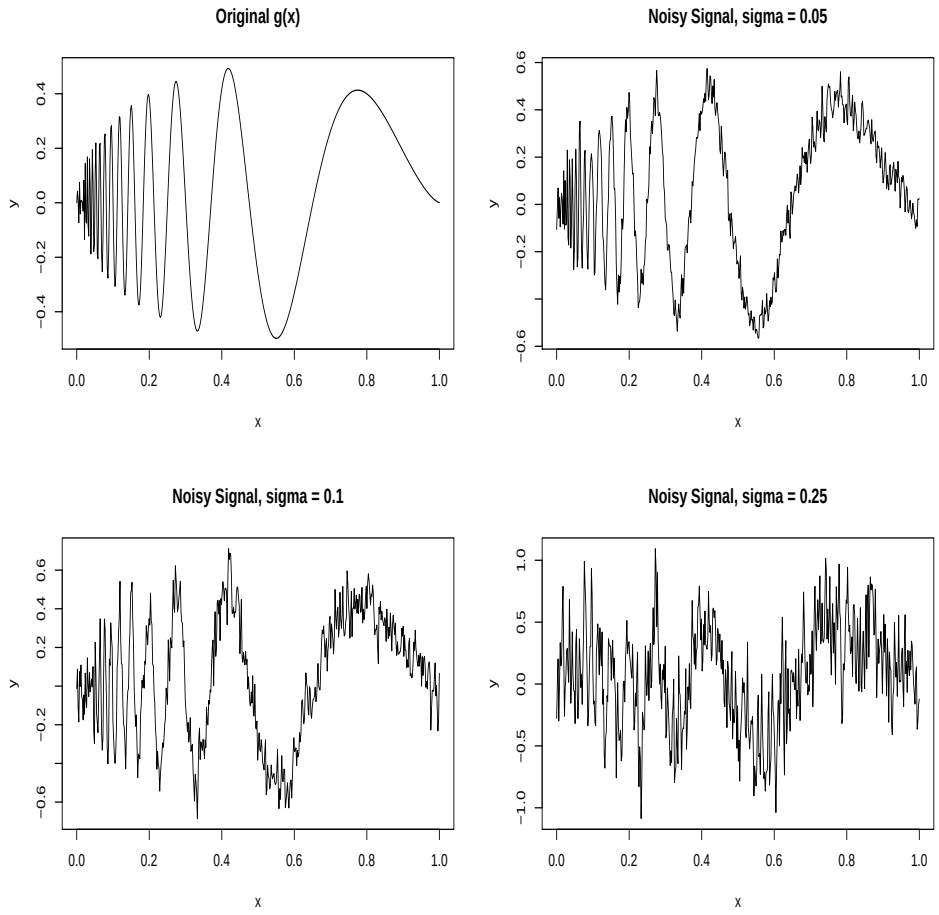

Figure 6.4: Doppler function and dependent signals

model, for various reasons.

The problem of model selection is historically dealt with, and there is also very recent research on the choice of which variables to include in a model (Efron *et al.* (2004), Barbieri and Berger (2004); and references therein). We focus here on a much easier problem: starting from a given model, determine if a single particular covariate of interest is in a significant relationship with the response. There are many applications, especially in biology, in which the researcher is more interested in the relationship between the response and a particular covariate than in a good fit. It is intuitive that, in that case, an MTP is to be applied to get more confidence in the significance statement. Note that in all cases the intercept is to be left out of the MTP, since it is clearly significant, or clearly not significant if the data are standardized. An example is given in the next subsection to clarify the idea.

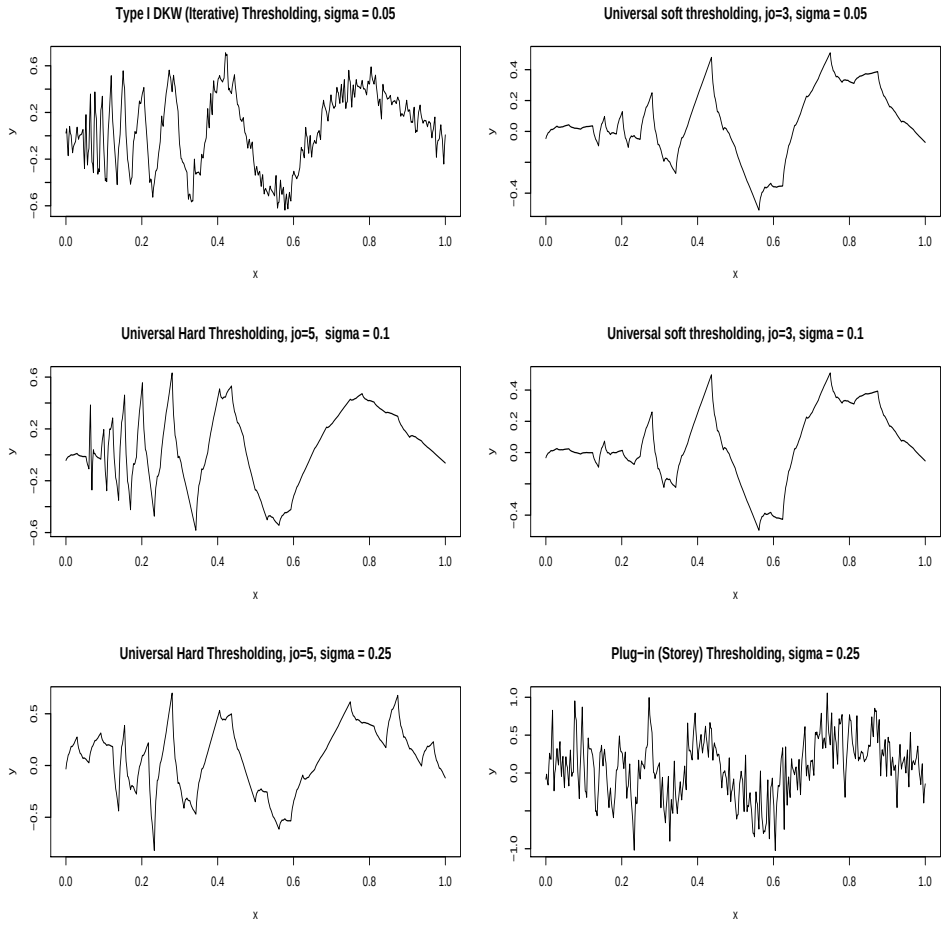

Figure 6.5: Best and worst estimated signals, under dependence

6.3.1 Determinants of Retinol levels in plasma

In Nierenberg *et al.* (1989) a large study is conducted on the determinants of plasma levels of beta-carotene and Retinol, which are known to reduce the risk of developing certain types of cancer. Hence, if a covariate is found to be in significant relationship with plasma levels of Retinol or beta-carotene, it contributes to reduce the risk of certain types of cancer. A linear regression is conducted with a large number of biological covariates. A question of interest is whether smoking can reduce the concentration of plasma levels of beta-carotene and Retinol.

In cases like this, it is intuitive that an MTP is to be preferred over uncorrected testing, to have a more reliable statement on the significance of the particular covariate of interest. In particular, for the Nierenberg *et al.* (1989) data, smoke (together with other five covariates) is found significant at level 0.05, while using a Bonferroni correction only age and dietary intake of beta-carotene are found to be significant on plasma levels of Retinol. BH and Plug-in (Iterative and Storey) methods find four significant covariates, but smoke (which is the fifth in p-value ordering) is not declared significant. DKW,

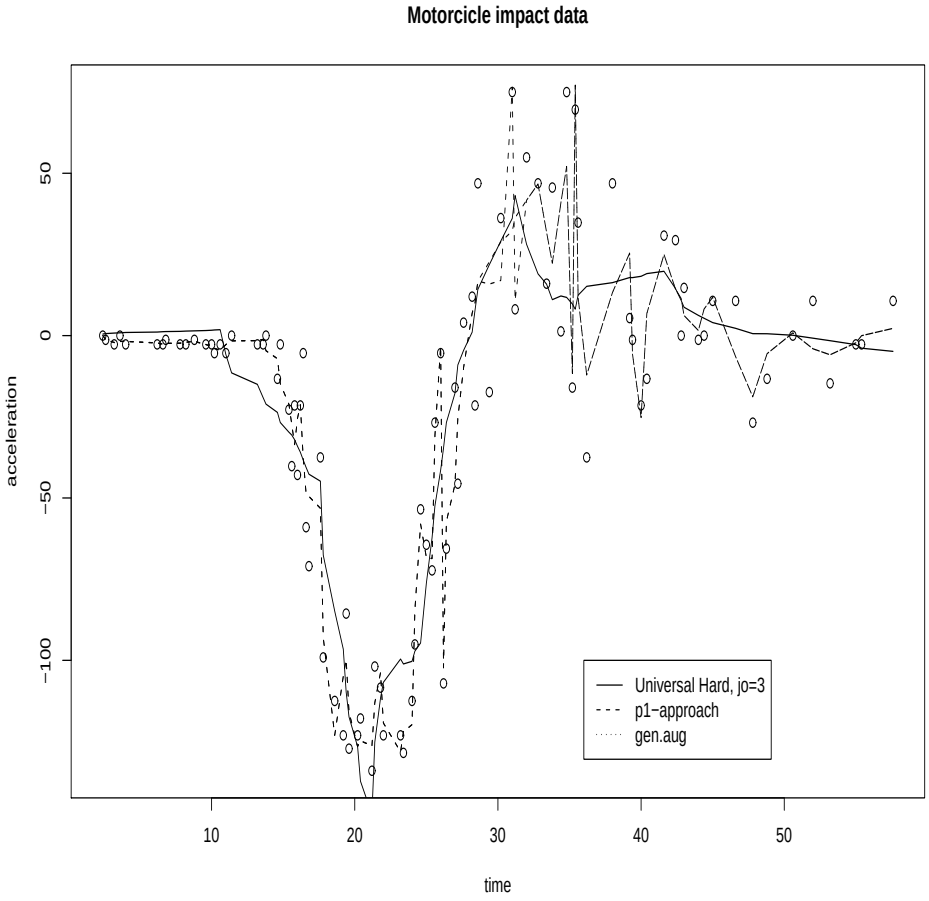

Figure 6.6: Motorcycle data and wavelet estimates

generalized augmentation and $p_{(1)}$ approaches find two significant covariates. Hence, from these data it is not sensible to conclude that smoking is related to decreased plasma levels of Retinol.

Figure 6.7: Sample Image

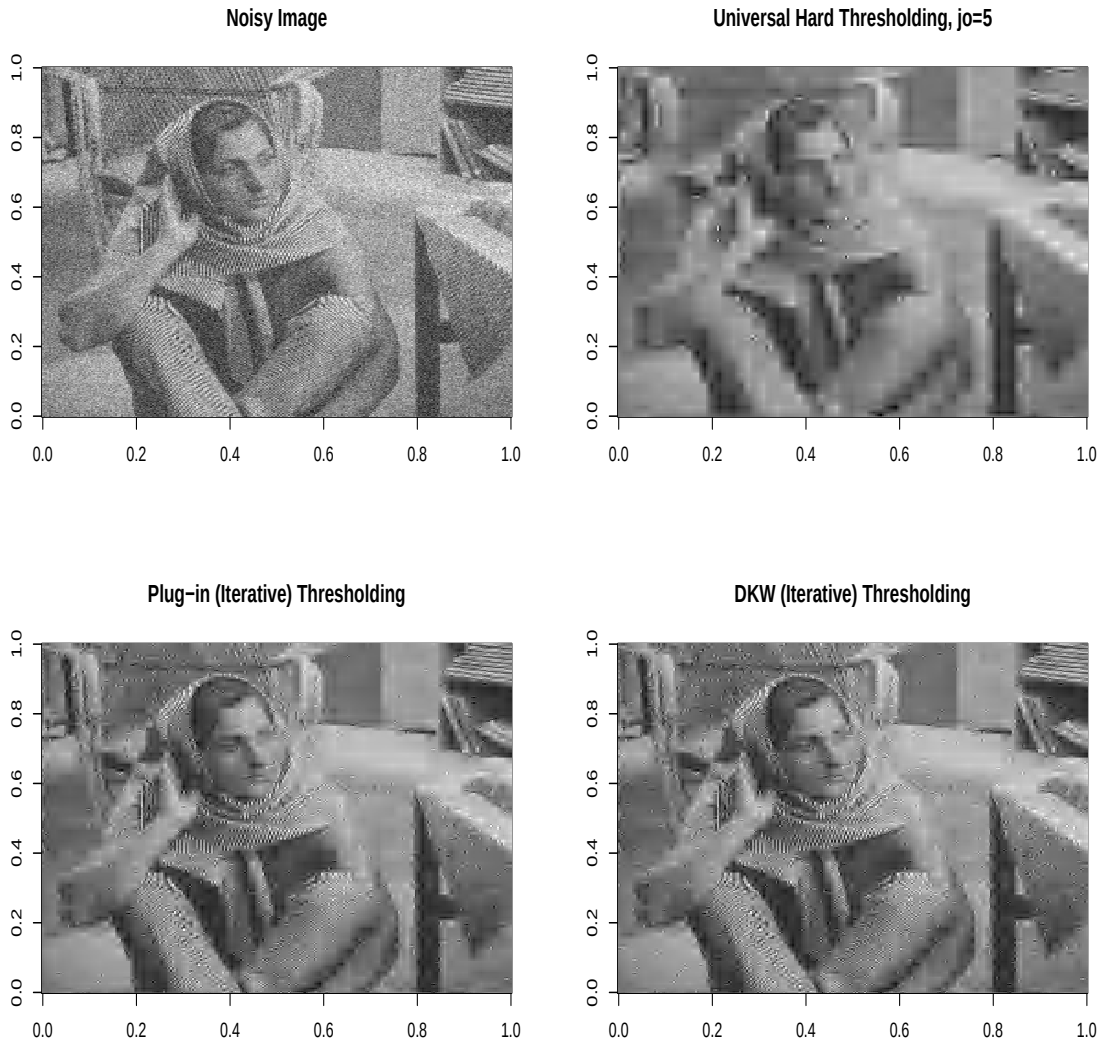

Figure 6.8: Reconstruction of Noisy Image with independent noise

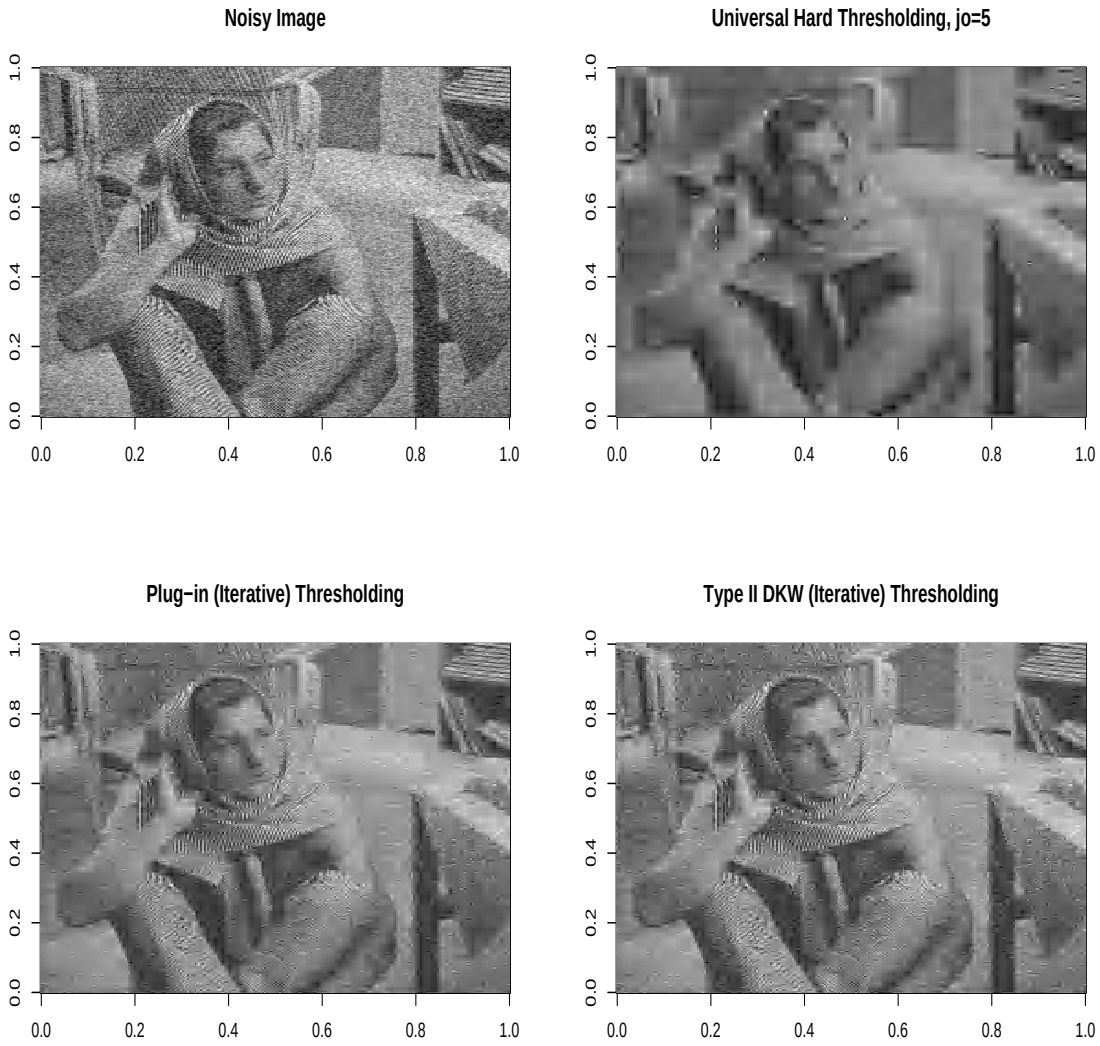

Figure 6.9: Reconstruction of Noisy Image with dependent noise

Appendix A

Proofs from Chapter 3

A.1 Proof of Lemma 3.1.3

$\forall\ (s,t) \in \mathcal{R}^2$ we have that

$$|\Lambda_j(t) - \Lambda_j(s)| \leq |\widehat{G}(t) - \widehat{G}(s)|.$$

This is easily seen. For instance, let WLOG $t > s$ and $j = 1$. We have that

$$\frac{1}{m}\sum H_i[1_{\{p_i<t\}} - 1_{\{p_i<s\}}] \leq \frac{1}{m}\sum[1_{\{p_i<t\}} - 1_{\{p_i<s\}}],$$

since H_i can either be 0 or 1.

Hence we can prove the asymptotic equicontinuity of $\Lambda_j(t)$ by applying the definition:

$$\limsup_n Pr^*(\sup_i \sup_{s,t\in T_i} |\Lambda_j(t) - \Lambda_j(s)| \geq \varepsilon) \leq \limsup_n Pr^*(\sup_i \sup_{s,t\in T_i} |\widehat{G}(t) - \widehat{G}(s)| \geq \varepsilon) \leq \eta$$

where T_i is an opportune partition of $[0,1]$.

A.2 Proof of Lemma 3.1.4

WLOG let $||\Lambda|| = |\Lambda_0(t)| + |\Lambda_1(t)|$. Let T_i be an opportune partition of $[0,1]$

We have that

$$\begin{aligned}
\sup_i \sup_{s,t\in T_i} ||\Lambda(t) - \Lambda(s)|| &= \\
\sup_i \sup_{s,t\in T_i} |\Lambda_0(t) - \Lambda_0(s)| + |\Lambda_1(t) - \Lambda_1(s)| &\leq \\
\sup_i \sup_{s,t\in T_i} |\Lambda_0(t) - \Lambda_0(s)| + \sup_i \sup_{s,t\in T_i} |\Lambda_1(t) - \Lambda_1(s)|;&
\end{aligned}$$

hence

$$\begin{aligned}
\limsup_n Pr(\sup_i \sup_{s,t\in T_i} ||\Lambda(t) - \Lambda(s)|| \geq \varepsilon) \ &\leq \ \limsup_n Pr(\sup_i \sup_{s,t\in T_i} |\Lambda_0(t) - \Lambda_0(s)| \geq \varepsilon/2) \\
&+ \ Pr(\sup_i \sup_{s,t\in T_i} |\Lambda_1(t) - \Lambda_1(s)| \geq \varepsilon/2) \leq 2\eta.
\end{aligned}$$

A.3 Proof of Lemma 3.1.5

We will use the Cramer-Wold device (see Van der Vaart (1998)) to prove the convergence of the vector. We will assume any of the mixing conditions holds. If association holds, it is straightforward to prove the results are the same. Let $(c_0, c_1) \in \mathcal{R}^2$. It is easy to see that $c_0\Lambda_0(t) + c_1\Lambda_1(t)$ is equal to

$$\frac{1}{m}\sum(c_0 + H_i(c_1 - c_0))1_{\{p_i<t\}}.$$

Define the sequence $\{\xi_i\}_{i\in\mathcal{N}}$ to be:

$$\xi_i = \begin{cases} H_i \;\; if \;\{i \mod 2\} = 0 \\ p_i \;\; if \;\{i \mod 2\} = 1. \end{cases}$$

By definition, the sequence will be $\{p_1, H_1, p_2, H_2, p_3, \ldots\}$.

We have then that $(c_0 + H_i(c_1 - c_0))1_{\{p_i<t\}} = (c_0 + \xi_{2i}(c_1 - c_0))1_{\{\xi_{2i-1}<t\}}$.

It is easy to see that the ξ sequence is still mixing, and the same conditions will be true. In particular, called $\alpha'(n)$ the mixing coefficients for the ξ_is at lag n, under Assumption 2 we have that $\sum\sqrt{\alpha'(n)} \leq \sqrt{C'}\sum n^{\frac{-3-\delta'}{2}} < +\infty$. Hence, the series of the partial sums $\sum\xi_i$ will be weakly convergent (see for instance Billingsley (1999)). Same results are immediately proved under the other conditions.

Since $(c_0 + \xi_{2i}(c_1 - c_0))1_{\{\xi_{2i-1}<t\}}$ is a measurable function which depends only on finitely many coordinates of the vector $\{\xi_i\}$, the same results on the mixing coefficients will hold for it (again, see Billingsley (1999)).

Hence $\frac{1}{m}\sum(c_0 + H_i(c_1 - c_0))1_{\{p_i<t\}}$, opportunely rescaled, will converge in distribution. By Cramer Wold device, also the two dimensional vector $(W_0(t), W_1(t))$ will converge in distribution.

A.4 Proof of Theorem 3.1.6

By Condition 1 (Van der Vaart (1998)), Condition 2 (Yoshihara (1975)), Condition 3 (Yu (1993)), Condition 4 or 5 (Oliveira and Suquet (1995)), the empirical process $\sqrt{m}(\widehat{G}(t) - G(t))$ will be convergent to a centered Gaussian random process[1], with covariance kernel given by:

$$K(s,t) = G(s \wedge t) - G(s)G(t) + 2\sum_{k=2}^{+\infty}[P(p_1 < s, p_k < t) - G(s)G(t)].$$

Since the empirical process is convergent, $\widehat{G}(t)$ is asymptotically equicontinouos and $\forall\, t_1, \ldots, t_k$ the vector $(\widehat{G}(t_1), \ldots, \widehat{G}(t_k))$ will converge in distribution.

[1]The convergence is to something very close to a Brownian bridge on the scale of G. The only difference will be given by the covariance kernel, which will be the usual kernel to which is added a convergent series.

For Lemma 1, then also $\Lambda_j(t),\ j = 0, 1$ will be asymptotically equicontinuous and by Lemma 2 $(\Lambda_0(t), \Lambda_1(t))$ will be asymptotically equicontinouos.

Moreover, by Lemma 3, $\forall\ t_1, \ldots, t_k$ the vector $[(\Lambda_0(t_1), \Lambda_1(t_1)), \ldots, (\Lambda_0(t_k), \Lambda_1(t_k))]$ will converge in distribution[2].

Then, it happens that (W_0, W_1) will converge to a centered two-dimensional Gaussian process. The covariance Kernel is obtained by direct calculation of $E[W_i(s)W_j(t)]$, $i = 0, 1$, $j = 0, 1$, and is given by

$$K_2(s,t) = \begin{bmatrix} K_{0,0}(s,t) & K_{0,1}(s,t) \\ K_{1,0}(s,t) & K_{1,1}(s,t) \end{bmatrix},$$

where

$$K_{0,0}(s,t) = (1-a)(s\wedge t) - (1-a)^2 st + 2\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 0, H_k = 0) - (1-a)^2 st],$$

$$K_{0,1}(s,t) = -(1-a)saF(t) + 2\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 0, H_k = 1) - (1-a)saF(t)],$$

$$K_{1,0}(s,t) = -(1-a)taF(s) + 2\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 1, H_k = 0) - (1-a)taF(s)]$$

and

$$K_{1,1}(s,t) = aF(s\wedge t) - a^2F(s)F(t) + 2\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 1, H_k = 1) - a^2F(s)F(t)].$$

The computations are tedious, not very complex and omitted for brevity.

Note now that $\Gamma(t) = \frac{\Lambda_0(t)}{\Lambda_0(t)+\Lambda_1(t)} = r(\Lambda_0(t), \Lambda_1(t))$; where

$$r(\cdot,\cdot) : l^\infty\ x\ l^\infty \to l^\infty.$$

For technical reasons, we restrict the $\Gamma(t)$ process in $[\delta, 1]$, for any $\delta > 0$. The variance of $\Gamma(t)$ is in fact infinite at $t = 0$. Let $Q(t) = (1-a)t/G(t)$ (refer to Section 3.1.2 for more details).

The function $r(\cdot,\cdot)$ is such that $r((1-a)U, aF) = Q$, and it is also Fréchet differentiable at that point, with derivative:

$$r'_{(1-a)U,aF}(V_0, V_1) = \frac{aFV_0 - (1-a)UV_1}{G^2}.$$

Hence, by functional δ-method,

$$\sqrt{m}(\Gamma(t) - Q(t))$$

[2]It is straightforward, in fact, to extend the result of Lemma 3 to the vector case.

will converge to the process defined by the evaluation of $r'_{(1-a)U,aF}(V_0, V_1)$ at the limit of (W_0, W_1). Since this is nothing but a linear combination of the two elements of the vector, with coefficients aF/G^2 and $-(1-a)U/G^2$, it will be a Gaussian process on $(0,1]$, with mean 0 and covariance kernel $K_3(s,t)$ given by:

$$\begin{aligned}
&\frac{a(1-a)}{G^2(s)G^2(t)}[(1-a)stF(s\wedge t) + aF(s)F(t)(s\wedge t)] + \\
\frac{2}{G^2(s)G^2(t)}[a^2F(s)F(t)&\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 0, H_k = 0) - (1-a)^2 st] + \\
- a(1-a)sF(t)&\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 0, H_k = 1) - (1-a)saF(t)] + \\
- a(1-a)tF(s)&\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 1, H_k = 0) - (1-a)taF(s)] + \\
&(1-a)^2 st\sum_{k=2}^{+\infty}[\Pr(p_1 < s, p_k < t | H_1 = 1, H_k = 1) - a^2F(s)F(t)].
\end{aligned}$$

Again, the long computations are omitted. Hence, the FDR process centered at $Q(t)$ has a centered Gaussian limiting distribution with covariance kernel $K_3(s,t)$.

Applying again the functional δ-method we see that

$$\sqrt{m}(\widehat{Q}(t) - Q(t)),$$

where $\widehat{Q}(t) = (1-a)t/\widehat{G}(t)$, will be convergent to a Gaussian process on $(0,1]$ with covariance kernel

$$K_4(s,t) = \frac{Q(s)Q(t)}{G(s)G(t)}(G(s\wedge t) - G(s)G(t) + 2\sum_{k=2}^{+\infty}(P(p_1 < s, p_k < t) - G(s)G(t))),$$

and then also

$$\sqrt{m}(\widehat{Q}^{-1}(u) - Q^{-1}(u)) \tag{A.1}$$

is convergent to a Gaussian process with covariance kernel

$$K_5(s,t) = \frac{K_4(Q^{-1}(s), Q^{-1}(t))}{Q'(Q^{-1}(s))Q'(Q^{-1}(t))}.$$

Since $T_{PI} = \widehat{Q}^{-1}(\alpha)$, by applying the δ-method to (A.1) it can be seen that

$$\sqrt{m}(T_{PI} - Q^{-1}(\alpha))$$

will converge in distribution to a $N(0,V)$, with $V = K_5(Q^{-1}(\alpha), Q^{-1}(\alpha))$; and finally that

$$\sqrt{m}(Q(T_{PI}) - \alpha)$$

will converge in distribution to a $N(0, (Q'(Q^{-1}(\alpha)))^2 V)$. We cannot conclude now that $FDR \to \alpha$, since FDR is the expected value of the stochastic process $\Gamma(t)$ evaluated at the *random* point T_{PI}. We need to make some further considerations: let now $0 < \delta < Q^{-1}(\alpha)$, and note that

$$\begin{aligned}
|\Gamma(T_{PI}) - \alpha| &\leq |\Gamma(T_{PI}) - Q(T_{PI})| + |Q(T_{PI}) - \alpha| \\
&\leq \sup_t |\Gamma(t) - Q(t)| 1_{\{T_{PI} < \delta\}} + \sup_t |\Gamma(t) - Q(t)| 1_{\{T_{PI} \geq \delta\}} + |Q(T_{PI}) - \alpha| \\
&\leq 1_{\{T_{PI} < \delta\}} + \frac{1}{\sqrt{m}} \sup_{t > \delta} |\sqrt{m}(\Gamma(t) - Q(t))| + |Q(T_{PI}) - \alpha| \\
&= O_P(m^{-1/2})
\end{aligned}$$

Since $0 < \Gamma(t) < 1$ for any m, the sequence is uniformly integrable.
Hence, $E[\Gamma(T_{PI})] = \alpha + o(1)$.

Appendix B

Proofs from Chapter 5

(Some Famous Inequalities Under Dependence)

B.1 Hoeffding Inequality Under Dependence

The key part in the proof of Hoeffding Inequality under dependence is to see the thesis of Lemma B.1.1, i.e., that $E(e^{t\sum X_i}) \leq \prod E(e^{tX_i})$ for any $t > 0$. Note that under association the reverse inequality is proved. A possible path is to request negative association, as in Definition 5.2.1.

Lemma B.1.1. *Suppose $X_1, \ldots X_n$ is negatively associated. Then $E(e^{t\sum X_i}) \leq \prod E(e^{tX_i})$ for any $t > 0$.*

Proof. Any non decreasing function of a vector of negatively associated random variables is negatively associated. Hence, $Cov(e^{tX_1}, \prod_{i=2}^{n} e^{tX_i}) \leq 0$, which implies $E(e^{t\sum X_i})) \leq E(e^{tX_1})E(\prod_{i=2}^{n} e^{tX_i})$. Iteration of the argument yields the thesis. □

Theorem B.1.2 (Hoeffding Inequality)**.** *Let $X_1, \ldots, X_n$ be a negatively associated sequence of random variables. Let $a_i < X_i < b_i$, $E(X_i) = 0$. Let $\varepsilon > 0$. Then, for any $t > 0$,*

$$P(\sum X_i \geq \varepsilon) \leq e^{-t\varepsilon} \prod e^{t^2(b_i - a_i)^2/8}$$

Proof. By Markov Theorem and by Lemma B.1.1,

$$\begin{aligned} P(\sum X_i \geq \varepsilon) &= P(t\sum X_i \geq t\varepsilon) \\ &= P(e^{t\sum X_i} \geq e^{t\varepsilon}) \\ &\leq e^{-t\varepsilon} E(e^{t\sum X_i}) \\ &\leq e^{-t\varepsilon} \prod E(e^{tX_i}) \end{aligned}$$

The rest of the proof is analogue to the proof of Hoeffding inequality for independent random variables. □

B.2 Vapnik-Cervonenkis Inequality Under Dependence

We begin by proving an extension of the symmetrization lemma under dependence:

Definition B.2.1 (Separability)**.** *Let $(Y(u), u \in \mathcal{U})$ be a family of random variables on a probability space $(\Omega, \mathcal{F}, P)$. The family is called separable if there exists a countable set $U_0 \subseteq \mathcal{U}$ and a set $E \in \mathcal{F}$ such that*

1. *$P(E) = 1$,*
2. *for any $\omega \in E$ and for any $u \in \mathcal{U}$ there exists a sequence $(u_j, j \geq 1)$ in U_0 such that $Y(u_j, w) \to Y(u, w)$ for $j \to \infty$.*

Lemma B.2.2 (Simmetrization Lemma)**.** *Let $(Y(u), u \in \mathcal{U})$ be a family of separable random variables, and $(Y'(u), u \in \mathcal{U})$ an independent copy of $(Y(u), u \in \mathcal{U})$ with the same joint distribution for any $u_1, \ldots, u_n$ (i.e., with the same dependency structure). Let $P(|Y'(u)| > \varepsilon/2) \leq 1/2$ for any $u \geq \mathcal{U}$. Then:*

$$P(\sup_u |Y(u)| > \varepsilon) \leq 2P(\sup_u |Y(u) - Y'(u)| > \varepsilon/2),$$

for any $\varepsilon > 0$.

Proof. If $(Y(u), u \in \mathcal{U})$ is separable, then also $(Y'(u), u \in \mathcal{U})$ is. Moreover, there exists a countable set $U_0 \subseteq \mathcal{U}$ such that $\sup_{u \in \mathcal{U}} |Y(u)| = \sup_{u \in U_0} |Y(u)|$. Let u_i be the i-th element of U_0. Let $A_1 = \{|Y(u_1)| > \varepsilon\}$, and $A_i = \{Y(u_1)| \leq \varepsilon, \ldots, |Y(u_{i-1})| \leq \varepsilon, |Y(u_i)| > \varepsilon\}$ for $i \geq 2$. Note that if $|Y(u_i)| > \varepsilon$ and $|Y'(u_i)| \leq \varepsilon/2$ then $|Y(u_i) - Y'(u_i)| > \varepsilon/2$. We have:

$$\begin{aligned}
1/2P(\sup_{u \in \mathcal{U}} |Y(u)| > \varepsilon) &= 1/2 \sum_{i \in U_0} P(A_i) \\
&\leq \sum P(A_i) P(|Y'(u_i)| \leq \varepsilon/2) \\
&= \sum P(A_i, |Y'(u_i)| \leq \varepsilon/2) \\
&\leq \sum P(A_i, |Y(u_i) - Y'(u_i)| > \varepsilon/2) \\
&\leq \sum P(A_i, \sup_{u \in U_0} |Y(u) - Y'(u)| > \varepsilon/2) \\
&\leq P(\sup_{u \in \mathcal{U}} |Y(u) - Y'(u)| > \varepsilon/2).
\end{aligned}$$

□

We will now prove a special case of Vapnik-Cervonenkis inequality under dependence. We will restrict to a class of sets such that $P(X_i \in A, X_j \in A) \leq P(X_i \in A)P(X_j \in A)$ for any A in the class. A class of this kind will be for instance the class of sets of the form $(-\infty, z]$, for z real, under negative dependence for the random variables. This class we will consider to derive our DKW-Type inequality.

Lemma B.2.3 (Special Case of Vapnik-Cervonenkis Inequality)**.** *Let* $\mu(A) = Pr(X_i \in A)$ *and* $\mu_n(A) = 1/n \sum 1_{x_i \in A}$*. Let* $\mathcal{A}$ *be a class of sets such that* $P(X_i \in A, X_j \in A) \leq P(X_i \in A)P(X_j \in A)$ *for* $A \in \mathcal{A}$*, and that* $(\mu_n(A), A \in \mathcal{A})$ *is separable. We have that*

$$\Pr\{\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon\} \leq 4 S_{\mathcal{A}}(2n) e^{-n\varepsilon^2/8}$$

for any $\varepsilon > 0$*, where* $S_{\mathcal{A}}(2n)$ *is the shatter coefficient, i.e.,*
$S_{\mathcal{A}}(n) = \max_{x_1,\ldots,x_n \in \mathcal{A}} |\{\{x_1, \ldots, x_n\} \bigcap A; A \in \mathcal{A}\}|$.

Proof. Let $X_1', \ldots, X_n'$ be an independent copy of $X_1, \ldots, X_n$, i.e. a vector of random variables independent from the first one but with the same dependence structure. Let $\sigma_1, \ldots, \sigma_n$ be n sign variables. Let $\mu_n'(A) = 1/n \sum 1_{X_i' \in A}$. Main steps of the proof are as follows:

1. *Symmetrization.* It is easy to see that under the assumptions

$$V[\mu_n(A) - \mu(A)] \leq \frac{\mu(A)(1 - \mu(A))}{n}. \tag{B.1}$$

We have now that $E[\mu_n(A) - \mu(A)] = 0$, which implies $V[\mu_n(A) - \mu(A)] = E[(\mu_n(A) - \mu(A))^2]$. We have:

$$\begin{aligned} E[(\mu_n(A) - \mu(A))^2] &= E[1/n^2 \sum_{ij} 1_{X_i \in A} 1_{X_j \in A}] - \mu^2(A) \\ &= 1/n^2 \sum_{ij} E[1_{X_i \in A} 1_{X_j \in A}] - \mu^2(A) \\ &= 1/n^2 \sum_{ij} P(X_i \in A, X_j \in A) - \mu^2(A) \\ &\leq 1/n^2 \sum_{ij} P(X_i \in A) P(X_j \in A) - \mu^2(A), \end{aligned}$$

where we used inequality (B.3) in the last step. Last expression is easily seen to be equal to $\mu(A)(1 - \mu(A))/n$, as desired.

We can now apply Chebyshev inequality to random variable $(\mu_n(A) - \mu(A))$, together with inequality (B.1), to see:

$$\begin{aligned} P(|\mu_n(A) - \mu(A)| > \frac{\varepsilon}{2}) &\leq \frac{4}{\varepsilon^2} V[(\mu_n(A) - \mu(A))] \\ &\leq \frac{4}{\varepsilon^2} \frac{\mu(A)(1 - \mu(A)}{n} \\ &\leq \frac{1}{n\varepsilon^2} \\ &\leq 1/2 \ \ \forall\ n \geq \frac{2}{\varepsilon^2}. \end{aligned}$$

We can then apply Lemma B.2.2, since we have separability. Hence, for $n \geq 2/\varepsilon^2$,

$$P(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| > \varepsilon) \leq 2P(\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu_n'(A)| > \varepsilon/2). \tag{B.2}$$

2. *Randomization by sign variables.* We have now that

$$P(\sup_{A\in\mathcal{A}}|\mu_n(A)-\mu_n'(A)|>\varepsilon/2)=P(1/n\sup_A|\sum\sigma_i(1_{X_i\in A}-1_{X_i'\in A})|>\varepsilon/2);$$

since $(X_i \perp X_i',\ X_i \stackrel{d}{=} X_i')$; which is the key in proving the fact that

$$[(1_{X_1\in A}-1_{X_1'\in A}),\ldots,(1_{X_n\in A}-1_{X_n'\in A})]\stackrel{d}{=}[\sigma_1(1_{X_1\in A}-1_{X_1'\in A}),\ldots,\sigma_n(1_{X_n\in A}-1_{X_n'\in A})],$$

since the symmetry is not altered by dependence.

More in detail, suppose $n=2$. Let $k_i=\{-1,0,1\}$. Let $Y_i=1_{X_i\in A}-1_{X_i'\in A}$. We have

$$\Pr(\sigma_1Y_1=k_1,\sigma_2Y_2=k_2)=1/4\sum_{i_1=-1,1}\sum_{i_2=-1,1}\Pr(i_1Y_1=k_1,i_2Y_2=k_2|\sigma_1=i_1,\sigma_2=i_2).$$

We have equality in distribution if $\Pr(Y_2=1,Y_1=1)=\Pr(Y_2=-1,Y_1=1)=\Pr(Y_2=-1,Y_1=-1)$, $\Pr(Y_2=0,Y_1=1)=\Pr(Y_2=0,Y_1=-1)$ and $\Pr(Y_2=1,Y_1=0)=\Pr(Y_2=-1,Y_1=0)$.

We will prove only that $\Pr(Y_2=1,Y_1=1)=\Pr(Y_2=-1,Y_1=-1)$. The other equalities follow from the same kind of calculations.

$$\begin{aligned}
\Pr(Y_2=1,Y_1=1) &= \Pr(Y_2=-1,Y_1=-1)\iff\\
\Pr(Y_2=1|Y_1=1) &= \Pr(Y_2=-1|Y_1=-1)\iff\\
\Pr(X_2\in A,X_2'\notin A|X_1\in A,X_1'\notin A) &= \Pr(X_2\notin A,X_2'\in A|X_1\notin A,X_1'\in A)\iff\\
\Pr(X_2\in A|X_1\in A)\Pr(X_2'\notin A|X_1'\notin A) &= \Pr(X_2\notin A|X_1\notin A)\Pr(X_2'\in A|X_1'\in A),
\end{aligned}$$

And the last equality is true since we took $X_1',\ldots,X_n'$ to be a copy of $X_1,\ldots,X_n$ with the same dependency structure.

Now, by induction, this can be extended to arbitrary n.

3. *Conditioning.* We will now condition on X_i and X_i': fix $X_i=x_i$, $X_i'=x_i'$, $i=1,\ldots,n$; and note that the random variable $\sum\sigma_i(1_{X_i\in A}-1_{X_i'\in A})$ is constant on the sets $A\in\mathcal{A}$ having same intersection with $\{x_1,\ldots,x_n,x_1',\ldots,x_n\}$ (i.e., it is now a discrete random variable with a finite number of possible values). Let $\mathcal{A}^*\subseteq\mathcal{A}$ be a finite subclass of $\mathcal{A}$ such that

$$\{A\cap\{x_1,\ldots,x_n,x_1',\ldots,x_n\};A\in\mathcal{A}\}=\{A\cap\{x_1,\ldots,x_n,x_1',\ldots,x_n\};A\in\mathcal{A}^*\},$$

and that as A varies in $\mathcal{A}^*$ the elements $\{A\cap\{x_1,\ldots,x_n,x_1',\ldots,x_n\}\}$ are *distinct*. We clearly have $|\mathcal{A}^*|\le S_{\mathcal{A}}(2n)$ and

$$\sup_{A\in\mathcal{A}}\sum\sigma_i(1_{X_i\in A}-1_{X_i'\in A})=\max_{A\in\mathcal{A}^*}\sum\sigma_i(1_{X_i\in A}-1_{X_i'\in A}).$$

We can now apply Hoeffding inequality to $\sigma_i(1_{X_i \in A} - 1_{X_i' \in A})$, $i = 1, \ldots, n$, since, conditionally on X_i and X_i', these are independent random variables with mean zero and taking values in $[-1, 1]$:

$$P(|1/n \sum_i \sigma_i(1_{X_i \in A} - 1_{X_i' \in A})| > \varepsilon/2 | X_i = x_i, X_i' = x_i', i = 1, \ldots, n) \leq 2e^{-n\frac{\varepsilon^2}{8}} \quad \forall\ A \in \mathcal{A}^*.$$

We can then note that:

$$\begin{aligned}
&P(\max_{A \in \mathcal{A}^*} |1/n \sum \sigma_i(1_{X_i \in A} - 1_{X_i' \in A})| > \varepsilon/2 | X_i = x_i, X_i' = x_i', i = 1, \ldots, n) \quad = \\
&P(\bigcup_{A \in \mathcal{A}^*} |1/n \sum \sigma_i(1_{X_i \in A} - 1_{X_i' \in A})| > \varepsilon/2 | X_i = x_i, X_i' = x_i', i = 1, \ldots, n) \quad \leq \\
&\sum_{A \in \mathcal{A}^*} P(|1/n \sum \sigma_i(1_{X_i \in A} - 1_{X_i' \in A})| > \varepsilon/2 | X_i = x_i, X_i' = x_i', i = 1, \ldots, n) \quad \leq \\
&\qquad\qquad 2S_{\mathcal{A}}(2n)e^{-n\frac{\varepsilon^2}{8}}.
\end{aligned}$$

4. *Marginalization.* Since the last inequality is true for any x_i and x_i', it is true also unconditionally. It is immediate to combine this result with result (B.2) to get the thesis for $n \geq 2/\varepsilon^2$. If $n < 2/\varepsilon^2$ it happens that $4e^{-n\varepsilon^2/8} \geq 4e^{-1/4} > 1$, so the thesis is proved for any ε.

□

Theorem B.2.4 (DKW-Type Inequality)**.** *Let $X_1, \ldots, X_n$ be a sequence of identically distributed negatively associated random variables. Let $F(z)$ be the CDF of X_1, and $\widehat{F}(z)$ the empirical distribution of the sequence $X_1, \ldots, X_n$. Then,*

$$\Pr\{\sup_{z \in \mathcal{R}} |F(z) - \widehat{F}(z)| > \varepsilon\} \leq 4(2n + 1)e^{-n\varepsilon^2/8}.$$

Proof. Let $\mathcal{A}$ be the class of sets of the form $(-\infty, z]$ for z real. Remember in fact that under negative association

$$P(X_i \leq x_i \cap X_j \leq x_j) \leq P(X_i \leq x_i)P(X_j \leq x_j). \tag{B.3}$$

For this reason, we can can simply apply Lemma B.2.3, since moreover the resulting class is separable.

A good upper bound for $S_{\mathcal{A}}(2n)$ is easily seen to be $2n+1$, from which the thesis. □

Note that the result is far from being optimal. Both the exponent $n\varepsilon^2/8$ and more importantly the rate can probably be improved under the same assumptions.

B.3 Vapnik-Cervonenkis Theorem Under Dependence

In this section we will derive another DKW-Type Inequality, by extension of Vapnik-Cervonenkis theorem under dependence.

Lemma B.3.1 (Vapnik-Cervonenkis Theorem)**.** *Suppose $X_1, \dots X_n$ is a vector of negatively associated identically distributed random variables. Let $\mu(A) = Pr(X_i \in A)$ and $\mu_n(A) = 1/n \sum 1_{x_i \in A}$. Let $\mathcal{A}$ be a class of subsets of R^d. We have that*

$$E\{\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\} \leq 24/\sqrt{n} \max_{x_1,\dots,x_n \in R^d} \int_0^1 \sqrt{log2 * N(r, A(x_1^n))}\, dr,$$

where $A(x_1^n) = \{b = (b_1, \dots, b_n) \in \{0,1\}^n : \exists A \in \mathcal{A} : b_i = 1_{[x_i \in A]}\}$ and $N(r, B)$ is the covering number.

Proof. Let $X_1', \dots, X_n'$ be an independent copy of $X_1, \dots, X_n$, i.e. a vector of random variables independent from the first one but with the same dependence structure. Let $\sigma_1, \dots, \sigma_n$ be n sign variables. Let $\mu_n'(A) = 1/n \sum 1_{X_i' \in A}$.

$$\begin{aligned}
E[\sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|] &\leq \\
\text{(Jensen inequality plus law of iterated expectation)} & \\
E[\sup E[|\mu_n(A) - \mu_n'(A)||X_1, \dots, X_n]] &\leq \\
E[\sup |\mu_n(A) - \mu_n'(A)|] &= \\
\text{(see below)} & \\
1/nE[\sup |\sum \sigma_i (1_{X_i \in A} - 1_{X_i' \in A})|] &\leq \\
1/nE[\sup |\sum \sigma_i 1_{X_i \in A}|] + 1/nE[\sup |\sum \sigma_i 1_{X_i' \in A}|] &= \\
2/nE[\sup |\sum \sigma_i 1_{X_i \in A}|] &= \\
2/nE[E[\sup |\sum \sigma_i 1_{X_i \in A}||X_1, \dots, X_n]] &
\end{aligned}$$

The fourth step is true because

$$[(1_{X_1 \in A} - 1_{X_1' \in A}), \dots, (1_{X_n \in A} - 1_{X_n' \in A})] \stackrel{d}{=} [\sigma_1(1_{X_1 \in A} - 1_{X_1' \in A}), \dots, \sigma_n(1_{X_n \in A} - 1_{X_n' \in A})],$$

since the symmetry is not altered by dependence, as we saw in the proof of Lemma B.2.3.

Now fix $X_i = x_i$ and study $E[\sup_{A\in\mathcal{A}} |\sum \sigma_i 1_{x_i \in A}|] = E[\max_{b \in A(x_1^n)} |\sum \sigma_i b_i|]$.

The rest of the proof is analogue to the proof in Devroye and Lugosi (2001), pag. 21. Lemma 2.2 in Devroye and Lugosi (2001) is directly extendable to any kind of dependent random variables since it relies only on Jensen inequality and basic algebra inequalities. □

Theorem B.3.2 (DKW Theorem)**.** *Let $X_1, \ldots, X_n$ be a sequence of identically distributed negatively associated random variables. Let $F(z)$ be the CDF of X_1, and $\widehat{F}(z)$ the empirical distribution of the sequence $X_1, \ldots, X_n$. Then,*

$$E\{\sup_{z\in\mathcal{R}} |F(z) - \widehat{F}(z)|\} \leq \frac{C}{\sqrt{n}},$$

where C is a universal constant less than or equal to $24\sqrt{2\pi}$.

Proof. Let $\mathcal{A}$ be the class of sets of the form $(-\infty, z]$ for z real.

We will simply apply Lemma B.3.1.

A good upper bound for $N(r, \mathcal{A}(x_1^n))$ is easily seen to be $1 + 1/r^2$ for any $r \in (0, 1)$.

Finally, with direct calculation it is seen that $\int_0^{+\infty} \sqrt{log 2N(r, \mathcal{A}(x_1^n))}\, dr \leq \sqrt{2\pi}$; which proves, by an application of Lemma B.3.1, that $E[\sup_{z\in\mathcal{R}} |F(z) - \widehat{F}(z)|] < C/\sqrt{n}$, with $C \leq 24 * \sqrt{2\pi}$. □

We now have an inequality for the expected supremum of the distance between the empirical distribution and the marginal. We want to give an inequality for the tail of this random variable.

A good way to convert inequalities for the expected value to exponential tail inequalities is the following theorem, which we prove under our assumptions for dependence[1]:

Theorem B.3.3 (Bounded Difference Inequality)**.** *Suppose $g(\cdot)$ satisfies the bounded difference assumption:*

$$\sup_{x_1,\ldots,x_n;x_i'\in A} |g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)| \leq c_i, \tag{B.4}$$

for $1 \leq i \leq n$, and any set A. Suppose the assumptions of Theorem B.1.2 are satisfied. Then, for all $t > 0$,

$$\Pr\{|g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)]| \geq t\} \leq 2e^{2t^2/\sum_{i=1}^n c_i^2}.$$

Proof. We will prove that

$$\Pr\{g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] \geq t\} \leq e^{2t^2/\sum_{i=1}^n c_i^2}. \tag{B.5}$$

Similarly it can be proved that

$$\Pr\{E[g(X_1, \ldots, X_n)] - g(X_1, \ldots, X_n) \geq t\} \leq e^{2t^2/\sum_{i=1}^n c_i^2}.$$

Combination of these two results yields the thesis.

We will use Theorem B.1.2 in this straightforward extension: let V and Z be such that $E[V|Z] = 0$, and for some $h(\cdot)$ and $c > 0$,

$$h(Z) \leq V \leq h(Z) + c.$$

[1]The proof for independent random variables was derived in McDiarmid (1989).

Then, for all $s > 0$,

$$E[e^{sV}|Z] \leq e^{s^2c^2/8}. \tag{B.6}$$

Call now $V = g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)]$.
Call $H_i(X_1, \ldots, X_i) = E[g(X_1, \ldots, X_n)|X_1, \ldots, X_i]$, and for any i define

$$V_i = H_i(X_1, \ldots, X_i) - H_{i-1}(X_1, \ldots, X_{i-1}).$$

Clearly, $V = \sum V_i$ and $H_{i-1}(X_1, \ldots, X_{i-1}) = \int H_i(X_1, \ldots, X_{i-1}, x)\ F_i(dx)$. Define moreover

$$W_i = \sup_u H(X_1, \ldots, X_{i-1}, u) - H_{i-1}(X_1, \ldots, X_{i-1}),$$

and

$$Z_i = \inf_u H(X_1, \ldots, X_{i-1}, u) - H_{i-1}(X_1, \ldots, X_{i-1}).$$

Clearly, $Z_i \leq V_i \leq W_i$ with probability one, and

$$W_i - Z_i = \sup_u \sup_v (H(X_1, \ldots, X_{i-1}, u) - H(X_1, \ldots, X_{i-1}, v)) \leq c_i,$$

by the bounded difference assumption. Therefore, by (B.6), for any i:

$$E[e^{sV_i}|X_1, \ldots, X_{i-1}] \leq e^{s^2c_i^2/8}.$$

Finally, by Chernoff bound, for any $s > 0$:

$$\begin{aligned}
\Pr\{g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] \geq t\} &\leq \frac{E[e^{s\sum_{i=1}^{n} V_i}]}{e^{st}} \\
&= \frac{E[e^{s\sum_{i=1}^{n-1} V_i} E[e^{sV_n}|X_1, \ldots, X_{n-1}]]}{e^{st}} \\
&\leq e^{s^2c_n^2/8} \frac{E[e^{s\sum_{i=1}^{n-1} V_i}]}{e^{st}} \\
&\leq e^{-st} e^{s^2 \sum c_i^2/8},
\end{aligned}$$

by repeating the same argument n times. Choosing $s = 4t/\sum c_i^2$ yields inequality (B.5). □

We are now ready to prove our DKW-Type Inequality:

Theorem B.3.4 (DKW-Type Inequality)**.** *Suppose the assumptions of Theorem B.1.2 are satisfied. Then,*

$$\Pr\{\sup_{z \in \mathcal{R}} |F(z) - \widehat{F}(z)| > \varepsilon + \frac{24\sqrt{2\pi}}{\sqrt{n}}\} \leq e^{-2n\varepsilon^2}.$$

Proof. Let $g(X_1, \ldots, X_n) = \sup_{z \in \mathcal{R}} |F(z) - \widehat{F}(z)|$. This function is easily seen to satisfy the bounded difference assumption (B.4): by changing one X_i, $g(\cdot)$ can change by at most $1/n$. This implies $c_i = 1/n$, and $\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} 1/n^2 = 1/n$.

Therefore, by Theorem B.3.3,

$$\Pr\{g(X_1, \ldots, X_n) - E[g(X_1, \ldots, X_n)] \geq \varepsilon\} \leq e^{-2n\varepsilon^2}.$$

By Theorem B.3.2,

$$E[g(X_1, \ldots, X_n)] \leq \frac{24\sqrt{2\pi}}{\sqrt{n}}.$$

These two results are easily combined to see the thesis. □

Appendix C

Some Concepts of Dependence

In this Appendix we will very briefly summarize three concepts of dependence used in the thesis.

C.1 Mixing

Main references on mixing are: Doukan (1994) and Billingsley (1999). Many different notions of mixing (α, β, ρ - mixing) are reviewed in Doukan (1994). We used in this thesis conditions on the α-mixing coefficients, defined in (3.1) as $\alpha(k) = \sup_j\{|P(E_1)P(E_2) - P(E_1 \cap E_2)| : \ E_1 \in \mathcal{M}_1^j, E_2 \in \mathcal{M}_{j+k}^{+\infty}\}$; for a sequence $(X_i)_{i\in\mathcal{N}}$, where $\mathcal{M}_i^j$ is the $\sigma-$algebra generated by the random variables $\{X_i, \ldots, X_j\}$.

Mixing coefficients measure how fast the dependence between the random variables decreases to zero as the *lag* between them increases. Conditions are in the form $\alpha(k) \to 0$, usually at a specific rate, and are mainly used to prove limit theorems (see Arcones and Yu (2000)).

Mixing is a general condition on dependence, often used in time series analysis, which is by the way implied by less general and more easily proved conditions, like m-dependence. A sequence of random variables is m-dependent of order h when X_i is independent of X_{i+h} for any i. For other equivalent conditions, see Bradley (1993).

C.2 Positive/ Negative Association

A sequence of random variables is said to be positively (negatively) associated if for all monotonically coordinate-wise non-decreasing functions g_1 and g_2,

$$Cov[g_1(X_1, \ldots, X_n), g_2(X_1, \ldots, X_n)] \geq (\leq)\ 0,$$

when it exists. Positive association is introduced in Esary *et al.* (1967) and negative association in Kumar and Proschan (1983). Another good reference is Tong (1980), and Block *et al.* (1982) for negative dependence. Main properties of a sequence of

(negatively) associated random variables are: $P(\bigcap_{i=1}^{n}\{X_i \leq z_i\}) \geq (\leq) \prod P(X_i \leq z_i)$ and similarly $P(\bigcap_{i=1}^{n}\{X_i > z_i\}) \geq (\leq) \prod P(X_i > z_i)$, for $z_i \in R$, $i = 1, \ldots, n$. Moreover, it is straightforward to prove that $E[\prod X_i] \geq (\leq) \prod E[X_i]$.

Multivariate normal random variables with all positive (negative) correlations are (negatively) associated. Independent random variables are both positively and negatively associated.

Examples of associated random variables

Multivariate exponential random variables, as defined in Marshall and Olkin (1967), are always associated. A set of random variables $X_1, \ldots, X_m$ is multivariate exponential if

$$\begin{aligned} F(x_1, \ldots, x_m) &= 1 - \exp[-\sum_{1}^{m} \lambda_i x_i - \sum_{i<j} \lambda_{ij} \max(x_i, x_j) + \\ &\quad - \sum_{i<j<k} \lambda_{ijk} \max(x_i, x_j, x_k) - \ldots - \lambda_{12\ldots m} \max(x_1, x_2, \ldots, x_m), \end{aligned}$$

where λ_I, $I \subseteq \{1, \ldots, m\}$ is an opportune parameter. As noted also by Benjamini and Yekutieli (2001), products of independent χ^2 random variables are associated. The most important example in our case is probably the one stating that t statistics arising from two-sided testing are associated if the covariances of the normal random variables in the numerator of the t statistic are all positive (again, see Benjamini and Yekutieli (2001)).

Examples of negatively associated random variables

Multinomial, Multivariate Hypergeometric, Dirichlet random variables are always negatively associated. t statistics are negatively associated when they arise from two-sided testing and the covariances of the normal random variables in the numerator are all negative. For other examples, refer to Kumar and Proschan (1983).

C.3 Block Dependence

In many applications, variables are dependent in blocks; i.e., they are dependent of at most a pre-specified number of r_b random variables, while independent of all the others. Formally, let $X = \{X_{i,b}\}$, $i = 1, \ldots, r_b$; $b = 1, \ldots, k$ be a sequence of random variables such that $X_{i,b}$ is independent of X_{j,b_1} for $b \neq b_1$ and for any i and j.

If, as k goes to infinity, r_b is bounded, then m-dependence is implied.

Bibliography

P. Abrahmsen (1997). A review of gaussian random fields and correlation functions. *Tech. Rep. 917*, Norwegian Computing Center.

F. Abramovich and Y. Benjamini (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis*, **22**, 351–361.

Affymetrix (1999). *Affymetrix Microarray Suite User Guide*. Affymetrix, Santa Clara, CA.

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

D. Amaratunga and J. Cabrera (2004). *Exploration and Analysis of DNA Microarray and Protein Array Data*. Wiley.

M.A. Arcones and B. Yu (2000). Limit theorems for empirical processes under dependence. In: Christian Houdre, ed., *Chaos Expansions, Multiple Wiener-Itô integrals and their applications*.

S. Banerjee, B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical modeling and analysis for spatial data*. Chapman & Hall/CRC.

M.M. Barbieri and J.O. Berger (2004). Optimal predictive model selection. *Annals of Statistics*, **32**, 870–897.

M.J. Bayarri and J.O. Berger (2000). p-values for composite null models. *Journal of the American Statistical Association*, **95**, 1127–1142.

M.J. Bayarri and J.O. Berger (2004). The interplay of Bayesian and frequentist analysis. *Statistical Science*, **19**, 58–80.

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

Y. Benjamini, A.M. Krieger, and D. Yekutieli (2004). Two staged linear step up FDR controlling procedures. *Tech. rep.*, Department of Statistics, Tel Aviv University.

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

Y. Benjamini and D. Yekutieli (2002). Hierarchical FDR testing of trees of hypotheses. *Tech. Rep. 02-02*, Department of Statistics, Tel Aviv University.

D.A. Berry and Y. Hochberg (1999). Bayesian perspectives on multiple comparisons. *Journal of Statistical Planning and Inference*, **82**, 215–277.

D.R. Bickel (2004). On "strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates": does a large number of tests obviate confidence intervals of the FDR? *Tech. rep.*, Office of Biostatistics and Bioinformatics, Medical College of Georgia.

P. Billingsley (1999). *Convergence of Probability Measures*. Wiley.

H.W. Block, T.H. Savits, and M. Shaked (1982). Some concepts of negative dependence. *The Annals of Probability*, **10**, 765–772.

S.R. Bolsover, J. Hyams, S. Jones, E.A. Shepard, and H.A. White (1997). *From Genes to Cells*. Wiley.

M.V. Boutsikas and M.V. Koutras (2000). A bound for the distribution of the sum of discrete associated or negatively associated random variables. *The Annals of Applied Probability*, **10**, 1137–1150.

H. Bovenhuis and R.J. Spelman (2000). Selective genotyping to detect quantitative trait loci for multiple traits in outbred populations. *Journal of Dairy Science*, **83, 1**, 173–180.

R.C. Bradley (1993). Equivalent mixing conditions for random fields. *Annals of Probability*, **21**, 1921–1926.

P.O. Brown and D. Botstein (1999). Exporing the new world of genome with DNA microarrays. *Nature Genetics*, **21**, 33–37.

S. Cabras (2004). *Control of the false discovery rate with frequentist p-values in Microarray data analysis*. Ph.D. thesis, Università degli studi di Firenze.

Louis H.J. Chen (1974). On the convergence of Poisson binomial to Poisson distributions. *The Annals of Probability*, **2**, 178–180.

Louis H.J. Chen (1975). Poisson approximation for dependent trials. *The Annals of Probability*, **3**, 534–545.

H. Chipman, E. Kolaczyk, and R. McCulloch (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, **92**.

K.C. Chou and L.P. Heck (1994). A multiscale stochastic modeling approach to the monitoring of mechanical systems. In: *Int. Symp. Time-Freq. Time-Scale Analysis*. IEEE.

T. Cover and P. Hart (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, **IT-13**, 21–27.

S. M. Crouse, D. R. Nowak, and G. R. Baraniuk (1998). Wavelet-based statistical signal processing using hidden Markov models. *IEEE transactions on signal processing*, **46**, 886–902.

I. Daubechies (1992). *Ten Lectures on Wavelets*. SIAM.

L. Devroye and G. Lugosi (2001). *Combinatorial Methods in Density Estimation*. Springer.

D.L. Donoho and I.M. Johnstone (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**, 425–455.

D.L. Donoho and I.M. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, **90**, 1200–1224.

D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard (1995). Wavelet shrinkage: asymtopia? (with discussion). *Journal of the Royal Statistical Society (Ser. B)*, **57**, 301–369.

P. Doukan (1994). *Mixing*. Lectures Notes in Statistics, 85; Springer-Verlag.

P. Doukhan and S. Louhichi (1999). A new weak dependence condition and applications to moment inequalities. *Stochastic processes and their applications*, **84**, 313–342.

E.I. Drigalenko and R.C. Elston (1997). False discoveries in genome scanning. *Genet. Epidemiol.*, **14**, 779–784.

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

S. Dudoit, M.J. van der Laan, and K.S. Pollard (2003b). Multiple testing. Part I. Single-step procedures for control of general type I error rates. *Tech. Rep. 138*, Division of Biostatistics, UC Berkeley.

D. Duggan, M. Bittner, Y. Chen, P. Meltzer, and J. Trent (1999). Expression profiling usign cDNA microarrays. *Nature Genetics*, **21**, 10–14.

O.J. Dunn (1958). Estimation of the means of dependent variables. *Annals of Mathematical Statistics*, **29**, 1095–1111.

B.P. Durbin and D.M. Rocke (2004). Variance stabilizing transformations for two-color microarrays. *Bionformatics*, **20**, 660–667.

A. Dvoretzky, J.C. Kiefer, and J. Wolfowitz (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, **33**, 642–669.

B. Efron, T. Hastie, I. Johnsone, and R. Tibshirani (2004). Least angle regression. *Annals of Statistics*, **31**, 407–499.

B. Efron, R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, **96**, 1151–1160.

S.P. Ellis, M.D. Underwood, and V. Arango (2000). Mixed models and multiple comparisons in analysis of human neurochemical maps. *Psychiat. Res-Neuroim.*, **9**, 111–119.

J.D. Esary, F. Proschan, and D.W. Walkup (1967). Association of random variables, with applications. *The Annals of Mathematical Statistics*, **38**, 1466–1474.

J. Fan (1994). Test of significance based on wavelet thresholding and Neyman's truncation. *Tech. Rep. 9415*, Institut de Statistique, University Catholique de Louvian.

H. Finner and M. Roters (2002). Multiple hypotheses testing and expected number of type I errors. *Annals of Statistics*, **30**, 220–238.

K.R. Gabriel (1959). The distribution of the number of successes in a sequence of dependent trials. *Biometrika*, **46**, 454–460.

R. H. Garret and C.M. Grisham (2002). *Principles of Byochemistry*. Brooks/Cole.

C.R. Genovese and L. Wasserman (2002). Operating characteristics and extensions of the FDR procedure. *Journal of the Royal Statistical Society (Ser. B)*, **64**, 499–518.

C.R. Genovese and L. Wasserman (2004a). Exceedance control of the false discovery proportion. *Tech. rep.*, Department of Statistics, Carnegie Mellon University.

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

E.I. George (2000). The variable selection problem. *Journal of the American Statistical Association*, **95, 452**, 1304–1308.

E.I. George and D.P. Foster (2000). Calibration and empirical Bayes variable selection. *Biometrika*, **87, 4**, 731–747.

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

P. J. Green and B. W. Silverman (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall.

S.B. Green and M.A. Babyak (1997). Control of type I errors with multiple tests constraints in structural equation modeling. *Multivar. Behav. Res.*, **32**, 39–51.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov (1998). *Wavelets, Approximation, and Statistical Applications*. Springer.

D.W. Heyen, J.I. Weller, and M. Ron (1999). A genome scan for QTL influencing milk production and health traits in dairy cattle. *Physiological Genomics*, **1, 3**, 165–175.

Y. Hochberg and Y. Benjamini (1990). More powerful procedures for multiple significance testing. *Statistics in Medicine*, **9**, 811–818.

Y. Hochberg and A.C. Tamhane (1987). *Multiple Comparisons Procedures*. Wiley.

S. Holm (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

I.A. Ibragimov (1962). Some limit theorems for stationary processes. *Theory Probab. Appl.*, **7**, 349–382.

I.A. Ibragimov (1975). A note on the central limit theorem for dependent random variables. *Theory Probab. Appl.*, **20**, 135–141.

E.H. Ip (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, **66, 1**, 109–132.

M.E. Johnson (1987). *Multivariate Statistical Simulation*. Wiley.

C.G. Khatri (1967). On certain inequalities for normal distributions and their applications to simultaneous confidence bounds. *Annals of Mathematical Statistics*, **38**, 1853–1867.

P. Khatri, M. Babyyak, and N.D. Croughwell (2001). Temperature during coronary artery bypass surgery affects quality of life. *Annals of Thoracical Surgery*, **71, 1**, 110–116.

R. Koenker and G. Bassett (1978). Regression quantiles. *Econometrica*, **46**, 33–50.

J.D. Kumar and F. Proschan (1983). Negative association of random variables with applications. *The Annals of Statistics*, **11**, 286–295.

D.W. Ladd (1975). An alghoritm for the binomial distribution with dependent trials. *Journal of the American Statistical Association*, **70**, 333–340.

L. Le Cam (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.*, **10**, 1181–1197.

N. Lee, Q. Huynh, and S. Schwarz (1996). New methods of linear time-frequency analysis for signal detection. In: *Int. Symp. Time-Freq. Time-Scale Analysis*. IEEE.

E.L. Lehmann (1966). Some concepts of dependence. *The Annals of Mathematical Statistics*, **37**, 1137–1153.

D.J. Lockhart, H. Dong, M.C. Byrne, M.T. Follettie, M.V. Gallo, M.S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E.L. Brown (1996). Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.*, **14**, 1675–1680.

S.G. Mallat (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. Pattn. Anal. Mach. Intell.*, **11**, 674–693.

A.W. Marshall and I. Olkin (1967). A multivariate exponential distribution. *Journal of the American Statistical Association*, **62**, 30–44.

P. Massart (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability*, **18**, 1269–1283.

C. McDiarmid (1989). On the method of bounded differences. In: Cambridge University Press, ed., *Surveys in Combinatorics 1989*, 148–188.

E.P. Merriam, C.R. Genovese, and C.L. Colby (2003). Spatial updating in human parietal cortex. *Neuron*, **39**, 361–373.

M.O. Mosig, E. Lipkin, G. Khutoreskaya, E. Tchourzyna, M. Soller, and A.A. Fridmann (2001). Whole genome scan for quantitative trait loci affecting milk protein percentage in Israeli-Holstein cattle, by means of selective milk DNA pooling in a daughter design, using an adjusted false discovery rate criterion. *Genetics*, **157**, 1683–1698.

G.P. Nason (1996). Wavelet shrinkage using cross-validation. *Journal of the Royal Statistical Society (Ser. B)*, **58**, 463–479.

G.P. Nason and B.W. Silverman (1994). The discrete wavelet transform in S. *J.Comp. Graph. Statist.*, **3**, 163–191.

K.L Neuhaus, R. Von Essen, U. Tebbe, A. Vogt, M. Roth, M. Riess, W. Niederer, F. Forycki, A. Wirtzfeld, W. Mauurer, P. Limbourg, W. Merx, and K. Haerten (1992). Improved thrombolysis in acute myocaridial infarction front-loaded administration of alteplase: results of the rt-PA-APSAC patency study (TAPS). *J.Am.Coll.Card.*, **19**, 885–891.

D.W. Nierenberg, T.A. Stukel, J.A. Baron, B.J. Dain, and E.R. Greenberg (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, **130**, 511–521.

P.A. Nze, P. Buhlmann, and P. Doukhan (2002). Weak dependence beyond mixing and asymptotics for nonparametric regression. *Annals of Statistics*, **30**, 397–430.

T. R. Ogden (1997). *Essential wavelets for statistical applications and data analysis*. Birkhauser.

P. Oliveira and C. Suquet (1995). $L^2(0,1)$ weak convergence of the empirical process for dependent variables. In: A. Antoniadis and G. Oppenheim, eds., *Wavelets and Statistics*, 331–344.

K.J. Ottenbacher (1998). Quantitative evaluation of multiplicity in epidemiology and public health research. *American Journal of Epidemiology*, **147**, 615–619.

A.B. Owen (2004). Variance of the number of false discoveries. *Tech. rep.*, Department of Statistics, Stanford University.

G. Parmigiani, E.S. Garret, R. Irizarry, and S.L. Zeger (2003). *The analysis of gene expression data: methods and software*. Springer.

M. Perone Pacifico, C. Genovese, I. Verdinelli, and L. Wasserman (2003). False discovery control for random fields. *Journal of the American Statistical Association*.

K.S. Pollard and M.J. van der Laan (2003a). A method to identify significant clusters in gene expression data. *Tech. Rep. 107*, Division of Biostatistics, UC Berkeley.

K.S. Pollard and M.J. van der Laan (2003b). Resampling-based multiple testing: Asymptotic control of type I error and applications to gene expression data. *Tech. Rep. 121*, Division of Biostatistics, UC Berkeley.

R. Reiter (2001). *Knowledge in Action: Logical Foundations for describing and implementing dynamical Systems*. Springer-Verlag.

C. P. Robert and G. Casella (1999). *Monte Carlo Statistical Methods*. Springer-Verlag.

R. Rocci and M. Vichi (2004). Multimode partitioning. Submitted.

S.K. Sarkar (2002). Some results on false discovery rate in stepwise multiple testing procedures. *Annals of Statistics*, **30**, 239–257.

C.M. Schaffer and P.E. Green (1998). Cluster-based market segmentation: some further comparisons of alternative approaches. *Journal Market Res. Soc.*, **40**, 155–163.

M. Schena (2000). *Microarray Biochip Technology*. Eaton.

M. Schena, D. Shalon, R.W. Davis, and P.O. Brown (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, **270**, 467–470.

M. Schlaeppi, K. Edwards, and R.W. Fuller (1996). Patient perception of the Diskus inhaler: A comparison with the Turbuhaler inhaler. *British Journal of Clinical Practice*, **50**, 14–19.

J.G. Scott and J.O. Berger (2003). An exploration of aspects of Bayesian multiple testing. *Tech. rep.*, Department of Statistics, Duke University.

T. Sellke, M.J. Bayarri, and J.O. Berger (2001). Calibration of p-values for testing precise null hypotheses. *The American Statistician*, **55**, 62–71.

Q.M. Shao (2000). A comparison theorem on moment inequalities between negatively associated and independent random variables. *Journal of theoretical probability*, **13**, 343–356.

G. R. Shorack and J. A. Wellner (1986). *Empirical Processes With Applications to Statistics*. Wiley.

Z. Sidak (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, **62**, 626–633.

Z. Sidak (1971). On probabilities of rectangles in multivariate Student distributions: their dependence on correlations. *Annals of Mathematical Statistics*, **42**, 169–175.

B.W. Silverman (1985). Some aspects of the spline smoothing approach to non-parametric curve fitting. *Journal of the Royal Statistical Society (Ser. B)*, **47**, 1–52.

R.J. Simes (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, **73**, 751–754.

E. P. Simoncelli and E.H. Adelson (1996). Noise removal via Bayesian wavelet coring. In: *IEEE int. conf. image processing*. ICIP.

R. Smith (2001). *Environmental Statistics*. http://www.math.uio.no/~glad/envnotes.ps.

S.Y.T. Soon (1996). Binomial approximation for dependent indicators. *Statistica Sinica*, **6**, 703–714.

C.M. Stein (1971). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In: *Proc. Sixth Berkeley Symp. Math. Statist. Probab.*, vol. 3, 583–602. University of California.

C.M. Stein (1986). *Approximate computation of expectations*. IMS.

J.D. Storey (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society (Ser. B)*, **64**, 479–498.

J.D. Storey (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, **31**, 2013–2035.

J.D. Storey, J.E. Taylor, and D. Siegmund (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. *Journal of the Royal Statistical Society (Ser. B)*, **66**, 187–205.

J.D. Storey and R. Tibshirani (2001). Estimating false discovery rates under dependence, with applications to DNA microarrays. *Tech. Rep. 2001-28*, Department of Statistics, Stanford University.

G. Strang (1989). Wavelets and dilation equations: a brief introduction. *SIAM rev*, **31**, 614–627.

J.W.H. Swanepoel (1999). The limiting behavior of a modified maximal symmetric $2s-$spacing with applications. *Annals of Statistics*, **27**, 24–35.

R. Tibshirani and E. Bair (2004). Improved detection of differential gene expression through the singular value decomposition. Preprint.

Y. L. Tong (1980). *Probability Inequalities in Multivariate Distributions*. Academic Press.

G. Tseng, M. Oh, L. Rohlin, J. Liao, and W. Wong (2001). Issues on cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research*, **29**, 2549–2557.

M.J. van der Laan and J. Bryan (2000). Gene expression analysis with the parametric bootstrap. *Biostatistics*, **1**, 1–19.

M.J. van der Laan, S. Dudoit, and K.S. Pollard (2003a). Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. *Tech. Rep. 139*, Division of Biostatistics, UC Berkeley.

M.J. van der Laan, S. Dudoit, and K.S. Pollard (2003b). Multiple testing. Part III. Procedures for control of the generalized family-wise error rate and proportion of false positives. *Tech. Rep. 141*, Division of Biostatistics, UC Berkeley.

A. W. van der Vaart and J. A. Wellner (1996). *Weak Convergence and Empirical Processes, with Applications to Statistics*. Springer-Verlag.

A.W. Van der Vaart (1998). *Asymptotic Statistics*. Cambridge University Press.

M. Vannucci and F. Corradi (1999). Covariance structure of wavelet coefficients: theory and models in a Bayesian perspective. *Journal of the Royal Statistical Society (Ser. B)*, **61**, 971–986.

K. Vedantham, A. Brunet, and R. Boyer (2001). Post-traumatic stress disorder, trauma exposure, and the current health of canadian bus drivers. *Canadian Journal of Psychiatrics*, **46, 2**, 149–155.

V.E. Velculescu, L. Zhang, B. Vogelstein, and K.W. Kinzler (1995). Serial analysis of gene expression. *Science*, **270**, 484–487.

X. Wang, S. Ghosh, and S.W. Guo (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research*, **29**.

X. Wang, M.J. Hessner, J. Wu, N. Pati, and S. Ghosh (2003). Quantitative quality control in microarray experiments and the application in data filtering, normalization and false positive rate prediction. *Bionformatics*, **19**, 1341–1347.

J.I. Weller, J.Z. Song, and D.W. Heyen (1998). A new approach to the problem of multiple comparisons in the genetic dissection of complex traits. *Genetics*, **150, 4**, 1699–1706.

P. H. Westfall and S. S. Young (1993). *Resampling-based Multiple Testing: Examples and Methods for p-value Adjustment*. Wiley.

M. Woodroofe and J. Sun (1999). Testing uniformity versus a monotone density. *The Annals of Statistics*, **27**, 338–360.

K.J. Worsley, S. Marrett, P. Neelin, A.C. Vandal, K.J. Friston, and A.C. Evans (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, **4**, 58–73.

Y. H. Yang, M. J. Buckley, S. Dudoit, and T. P. Speed (2002a). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics*, **11**, 108–136.

Y.H. Yang, S. Dudoit, P. Luu, D.M. Lin, V. Peng, J. Ngai, and T.P. Speed (2002b). Normalization for cDNA microarray data: a robust composite method addressig single and multiple slide systematic variation. *Nuclic Acids Research*, **30**.

Y.H. Yang, S. Dudoit, P. Luu, and T. P. Speed (2001). *Normalization for cDNA Microarray Data*. SPIE BiOS 2001.

D. Yekutieli and Y. Benjamini (1999). Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, **82**, 171–196.

K. Yoshihara (1975). Billingsley's theorems on empirical processes of strong mixing sequences. *Yokohama Math. J.*, **23**, 77–83.

H. Yu (1993). A Glivenco-Cantelli lemma and weak convergence for empirical processes of associated sequences. *Probab. Theorey Related Fields*, **95**, 357–370.

G. Zweiger (2001). *Transducing the genome: information, anarchy and revolution in the biomedical sciences*. McGraw-Hill.

Acknowledgements

Rivolgo il ringraziamento piú grande a Fulvio De Santis, che mi ha spinto a continuare a studiare dopo la laurea, sempre attivamente interessandosi ai miei progressi, abbondando di consigli, indicazioni e non ultimo spedendomi negli USA al momento piu' opportuno. Ringrazio Luca Tardella, che ha partecipato con Fulvio al processo di revisione di questo scritto; e lo ringrazio anche per il supporto extra-tesi e l'intensa collaborazione scientifica. Vorrei ringraziare anche Larry Wasserman, che non ha mai smesso di spingermi a cercare risultati nuovi, indicarmi nuovi problemi, suggerirmi possibili strade. Il Capitolo 5 è svolto in collaborazione con lui. Naturalmente, eventuali imprecisioni sono solo mie. Ringrazio Chris Genovese, con cui è in collaborazione il Capitolo 3. Vorrei ringraziare zia Isa Verdinelli, che ha saputo starmi vicino professionalmente e umanamente, salvandomi anche la vita almeno una volta in quel di Pittsburgh. Per non parlare della sua pasta e fagioli. Ringrazio molto il Prof. Conti, che con grande disponibilità ha ricontrollato piú volte le dimostrazioni in Appendice B e ha suggerito possibili miglioramenti. Per lo stesso motivo ringrazio il Prof. Orsingher. Un ringraziamento va al dipartimento, e al coordinatore del dottorato Prof. Coppi, che ha pazientemente fornito strumenti e supporto durante questi tre anni. Ringrazio anche il personale: Cinzia, Adriana, Paolo, Sante, e tutti gli altri. Ringrazio inoltre le persone del dipartimento con cui ho avuto il piacere di interagire professionalmente: il Prof. Giorgi, il Prof. Piccinato, la Prof. Jona-Lasinio, la Prof. Salinetti, Marco Alfó, Marco Perone Pacifico, il Prof. Vichi e il Prof. Dell'Olmo. Ringrazio i colleghi dottorandi, in particolare l'onorevole Sara Antignani, e gli assegnisti, che non hanno mai fatto mancare un pronto aiuto, appassionanti discussioni di Statistica, e una piacevole compagnia per staccare e farsi due risate. Un ringraziamento speciale va a chi si è letto 'sto malloppo, e a chi, come fanno tutti, lo leggerà dopo i ringraziamenti. Ringrazio chi mi sono scordato di ringraziare e se lo sarebbe meritato, ringrazio in ordine sparso gli Iron Maiden, i Modena City Ramblers, Stevie Ray Vaughan, i Children of Bodom per la loro versione di "The Trooper", i Dire Straits, Santana, La Famiglia, i Counting Crows, i System of a Down, i Pearl Jam, Gary Moore, i Led Zeppelin, i 99 Posse, Django Reinhardt, Ben Harper e tutti coloro la cui musica ascoltare o suonare è stato di indispensabile supporto durante la stesura di questo scritto.