N° ordre : 2010-21
N° Série : G-7

THESIS / AGROCAMPUS OUEST

European University of Brittany,

to get the diploma of :

DOCTOR OF THE *INSTITUT SUPERIEUR DES SCIENCES AGRONOMIQUES, AGRO-ALIMENTAIRES, HORTICOLES ET DU PAYSAGE*

Speciality : Applied Mathematics

Doctoral school : MATISSE

presented by :

Chloé FRIGUET

IMPACT OF DEPENDENCE IN LARGE SCALE MULTIPLE TESTING PROCEDURES

Defended on September 2010, 24th

Jury :

Christophe AMBROISE	President
Stéphane ROBIN	Reviewer
John STOREY	Reviewer
David CAUSEUR	Supervisor
Anne SIEGEL	Examiner

CONTENTS

FOREWORD

This document is an English version of my Ph.D. thesis, initially written in French.

This is not a literal translation from the whole original document. Nevertheless, INTRODUCTION, CHAPTER 1, which sums up the state of the art of multiple testing, and the general CONCLUSION of this work are fully translated.

Comments and discussions in CHAPTER 2 and CHAPTER 3 are those of three articles published recently. They have simply been organized as in the French document. Note that these results are the core of my thesis and, of course, the French version has been fully inspired by these three articles so that the two versions of these chapters are quite similar.

CHAPTER 4 introduces on-going work on Factor Analysis, and deals more particularly with the implementation of this method in a high-dimensional setting. The English version of this chapter sums up the main results on this subject whereas the French version gives in addition more bibliographical details.

The proposed method to deal with dependence and high-dimension in multiple testing has been assessed on several real datasets. The French document presents application on two examples (a public dataset and a current study led by the Genomics department, INRA, Rennes). Here, the application of the method on real data is described through the tutorial of the associated `R` package (`FAMT`), which has been recently submitted to the Journal of Statistical Software.

The French version is available on TEL website: http://tel.archives-ouvertes.fr/tel-00539741/fr/

CHAPTER 1

ABSTRACT

Motivated by issues raised by the analysis of gene expressions data, this thesis focuses on the impact of dependence on the properties of multiple testing procedures for high-dimensional data. We propose a methodology based on a Factor Analysis model for the correlation structure. Model parameters are estimated thanks to an EM algorithm and an *ad hoc* methodology allowing to determine the model that fits best the covariance structure is defined.

Moreover, the factor structure provides a general framework to deal with dependence in multiple testing. Two main issues are more particularly considered: the estimation of π_0, the proportion of true null hypotheses, and the control of error rates. The proposed framework leads to less variability in the estimation of both π_0 and the number of false-positives. Consequently, it shows large improvements of power and stability of simultaneous inference with respect to existing multiple testing procedures.

These results are illustrated by real data from microarray experiments and the proposed methodology is implemented in a `R` package called `FAMT`.

Key words *Multiple testing, Dependence, Factor Analysis, Proportion of null hypotheses, FDR,* `R` *package* `FAMT`

INTRODUCTION

Extending the well-assessed theory of hypothesis testing, issues raised by multiple testing, and more generally by simultaneous inference, have been widely discussed in the statistical literature for a long time. Indeed, Fisher firstly proposed in the 1930's testing procedures to test several linear contrasts in analysis of variance. These procedures are deduced from the general tests theory introduced at the beginning of the XX^{th} century by Fisher, Student, Neymann, K. and E. Pearson, itself based on Laplace, Demoivre and Bernoulli works on the Error Theory by the end of the XIX^{th} century (for more historical details, see Salsburg [2002]).

The decision of a statistical test requires to choose between two hypotheses: the null hypothesis (H_0) and the alternative one (H_1). The goal of a test procedure is to control the risk of wrongly reject H_0 (type-I error). Naturally, extending this approach to multiple tests consists initially in controlling the risk of wrongly rejecting H_0 at least once. In the parametric setting, there is a parallel between the univariate test theory and parameter estimation as determining the critical region of the test is similar to determining the tested parameter's confidence interval. A parallel can also be made in the case of multiple tests with the simultaneous confidence interval for the tested parameters.

The univariate approach of test theory focus on optimal procedures. Optimality is achieved when, while controlling the type-I error, power is maximized. When testing several hypotheses at the same time, such definition of optimality does not arise and finding the best multiple testing procedure is still an open question [Shaffer, 1995].

Multiplicity has been an abounding issue and many methods to deal with the number of tests are available. We can quote among them post-hoc tests in ANOVA by Duncan, Dunnett, Scheffé and Tukey. Choosing one method or another depends on the context: in biology when one studies the effet of a treatment on several response variables, in medicine when one studies the effect of a dose of drug at different clinical trial steps, or even in food-industry, when sensory studies are conducted to characterize several products thanks to different descriptors.

Basically, multiple testing procedures rely on the choice of a threshold on the p-values associated to the individual tests, differences between them being due to the way of finding the threshold. Comparative studies focus on the control of the type-I error, determined at the level of the whole set of tests, namely the risk to wrongly reject H_0 at least once (called the Family-Wise Error Rate, FWER). Controlling type-I error is of most importance in these contexts where a moderate number of tests are simultaneously performed, and power issue is often set apart. procedures such as Bonferroni [1936] or Sidak [1967] that control the FWER become highly conservative as the number of tests increases. As a matter of fact, the number of truly rejected null hypotheses is very low.

Besides, the assumption of independence on which most of these procedures are based is discussed very seldom in the literature. In post-hoc tests, dependence is derived from the experimental design and from the tested contrasts. It can be argued that a good design can limit and balance the effect of dependence. General approaches to take into account dependence are often computationally expensive, when the number of tested hypotheses increases.

Large scale multiple testing For the last two decades, innovative improvements have been made to face new scientific challenges. Particularly, high throughput technologies result in huge volume of data that allows the global analysis of complex systems. We can explore for example brain activity thanks to functional Magnetic Resonance Imaging (fMRI), clusters of elementary particles thanks to imaging in astrophysics, the capital market through commercial flows in finance, or the genome thanks to functional genomics in life sciences.

Understand, analyse and predict the working of such complex systems require to take into account both the heterogeneity and the dimension of the data. In most situations, the number of measured variables is close to several thousands, whereas the sample size is about some tens at most. Data are then said in high-dimension.

At the root of the main issue of this thesis lies questions raised by the analysis of DNA microarray data. DNA microarrays are a biotechnology that allows the simultaneous measurement of gene expressions, at the level of the whole genome. Such data can be used, for example, to diagnose tumors, to profile drug-effect, or to group genes with similar expression patterns associated to common biological processes. This biological context has markedly contributed to the development of the statistical methodology for multiple testing in high dimensional data [Efron et al., 2001, Storey, 2002, Dudoit et al., 2003]. Indeed, an important question in microarray experiments is the identification of differentially expressed genes i.e genes whose expression levels differ with respect to a covariate of interest, that can be either categorical, such as treatment/control status, or continuous such as a drug dose. The biological question of differential analysis is then restated as a multiple hypotheses testing issue, considering the simultaneous tests for each gene of the null hypothesis: H_0 : "*there is no association between the expression levels and the covariate*".

Contexts evoked previously induce thousands of simultaneous tests. Procedures controlling the FWER have appeared unsuitable, as they lead to conservative decision in a high-dimensional setting. An approach that has turned out to be more appropriate in high dimension is to control the False

Discovery Rate (FDR) [Benjamini and Hochberg, 1995], which is the expected proportion of false positives among the rejected hypotheses. This approach is useful in exploratory analyses, where one aims at maximizing the discoveries of true positives, rather than guarding against one or more false positives. Many methods have been proposed to control the FDR, the most famous being due to Benjamini and Hochberg [1995] and called hereafter the BH procedure.

Most of such procedures assume that the p-values are independently distributed according to a two-component mixture model [Efron et al., 2001], characterising p-values distributions under the true null hypothesis and under the alternative one, respectively. The mixing parameter of this model, denoted π_0, is defined as the fraction of null hypotheses among the tests. In addition to being an interesting quantity in itself, for its biological interpretation, π_0 is a key parameter in assessing or controlling error rates. Black [2004] or more recently Kim and Van de Wiel [2008] showed that a more accurate estimation of π_0 would improve the power of multiple testing procedures. Adaptative procedures, including π_0 estimation, has then emerged in the literature [Storey et al., 2004, Benjamini et al., 2006].

Multiple testing and dependence Among the topics that are recently discussed in the literature on multiple testing in large-scale data, the impact of dependence between the variables has elicited an increasing interest. Indeed, dependence between tests in directly deduced from dependence between the involved response variables. The true signal and several confusing factors are often observed at the same time. These factors lead to misleading conclusion on tests decisions. In microarray data analysis, dependence between gene expression may comes from some biological gene interactions, in which the studied biological process is not necessarily involved but which impact the level of gene expressions as well. Technological bias can also affect gene expressions, even if some pre-processing treatments of the data such as normalisation aim at limiting their impacts. Dependence is therefore complex and hard to model but its impact on procedures properties is far from negligible.

Methodology for multiple testing in high dimension under dependence Many papers have especially focused on the control of the FDR under various patterns of dependence between test statistics. An important contribution to this point was given by Benjamini and Yekutieli [2001]. They showed the BH procedure still controls the FDR under assumption of a certain class of dependence called positive. Extending the initial condition of the BH procedure was also the point of view of Storey et al. [2004] or Blanchard and Roquain [2008]. Some authors [Storey et al., 2007] proposed to modify the test statistic and recent proposals also suggest to modify the theorical null distribution [Efron, 2004]. In fact, the general message seems to be that, for a high amount of dependence, the BH thresholding method tends to over-control the FDR, leading to more conservative rules than expected under the assumption of independence. Consequently, this also means that dependence affects the power of the BH procedure and its stability.

Taking into account dependence casts doubt on multiple testing procedures as a whole. The models that link each response variable and the covariate are not independent, but independent conditionally

to the factors of heterogeneity. Thus, a common idea in many recent papers is that dependence between test statistics should be taken into account by borrowing information across the variables rather than treating them as independent [Leek and Storey, 2008]. This can be achieved by modeling the common information and taking advantage of the information shared between variables.

The results presented in this thesis are in the continuation of this idea, and we focus on the properties of multiple testing procedures under dependence. More particularly, we propose to stabilize multiple testing procedures considering a Factor Analysis model for the dependence structure. Factor Analysis (FA) is an analytic tool used for many years in economics, social sciences and psychometrics, originally in the field of intelligence research [Spearman, 1904], and has only appeared recently in the study of the dependence structure in high dimensional datasets provided by microarray technology [Pournara and Wernisch, 2007, Kustra et al., 2006]. It describes the covariance relations between observed variables in terms of a meaningful small set of latent variables, called "common factors". FA model resembles a latent variables model, so an EM algorithm is used to estimates its parameters [Rubin and Thayer, 1982].

The main goal of this thesis is to study the statistical properties of simultaneous inference under dependence. First of all, we focus on the impact of dependence on the control of error rates and on the power of procedures. Then, this study leads to th development of an adaptative strategy to deal with dependence in multiple testing.

CHAPTER 1 sums up the state of the art of multiple testing in high-dimension. It essentially describes the underlying hypotheses of usual multiple testing procedures, introducing error rates, decision rules and the estimation of a key parameter of most procedures, the proportion of true null hypotheses.

In CHAPTER 2, a general framework to take into account dependence in multiple testing is presented. In this framework, we study the impact of a deviation to the assumption of independence. The impact on tests statistics and p-values distributions, π_0 estimation and error rates control are successively considered. The main result is that the proposed framework allows to derive the exact expression of π_0 variance as well as the number of false-positives variance, under general dependence.

CHAPTER 3 introduce Factor Analysis as a model for dependence in multiple testing. More precisely, we propose an approach based on conditionally independent test statistics to reduce the impact of dependence. The proposed procedure is called FAMT, for Factor Analysis for Multiple Testing. More over, conditional estimators of π_0 and FDR are proposed.

CHAPTER 4 focuses on the estimation of the model parameters. A Maximum Likelihood estimation is proposed, based on an EM algorithm. The choice of the number of factors to extract is studied, and method to is proposed to determine the optimal number.

Finally, the proposed method is illustrated in CHAPTER 5 thanks to an application to gene expressions data. This chapter is both a case study and a tutorial for the `R` package that implements FAMT. The aim of the case study is to show the improvement brought by our approach in a larger biological study, such as factor interpretation, QTL identification or even gene networks inference. The examples studied are different on the French and in the English version. Here, this chapter is the article submitted to the *Journal of Statistical Software* presenting the `R` package `FAMT`.

CHAPTER 2

LARGE-SCALE MULTIPLE TESTING

Abstract High-throughput experiments have markedly contributed to the development of the statistical methodology for multiple testing in high-dimensional data. First of all, this chapter presents the classical framework of large-scale multiple testing, in particular usual hypotheses on p-values distribution. We put the emphasis on essential concepts of multiple testing procedures, such as error rates and the true null hypotheses proportion.

Sommaire

Introduction: statistical context

Statistical modeling For $k = 1, \ldots, m$, let Y_k denotes the kth response variable among m. In high dimensional frameworks, m can be much larger than the number n of independent observations of $Y = [Y_1, Y_2, \ldots, Y_m]$. For each response Y_k, the link with p explanatory variables is explicitly defined by the following regression model:

$$Y_k = m_k(x) + \epsilon_k \quad \forall k \in [1; m] \equiv \mathcal{M} \tag{2.1}$$

where x is the $p-$vector of covariates, m_k is an unspecified regression function and ϵ_k is a random error term with density function φ_k. Furthermore, it is assumed that the density functions φ_k are the same up to a scaling factor σ_k: $\forall k, \varphi_k(\epsilon) = \varphi(\epsilon/\sigma_k)/\sigma_k$ where φ is the common standardized density function with mean 0 and standard deviation 1. In practice, this means that the response variables have homogeneous distributions.

Multiple testing For $k \in \mathcal{M}_0 \subset \mathcal{M}$ with $\#\mathcal{M}_0 = m_0$, $m_k(x) = m_k^{(0)}(x)$, where $m_k^{(0)}$ is an arbitrary function of interest and for $k \notin \mathcal{M}_0$, $m_k(x) \neq m_k^{(0)}(x)$. Multiple testing aims at finding out the response variables for which $\mathrm{H}_0^k : \ m_k(x) = m_k^{(0)}(x)$ is not true.

The test statistic is denoted $T_k = s_k(Y_k)$. Under the true null hypothesis, its distribution $F_0^k(T)$ is known and we define the p-value for each test:

$$p_k \ = \ 1 - F_0^k(T_k) \tag{2.2}$$

Multiple hypotheses testing issues consider the simultaneous tests of the null hypotheses H_0^k, one test for each response variable. Most procedures can be split into two steps:

1. Computing a test statistic for each response variable, and deduce the associated p-value
2. Applying a thresholding procedure on the p-values of the individual tests to determine which null hypotheses have to be rejected

In the first step, the choice of an appropriate test statistic only depends on the experimental design and the type of response and covariate. We consider that the test statistic is correctly chosen with respect to the statistical context. The second step is the main concern of the following as the threshold on p-values can not be determined as in the univariate issue [Dudoit et al., 2002]. More particularly, the choice of the threshold influence the number of errors in tests decisions.

	declared non significant	declared significant	Total
H_0	U_t	V_t	m_0
H_1	T_t	S_t	m_1
Total	$m - R_t$	R_t	m

Table 2.1: Numbers of errors in a multiple testing procedure

For a given t, the number of possible errors in a multiple testing procedure are summarized in TABLE 2.1, with the same notations as in Benjamini and Hochberg [1995].

m is the known number of tested hypotheses. m_0 and m_1, respectively the number of true null and true alternative hypotheses, are unknown parameters. For a given threshold t for the p-values, $R_t = \sum_{k \in \mathcal{M}} \mathbb{1}_{p_k \leq t}$, the total number of significant tests, is an observed random variable. On the contrary, U_t and S_t on the one hand, and T_t and V_t on the other hand, respectively the number of right and wrong decisions, are unobserved random variables.

Ideally, a multiple testing procedure would minimize both the number V_t of false positives (type-I errors) and the number T_t of false negatives (type-II errors). A standard approach in practice in the univariate setting is to minimize the type-II error rate, that is to say maximize power, for a given acceptable level α for the type-I error rate. More false positives can occur when the number of tests increase. Multiplicity necessitate to clearly define global type-I error rates, at the level of the whole set of tests instead of the level of individual tests. More over, statistical significance is more complex to manage in the multiple testing setting and dedicated procedures are settled to deal with multiplicity.

Many recent articles and books describe the general framework of multiple testing [Efron et al., 2001, Storey, 2003, Dudoit et al., 2002, Dudoit and VanDerLaan, 2008]. This usual setting for multiple testing is defined in the first section of this chapter. Then, essential concepts are introduced, such as the true null hypotheses proportion and error rates. Finally, major multiple testing procedures are presented.

2.1. A mixture model for p-value density

2.1.1 General framework

Under the assumption of identically distributed standardized error terms ϵ_k/σ_k, the test statistics marginal distributions $F^k(T)$ only differ by the scaled regression function $\tau_k = m_k(x)/\sigma_k$. In the following, $G(t) = \mathbb{P}(p_k \leq t)$ stands for the probability distribution function of p_k.

Most of existing methodological development rely on the following assumption:

Hypothèse 1. (P-values distribution)

1. *$\forall k$, $F(T; m_k^{(0)}(x)/\sigma) = F_0(T)$, so that the null distribution is the same for all tests. The p-values are now defined as $p_k = 1 - F_0(T_k)$. For $k \in \mathcal{M}_0$, p_k is therefore distributed according to the uniform distribution: $\forall k \in \mathcal{M}_0, G_0^k(t) = G_0(t) = \begin{cases} t \text{ if } p \in [0;1] \\ 0 \text{ otherwise} \end{cases}$*
2. *Under H_1, p-values are identically distributed: $\forall k \in \mathcal{M}_1, G_1^k(t) = G_1(t)$. The non-centrality parameter is then assumed to be the same: $\forall k \in \mathcal{M}_1\ \tau_k = \tau$.*

This leads to the following two-component mixture model [Efron et al., 2001, Storey, 2002]:

$$G(t) \quad = \quad \pi_0 G_0(t) + (1 - \pi_0) G_1(t) \tag{2.3}$$

where $\pi_0 = \frac{m_0}{m}$ is the unknown proportion of true null hypotheses (see SECTION 2.2 for π_0 estimation). Let's g denotes the p-values density.

$$g(p) = \pi_0 g_0(p) + (1 - \pi_0) g_1(p) \tag{2.4}$$

where $g_0(p) = 1, \forall p$. Further conditions are necessary to ensure identifiability of π_0, which can be obtained for instance by assuming that g_1 is a decreasing function of the p-values with $g_1(1) = 0$ [Genovese and Wasserman, 2002].

2.1.2 Linear model

The linear model, with $m_k(x) = x\theta_k$, is a very usual special case of model (2.1) for which $\varphi_k(\epsilon) \equiv \mathcal{N}(0; \sigma_k^2)$. $m_k^{(0)}$ is obtained by restrictions on m_k such as $c'\theta_k = 0$, where the $p-$vector c defines a given linear contrast. Testing H_0^k relies on the following t-test statistic:

Définition 2.1.1 (Student test statistic)**.**

$$T_k \quad = \quad \frac{c'\hat{\theta}_k}{\sqrt{\hat{\mathbb{V}}(c'\hat{\theta}_k)}} = \frac{c'\hat{\theta}_k}{\sqrt{\hat{\sigma}_k^2 n^{-1} c' S_x^{-1} c}} \sim \mathcal{T}_{\tau_k}(df = n - 2)$$

where S_x is the empirical covariance matrix of explanatory variables, and τ_k is the non-centrality parameter of the Student distribution with df degrees of freedom.

Note that $\tau_k = 0$ if $k \in \mathcal{M}_0$ and $\tau_k = \tau \neq 0$, if $k \in \mathcal{M}_1$.

Exemple 1. We first consider $m = 500$ independent variables and $n = 60$ observations such as $Y_{n\times m} \sim \mathcal{N}_m(\mu; \mathbb{I}_m)$. The multiple testing procedure aims at finding out which among these m

variables have different expectations in two groups with equal sample size $n = 30$. For $m_1 = 100$ variables having different expectations in each group, the difference δ is chosen so that the usual t-tests have a variable-by-variable power of $0,8$ and for the remaining $m_0 = 400$ variables, the difference is set to 0. The non-centrality parameter is here defined by $\tau = \delta/\sqrt{2/n}$.

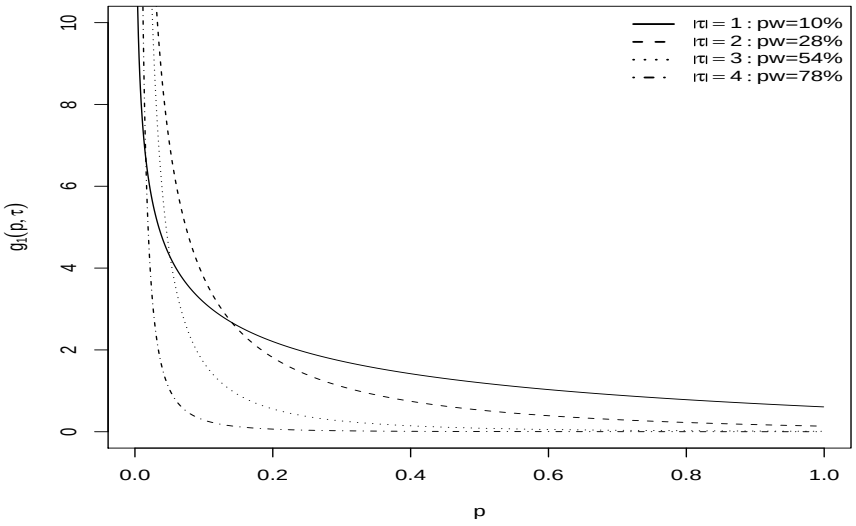


Figure 2.1: Graphical representation of g_1 for different values for τ - pw: individual power of t-test

ϕ_0 and ϕ_τ denote Student density ($df = n - 2$) with non-centrality parameters 0 and τ respectively. If $q_a = \phi_0^{-1}(a)$, $g_1(p) = \frac{\phi_\tau(q_{1-p/2}) + \phi_\tau(q_{p/2})}{2 \times \phi_0(q_{1-p/2})}$ (see FIGURE 2.1). If the individual test power is high, then $g_1(1)$ is close to 0. On the contrary, under the true null and the alternative hypotheses, distributions are not well separated, which can induce problems for the identification of the mixture components (see SECTION 2.2).

2.1.3 Semi-parametric approach

EXEMPLE 2. ASSUMPTION 1 is now illustrated using real microarray data which were primarily analyzed by Golub Golub et al. [1999] in order to identify genes that are differentially expressed in patients with two types of leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). The data are summarized by a 38 matrix $Y = [Y_{ij}]$, where Y_{ij} denotes the expression level for gene j in tumor sample i. The dataset comprises $n = 38$ samples, 27 ALL cases and 11 AML cases, and $m = 3051$ gene expressions. Preprocessing steps were applied to raw data (available on the website `http://www.broadinstitute.org/cgi-bin/cancer/datasets.cgi`). Normalized data are then available from the `R` package `multtest` [Pollard et al.]. A t-test is performed for each gene expression.

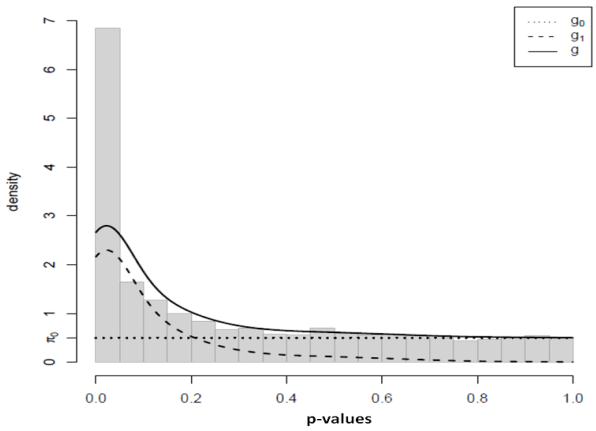


Figure 2.2: P-values distribution (Golub data): estimated density and identification of mixture components

FIGURE 2.2 shows the decomposition of the p-values' distribution into the two-component mixture model introduced in (2.4). The semi-parametric approach implemented in the `R` package `kerfdr` [Guedj et al., 2007] is used. In model (2.4), we need to estimate π_0 and g_1. Many methods can achieve π_0 estimation and are detailed in the following section. In this example, its estimation with `kerfdr`'s dedicated function is $0,499$. We consider a kernel approach for g_1 estimation [Silverman, 1986, Robin et al., 2007]:

$$\hat{g}_1(p) = \frac{1}{m_1 \omega} \sum_{k \in \mathcal{M}_1} K\left(\frac{p - p_k}{\omega}\right) \tag{2.5}$$

$\mathcal{M}_1$ is unknown. A solution is to weight each observation with its posterior probability: $\eta(p_k) = \frac{(1-\pi_0) g_1(p_k)}{g(p_k)}$ [Efron et al., 2001]. The estimation of the p-values density under the alternative hypothesis is:

$$\hat{g}_1(p) = \frac{1}{\omega \sum_{k \in \mathcal{M}} \eta(p_k)} \sum_{k \in \mathcal{M}} \eta(p_k) K\left(\frac{p - p_k}{\omega}\right) \tag{2.6}$$

And then one iterates g_1 estimation steps and η update steps [Robin et al., 2007]. For the choice of the bandwidth ω, see for example Silverman [1986].

2.2. Estimation of the proportion of true null hypotheses

In addition to being an interesting quantity in itself, for its interpretation in the studied context, π_0 is a key parameter in assessing or controlling error rates.

Various methods for π_0 estimation have been developed in the literature. Most of them rely on the assumption of independent p-values distributed according to the two-component mixture model (2.4), with a uniform distribution for null p-values, and taking advantage of the dominance of the null component $\pi_0 g_0$ of g for large p-values.

2.2.1 Empirical estimator

The main approach, initially due to Schweder and Spjotvoll [1982], consists in estimating π_0 by the density of p-values exceeding a tuning parameter λ:

DÉFINITION 2.2.1 (empirical estimator of π_0 [Schweder and Spjotvoll, 1982]).

$$\hat{\pi}_0(\lambda) = \frac{\#(p_k > \lambda)}{(1-\lambda)m} = \frac{W_\lambda}{(1-\lambda)m}$$

where W_λ denotes the number of p-values larger than λ. Note that W_λ can be decomposed into the sum of two independent binomial variables: $W_\lambda = \sum_{k=1}^{m} \mathbb{1}_{p_k>\lambda} = \sum_{k\in\mathcal{M}_0} \mathbb{1}_{p_k>\lambda} + \sum_{k\notin\mathcal{M}_0} \mathbb{1}_{p_k>\lambda}$. Under ASSUMPTION (1), $U_\lambda = \sum_{k\in\mathcal{M}_0} \mathbb{1}_{p_k>\lambda} \sim \mathcal{B}in(m_0, 1-\lambda)$ and $T_\lambda = \sum_{k\notin\mathcal{M}_0} \mathbb{1}_{p_k>\lambda} \sim \mathcal{B}in(m - m_0, 1 - G_1(\lambda))$. The expectation and variance of $\hat{\pi}_0(\lambda)$ under assumption of independent p-values are deduced:

PROPOSITION 2.2.1.

$$\mathbb{E}\left(\hat{\pi}_0(\lambda)\right) = \pi_0 + \frac{1-G_1(\lambda)}{1-\lambda}(1-\pi_0) \tag{2.7}$$

$$\mathbb{V}\left(\hat{\pi}_0(\lambda)\right) = \frac{\lambda\pi_0}{m(1-\lambda)} + \frac{G_1(\lambda)(1-G_1(\lambda))(1-\pi_0)}{m(1-\lambda)^2} \tag{2.8}$$

It follows from the above expression of the bias of $\hat{\pi}_0(\lambda)$ that dominance of the null component in the mixture of distributions should reasonably be specified by assuming that $[1-G_1(\lambda)]/(1-\lambda)$ is a decreasing function of λ. In this case, the minimum bias is $(1-\pi_0)g_1(1)$, which is obtained for $\lambda = 1$. As illustrated by FIGURE 2.3, it can be checked in the special case of t-tests (see EXAMPLE 2) that $(1-G_1(\lambda))/(1-\lambda)$ is actually a positive decreasing function of λ with a lower bound at $\lambda = 1$ given by $\phi_\tau(0)/\phi_0(0)$. However, for small values of $|\tau|$, the minimum bias can be non-negligible, even for large values of λ: $\mathbb{E}(\hat{\pi}_0(\lambda)) - \pi_0 \geq (1-\pi_0)\phi_\tau(0)/\phi_0(0) \geq 0$.

On the contrary, variance of $\hat{\pi}_0$ increases when λ tend to 1.

Choice of λ A relevant choice of the tuning parameter λ should result from a bias-variance trade-off for $\hat{\pi}_0(\lambda)$. Several techniques have been proposed to achieve a good compromise between bias and variance [Langaas et al., 2005]. We only mention here the minimization of a bootstrap estimation of the Mean Square Error [Storey et al., 2004] which is one of the most used in practice.

Smoothing method Note that $\hat{\pi}_0(t)$ can also be expressed as follows: $\hat{\pi}_0(\lambda) = (1-\hat{G}(\lambda))/(1-\lambda)$, where $\hat{G}$ is the empirical estimate of the probability distribution function G of the p-values. It is deduced from the previous expression that, for λ close to 1, $\hat{\pi}_0(\lambda)$ can be approximated by $\hat{g}(1)$, where $\hat{g}$ is a consistent estimate of the density function g. This motivates the estimation of π_0, using smoothing techniques, by the limiting value of $\hat{\pi}_0(\lambda)$ for $\lambda = 1$ [Storey and Tibshirani, 2003].

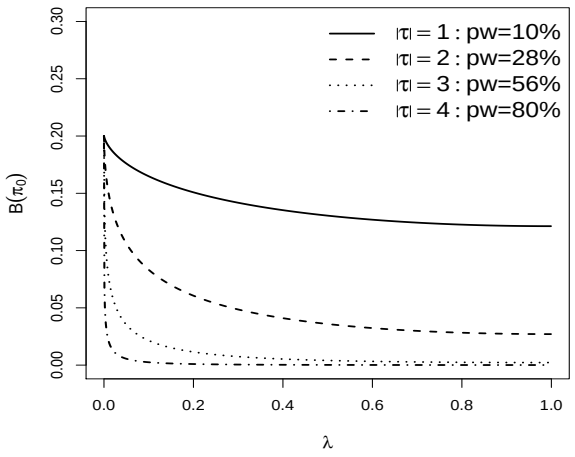


Figure 2.3: Bias of $\hat{\pi}_0$ with respect to the tuning parameter λ, considering different levels of power (pw) for individual tests (non-centrality parameter $|\tau|$)

2.2.2 Density-based estimators

Furthermore, alternative estimators of the form $\hat{\pi}_0 = \hat{g}(1)$ have recently been proposed. Estimation procedures essentially differ by the choice of a non-parametric estimate for g, depending on the underlying assumptions on the regularity and shape of g. We mention here kernel methods (see example 2) and an algorithm dedicated to the estimation of a decreasing convex function [Langaas et al., 2005]

A comprehensive comparative study provided by Langaas et al. [2005] concludes to a better robustness of this kind of density-based estimates to departures from independence of the p-values. Nevertheless, the empirical estimator with a bootstrap approach to choose the tuning parameter λ is widely used in practice.

2.3. Error rates

When testing a single hypothesis, the probability of a type-I error is usually controlled at some designated level α. This can be achieved by choosing a critical value t for the p-values such that $\mathbb{P}(p_k < t|H_0) < \alpha$. The multiple testing framework is concerned with several tests simultaneously and statistical significance is more complex to define. A variety of generalizations to the multiple testing situation are possible for error rates: the type-I and type-II error rates described in this section are the most standard [Dudoit and VanDerLaan, 2008].

2.3.1 Type-I error rates

Two error rates associated with false rejections of null hypotheses (type-I errors) are manly considered:

- The family-wise error rate (FWER): $FWER_t = \mathbb{P}(V_t \geq 1) = 1 - \mathbb{P}(V_t = 0)$
- The false discovery rate (FDR): $FDR_t = \mathbb{E}(FDP_t | R_t > 0).\mathbb{P}(R_t > 0)$, where $FDP_t = \frac{V_t}{R_t}$.

The FWER of a multiple testing procedure is the probability of falsely rejecting at least one true null hypothesis and the FDR is the expected ratio of the number of erroneously rejected null hypotheses to the total number of rejected null hypotheses.

When the number of tests m is large, $\mathbb{P}(R_t > 0) \overset{m \to \infty}{\to} 1$. Storey et al. [2004] therefore define the *positive*-FDR by:

$$pFDR = \mathbb{E}(\frac{V_t}{R_t} | R_t > 0) = \frac{FDR_t}{\mathbb{P}(R_t > 0)}$$

The Family Wise Error Rate (FWER) is historically the controlled error rate (see INTRODUCTION).

However, these procedures may be very conservative, mainly when the number of hypotheses is very large. An approach that has turned out to be more appropriate in high dimension is to control the False Discovery Rate (FDR). This approach is useful in exploratory analyses, when one aims at maximizing the discoveries of true positives, rather than guarding against one or more false positives.

2.3.2 Type-II error rates

Two error rates associated with false non-rejections of null hypotheses (type-II errors) can be considered:

- Expectation of the proportion of false-negatives $FNP_t = \frac{T_t}{m_1}$: $NDR_t = \mathbb{E}(FNP_t)$
- Probability that at least one true-positive occurs: $\mathbb{P}(S_t \geq 1) = \mathbb{P}(T_t \leq m_1 - 1)$

2.4. Multiple testing procedures

Multiple testing procedures account for the multiplicity of the tests by a relevant thresholding technique: t is chosen so that all the null hypotheses H_0^k for which $p_k \leq t$ are rejected, where p_k stands for the p-value of the k^{th} test. The threshold t can be fixed or data-dependent. A multiple testing procedure is said to control a particular type-I error rate at level α, if this error rate is less than or equal to α when the procedure is applied. Strong control refers to control of the type-I error rate under any combination of true and false null hypotheses, and weak control refers to control of the type-I error rate only when all the null hypotheses are true ($\pi_0 = 1$).

In general, the complete null hypothesis is not realistic and weak control is unsatisfactory. In many realm of applications, not all null hypotheses may be true, but the subset $\mathcal{M}_0$ is unknown. Strong control ensures that the type-I error rate is controlled in this case. In practice then, where it is very unlikely that no test is positive, it seems particularly important to consider strong control of the type-I error rate.

A detailed review of multiple testing procedure is available in Dudoit and VanDerLaan [2008]. The following section present the procedures mostly used in practice and usually implemented in the leading statistical software. The different steps of multiple testing procedures are simplified on FIGURE 2.4.

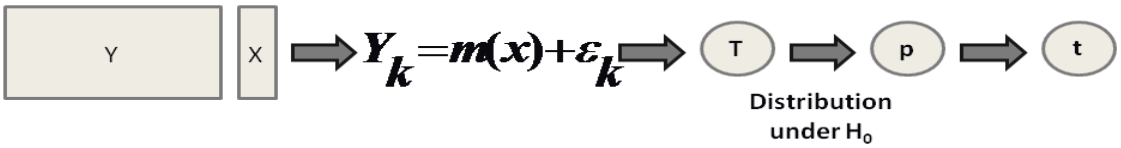


Figure 2.4: Multiple testing procedures steps

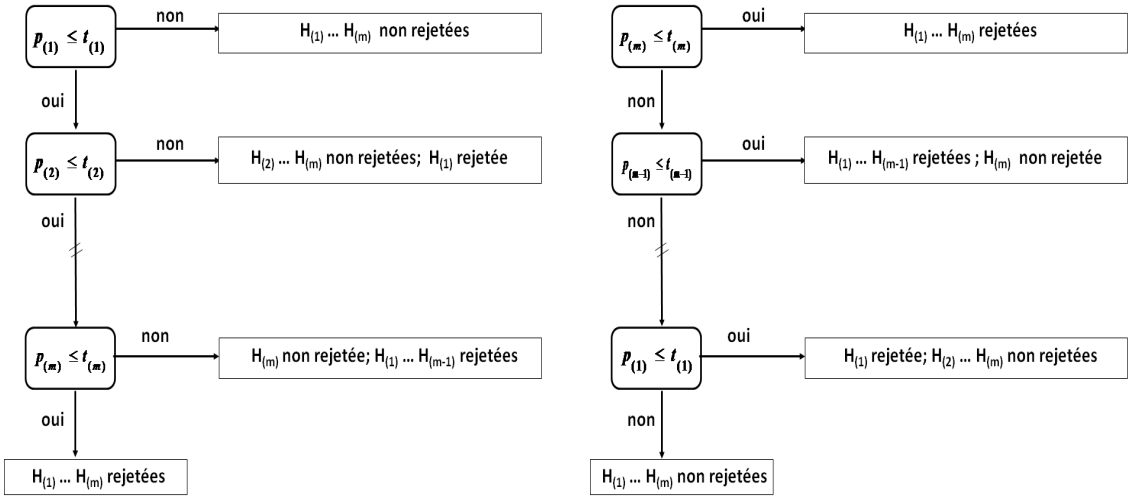
2.4.1 Definitions and principle

Usually, one distinguishes among two types of multiple testing procedures:

- **Single-step procedures** Equivalent multiplicity adjustments are performed for all hypotheses, regardless of the ordering of the p-values. Each hypothesis is evaluated considering a threshold t for the p-values that is independent of the results of other tests.
- **Step-wise procedures** Hypotheses that correspond to the most significant p-values (resp. least significant) are considered successively in step-down (resp. step-up) procedures, with further tests dependent on the outcomes of earlier ones. In step-down procedures, as soon as one fails to reject a null hypothesis, no further hypotheses are rejected - see FIGURE 2.5(a). In step-up procedures, as soon as one null hypothesis is rejected, all further hypotheses are also rejected - see FIGURE 2.5(b).

In step-wise procedures, rejection of a particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. The threshold t is therefore redefined at each step.

A large variety of procedures can be used to guarantee that the number of erroneous rejections of the null hypothesis is maintained under a pre-specified level. Each of them provides a more or less conservative trade-off strategy between rejecting true null hypotheses (false-positives) and accepting true alternative hypotheses (false-negatives).



(a) Step-down procedures

(b) Step-up procedures

Figure 2.5: Principle of step-wise procedures. $p_{(k)}$ represents the k^{th} p-value

2.4.2 Control of the FWER

This section discusses two procedures for control of the FWER. The first one is proposed by Bonferroni [1936], based on the following threshold $t = \alpha/m$:

$$\begin{aligned}\mathbb{P}(V_t > 0) &= \mathbb{P}\left(\sum_{k\in\mathcal{M}_0} \mathbb{1}_{p_k<\alpha/m} > 0\right) = \mathbb{P}\left(\bigcup_{k\in\mathcal{M}_0} p_k < \alpha/m\right) \\ &\leq \sum_{k\in\mathcal{M}_0} \mathbb{P}\left(p_k < \alpha/m\right) = m_0\frac{\alpha}{m} \leq \alpha\end{aligned}$$

from Boole's inequality.

Sidak [1967]'s procedure assumes independence for the null p-values. If $t = 1-(1-\alpha)^{1/m}$ then:

$$\begin{aligned}\mathbb{P}(V_t > 0) &= 1 - \mathbb{P}(V_t = 0) = 1 - \mathbb{P}\left(\bigcap_{k\in\mathcal{M}_0} p_k > 1-(1-\alpha)^{1/m}\right) \\ &= 1 - \prod_{k\in\mathcal{M}_0} \mathbb{P}\left(p_k > 1-(1-\alpha)^{1/m}\right) = 1-(1-\alpha)^{m_0/m} \leq \alpha\end{aligned}$$

For small value of α and large m, the two thresholds lead to similar control level as $1-(1-\alpha)^{1/m} \approx \alpha/m$.

2.4.3 Control of the (p)FDR

Let's call $\mathcal{R}_t = \{k|p_k \leq t\}$, $\mathcal{R}_t \subset \mathcal{M}$, of size R_t. Intuitively, the aim is to define the threshold t so that $\mathcal{R}_t$ is as large as possible while $FDR_t \leq \alpha$. For a given t, an approximation of FDR_t is

given considering the ratio of V_t and R_t expectancies: $FDR_t \approx \frac{\mathbb{E}(V_t)}{R_t} = \frac{\sum_{k\in\mathcal{M}_0}\mathbb{P}(p_k\leq t)}{R_t} \leq \frac{m_0 t}{R_t} \leq \frac{mt}{R_t}$. Therefore, $t \leq \alpha R_t/m \Rightarrow FDR_t \leq \alpha$.

More generally, determining the subset $\mathcal{R}_t$ should check the condition $\mathcal{R}_t \subset \{k|p_k \leq \alpha\beta(R_t)/m\}$, called self-consistency condition [Blanchard and Roquain, 2008], where $\beta(x)$ is an increasing function. In particular, step-wise procedures are self-consistent.

The step-up procedure associated to the linear shape function $\beta(x) = x$ is the well-known linear step-up procedure of Benjamini and Hochberg [1995], hereafter called BH procedure. If $p_{(1)}, \ldots, p_{(m)}$ are the ordered p-values, the BH procedure threshold is: $t = p_{(k^*)}$, where $k^* = \text{argmax}\left(k \mid p_{(k)} < \frac{k}{m}\alpha\right)$. The algorithm of the BH procedure was initially demonstrated to control FDR thanks to Simes's inequality [Simes, 1986].

For a given level α, the BH procedure aims at finding the (data-dependent) threshold t so that FDR_t is less than the α level. An other approach is often considered in practice when determining multiple testing significance. For a given threshold t on the p-values, we can form the associated FDR [Storey, 2002]:

$$t_\alpha = \underset{t\in[0;1]}{\text{argmax}}\{\widehat{FDR}_t(\lambda) \leq \alpha\}$$

where $\widehat{FDR}_t(\lambda) = \frac{\hat{m}_0(\lambda)t}{R_t \vee 1}$, and $\hat{m}_0(\lambda) = m\hat{\pi}_0(\lambda)$. The choice of the λ parameter has been evoked previously (see SECTION 2.2). Both point of view are equivalent, provided m is large and if m is replaced by $\hat{m}_0(\lambda)$ in the BH procedure [Storey et al., 2004].

The p-value measures the significance in the test of a single hypothesis. Similarly, the *q-value* [Storey, 2003] is the FDR based measure of significance in the multiple testing setting. It is defined as: $q_k = \text{argmin}_{t>p_k}\{pFDR_t\}$ and estimated in practice by $\hat{q}_k = \frac{\hat{m}_0(\lambda)p_{(k)}}{k}$ where $\hat{m}_0(\lambda) = \frac{\#\{p_k>\lambda\}}{(1-\lambda)}$.

Conclusion: improving multiple testing procedures

In this chapter, we have presented a usual framework of high-dimensional multiple testing and introduced the main concepts. This framework relies on the assumption of independent p-values distributed according to the two-component mixture model (2.4), with a uniform distribution for null p-values. Special attention is paid to the estimation of the proportion of true null hypotheses and to the definition of error rate in the multiple testing setting.

An area of current research is aimed at improving the power of multiple testing procedures. For a given level of error rate control, the goal is to declare positive a larger number of tests.

First, improvement in power may be achieved by step-wise procedures with respect to single-step ones, in which rejection of a particular hypothesis is based not only on the total number of hypotheses, but also on the outcome of the tests of other hypotheses. For example, step-wise version of Bonferroni and Sidak procedures have therefore been developed [Holm, 1979] and result in less conservative decisions.

In multiple hypotheses testing, the control of error rates is a crucial point of interest. However, trade-off is to be made between controlling error rates and maximizing power of the procedures. Besides, the control of these quantities is not exact and depends on the proportion of true null hypotheses, which is of course unknown. A challenging issue is therefore to provide an accurate estimation (small bias and small variance) of the proportion of null hypotheses among the whole set of tests. Recent articles Storey [2002], Black [2004], Kim and Van de Wiel [2008] showed that a more accurate estimation of π_0 would improve the power of multiple testing procedures.

Most of the existing procedures involve assumptions about the p-values being independent. Recent studies, many of them listed by Gordon et al. [2007], also suggest that high correlations among test statistics affect a strong control of the actual proportion of false discoveries. Indeed, although current methods of simultaneous testing are generally shown to control expected type-I error rates, they suffer from high instability in the presence of correlation. The impact of dependence on the procedure is also, and from the very beginning, a questioning issue. Many papers have especially focused on the control of the FDR under various patterns of dependence between test statistics. An important contribution to this point was given by Benjamini and Yekutieli [2001]. They showed the BH procedure still controls the FDR under assumption of a certain class of dependence called positive. In fact, the general message seems to be that, for a high amount of dependence, the BH thresholding method tends to over-control the FDR, leading to more conservative rules than expected under the assumption of independence. Consequently, this also means that dependence affects the power of the BH procedure. These last ten years, many procedures have therefore been inspired by the BH method, focusing on improvements of the thresholding technique, but without modifications of the individual test statistics.

Some authors [Lönnstedt and Speed, 2002, Smyth, 2004] proposed moderated versions of the t-statistics where the variable-specific variance estimator in the denominator is augmented by a constant, which is derived from the whole set of variables. More recently, Storey et al. [2007] proposed the so-called Optimal Discovery Procedure in which the idea of test statistics combining common information across the variables is exploited thoroughly.

Recent proposals also suggest to modify the theorical null distribution [Efron, 2004].

Resampling methods are sometimes suggested to capture the dependence on the joint null distribution of the test statistics [Westfall and Young, 1993]. However, some shortcomings can be highlighted for resampling-based procedures, especially concerning the control of the type-I error rate [Yifan et al., 2006] and computation time.

In the following chapter, we focus on the impact of dependence in multiple testing procedures. A common idea in many recent papers is also that dependence between test statistics should be taken into account by borrowing information across the variables rather than treating them as independent [Kendziorski et al., 2003, Leek and Storey, 2008]. This can be achieved by modeling the common information and taking advantage of the shared information between variables.

This approach is studied in CHAPTER 3.

CHAPTER 3

MULTIPLE TESTING UNDER DEPENDENCE

Abstract The impact of dependence is currently one of the most discussed topics in the literature about multiple testing for high-dimensional data. High correlations among test statistics affect a strong control of the actual proportion of false discoveries and the estimation of key parameters of procedures such as the proportion of true null hypotheses. In many areas, dependence can be explained by an underlying structure of unobserved factors. Modeling this structure by a Factor Analysis model provides a framework to study the impact of dependence on multiple testing procedures. It is shown that the variance of the number of false discoveries increases along with the fraction of common variance. The same results are obtained for the empirical estimator of the proportion of true null hypotheses.

Sommaire

Introduction

In multiple hypotheses testing, the control of error rates under dependence is a crucial point of interest. Most proposals are concerned with the modification of initial algorithms to extend the assumptions on p-values distribution [Benjamini and Yekutieli, 2001] or with the modification of the theorical null distribution [Efron, 2004].

In the case of t-tests, the test statistics are then jointly distributed according to a non-central multivariate Student distribution $\mathcal{T}_{df}(\tau; R)$, where $\tau = (\tau_1; \ldots; \tau_m)$ is the $m-$vector of non-centrality parameters, *df* is the number of residual degrees of freedom and R is the residual correlation matrix R. For $k = 1, \ldots, m$, $\tau_k = \sqrt{n}\lambda'\beta^{(k)}/(\sigma_k\sqrt{\lambda'\mathbf{S}_{xx}^{-1}\lambda})$, which equals 0 for $k \in \mathcal{M}_0$. In this case, correlation across the test statistics is exactly the correlation between the response variables. More generally, dependence among the p-values is also straightforward inherited from dependence among the data.

We suppose now that the residual random vector $E = [\epsilon_1, \ldots, \epsilon_m]$ has a variance matrix Σ which is not necessarily a diagonal matrix.

The following proposition proved by Leek and Storey [2008] defines a general framework for multiple testing dependence.

Proposition 3.0.1 (see Leek and Storey [2008]). *Under assumption* (2.1)*, suppose that for each ϵ_k, there is no Borel measurable function g such that $\epsilon_k = g(\epsilon_1, \ldots, \epsilon_{k-1}, \epsilon_{k+1}, \ldots, \epsilon_m)$ almost surely. Then, there exists a random $Q-$vector Z, with $0 \leq Q \leq m$ and, for all $k = 1, \ldots, m$, there exist $Q-$vectors b_k such that,*

$$Y_k = m_k(x) + Zb_k' + \varepsilon_k, \tag{3.1}$$

where $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_m)$ is a random vector with independent components.

The above result establishes the existence of latent variables Z which capture the dependence among the variables in a $Q-$dimensional linear space. Therefore, model (3.1) can be viewed as a Factor Analysis model and the variables Z will henceforth be called factors. If it is furthermore assumed that the factors have means 0 and variance I_q as in the exploratory Factor Analysis model [Mardia et al., 1979], the mixed-effects regression models (3.1) are equivalently defined as fixed-effects regression models which residual variance Σ can be decomposed into the sum of two components: a diagonal matrix Ψ of specific variances $\psi_k^2 = \mathbb{V}(\varepsilon_k)$ and a common variance component $B'B$, where the kth

row of B is b_k:

$$\Sigma = BB' + \Psi \tag{3.2}$$

EXEMPLE 3. Impact of dependence is first illustrated by simulation scenarios with increasing amounts of dependence among data: 10 levels of dependence are considered, form independence (scenario 0) to highly correlated data (scenario 9). The proportion $\text{tr}(BB')/\text{tr}(\Sigma)$ of common variance increases along with the scenarios: In each case, 1 000 datasets are simulated according

Scenario	0	1	2	3	4	5	6	7	8	9
Common variability (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19

Table 3.1: Common variability (%) for the 10 simulated scenarios

to a multivariate normal distribution. Each dataset is composed of $m = 500$ variables and $n = 60$ observations such as $Y_{n\times m} \sim \mathcal{N}_m(0; \Sigma_s)$. Besides, let's consider a binary variable X such that the observations are split into two groups of size $n/2$. For each dataset, the p-values of the usual t-tests for the comparison of means are calculated.

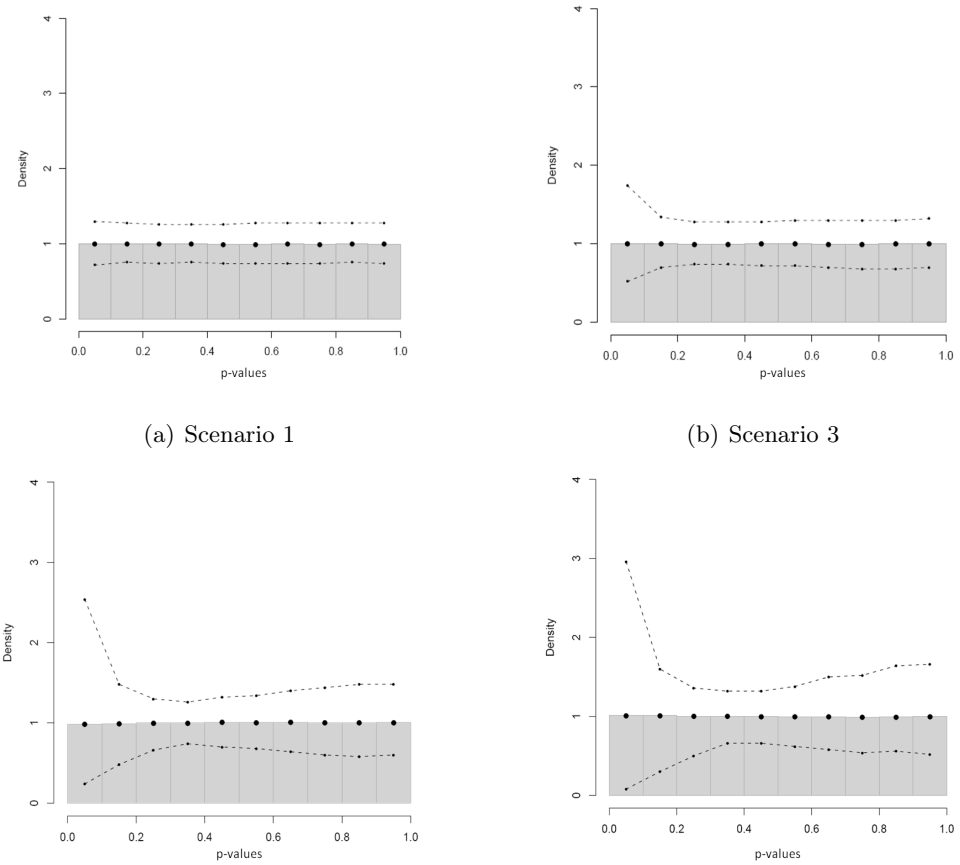
3.1. Impact of dependence on p-values distribution

FIGURE 3.1 reproduces the mean histograms of p-values with 95% confidence intervals in situations of independence (scenario 1), intermediate (scenarios 3 and 6) and high level of dependence (scenario 9). Obviously, this shows that uniformity of the distribution of the null p-values is true on average, but in case of dependent data, the histograms can show marked departures from uniformity.

FIGURE 3.2 represents the examples of histograms from two simulated datasets from scenario 9. Dependence can lead either to a much larger representation of the p-values close to 0 (and consequently an under-representation of the p-values close to 1) or inversely much lesser small p-values than expected under uniformity. This violation of the uniformity of the null distribution is also mentioned in Efron [2007], which reports that correlation can widen or narrow down the distribution of Z-scores with respect to the theoretical null distribution.

3.2. Impact of dependence on the estimation of the proportion of true null hypotheses (π_0)

The immediate consequence of the above comments is a less accurate estimation of the proportion of null hypotheses. We now slightly modify the simulation scheme of EXAMPLE 3:



(a) Scenario 1

(b) Scenario 3

(c) Scenario 6

(d) Scenario 9

Figure 3.1: P-values distribution - Mean histogram over 1000 simulations - dotted lines: 95% confidence interval

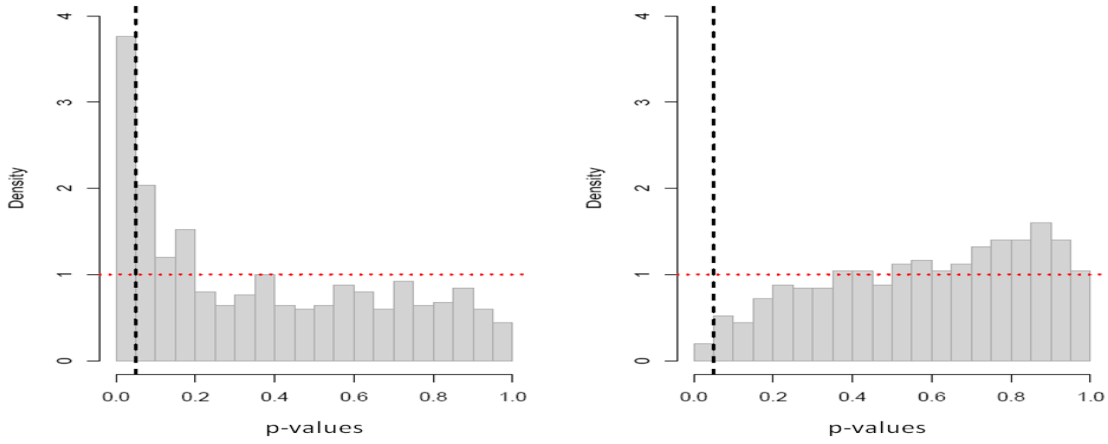


Figure 3.2: P-values distribution under H_0: examples of two datasets - scenario 9

EXEMPLE 4. For each dependence scenario, for each dataset, we now set different expectations between groups of X for $m_1 = 100$ variables. The difference is chosen so that the usual t-tests have a variable-by-variable power of 0.8 and for the remaining $m_0 = 400$ variables, the difference is set to 0. The true value of π_0 is therefore $0,80$. The p-values of the usual t-tests for the comparison of means are then calculated.

In the simulated situations of dependence of EXAMPLE 4, π_0 is estimated using the different methods presented in CHAPTER 2. The results are presented for two of them: empirical estimator with bootstrap choice of λ [Storey et al., 2004] and considering the density estimation assuming it is a convex decreasing function [Langaas et al., 2005]).

FIGURE 3.3 shows the estimated proportion of null hypotheses, along with the Δ criterion. This simple criterion is just used here to characterize departures of the null distribution of p-values from uniformity. It is positive if close-to-0 p-values are over represented (FIGURE 3.2, left)and negative if close-to-1 p-values are over represented (FIGURE 3.2, right).

FIGURE 3.3 shows that dependence induces a kind of local bias, in the sense that concave densities of null p-values lead to an overestimation of π_0 and inversely, convex densities to an underestimation of π_0.

π_0 estimation under dependence In the following proposition, for $k \neq k'$, $G^k(t) = \mathbb{P}(p_k \leq t)$ and $G^{kk'}(t) = \mathbb{P}(p_k \leq t; p_{k'} \leq t)$. $G^k(Z,t) = \mathbb{P}(p_k \leq t|Z)$ is the conditional distribution of p-values with respect to factor Z.

PROPOSITION 3.2.1.

$$\begin{aligned} \mathbb{E}(G^k(Z,t)) &= G^k(t) \\ \mathrm{Cov}(G^k(Z,t); G^{k'}(Z,t)) &= G^{kk'}(t) - G^k(t)G^{k'}(t) \end{aligned}$$

More particularly, if $k \in \mathcal{M}_0$: $\mathbb{E}(G^k(Z,t)) = G_0(t) = t$

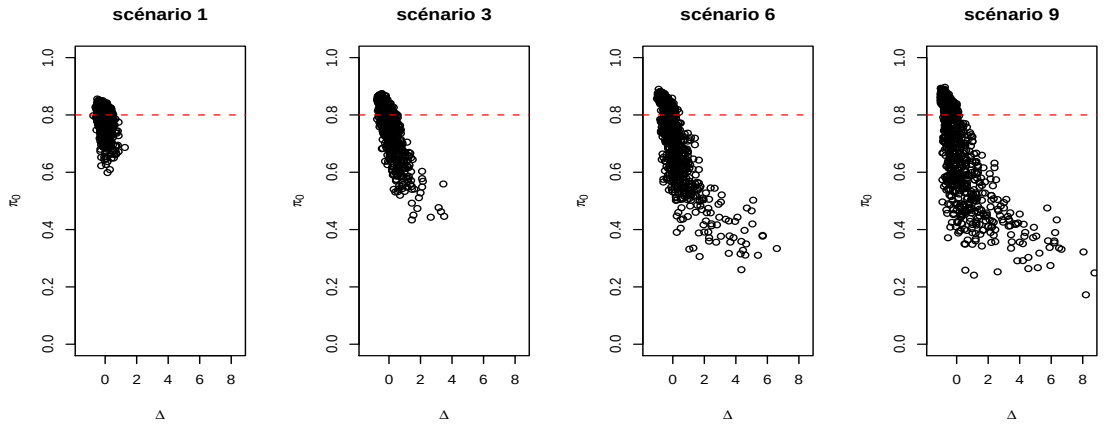
Proof. $G^k(Z,t) = \mathbb{P}(p_k \leq t|Z) = \mathbb{E}(\mathbb{1}_{p_k \leq t}|Z)$.
Then $\mathbb{E}(G^k(Z,t)) = \mathbb{E}(\mathbb{1}_{p_k \leq t}) = G^k(t)$ □

We also define a general function called $D^{kk''}(t)$ as:

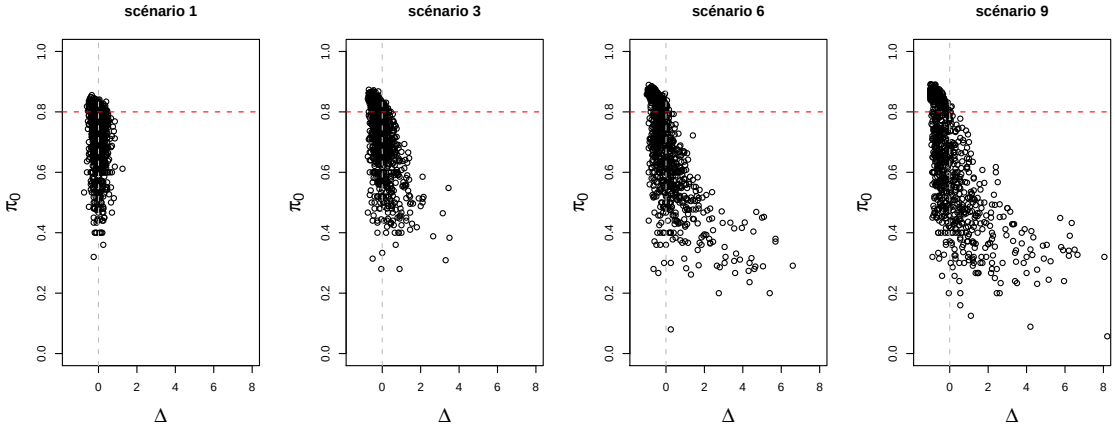
DÉFINITION 3.2.1.

$$D^{kk'}(t) = \frac{G^{kk'}(t) - G^k(t)G^{k'}(t)}{t(1-t)}$$

Moreover, $\bar{G_1}(\lambda) = \frac{1}{m_1} \sum_{k \in \mathcal{M}_1} \left(G_1^k(\lambda)\right)$ where $G_1^k(\lambda)$ represents the p-values distribution under the alternative hypothesis.



(a) Density estimation assuming it is a convex decreasing function



(b) Empirical estimator with bootstrap choice of λ

Figure 3.3: Estimation of π_0 along with the Δ criterion that characterizes the density of null p-values around 0 - $\pi_0 = 0,80$

Proposition 3.2.2.

$$\begin{aligned}
\mathbb{E}(\hat{\pi}_0(\lambda)) &= \pi_0 + (1-\pi_0)\frac{1-\bar{G}_1(\lambda)}{1-\lambda} \\
\mathbb{V}(\hat{\pi}_0(\lambda)) &= \frac{\lambda\pi_0}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}_1}\left[G_1^k(\lambda)(1-G_1^k(\lambda))\right]}{m^2(1-\lambda)^2} + \frac{\lambda}{1-\lambda}\frac{1}{m^2}\sum_{k\neq k'\in\mathcal{M}} D^{kk'}(\lambda)
\end{aligned}$$

The proof of proposition 3.2.2 rely on the following lemmas about the conditional properties of U_t and T_t.

Lemme 3.2.1.

$$\begin{aligned}
\mathbb{E}(U_\lambda) &= m_0(1-\lambda) \\
\mathbb{V}(U_\lambda) &= \Big[m_0 + \sum_{k\neq k'\in\mathcal{M}_0} D^{kk'}(\lambda)\Big]\lambda(1-\lambda)
\end{aligned}$$

Proof.

$$\begin{aligned}\mathbb{E}(U_\lambda|Z) &= \sum_{k\in\mathcal{M}_0}\mathbb{P}\left(p_k\geq\lambda|Z\right)=\sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda)\Big)\\ \mathbb{V}(U_\lambda|Z) &= \sum_{k\in\mathcal{M}_0}\mathbb{P}\left(p_k\geq\lambda|Z\right)\left(1-\mathbb{P}\left(p_k\geq\lambda|Z\right)\right)=\sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda,Z)\Big)G^k(\lambda,Z)\end{aligned}$$

$card(\mathcal{M}_0) = m_0$ and $\mathbb{E}(G^k(\lambda,Z)) = G^k(\lambda) = \lambda$, $\forall k\in\mathcal{M}_0$ from PROPOSITION 3.2.1. Therefore, $\mathbb{E}(U_\lambda) = \mathbb{E}\left(\mathbb{E}(U_\lambda|Z)\right) = m_0(1-\lambda)$.
The variance of U_λ is deduced from: $\mathbb{V}(U_\lambda) = \mathbb{E}\left(\mathbb{V}(U_\lambda|Z)\right) + \mathbb{V}\left(\mathbb{E}(U_\lambda|Z)\right)$.

$$\begin{aligned}\mathbb{V}[\mathbb{E}(U_\lambda|Z)] &= \mathbb{V}\left[\sum_{k\in\mathcal{M}_0}(1-G^k(\lambda,Z))\right]=\mathbb{V}\left[\sum_{k\in\mathcal{M}_0}(G^k(\lambda,Z))\right]=\sum_{k\in\mathcal{M}_0}\mathbb{V}\Big(G^k(\lambda,Z)\Big)+\sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)-\sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)\Big)^2+\sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)-m_0\lambda^2+\sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ \mathbb{E}[\mathbb{V}(U_\lambda|Z)] &= \mathbb{E}\left[\sum_{k\in\mathcal{M}_0}\Big(1-G^k(\lambda,Z)\Big)G^k(\lambda,Z)\right]\\ &= \sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)\Big)-\sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)\\ &= m_0\lambda-\sum_{k\in\mathcal{M}_0}\mathbb{E}\Big(G^k(\lambda,Z)^2\Big)\end{aligned}$$

Using the $D^{kk'}(\lambda)$ function introduced previously:

$$\begin{aligned}\Rightarrow\mathbb{V}(U_\lambda) &= m_0\lambda(1-\lambda)+\sum_{k\neq k'\in\mathcal{M}_0}\mathrm{Cov}\Big(G^k(\lambda,Z);G^{k'}(\lambda,Z)\Big)\\ &= \Big[m_0+\sum_{k\neq k'\in\mathcal{M}_0}D^{kk'}(\lambda)\Big]\lambda(1-\lambda)\end{aligned}\tag{3.3}$$

□

LEMME 3.2.2. *Let's denote* $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$.

$$\begin{aligned}\mathbb{E}(T_\lambda) &= (m-m_0)(1-\bar{G}_1^k(\lambda))\\ \mathbb{V}(T_\lambda) &= \sum_{k\in\mathcal{M}_1}\Big[G_1^k(\lambda)(1-G_1^k(\lambda))\Big]+\Big[\sum_{k\neq k'\in\mathcal{M}_1}D^{kk'}(\lambda)\Big]\lambda(1-\lambda)\end{aligned}$$

Proof.

$$\begin{aligned}\mathbb{E}(T_\lambda|Z) &= \sum_{k\in\mathcal{M}_1}\mathbb{P}\left(p_k\geq\lambda|Z\right)=\sum_{k\in\mathcal{M}_1}\Big(1-G^k(\lambda,Z)\Big)\\ \mathbb{V}(T_\lambda|Z) &= \sum_{k\in\mathcal{M}_1}\mathbb{P}\left(p_k\geq\lambda|Z\right)\left(1-\mathbb{P}\left(p_k\geq\lambda|Z\right)\right)=\sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))G^k(\lambda,Z)\end{aligned}$$

$card(\mathcal{M}_1) = m - m_0$ and $\mathbb{E}(G^k(\lambda, Z)) = G_1^k(\lambda)$, $\forall k \in \mathcal{M}_1$ from PROPOSITION 3.2.1.

Let's call $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$. $\mathbb{E}(T_\lambda) = \mathbb{E}\left(\mathbb{E}(T_\lambda|Z)\right) = (m-m_0)(1-\bar{G}_1^k(\lambda))$.
The variance of T_λ is deduced from: $\mathbb{V}(T_\lambda) = \mathbb{E}\left(\mathbb{V}(T_\lambda|Z)\right) + \mathbb{V}\left(\mathbb{E}(T_\lambda|Z)\right)$. On a :

$$\begin{aligned}
\mathbb{V}[\mathbb{E}(T_\lambda|Z)] &= \mathbb{V}\left[\sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))\right]\mathbb{V}\left[\sum_{k\in\mathcal{M}_1}(G^k(\lambda,Z))\right] = \sum_{k\in\mathcal{M}_1}\mathbb{V}\left(G^k(\lambda,Z)\right) + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\left(G^k(\lambda,Z);G^{k'}(\lambda,Z)\right)\\
&= \sum_{k\in\mathcal{M}_1}\mathbb{E}\left(G^k(\lambda,Z)^2\right) - \sum_{k\in\mathcal{M}_1}\mathbb{E}\left(G^k(\lambda,Z)\right)^2 + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\left(G^k(\lambda,Z);G^{k'}(\lambda,Z)\right)\\
\mathbb{E}[\mathbb{V}(T_\lambda|Z)] &= \mathbb{E}\left[\sum_{k\in\mathcal{M}_1}\left(1-G^k(\lambda,Z)\right)G^k(\lambda,Z)\right]\\
&= \sum_{k\in\mathcal{M}_1}\mathbb{E}\left(G^k(\lambda,Z)\right) - \sum_{k\in\mathcal{M}_1}\mathbb{E}\left(G^k(\lambda,Z)^2\right)
\end{aligned}$$

Using the $D^{kk'}(\lambda)$ function introduced previously:

$$\begin{aligned}
\Rightarrow \mathbb{V}(T_\lambda) &= \sum_{k\in\mathcal{M}_1}\left[\mathbb{E}\left(G^k(\lambda,Z)\right) - \mathbb{E}\left(G^k(\lambda,Z)\right)^2\right] + \sum_{k\neq k'\in\mathcal{M}_1}\mathrm{Cov}\left(G^k(\lambda,Z);G^{k'}(\lambda,Z)\right)\\
&= \sum_{k\in\mathcal{M}_1}\left[G_1^k(\lambda)(1-G_1^k(\lambda))\right] + \sum_{k\neq k'\in\mathcal{M}_1}\left[G^{kk'}(\lambda) - G^k(\lambda)G^{k'}(\lambda,Z)\right]\\
&= \sum_{k\in\mathcal{M}_1}\left[G_1^k(\lambda)(1-G_1^k(\lambda))\right] + \Big[\sum_{k\neq k'\in\mathcal{M}_0}D^{kk'}(\lambda)\Big]\lambda(1-\lambda) \qquad (3.4)
\end{aligned}$$

□

LEMMAS 3.2.1 and 3.2.2 are now used for the proof of PROPOSITION 3.2.2.

Proof. Let's call $\bar{G}_1(\lambda) = \frac{1}{m_1}\sum_{k\in\mathcal{M}_1}\left(G_1^k(\lambda)\right)$. Expectation of $\hat{\pi}_0$ is deduced:

$$\begin{aligned}
\mathbb{E}(\hat{\pi}_0(\lambda)) &= \frac{\mathbb{E}(W_\lambda)}{m(1-\lambda)} = \frac{\mathbb{E}(U_\lambda)+\mathbb{E}(T_\lambda)}{m(1-\lambda)}\\
&= \frac{m_0(1-\lambda)+m_1(1-\bar{G}_1(\lambda))}{m(1-\lambda)} = \pi_0 + (1-\pi_0)\frac{(1-\bar{G}_1(\lambda))}{(1-\lambda)}
\end{aligned}$$

Moreover, $\mathbb{V}(W_\lambda) = \mathbb{V}(U_\lambda) + \mathbb{V}(T_\lambda) + 2\,\mathrm{Cov}(U_\lambda;T_\lambda)$, where $\mathbb{V}(U_\lambda)$ and $\mathbb{V}(T_\lambda)$ are given in LEMMAS 3.2.1 and 3.2.2. $\mathrm{Cov}(U_\lambda;T_\lambda) = \mathbb{E}\left(\mathrm{Cov}(U_\lambda;T_\lambda|Z)\right) + \mathrm{Cov}(\mathbb{E}(U_\lambda|Z);\mathbb{E}(T_\lambda|Z))$. Conditional independence in model (3.1) gives $\mathrm{Cov}(U_\lambda;T_\lambda|Z) = 0$.

$$\begin{aligned}
\mathrm{Cov}(\mathbb{E}(U_\lambda|Z);\mathbb{E}(T_\lambda|Z)) &= \mathrm{Cov}\left(\sum_{k\in\mathcal{M}_0}(1-G^k(\lambda,Z));\sum_{k\in\mathcal{M}_1}(1-G^k(\lambda,Z))\right)\\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\mathrm{Cov}(G^k(\lambda,Z);G^{k'}(\lambda,Z))\\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\left[G^{kk'}(\lambda) - G^k(\lambda,Z)G^{k'}(\lambda,Z)\right]\\
&= \sum_{k\in\mathcal{M}_0}\sum_{k'\in\mathcal{M}_1}\left[D^{kk'}(\lambda)\right]\lambda(1-\lambda)
\end{aligned}$$

The variance of $\hat{\pi}_0(\lambda)$ is deduced:

$$\mathbb{V}(\hat{\pi}_0(\lambda)) = \frac{\lambda\pi_0}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}_1}\left[G_1^k(\lambda)(1-G_1^k(\lambda))\right]}{m^2(1-\lambda)^2} + \frac{\lambda}{1-\lambda}\frac{1}{m^2}\sum_{k\neq k'\in\mathcal{M}} D^{kk'}(\lambda)$$

□

In the case of independent p-values, with identical distribution under the alternative hypothesis, $\bar{G}_1(\lambda) = G_1(\lambda)$ and $D(\lambda) = 0$, which gives the bias and the variance as provided in expression 2.2.1 under assumption of a two-component mixture of distributions.

Minimising MSE to choose the tunig parameter PROPOSITION 3.2.2 also shows how the variance of the estimator depends on the correlation matrix R. In order to illustrate the impact of dependence both on the choice of an optimal threshold and on the variability of the estimation, the bias, variance and RMSE of $\hat{\pi}_0(\lambda)$ are calculated in the three multiple testing situations introduced for the simulation study in EXAMPLE 4. In a completely nonparametric framework, the bivariate probability $G^{kk'}(\lambda)$, which appears in the expression of the variance of $\hat{\pi}_0(\lambda)$, can be estimated using permutation techniques, provided it can be assumed that $G_{kk'}(\lambda)$ is the same for all (k, k'). In the present situation of Student's tests, a closed-form expression of $G^{kk'}(\lambda)$ is used, given by $\Phi^{(2)}([-u_\lambda; u_\lambda]^2; \rho_{kk'}, \tau_k, \tau_{k'}, n-2)$, where $\rho_{kk'}$ is the correlation between T_k and $T_{k'}$, τ is the non-centrality parameter of the non-null distribution of the test statistics, u_λ is the $(1-\lambda/2)-$quantile of the Student distribution with $n-2$ degrees of freedom and $\Phi^{(2)}(A; \rho, \tau, \tau', df)$ stands for the probability that a bivariate Student vector with non-centrality parameter (τ, τ'), degree of freedom df and correlation ρ belongs to $A \subseteq \mathbb{R}^2$.

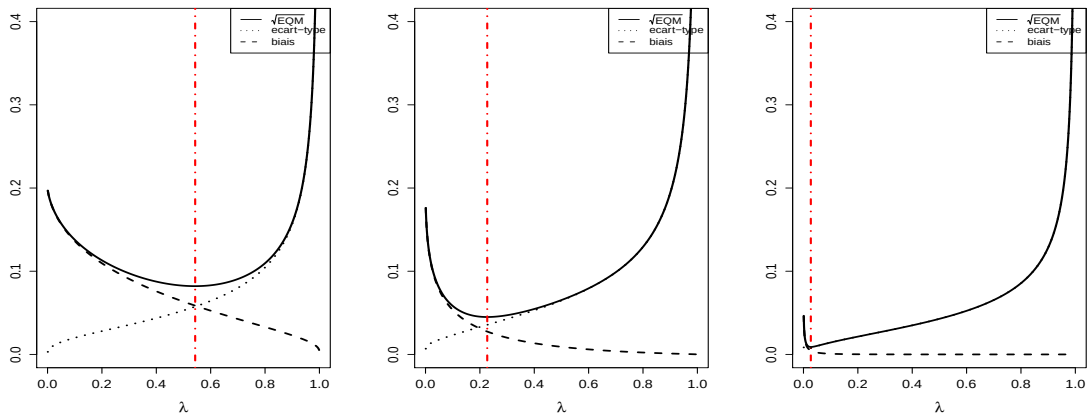
FIGURE 3.4 compares the RMSE curves of $\hat{\pi}_0(\lambda)$ for different levels of dependence among test statistics and different values for τ. The optimal choice for the threshold is given. It essentially shows that ignoring dependence leads to an overestimation of this optimal threshold, and simultaneously to an underestimation of the true variability of $\hat{\pi}_0(\lambda)$.

3.3. Impact of dependence on error rates

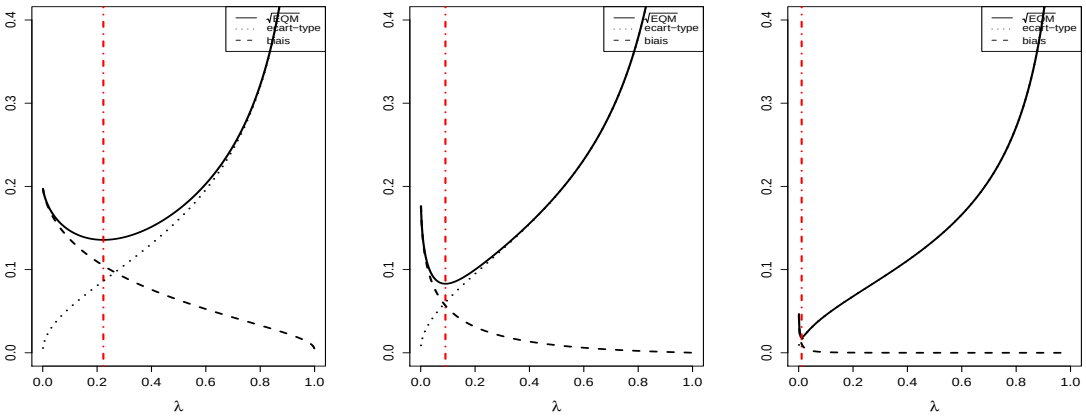
3.3.1 Impact of dependence on the number of false-positives (V_t)

For each simulated dataset in EXAMPLE 4), V_t is calculated considering a fixed threshold $t = 0,05$ on the p-values. results are given in TABLE 3.2.

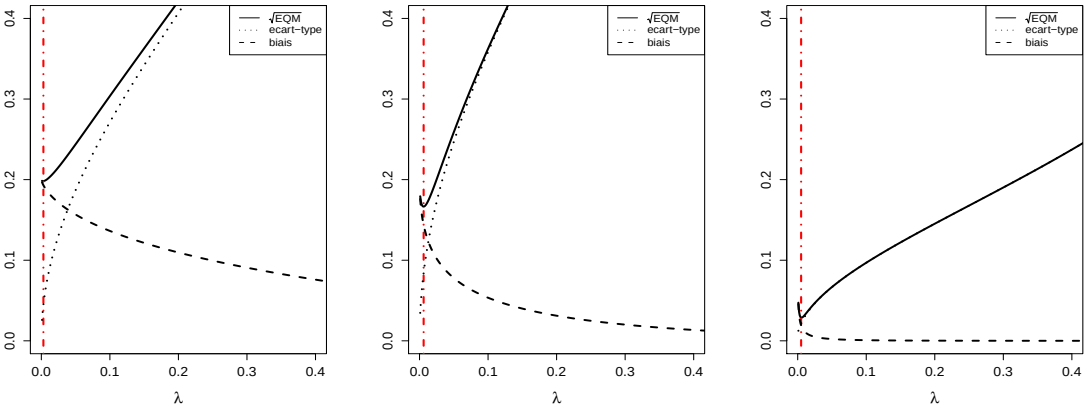
FIGURE 3.5 represents the distribution of V_t for four scenarios of increasing level of dependence. Both TABLE 3.2 and FIGURE 3.5 show that the mean of V_t does not seem to be affected by dependence. On the contrary, the variance of false-positives is increased in case of dependent data. These comments are confirmed by the following proposition.



(a) Scenario 1



(b) Scenario 4



(c) Scenario 8

Figure 3.4: Bias, variance and RMSE of $\hat{\pi}_0(\lambda)$ under different scenarios of dependence (on each plot, the vertical line locates the threshold for which the RMSE is minimized) - left panel: $\tau = 1$ (power: 17%), middle panel: $\tau = 2, 8$ (power: 80%), right panel: $\tau = 4$ (power: 97%)

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
min	6.00	5.00	6.00	5.00	1.00	1.00	0.00	0.00	0.00	0.00
$q_{0,25}$	17.00	17.00	15.00	13.00	11.00	8.00	8.00	6.00	6.00	4.00
median	20.00	20.00	19.00	17.00	16.00	14.00	14.00	12.00	12.00	11.00
mean	**19.96**	**20.12**	**20.03**	**19.44**	**19.48**	**19.94**	**19.63**	**21.12**	**21.29**	**20.22**
$q_{0,75}$	23.00	23.00	23.00	23.00	23.00	24.25	24.00	25.00	26.25	25.25
max	36.00	45.00	72.00	90.00	157.00	169.00	152.00	206.00	206.00	195.00
std	**4.43**	**4.96**	**7.33**	**10.43**	**14.26**	**18.73**	**19.83**	**26.19**	**26.14**	**26.35**

Table 3.2: Descriptives statistics of V_t for the 10 different scenarios of dependence

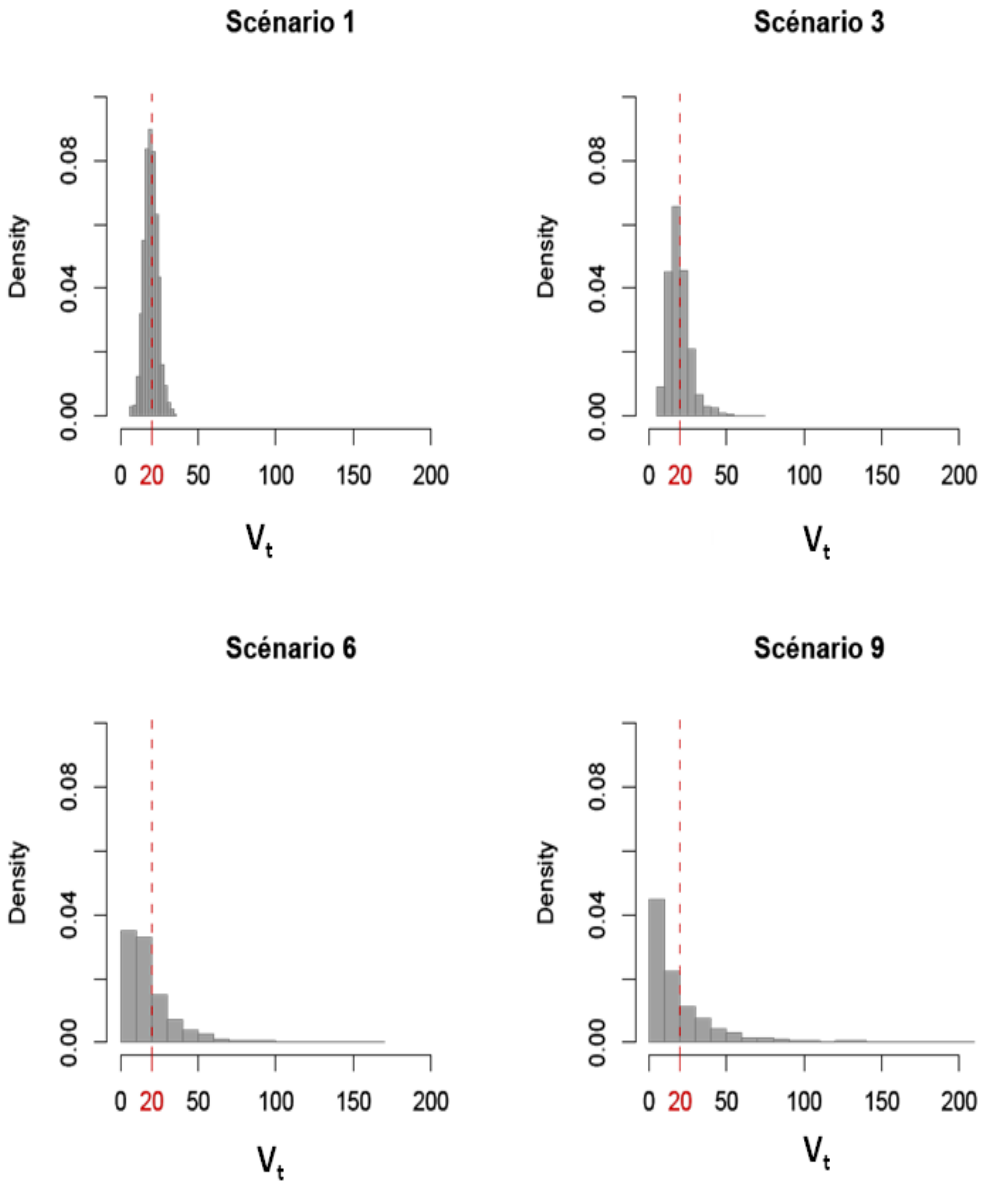


Figure 3.5: Distribution of V_t for different level of dependence.

For a given threshold t, V_t is defined as the number of erroneous rejections of the null hypotheses. For independent test statistics, V_t is distributed according to a binomial distribution: $V_t \sim \mathcal{B}in(m_0, t)$. This random variable has mean $m_0 t$ and variance $m_0 t(1-t)$. Under general dependence, the following proposition holds:

Proposition 3.3.1.

$$\begin{aligned} \mathbb{E}(V_t) &= m_0 t \\ \mathbb{V}(V_t) &= \Big[m_0 + \sum_{k \neq k' \in \mathcal{M}_0} D^{kk'}(t)\Big] t(1-t) \end{aligned}$$

Proof. The proof is deduced from the proof of LEMMA 3.2.1. Indeed, $V_t = m_0 - U_t$. Therefore, $\mathbb{E}(V_t) = m_0 - m_0(1-t) = m_0 t$ et $\mathbb{V}(V_t) = \mathbb{V}(m_0 - U_t) = \mathbb{V}(U_t) = \Big[m_0 + \sum_{k \neq k' \in \mathcal{M}_0} D^{kk'}(t)\Big] t(1-t)$. □

Therefore, by comparison with the binomial variance $m0t(1-t)$, the impact of dependence on the variance of V_t can be measured by $M_0(R) = \sum_{k \neq k' \in \mathcal{M}_0} D^{kk'}(t)$, where R is the correlation between the test statistics. PROPOSITION 3.3.1 shows that correlation modifies the distribution of V_t by increasing its dispersion, leaving its expectation unchanged. Therefore, correlation shall have an important impact on the tail of the distribution, which is directly involved in the calculation of the error rates.

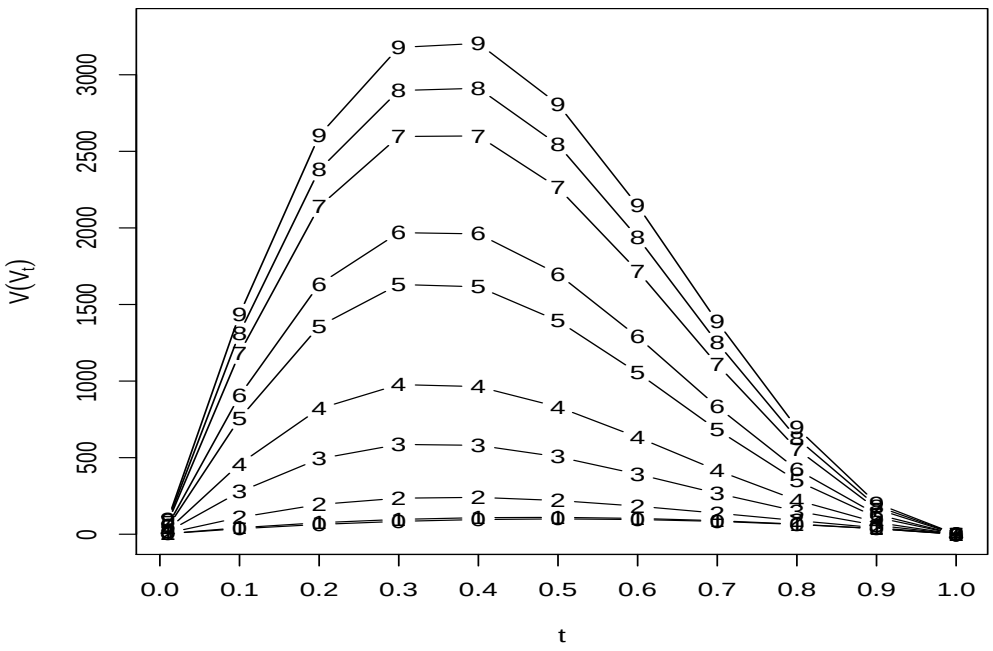


Figure 3.6: Variance of V_t along with the threshold t, for each dependence scenario

FIGURE 3.6 confirms that a high amount of conditional correlation leads to a very unstable distribution of the number of false discoveries. Although it seems to show that the impact of dependence becomes smaller when the threshold tends to zero, this have to be tempered by the fact that the expected number of false discoveries also decreases along with the threshold.

FIGURE 3.7 shows that, for any preset t, $Dt(\rho)$ is a U-shaped function that is close to some equivalent term appearing in Owen [2005]'s formula for the variance of the number of false discoveries and in Efron [2007], where the bivariate normal probability function is also involved in the expressions of the variance inflation.

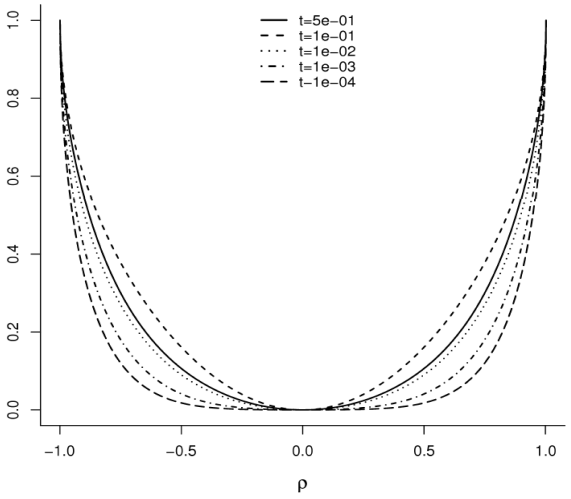


Figure 3.7: $D_t(\rho)$ for various values of the threshlod t

3.3.2 Impact of dependence on the FWER

The Sidak procedure [Sidak, 1967] is now applied on the datasets of the simulation study presented in EXAMPLE 4.

Implementation is done using the `R` package `multtest` [Pollard et al.]. TABLE 3.3 displays the frequencies of the number of false-positives V_t and the estimated FWER using the simulated datasets. It shows that the procedure based on the Student's tests controls the FWER at a lower level than $\alpha = 0,05$. FIGURE 3.8 reproduces multiple boxplots of the distributions of the non discovery proportion $NDP_t = \#\{k \notin \mathcal{M}_0, H_0^{(k)} \text{ not rejected}\}/m1$. It shows that the fraction of common variance generates instability in the distribution of NDP_t. Moreover, the mean NDP_t, which can be viewed as a type-II error rate, remains high whatever the level of dependence.

3.3.3 Impact of dependence on the FDR

Let's consider two estimators: the empirical estimator $\widehat{FDR_t} = \frac{m_0 t}{R_t}$ and the one resulting from Storey et al. [2004]'s procedure where the FDR is estimated considering $t = p_{(k^*)}$, with $k^* =$

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
counts										
0	957	958	970	973	966	971	969	964	970	968
1	42	41	29	24	33	26	27	24	22	25
2		1	1	3		2	3	4	3	3
3	1				1			5	3	2
4						1	1	1		
5								2		
6									1	
7										
8										2
9										
10										
11										
12										
13										
14										
15									1	
FWER										
$\#\{V_t > 0\}/m$	4,3	4,2	3,0	2,7	3,4	2,9	3,1	3,6	2,9	3,2

Table 3.3: V_t counts and $FWER_t$ estimated from the simulated datasets along with the proportion of common variance, for the procedures based on t-tests p-values with Sidak correction - threshold: $t = 1,0258.10^{-4}$)

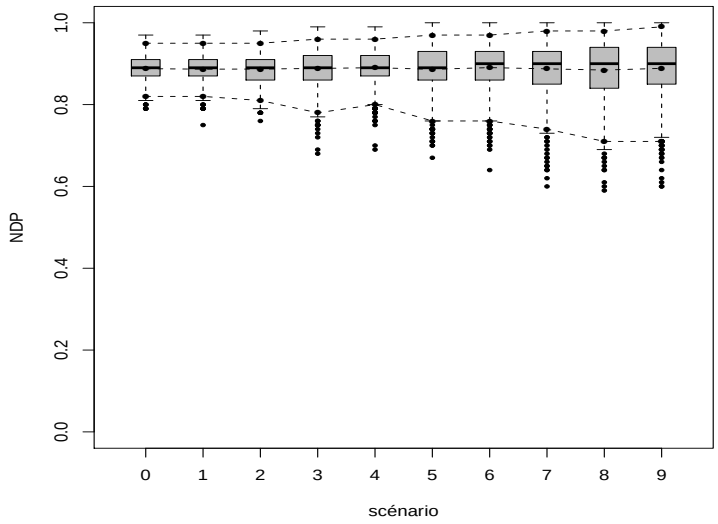


Figure 3.8: Distributions of the Non Discovery Proportion for the Sidak procedure based on the Student's tests along with the different scenarios of dependence

$\text{argmax}_k \left(p_{(k)} < \frac{\alpha}{m_0} k \right)$. Both estimator are similar, but the second one considers a data-dependent threshold.

PROPOSITION 3.3.2 (from Storey et al. [2004]). *Under* ASSUMPTION *1, for a fixed* $\lambda \in [0; 1[$*:*

$$\mathbb{E}(\widehat{FDR}_t(\lambda)) \geq FDR_t$$

FDR estimator is biased. But this bias is positive, so procedures control the actual FDR at a lower bound than estimated. The immediate consequence in practice is a lower power of procedures than expected [Sarkar, 2008].

The BH procedure [Benjamini and Hochberg, 1995] and the q-value procedure [Storey, 2003] are now applied on the datasets of the simulation study presented in exemple 4. Implementation is done using the `R` package `multtest` [Pollard et al.] for the first method and the `q-value` package [Dabney et al., 2009] for the second one. TABLE 3.3 displays the frequencies of the number of false-positives V_t and the estimated FWER using the simulated datasets. It shows that the procedure based on the t-tests control the FWER at a lower level than $\alpha = 0,05$. FIGURE 3.8 reproduces multiple boxplots of the distributions of the non discovery proportion $NDP_t = \#\{k \notin \mathcal{M}_0, H_0^{(k) \text{ not rejected}}\}/m1$. It shows that the fraction of common variance generates instability in the distribution of NDP_t. Moreover, the mean NDP_t, which can be viewed as a type-II error rate, remains high whatever the level of dependence. To avoid discussions about the estimation of π_0, this parameter is supposed to be known in this comparative study.

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_{t=0.05}$										
min	17,09	15,62	13,33	12,20	8,85	8,55	9,22	7,30	7,38	7,75
$q_{0,25}$	19,23	19,23	19,05	19,23	19,23	19,42	19,61	19,23	19,23	19,42
median	20,00	20,00	20,00	20,41	20,62	20,83	21,05	21,05	21,05	21,28
mean	**20,11**	**20,02**	**20,08**	**20,20**	**20,33**	**20,34**	**20,52**	**20,46**	**20,51**	**20,65**
$q_{0,75}$	20,83	20,83	21,05	21,28	21,74	21,98	21,98	22,47	22,47	22,47
max	25,32	24,10	24,69	25,00	25,64	25,00	27,40	26,67	28,17	30,30
sd	**1,23**	**1,25**	**1,47**	**1,78**	**2,11**	**2,39**	**2,57**	**3,08**	**3,12**	**3,19**
q-value										
min	14,69	14,88	15,99	15,94	15,83	15,18	13,37	15,07	13,63	12,99
$q_{0,25}$	19,16	19,17	19,16	19,10	19,12	19,15	19,09	19,10	19,11	19,02
median	19,59	19,59	19,59	19,55	19,62	19,59	19,59	19,56	19,60	19,56
mean	**19,41**	**19,40**	**19,42**	**19,37**	**19,38**	**19,38**	**19,33**	**19,34**	**19,34**	**19,27**
$q_{0,75}$	19,84	19,83	19,85	19,82	19,85	19,83	19,82	19,84	19,84	19,84
max	20,29	20,33	20,54	20,44	20,47	20,36	20,28	20,58	20,51	20,41
sd	**0,63**	**0,63**	**0,60**	**0,63**	**0,67**	**0,68**	**0,76**	**0,73**	**0,78**	**0,87**

Table 3.4: Descriptive statistics for FDR estimation along with the different scenarios of dependence (%)

For each scenario, TABLE 3.4 gives the descriptive statistics for both estimators. Whatever the level of dependence, the mean estimation is steady in each case. The variability of FDR estimation is lower with the data-dependent threshold.

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
FDP ($t = 0,05$)										
min	7,50	5,95	6,59	5,32	1,16	1,09	0,00	0,00	0,00	0,00
$q_{0,25}$	17,53	17,17	15,83	13,33	11,83	9,19	8,51	6,67	6,38	4,88
median	19,80	20,00	19,01	17,53	16,13	14,74	14,44	13,16	13,02	11,52
mean	**19,86**	**19,90**	**19,67**	**18,87**	**18,48**	**18,31**	**17,88**	**17,97**	**18,23**	**17,74**
$q_{0,75}$	22,43	22,45	22,55	22,93	22,25	23,79	23,75	24,53	26,18	24,82
max	31,58	35,16	48,00	58,07	69,47	72,22	70,05	76,37	76,75	76,03
sd	**3,64**	**4,04**	**5,70**	**7,82**	**9,69**	**12,36**	**12,79**	**15,13**	**15,80**	**16,08**
FDP (threshold *q-value*)										
min	4,17	4,94	4,76	3,49	1,16	0,00	0,00	0,00	0,00	0,00
$q_{0,25}$	17,17	16,83	15,05	12,37	10,78	8,05	7,30	5,60	5,32	3,90
median	19,82	20,21	19,00	17,20	15,53	14,00	13,40	12,07	11,43	10,01
mean	**19,96**	**20,02**	**19,82**	**19,03**	**18,78**	**18,60**	**18,07**	**18,34**	**18,68**	**17,62**
$q_{0,75}$	22,86	23,17	23,15	23,48	23,00	24,16	24,30	25,05	26,6	25,60
max	32,80	36,50	54,26	63,59	74,83	76,27	76,38	78,07	79,18	79,75
sd	**4,38**	**4,84**	**6,91**	**9,33**	**11,65**	**14,55**	**14,98**	**17,64**	**18,37**	**18,59**

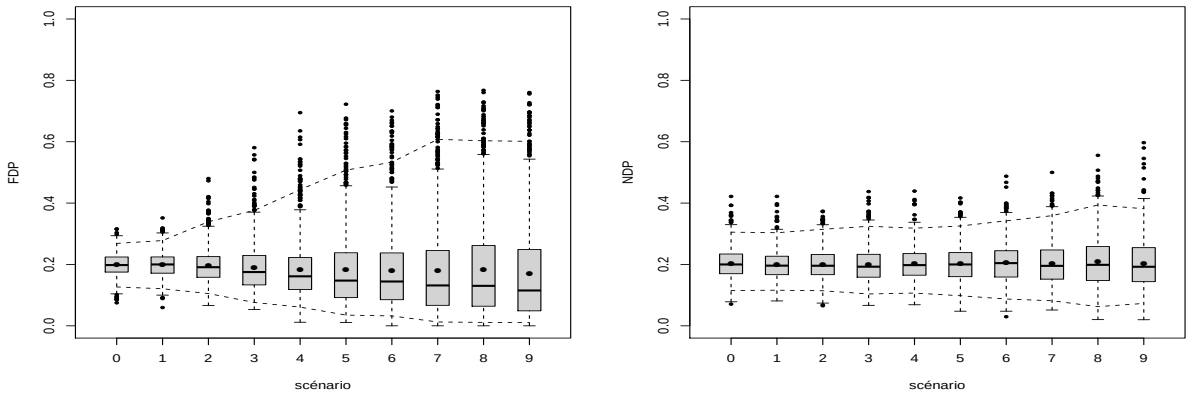
Table 3.5: Descriptive statistics for False Discovery Proportion along with the different scenarios of dependence (%)

TABLE 3.5 gives descriptive statistics for the true proportion of false-positives (FDP). Mean of FDP, which is FDR, is steady for all scenarios, but its variability sharply increases along with the proportion of common variability (see FIGURE 3.9(a)).

FIGURE 3.9(b) shows that the fraction of common variance generates slight instability in the distribution of NDP_t. The mean NDP_t, which can be viewed as a type-II error rate, remains steady whatever the level of dependence.

We now compare FDR estimation and the true proportion of false-positives (NDP) in the simulation study. Results are presented on FIGURE 3.10 for four different scenarios and TABLE 3.6 gives the regression coefficients between the two FDR estimators and FDP.

As already observed by Efron [2007], the empirical estimate is negatively correlated with the observed FDP, which can result in strongly misleading estimations especially when FDP is high. FIGURE 3.10 shows that this concern is particularly clear for large fractions of shared variance. When estimating FDR with a data-dependent threshold, correlation is not negative anymore.



(a) FDP (b) NDP

Figure 3.9: False Discovery Proportion (FDP) and Non Discovery Proportion (NDP), along with the 10 scenarios

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_{t=0.05}$	-0,197	-0.196	-0,189	-0,176	-0.187	-0,171	-0,175	-0,182	-0,169	-0,171
q-value	0,025	0,026	0,017	0,016	0,015	0,013	0,017	0,013	0,014	0,016

Table 3.6: Regression coefficients between estimated FDR and the true proportion of false positives (FDP)

Conclusion

Dependence induces variability that interferes in particular with p-values distribution that can sharply deviates from the theorical null distribution when the level of common variability between variables is high. Consequently, the variability of false-positives increases in presence of dependence, leading to high instability in multiple testing procedures. Moreover, π_0 estimation is biased when the level of dependence is high. This bias depends on the impact of dependence on p-values distribution, whether it yields to more small p-values or to more close-to-one p-values.

Considering the factor modeling of dependence, which describes the common structure through latent variables as in (3.1), the variance of the number of false-positives and the variance of π_0 are both derived. Their expressions include a term which directly depends on the correlation between test statistics.

Next chapter describes a method that borrows the information shared by all the variables to improve multiple testing under dependence.

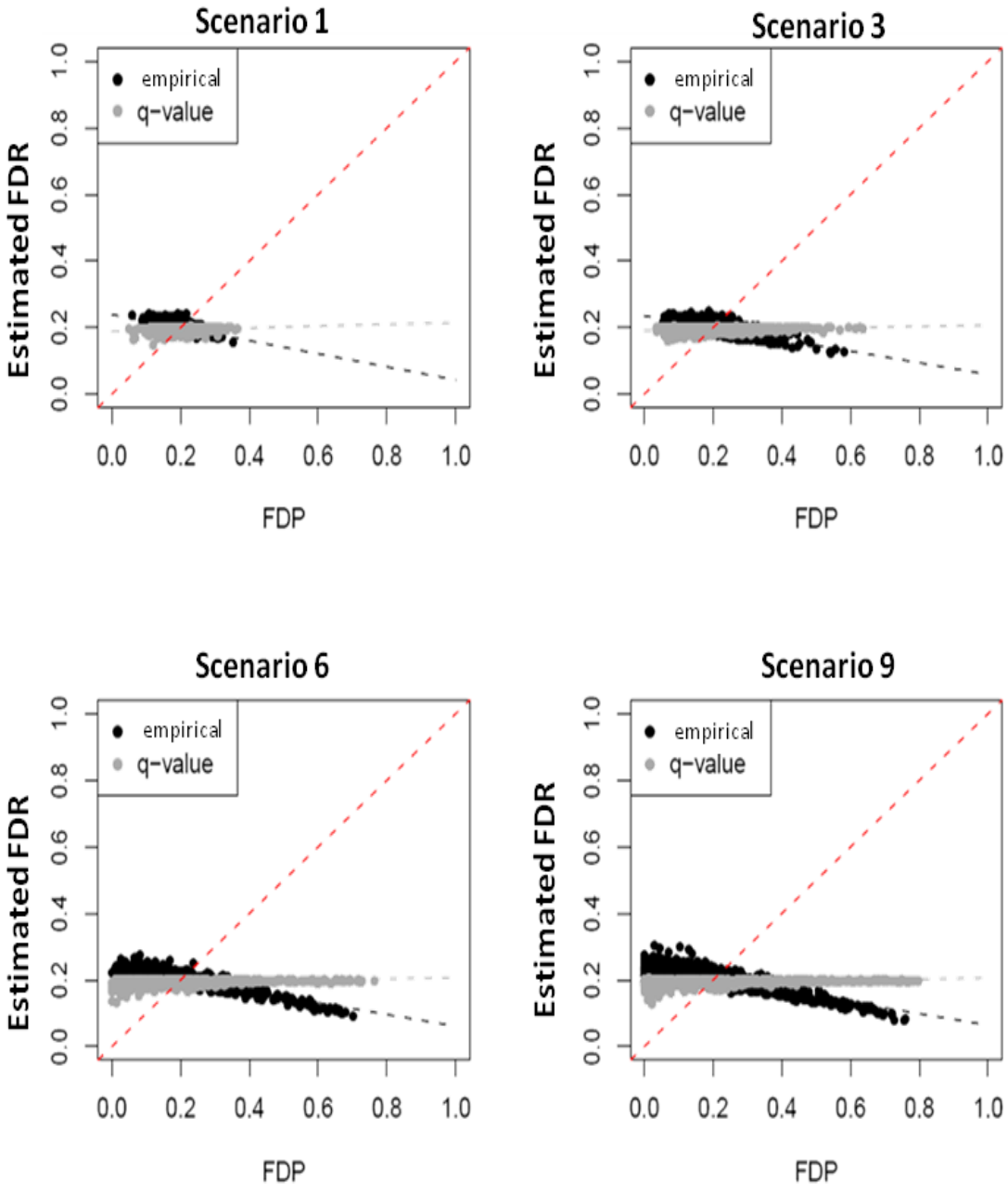


Figure 3.10: Estimated FDR versus the observed proportion of false positives with $t = 0,05$ (empirical FDR) and data-dependent threshold (q-value) for 4 levels of dependence

CHAPTER 4

CONDITIONAL APPROACH OF LARGE-SCALE MULTIPLE TESTING UNDER DEPENDENCE

Abstract Dependence between the responses is here modeled by a multiple factor structure which is exploited to define new test statistics. As data are independent conditionally on the factors, this framework allows to extend the results on error rates control and on π_0 estimation from the independent case to general dependence. This leads to less correlation among tests and shows large improvements of the power and stability of simultaneous inference. Conditional estimates are also deduced for FDR and π_0.

Sommaire

Introduction

There are two main issues in carrying out multiple testing procedures: estimating the proportion of true null hypotheses and controlling error rates. The classical approach is to perform simultaneous tests and to base the decision rule on the resulting p-values. As described in CHAPTER 2, most of procedures make independence assumptions on the p-values distributions. Taking into account dependence in large scale multiple testing has appeared to be a major concern. More particularly, even if error rates can still be controlled under some weaker assumption on the p-values distribution, each step of the procedures are impacted by the presence of dependence among tests.

Model (3.1) introduced in CHAPTER 3 assumes that dependence can be represented by a small set of random vectors Z associated to the information shared by the whole set of response variables:

$$Y_k \quad = \quad m_k(x) + Zb_k' + \varepsilon_k$$

where $E = [\varepsilon_1, \ldots, \varepsilon_m]$ has independent components.

In the first section, assuming such a covariance model for the conditional dependence of the responses given the predictors, we propose modified test statistics, based on improved estimators of the linear contrasts to be tested, by taking advantage of the factor structure. SECTION 4.2 presents a conditional approach to estimate π_0 and the FDR. SECTION 4.3 is dedicated to the presentation of a factor-adjusted multiple testing procedure showing large improvements on the usual methods, called FAMT.

4.1. Factor-adjusted data

4.1.1 Test statistics

Before presenting an improved multiple testing procedure which takes advantage of the factor structure, we start with a similar, yet much simpler, single-testing issue in which it can be assumed that the null hypothesis is true for some auxiliary covariates.

4.1.1.1 Likelihood-ratio test in the presence of covariates under H_0

Let us examine the following single-testing issue in the multivariate context: for a contrast of interest defined by the p-vector of coefficients c, our aim is to test the null hypothesis $H_0^{(m)} : c\theta_m = 0$ against

$H_1^{(m)} : c\theta_m \neq 0$ under the assumptions $H_0^{(j)} : c\theta_j = 0$, $\forall j \in [1; m-1]$. For example, this situation can be encountered in microarray data analysis, where Y_m is a gene expression of interest and $[Y_1; \ldots; Y_{m-1}]$ are expressions of so-called housekeeping genes. Such control genes, which expression has no biological reason to vary from an experimental condition to another, are introduced in some microarray experiments in order to estimate and to remove an eventual technological bias between microarrays. The above problem can be restated as a classical testing issue in a general linear model context. Let Y be the mn-vector obtained by concatenating the measurements of Y_j, $j \in [1; m]$, on the sample of size n. If $\theta = [\theta_1^0; \theta_1; \ldots; \theta_m^0; \theta_m]$ is the vector of unknown regression coefficients in the model relating Y to x, the test of $H_0^{(m)}$ can be viewed as a test for the significance of a particular linear combination of θ under linear restrictions which state the nullity of $c\theta_j$, $j \in [1; m-1]$. If the variance parameters are assumed to be known, it can be shown that the Likelihood-Ratio Test statistics resulting from application of the normal theory in this special case of general linear hypothesis testing is given by:

$$\tilde{T}_k = \frac{c'\hat{\theta}_m - c'\hat{B}^{(m-1)}\gamma^{(\hat{m-1})}}{\sqrt{\sigma^2_{m|m-1} c' S_x^{-1} c}} \tag{4.1}$$

where $B^{(m-1)}$ is the matrix containing the θ_j coefficients, $j \in [1; m-1]$, $\gamma^{(m-1)}$ is the vector of regression coefficients for $[Y_1; \ldots; Y_{m-1}]$ in the model relating Y_m to x and $[Y_1; \ldots; Y_{m-1}]$, and $\sigma^2_{m|m-1}$ is the residual variance of this model. Of course, if the conditional covariance between Y_m and $[Y_1; \ldots; Y_{m-1}]$, given x, is zero, T coincides with the classical Student's test statistics. Plugging-in the maximum likelihood estimators of $\gamma^{(m-1)}$ and $\sigma^2_{m|m-1}$ in expression (4.1) leads to an asymptotically optimal test statistics, which can show large improvements with respect to the classical Student's test.

4.1.1.2 General framework for high-dimensional data

Generally, apart from the special case mentioned above of control genes in microarray experiments, direct measurements of auxiliary covariates for which it could be assumed that H_0 is true are not available to improve large scale significance tests. However, in gene expression datasets for example, it can often be assumed that H_0 is true for a large fraction of variables, or in other words, that $\mathcal{M}_0$ is large. The method we propose hereafter consists in taking advantage of this unknown but large set of variables to derive new individual test statistics inspired by expression (4.1). A crucial issue in such an approach is the handling of the potentially huge size of $\mathcal{M}_0$. This can be addressed by the Factor Analysis modeling of the conditional variance of the variables. Assumption (3.2) can indeed be viewed as equivalent to the existence of latent factors $Z = (Z^{(1)}, \ldots, Z^{(q)})$, supposed to concentrate in a small dimension space the common information contained in the m responses: for $k = 1, \ldots, m$,

$$Y_k = m_k(x) + Zb_k' + \varepsilon_k \tag{4.2}$$

where b_k is the k^{th} row of B and $E = [\varepsilon_1; \ldots; \varepsilon_m]$ is a random $m-$vector, independent of Z, with mean 0 and variance-covariance Ψ. The kernel dependence ZB' is independent from the covariates

x. We consider the data centered with respect to this dependence kernel:

$$\tilde{Y}_k = Y_k - Zb'_k = m_k(x) + \varepsilon_k \tag{4.3}$$

Application of expression (4.1) using the factors as covariates results in the following factor-adjusted test statistics:

$$\tilde{T}_k \quad = \quad s(\tilde{Y}_k) = s(Y_k - Zb'_k) \tag{4.4}$$

The factor-adjusted p-values $\tilde{p}_k = 1 - F_0\big[s(Y_k - Zb'_k)\big] = 1 - F_0\big[s(m_k(x) + \varepsilon_k)\big]$, obtained using on the residual vector $Y_k - Zb'_k$ the same individual testing method as for the p-values p_k on the raw data Y_k, are independent. The following proposition gives the probability distribution function of these factor-adjusted p-values:

Proposition 4.1.1. *Let $\tilde{G}_k(t) = \mathbb{P}(\tilde{p}_k \leq t)$ denote the probability distribution function of the kth factor-adjusted p-value $\tilde{p}_k$. Assume that the density functions ϕ_k of the independent error terms ε_k in model 2.1 only differs by a scaling parameter ψ_k and have the same standardized density function as the random errors in model (3.1): $\phi_k(x) = \varphi(x/\psi_k)/\psi_k$. Then, for $k \in \mathcal{M}_0$, $\tilde{G}_k(t) = G_0(t) = t$ and for $k \notin \mathcal{M}_0$, $\tilde{G}_k(t) = G_1(t; m_k(x)/\psi_k)$.*

Proof. The probability distribution function of the factor-adjusted p-values is given by:

$$\tilde{G}^k(t) = \int \mathbb{1}_{[1-F_0(s(m_k(x)-\varepsilon_k))\leq t]}\phi_k(\varepsilon_k)d\varepsilon_k$$

Due to the scale-invariance of the test statistics and the identical distribution of the standardized residuals ε_k/ψ_k,

$$\tilde{G}^k(t) = \int \mathbb{1}_{[1-F_0(s(m_k(x)/\Psi_k-u_k))\leq t]}\varphi(u_k)du_k = G^k(t;\tilde{\tau}_k)$$

Moreover, $\tilde{\tau}_k = m_k(x)/\psi_k = \tau_k\sigma_k/\psi_k$. □

Note that ψ_k^2 is the conditional variance of ε_k given the factors, which implies that $\sigma_k > \psi_k$. Therefore, the non-centrality parameter of the non-null distribution, or equivalently the power of the individual tests, is larger for the factor-adjusted p-values than for the corresponding raw p-values.

4.1.1.3 Linear model

The present section focuses on t-tests because they are of major interest in various applied situations but the general conclusions are valid for other types of tests such as Fisher's analysis of variance tests for example.

Hereafter, the least-squares estimator of θ_k is denoted $\hat{\theta}_k$: $c'\hat{\theta}_k - c'\theta_k$ is normally distributed with expectation 0 and variance $\sigma_k^2 n^{-1} c' S_x^{-1} c$, where σ_k is the conditional standard deviation of $Y^{(k)}$ given x and S_x is the sample variance-covariance matrix of the explanatory variables x. Furthermore, for $k \neq k'$, $\mathrm{Cor}(c'\hat{\theta}_k, c'\hat{\theta}_k) = \rho_{kk'}$, where $\rho_{kk'}$ is the conditional correlation between Y_k and $Y_{k'}$ given x.

As mentioned above, the present section focuses on test statistics defined as normalized estimations of the linear contrasts: $T_k = \sqrt{n}c'\hat{\theta}_k/(\sigma_k\sqrt{c'S_x^{-1}c})$. This test statistic has the following properties:

PROPOSITION 4.1.2.

$$\begin{aligned} \mathbb{E}(T_k) &= \tau_k = \frac{\sqrt{n}c'\theta_k}{\sigma_k\sqrt{c'S_x^{-1}c}} \\ \mathbb{V}(T_k) &= 1 \\ \mathrm{Cov}(T_k, T_{k'}) &= \rho_{kk'}, k \neq k' \end{aligned}$$

Note that τ_k equals 0 if $k \in \mathcal{M}_0$

PROPOSITION 4.1.2 shows that the correlation structure between the test statistics is directly inherited from the correlation between the response variables. This property is generally not true for other types of tests. Therefore, specific relationships between both correlation structures should be taken into account in order to adapt the following results to other testing procedures.

The following proposition gives the conditional distribution of $T = (T_1; \ldots; T_m)$, given the factors.

PROPOSITION 4.1.3. *Conditionally on Z, T is normally distributed with, for $k = 1, \ldots, m$, $\mathbb{E}(T^{(k)} \mid Z) = \tau_k + b_k'\tau_Z/\sigma_k$, where τ_Z is the $Q-$vector defined by $\tau_Z = \sqrt{n}c\hat{\theta}_z/\sqrt{c'S_x^{-1}c}$ and $\hat{\theta}_z$ denotes the least-squares estimator of the $p \times Q$ matrix of the slope coefficients in the multivariate regression of Z onto the explanatory variables x. Moreover, $\mathbb{V}(T|Z) = diag(\psi_k^2/\sigma_k^2)$. Note that τ_Z is normally distributed with mean 0 and variance $\mathbb{I}_Q$*

Proof. The conditional independence between the test statistics is inherited from the conditional independence between the responses. The conditional expectation and variance are also deduced from the conditional moments of $\hat{\theta}^{(k)}$ given the factors: $\mathbb{E}(\hat{\theta}_k|Z) = \theta_k + \hat{\theta}_z b_k$ and $\mathbb{V}(\hat{\theta}_k|Z) = (\psi_k^2/n)S_x^{-1}$. □

Application of expression (4.1) using the factors as covariates results in the following factor-adjusted test statistics:

$$\tilde{T}_k = s(Y_k - Zb_k') = \frac{c'\hat{\theta}_k - \hat{b}_k'\hat{\theta}_z c}{\hat{\psi}_k\sqrt{c'S_x^{-1}c}} \tag{4.5}$$

where $\hat{b}_k$ is the k^{th} row of the matrix $\hat{B}$ of estimated loadings, $\hat{\psi}_k^2$ is the k^{th} diagonal element of the matrix $\hat{\Psi}$ of estimated uniquenesses and $\hat{\theta}_z$ is the least-squares estimator of the $p \times Q$ matrix of slope coefficients in the multivariate regression model relating the estimated factors $\hat{Z}$ and the explanatory variables x.

Factor-adjusted test statistics $\tilde{T}_k$ are defined as conditionally centered and scaled versions of the classical t-statistics T_k:

Définition 4.1.1.

$$\tilde{T}_k = \frac{\sigma_k}{\psi_k}\Big[T_k - \frac{b'_k}{\sigma_k}\tau_Z\Big].$$

The following proposition gives the distribution of $\tilde{T} = [\tilde{T}_1; \ldots; \tilde{T}_m]$.

Proposition 4.1.4. *Under assumption of a decomposition of the covariance matrix as in* (3.2)*, $\tilde{T}$ is normally distributed with, for all $k \in [1; m]$, $\mathbb{E}(\tilde{T}_k) = \tau_k/\sqrt{1-h_k^2}$, where $h_k^2 = b_k b'_k/\sigma_k^2$ is the communality of Y_k. Moreover, $\mathbb{V}(\tilde{T}) = \mathbb{I}_m$.*

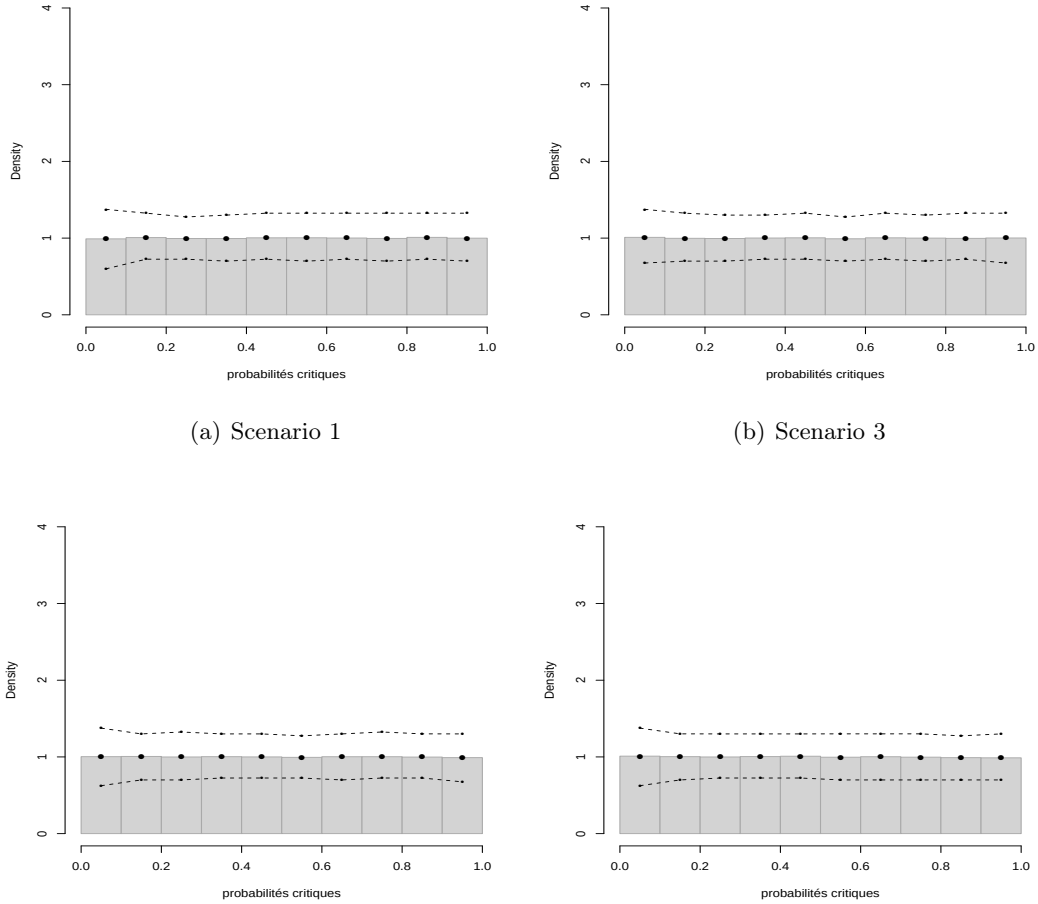
Proof. The factor-adjusted test statistics are linear functions of the residual terms $Y_k - Zb'_k$, which are, under assumption 3.1), normally distributed. Therefore, $\tilde{T}$ is also normally distributed. Moreover, it results from PROPOSITION 4.1.3 that $\mathbb{E}(\tilde{T}_k) = \frac{\sigma_k}{\psi_k}\tau_k = \tau_k/\sqrt{1-h_k^2}$ and $\mathbb{V}(\tilde{T}) = \mathbb{E}[\mathbb{V}(\tilde{T}|Z)] + \mathbb{V}[\mathbb{E}(\tilde{T}|Z)] = (\sigma_k^2/\psi_k^2)\mathbb{E}[\mathbb{V}(T|Z)] = \mathbb{I}_m$ □

The non-centrality parameter of $\tilde{T}_k$ being always larger than τ_k, the variable-by-variable power of the factor-adjusted tests are larger than for the t-tests. Furthermore, this non-centrality parameter, and consequently the power of the factor-adjusted tests, are increasing functions of the communality h_k^2, which confirms the idea that the multiple testing procedure can be improved by a correction of the individual test-statistics regarding their contribution to the common variability across variables. On the contrary, if the k^{th} variable does not contribute to the factor structure, $b_k = 0$ and $\tilde{T}_k$ coincides with the usual t-test T_k.

As proposed in CHAPTER 5, estimated factor-adjusted test statistics $\tilde{T}$ are obtained by plugging the ML estimates of the factor model's parameters in the test statistic defined in 4.1.1. As these estimators of the variance parameters are consistent, this does not affect the asymptotic distribution of the factor-adjusted test statistics. By analogy with the classical situation, we propose to take into account the effect of estimating the variance parameters in small-sample conditions by approximating the null distribution of $\hat{\tilde{T}}_k$ by a Student distribution with df_r degrees of freedom (see (5.9)).

P-values distribution under H_0 Factor-adjusted test statistics and associated p-values are first calculated in the simulation study of EXAMPLE 3. FIGURE 4.1 shows the mean distribution of adjusted p-values for four scenarios od dependence, as in SECTION 3.1. In each case, p-values distribution is uniform in average. In opposition to the distribution of usual p-values (FIGURE 3.1), there is low variability around this uniform distribution, whatever the level of dependence, as suggested by the 95% confidence interval.

FIGURE 4.2 shows the p-values distribution for two datasets from scenario 9 as in SECTION 3.1, for factor-adjusted data and for raw data (in blue). For highly correlated data, p-values distribution shows low depature from the theorical null distribution, with respect to the distribution of usual p-values.



(a) Scenario 1

(b) Scenario 3

(c) Scenario 6

(d) Scenario 9

Figure 4.1: P-values distribution - Mean histogram over 1000 simulations - dotted lines: 95% confidence interval

4.1.2 Estimation of the proportion of true null hypotheses

Factor-adjusted test statistics and associated p-values are now calculated in the simulation study of EXAMPLE 4. π_0 estimations are then performed with the different methods previously introduced. The results are presented for two of them: empirical estimator with bootstrap choice of λ [Storey et al., 2004] and considering the density estimation assuming it is a convex decreasing function [Langaas et al., 2005]).

FIGURE 4.3 shows the estimated proportion of null hypotheses, along with the Δ criterion, introduced in SECTION 3.2. The local bias induced by dependence observed previously when using usual p-values is strongly reduced.

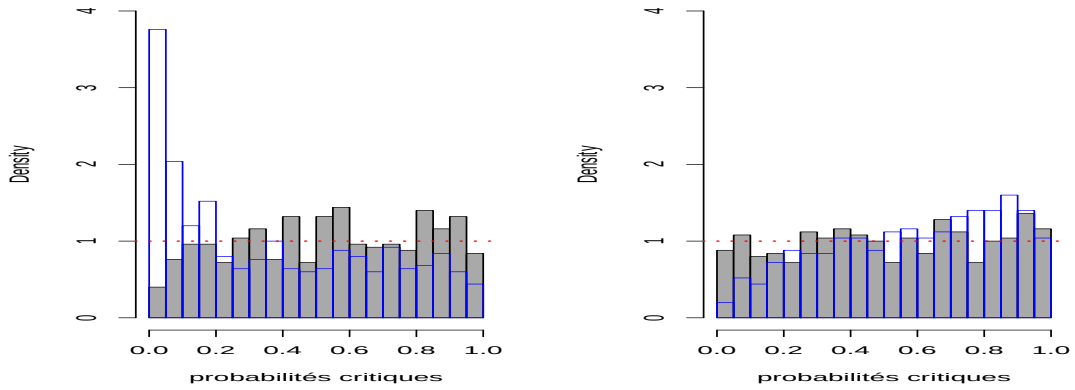


Figure 4.2: P-values distribution under H_0: examples of two datasets - scenario 9. In blue: usual p-values distribution from the same datasets (see FIGURE 3.2).

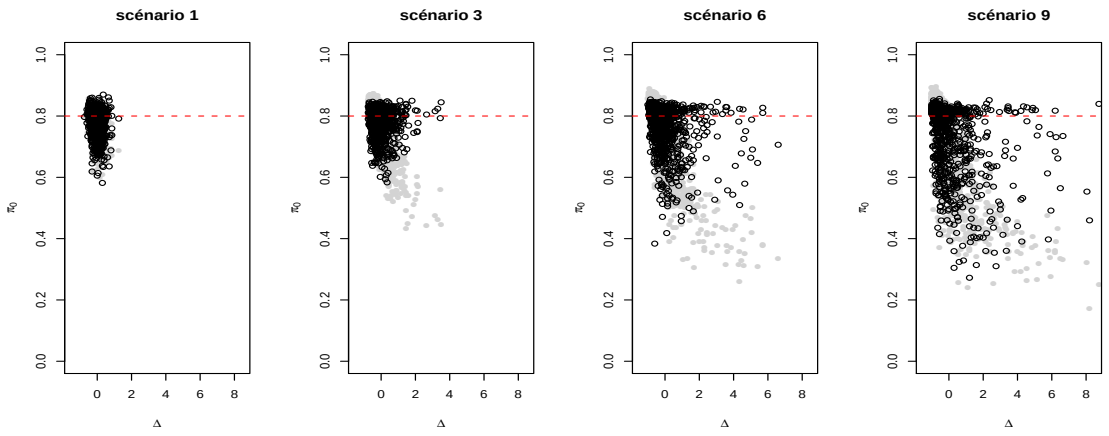
4.1.3 FWER and FDR control

PROPOSITION 3.3.1 shows that correlation modifies the distribution of V_t by increasing its dispersion, leaving its expectation unchanged. Therefore, correlation shall have an important impact on the tail of the distribution, which is directly involved in the calculation of the error rates.

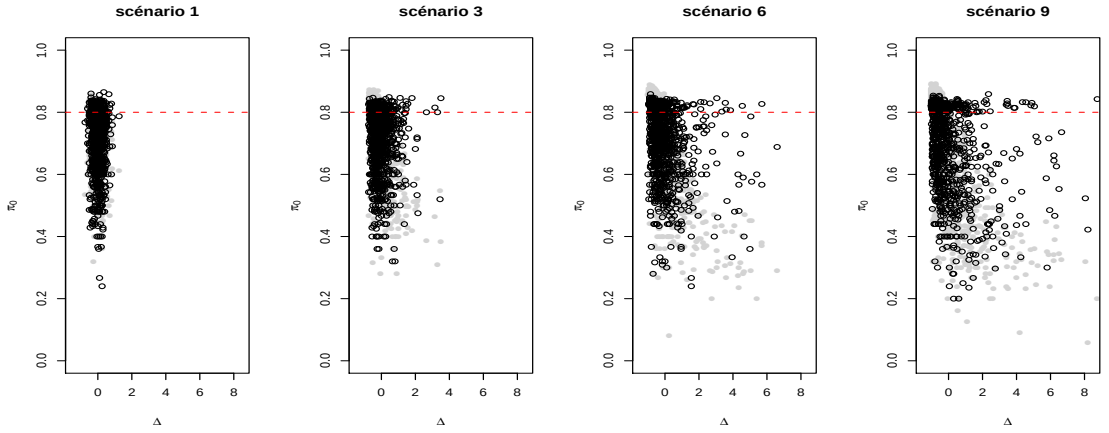
FWER Hsu [1992] gives an explicit expression of the FWER in the case of a one-factor approximation of the correlation between test statistics. This approximation is shown to be appropriate for the most usual experimental designs in univariate ANOVA situations where a set of linear contrasts are tested simultaneously. In the case of differential analysis for microarray data, the significance of only one contrast is tested for each response variable and multiplicity of the tests is due to the number of responses. Here, correlation between test statistics does not depend on the design but is directly inherited from the correlation between the responses. The one-factor approximation can therefore be too simple regarding the complexity of the dependence between gene expressions. Moreover, PROPOSITION 3.3.1 gives an exact expression for the first two moments of V_t which only involves a 2-d integration for the probability function of the bivariate normal distribution, whereas the extension of Hsu [1992]'s formula for the FWER to a larger number of factors leads probably to numerically untractable expressions.

The Sidak procedure [Sidak, 1967] is now applied on factor-adjusted p-values from the datasets of the simulation study presented in EXAMPLE 4. Implementation is done using the `R` package `multtest` [Pollard et al.]. TABLE 3.3 displays the frequencies of the number of false-positives V_t and the estimated FWER using the simulated datasets. It shows that the procedure based on the factor-adjusted test statistics controls the FWER at a level close to $\alpha = 0,05$, whatever the level of dependence.

FIGURE 4.4 reproduces multiple boxplots of the distributions of the non discovery proportion $NDP_t = \#\{k \notin \mathcal{M}_0, H_0^{(k)} \text{ not rejected}\}/m1$. It shows that the variability of the NDP_t distribution remains



(a) Density estimation assuming it is a convex decreasing function



(b) Empirical estimator with bootstrap choice of λ

Figure 4.3: Estimation of π_0 along with the Δ criterion that characterizes the density of null p-values around 0, from factor-adjusted p-values (in black) and usual p-values (in grey) - $\pi_0 = 0,80$

constant along with the proportion of common variance. In gray is recalled the distribution of NDP_t obtained for the usual t-tests' p-values. Moreover, the mean NDP_t, which can be viewed as a type-II error rate, decreases markedly with the factor-analytic procedure with respect to procedures based on usual t-tests. A striking property of our method, compared with the procedures based on t-tests, is the very important improvement of the global power (1-type-II error rate) of the multiple testing procedure in presence of dependence.

FDR FIGURE 4.5 shows multiple boxplots of the distributions of the False Discovery Proportion and the Non-Discovery Proportion. By comparison with FIGURE 3.9 for a BH procedure applied on classical t-tests (duplicated in gray on FIGURE 4.5), the distribution of the FDP is clearly stabilized: using the usual t-tests, the standard deviations of the FDP reaches up to four times the standard deviations obtained under independence whereas it remains controlled at almost the same level using the factor-adjusted method.

scenario	0	1	2	3	4	5	6	7	8	9
common var.(%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
counts										
0	957	954	965	951	960	966	967	963	960	955
1	42	45	34	49	39	32	30	35	36	34
2		1	1		1	2	3	1	3	9
3	1							1	1	1
4										1
FWER										
$\#\{V_t > 0\}/m$	4,3	4,6	3,5	4,9	4,0	3,4	3,3	3,7	4,0	4,5

Table 4.1: Estimated FWER along with the proportion of common variance (%) and observed counts of false-positives - Sidak procedure [Sidak, 1967] applied to factor-adjusted p-values

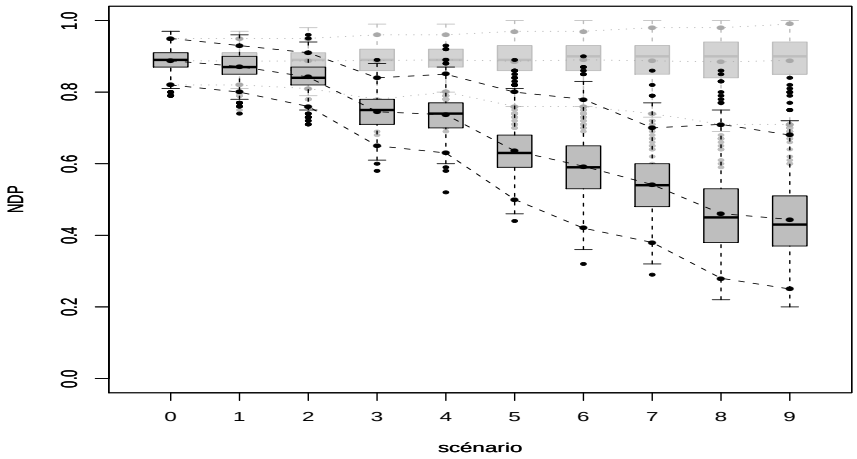
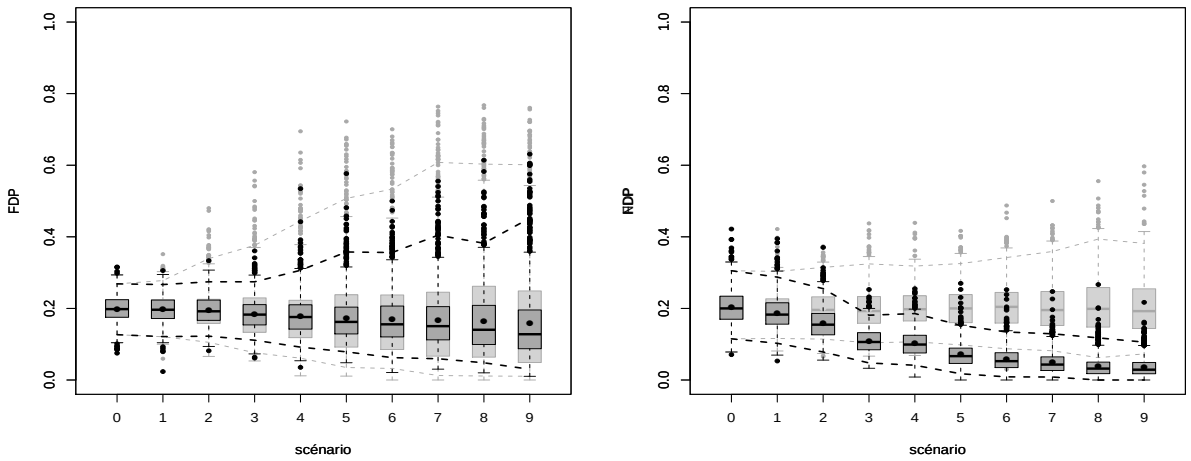


Figure 4.4: Non Discovery Proportion along with the level of dependence - Sidak procedure applied on factor-adjusted p-values. In gray: same procedure applied on usual t-tests p-values (FIGURE 3.8)

Another striking property of our method, as already noticed in the previous SECTION on FWER, is the very important improvement of the global power of the multiple testing procedure compared with the BH procedure based on t-tests, illustrated by FIGURE 4.5(b). This result probably illustrates the idea that dependence between the responses should not just be seen as a nuisance for controlling the FDR but also as a support to provide improved estimation of the effects of covariates.



(a) FDP (b) NDP

Figure 4.5: Distributions of FDP and NDP along with the proportion of common variance, using the factor-adjusted test statistics - BH procedure. In gray: same thresholding procedure applied to usual t-tests (FIGURE 3.9)

4.2. Conditional estimators

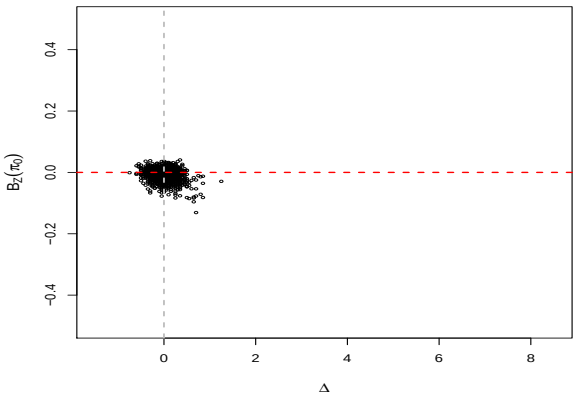
4.2.1 Conditional estimation of π_0

It is noticeable in the proof of PROPOSITION 3.2.2 that the variability of the empirical estimator of π_0 due to dependence essentially comes from the variance of the conditional bias given the factors:

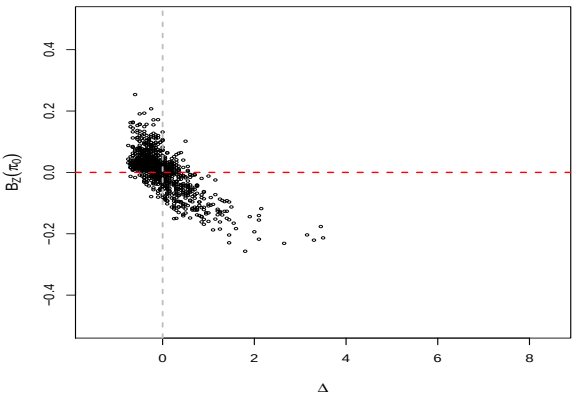
$$\begin{aligned}\mathbb{E}\left(\hat{\pi}_0(\lambda)|Z\right) &= \frac{\mathbb{E}\left(W_\lambda|Z\right)}{m(1-\lambda)} = \frac{\sum_{k\in\mathcal{M}}\left(1-G^Z(k;\lambda)\right)}{m(1-\lambda)} = \frac{\sum_{k\in\mathcal{M}}\left([G(k;\lambda)-G(k;\lambda)]+[1-G^Z(k;\lambda)]\right)}{m(1-\lambda)} \\ &= \frac{\sum_{k\in\mathcal{M}}\left(1-G(k;\lambda)\right)}{m(1-\lambda)} + \frac{\sum_{k\in\mathcal{M}}\left(G(k;\lambda)-G^Z(k;\lambda)\right]\right)}{m(1-\lambda)} \\ &= \mathcal{B}(\hat{\pi}_0) + \frac{\sum_{k\in\mathcal{M}}\left(G(k;\lambda)-G^Z(k;\lambda)\right]\right)}{m(1-\lambda)} \qquad (4.6)\end{aligned}$$

$\mathcal{B}(\hat{\pi}_0) = (1-\pi_0)\frac{1-\bar{G}_1(\lambda)}{1-\lambda}$ is the unconditional bias of $\hat{\pi}_0(\lambda)$ and $\mathcal{B}_Z(\hat{\pi}_0) = \frac{\sum_{k\in\mathcal{M}}\left(G(k;\lambda)-G^Z(k;\lambda)\right]\right)}{m(1-\lambda)} = \frac{\bar{G}(Z,\lambda)}{(1-\lambda)}$, is a random variable with mean 0.

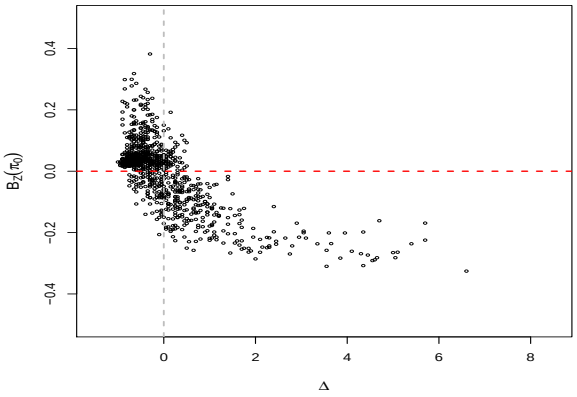
For each dataset in the simulation scenarios introduced in EXAMPLE 4, the random part of $\mathcal{B}(\hat{\pi}_0)$ is calculated with the optimal threshold λ previously used to obtain the estimated values displayed in FIGURE 3.3(b) and the classical Thompson method [Mardia et al., 1979] to estimate the factors Z. In FIGURE 4.6, the graphical representation of this random term along with the Δ criterion exhibits the same kind of patterns as observed on FIGURE 3.3(b). This suggests that $\bar{\Delta}G^Z(k;\lambda)/(1-\lambda)$ captures the local bias identified in SECTION 3.2 as due to dependence.



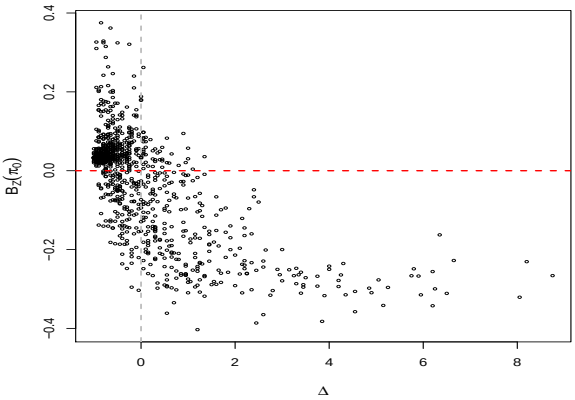
(a) Scenario 1



(b) Scenario 3



(c) Scenario 6



(d) Scenario 9

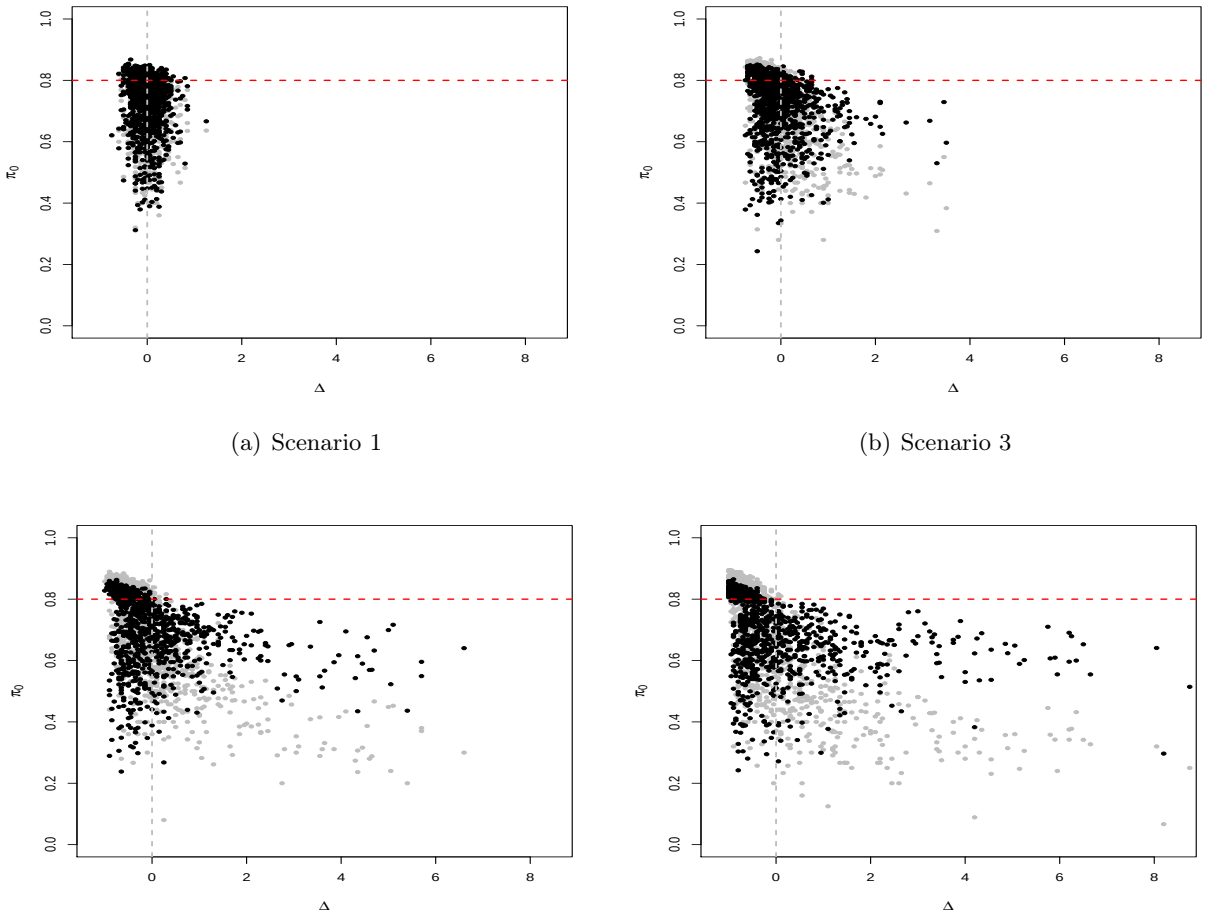
Figure 4.6: Graphical representation of $\mathcal{B}_Z(\hat{\pi}_0)$ along with the Δ criterion for different levels of dependence.

We therefore propose a conditionally bias-corrected version of $\hat{\pi}_0(\lambda)$:

Définition 4.2.1.

$$\check{\pi}_0(\lambda) = \hat{\pi}_0(\lambda) - \mathcal{B}_Z(\hat{\pi}_0)$$

where $\mathcal{B}_Z(\hat{\pi}_0) = \bar{\Delta} G^Z(k;\lambda)/(1-\lambda)$.



(a) Scenario 1

(b) Scenario 3

(c) Scenario 6

(d) Scenario 9

Figure 4.7: Conditional estimate $\check{\pi}_0$ along with the Δ criterion for different levels of dependence

4.2.2 Conditional estimation of FDR

The following expression gives the expectation of V_t conditionally on the factors.

$$\mathbb{E}(V_t|Z) \quad = \quad \sum_{k \in \mathcal{M}_0} \mathbb{P}\left(p_k \leq t|Z\right) = \sum_{k \in \mathcal{M}_0} G^Z(k,t) \tag{4.7}$$

$$\tag{4.8}$$

The above expression of $\mathbb{E}(V_t|Z)$ is now used to define a conditional estimate FDR_t^Z of the FDR, by analogy with the proposition made by Efron [2007], who defines FDR_t^A as $\mathbb{E}(V_t|A)/R_t$, where A is a random variable which value essentially differs according to the amount of dispersion among the correlations between the test statistics.

Définition 4.2.2 (FDR conditionnel)**.**

$$\begin{aligned} FDR_t^Z &= \frac{\mathbb{E}(V_t|Z)}{R_t} = \frac{\sum_{k\in\mathcal{M}_0}(G^Z(k,t))}{R_t} \\ &= \frac{m_0 t}{R_t} + \frac{\sum_{k\in\mathcal{M}_0}(G^Z(k,t))}{R_t} - \frac{m_0 t}{R_t} \\ &= \widehat{FDR}_t.\left[1 + \frac{\sum_{k\in\mathcal{M}_0}(G^Z(k,t)-t)}{m_0 t}\right] \end{aligned}$$

where $\widehat{FDR}_t = \frac{m_0 t}{R_t}$ *is the empirical estimator of FDR.*

The above expression of FDR_t^Z appears to be very close toEfron [2007]'s conditional FDR_t^A estimate:

$$FDR_t^A = \widehat{FDR}_t.\left[1 + A\frac{u_t\phi(u_t)}{\sqrt{2(1-\Phi(u_t))}}\right]$$

Both conditional FDR estimates are defined as corrections of the unconditional estimate, accounting for the correlation among the test statistics through factors for FDR_t^Z or through A for FDR_t^A.

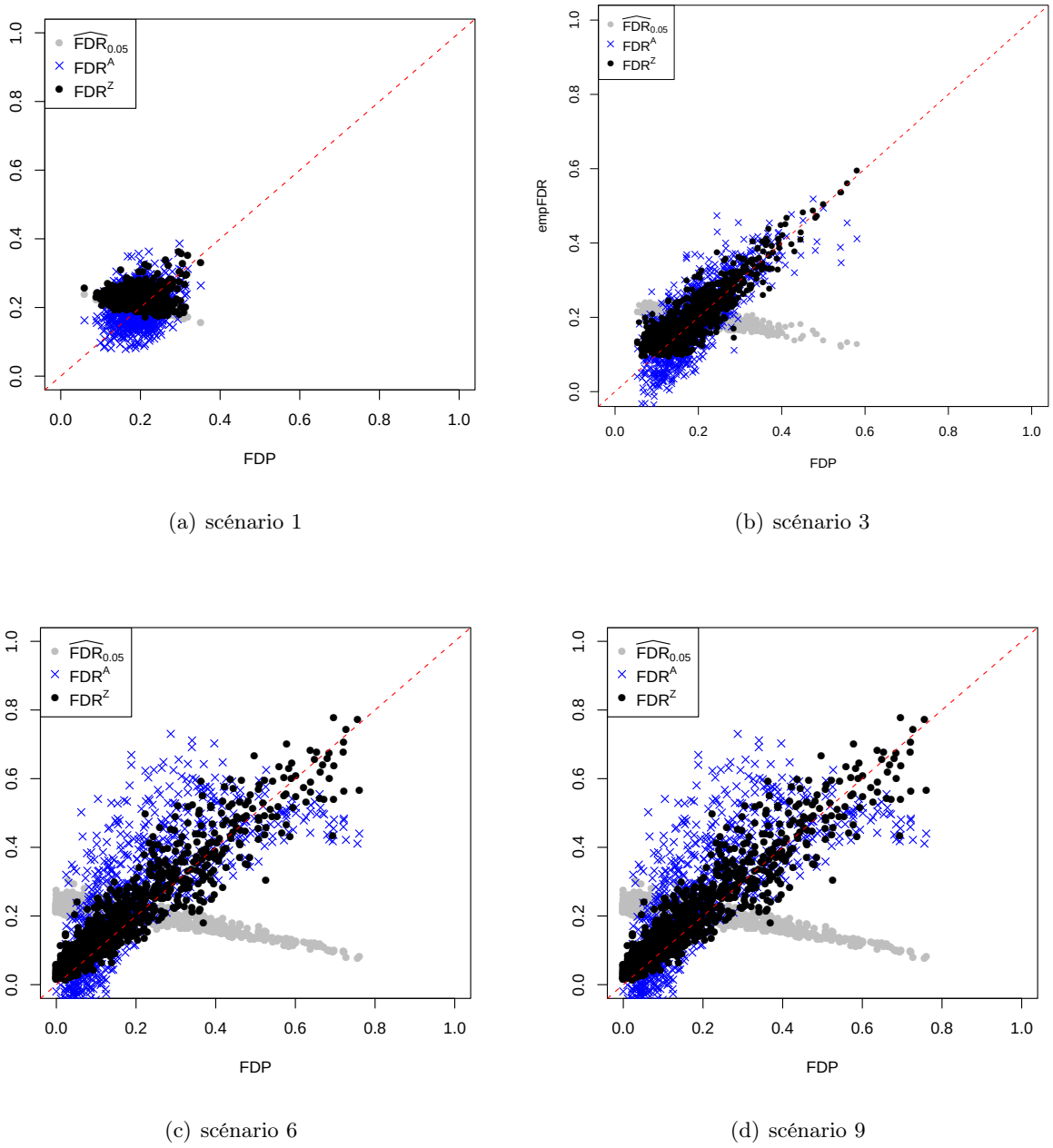
If m_0 is unknown, it is replaced by m (as in the expression of FDR_t^A given by Efron [2007]) or estimated (see SECTION 2.2). It is also interesting to note that estimation of both conditional FDR involves a preliminary approximation of the null distribution of the test statistics: this point is handled in Efron [2007] by focusing on the Z-scores which absolute value is less than 1 to estimate A and a rough estimation of $\mathcal{M}_0$ is also proposed later in SECTION 4.3 by $\widehat{\mathcal{M}}_0 = \{k : p_k \geq 0,05\}$.

The conditional estimates of the FDR are now compared by means of simulations involving 10 scenarios of conditional correlation matrices which differ by their proportion of common variance (EXAMPLE 4). For each dataset and for $t = 0,05$, the conditional estimate FDR_t^Z is estimated, together with FDR_t^A (following the suggestions in Efron [2007]) and the unconditional estimate $\widehat{FDR}_t$. To avoid discussions about the impact of the estimation of m_0 in this comparative study, $m_0 = 400$ is supposed to be known. For each scenario, TABLE 4.2 gives the regression coefficients between the observed false discovery proportion FDP_t and each of the estimates. Results for scenarios 1, 3, 6 and 9 are illustrated on FIGURE 4.2.2 by plots of the FDR estimates versus FDP_t. As already observed

scenario	0	1	2	3	4	5	6	7	8	9
common var. (%)	0	4,65	15,56	28,56	40,31	50,76	57,90	65,45	70,09	75,19
$\widehat{FDR}_t$	-0,197	-0,196	-0,189	-0,176	-0,187	-0,171	-0,175	-0,182	-0,169	-0,171
FDR_t^A	0,155	0,335	0,748	0,966	1,062	1,129	1,139	1,09	1,134	1,180
FDR_t^Z	-0,197	0,025	0,578	0,815	0,844	0,886	0,907	0,889	0,915	0,913

Table 4.2: Regression coefficients between FDR estimates and FDP - $t = 0,05$

by Efron [2007], the unconditional estimate $\widehat{FDR}_t$ is negatively correlated with the observed FDP_t, which can result in strongly misleading estimations especially when FDP_t is high. FIGURE 4.2.2 shows that this concern is particularly clear for large fractions of shared variance. For small fraction of shared variance, FDR_t^Z suffers from the same problem, essentially because the number of factors



(a) scénario 1

(b) scénario 3

(c) scénario 6

(d) scénario 9

Figure 4.8: Estimated FDR versus the observed FDP with $t = 0,05$, for different simulation scenarios

is most often estimated by zero (see FIGURE 5.3), in which case $FDR_t^Z = \widehat{FDR}_t$. From scenario 3 to 10, when the factor structure is clearer, both FDR_t^Z and FDR_t^A are positively correlated with FDP_t, confirming the importance of accounting for the correlation between test statistics in multiple testing procedures. FIGURE 4.2.2 and TABLE 4.2 both suggest that, for large values of common variance, FDR_t^Z provides much more precise estimations of FDP_t than FDR_t^A. The important dispersion appearing in the distribution of FDR_t^A is also revealed in the simulation experiment reported in Efron [2007].

4.3. Factor Analysis for Multiple Testing: FAMT

This section is dedicated to the presentation of a procedure called Factor Analysis for Multiple Testing (FAMT), from the name of the `R` package. This package is available on the R website (`http://cran.r-project.org/`). The package has also its own website (`http://famt.free.fr`). Next chapter is organiased as a tutorial. This section describes the different steps of the procedure, built on the factor-adjusted approach presented in the present chapter.

A factor-adjusted testing procedure is carried out on the datasets simulated in the 10 scenarios of dependence introduced in section 3:

1. **Estimation of $\mathcal{M}_0$** Classical t-tests are calculated for each variables and a first estimation of $\mathcal{M}_0$ is deduced by taking the indices of the p-values exceeding $\alpha = 0.05$;
2. **Choice of the number of factors** The number of factors is estimated by minimisation of the criterion given in PROPOSITION 5.4.1 in the following chapter;
3. **Estimation of the model's parameters**
4. **Calculating factor-adjusted p-values** The factor-adjusted test statistics $\tilde{T}$ and the corresponding p-values are calculated;
5. **Up-dating the estimation of $\mathcal{M}_0$** The estimation of $\mathcal{M}_0$ is updated by taking the indices of the factor-adjusted p-values exceeding $0,05$.
   STEPS 2 to 4 are performed again with this new estimation of $\mathcal{M}_0$;
6. **π_0 estimation** The estimation of this parameter is in general a crutial step of multiple testing procedures. It is here estimated according to one of the method introduced in SECTION 2.2, applied to the factor-adjusted p-values.
7. **Decision rule** A BH thresholding procedure at level $\alpha = 0,05$ is applied to the factor-adjusted p-values to decide which null hypotheses are rejected. The BH procedure is improved by estimating m_0 according to one of the method introduced in SECTION 2.2, applied to the factor-adjusted p-values.

Conclusion

This chapter introduces a general framework for high-dimensional multiple testing procedures. This framework makes the most of the modeling of dependence through a low dimensional set of latent variables, as described in the previous chapter.

As data are independent conditionally on the factors, this framework allows to extend the results on error rates control in the independent case to general dependence. Factor-adjusted data can be used to apply multiple testing procedures initially derived for independent variables. This setting leads

to less correlation among tests and shows large improvements of power and stability of simultaneous inference.

The impact of dependence on π_0 estimation is markedly reduced, as the variability of null p-values under dependence becomes stable.

Besides, this modeling of dependence allows to derive the exact expression of the variance of both false positives and π_0 (CHAPTER 3). Considering these properties, we introduce conditional estimates for FDR and π_0, that appear as corrections for the respective empirical estimates under dependence.

Finally, a procedure called FAMt is introduced. It describes the different steps implementing the proposed methodology for large-scale multiple testing under dependence. This procedure is illustrated thanks to the tutorial of the associated `R` package (`FAMT`). Beforehand, CHAPTER 5 focuses on model parameters estimation and on the differents methods to choose the number of factors to be included in the model.

CHAPTER 5

FACTOR ANALYSIS IN HIGH-DIMENSIONAL DATA

Abstract This chapter deals with the parameter estimation of the latent variable model introduced in the previous chapters. Two methods are discussed, and we opt for a Maximum Likelihood approach. An algorithm that implements this method in the high-dimensional setting is presented in this chapter. Moreover, a crucial step in conducting Factor Analysis is to determine the optimal number of factors. Different methods are discussed and a strategy suited to the multiple testing context is proposed.

Sommaire

5.1. Introduction

In all applied sciences that deal with large quantities of data, such as economics, marketing or even life sciences, we often need to determine a smaller set of synthetic variables that could explain the original set. Moreover, the dimensionality of the data may make one to assume that simple dependence structures, for instance independence between blocks of variables, or simple within-block correlation patterns, can be far from the observed structure. Factor Analysis aims at dealing with such issues: originally developed for reducing a large number of observed variables in term of a meaningful small set of latent variables, also called common factors, this method appears as a nice tool to investigate the dependence structure between a large set of variables. It describes the covariance relations between observed variables in terms of a meaningful small set of latent variables, called "common factors". It is an analytic tool used for many years in economics, social sciences and psychometrics, originally in the field of intelligence research [Spearman, 1904], and has only appeared recently in the study of the dependence structure in high dimensional datasets provided by microarray technology [Pournara and Wernisch, 2007, Kustra et al., 2006].

In Factor Analysis, we assume a model for the correlation matrix between the m observed variables to describe the data dependence structure that links factors, unobserved latent variables, to the original data. The influence of the latent variables can be identified considering factors that are common to the whole set of variables, and factors that are specific to each ones. The $n \times m$-data matrix $Y = [Y_1, ..., Y_m]$ can the be split into two parts, associated to the common and to the specific information as in (3.1):

$$Y_k = m_k(x) + b_k Z + \varepsilon_k$$

or in a matrix form:

$$Y = M_x + ZB' + E$$

Where B is the $m \times Q$-matrix of loadings, $Z = [Z_1, ..Z_Q]$ is the $n \times Q$-matrix of common factors, and $E = [E_1, ..E_m]$ is the $n \times m$-matrix of specific factors. Some constraints are assumed in the FA framework:

Hypothèse 2.

1. *Common factors Z are centred, scaled and independent random variables: $Z \sim \mathcal{N}(0; \mathbb{I}(Q))$*
2. *Specific factors E are centred and independent random variables: $E \sim \mathcal{N}(0; \Psi)$, where Ψ is diagonal*

3. Common and specific factors are uncorrelated:

$$\text{Cov}(Z_q, E_i) = 0, q = 1..Q, i = 1..m$$

Considering model (3.1), $\mathbb{V}(Y_k) = \sum_{q=1}^{Q} \left(b_{kq}^2\right) + \psi_k^2$, where b_{kq} is the loading for variable k and common factor q. $h_k^2 = \sum_{q=1}^{Q} \left(b_{kq}^2\right)$ is called "communality", it is the kth diagonal element of BB' and represents the part of variability of Y_k explained by the common factors. The kth diagonal element of Ψ, ψ_k^2, is the specific variability. We also deduce from (3.1) that $\text{Cov}(Y_k, Y_{k'}) = \sum_{q=1}^{Q} b_{kq} b_{k'q}$. We can the define a model for the covariance matrix Σ:

$$\Sigma = BB' + \Psi \tag{5.1}$$

A recurring question in Factor Analysis is to determine the minimal number of observation necessary to conduct the study and to get stable estimates [Preacher and MacCallum, 2002]. Many rules of thumb can sometimes be found in the literature, such as a minimum of 500 or a ratio of 10 times as many subjects as variables. Actually, the issue mostly depends on the studied structure, and more particularly on the communalities. Indeed, smaller sized sample can be considered if both the ratio m/Q and the communalities are high. Although the more observations the better, it is even more important to have high communalities and a small number of factors with respect to the number of variables (a high m/Q ratio) [MacCallum et al., 2008].

Considering high-dimensional data for Factor Analysis is the major concerns of this chapter. Indeed, in the settings evoked previously, $n << m$ then the rules on size sample needed to yield accurate estimates are strongly violated. After a recall on various methods for parameters estimation, we study the estimation bias of parameters, and propose a correction to almost remove it in the small sample case. In the next part, we consider the factor number determination issue, which is a fundamental question of the method.

Note that estimation algorithms consider centered data:

$$Y_k^* = Y_k - m_k(x) = b_k' Z + \varepsilon_k$$

5.2. Estimation of the Factor Analysis model

Factor Analysis first consists in estimating the number Q of common factors, which is a crucial point of the method. In this section, we will consider that Q is known, in order to focus on the estimation of model's parameters only. The issue of factor number determination is discussed latter in this paper.

Methods Let's consider that each variable Y_k, $k \in [1..m]$, can be expressed as a linear combination of common factors, Z_q, $q \in [1..Q]$ and a specific term E_k as in (3.1). The issue is then to determine $\hat{B}$ and $\hat{\Psi}$, respectively the estimation of the loadings, representing the weight of the considered variable

on the Z structure, and the estimation of the uniquenesses, so that: $S = \hat{B}\hat{B}' + \hat{\Psi}$. There are several methods to extract factors from the data such as Unweighted Least Squares (ULS), Generalized Least squares (GLS), Singular Value Decomposition (SVD), Principal Factoring (PF), Maximum Likelihood (ML) or even Alpha Factoring. Generally, two of them are evoked to deal with this issue and are implemented in the leading statistical software: Principal Factoring (PF) and Maximum Likelihood (ML).

PF consists in iterative PCA. B is then estimated by the Q first principal components and Ψ by the diagonal of $R - \hat{B}\hat{B}'$. Each step of PF algorithm involves the storage of the $m \times m$ correlation matrix and its decomposition that can be numerically complicated. Nevertheless, factorial approaches are commonly used in practice, and results are said correct provided the number of factors Q is well chosen.

The ML method has some nice statistical properties such as asymptotic efficiency, invariance under change of scale, and the existence of a test for additional factors. ML method is favored in the following for these interesting properties in an inference aim. Algorithms implementing this method usually constrain Ψ to be diagonal, so that all the items given in ASSUMPTION 2 are satisfied. This is a point of most interest as, as described in CHAPTER 4, the Ψ matrix plays an important role in the proposed inference framework.

ML Factor Analysis for high-dimensional data To avoid "Heywood cases" that can be brought by Newton-Raphson algorithms in the seek of the ML estimators, Rubin and Thayer [1982] proposed an EM algorithm. This class of algorithm is now a very popular tool for iterative ML estimation in issues involving missing or incomplete data. In the Factor Analysis framework, we aim at estimating the parameters of a multivariate normal model with missing data, where in this case, the missing data are the unobserved latent variables Z. The EM algorithm developed for Factor analysis is iterative, and has the same attractive properties as ML-based methods. An iteration consists of two steps, the Expectation Step (E-step) and the Maximization Step (M-step), increasing each time the (log)likelihood of the parameters. Each iteration is not computationally expensive, and only involves the decomposition of a $Q \times Q$-matrix, $Q << m$. Many extensions have been developed, for instance when closed forms can not be achieved for the E-step solution (Monte Carlo EM (MCEM), [Wei and Tanner, 1990]) or for the M-step solution (Generalized EM (GEM),[Dempster et al., 1977], Expectation-Conditional Maximisation (ECM), [Meng and Rubin, 1993]). Other extensions are also proposed to improve computational properties and convergence rate of the EM algorithm (Incremental EM (IEM), Sparse EM (SPEM), [Neal and Hinton, 1998]). See for instance McLachlan et al. [2004] for a larger review. In the following, we consider the issue of parameter estimation in the Factor Analysis framework, and speed of convergence improvements are kept aside.

Let's call $Y^{(i)}$ the i^{th} line of Y, associated to the observation i, of size $1 \times m$, and $Z^{(i)}$ the $1 \times Q$-vector of scores associated to the i^{th} observation. The EM algorithm is based on the log-likelihood

of the model (3.1) given in (5.2) and aims at maximizing its expectation (5.3):

$$\mathcal{L}(B,\Psi) = -\frac{nm}{2}ln(2\pi) + \frac{n}{2}ln\left|\Psi^{-1}\right| - \frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)} - 2Y^{(i)}\Psi^{-1}BZ'^{(i)} + tr\left(B'\Psi^{-1}BZ'^{(i)}Z^{(i)}\right)\right] \quad (5.2)$$

$$\mathbb{E}(\mathcal{L}|Y) = cst + \frac{n}{2}ln\left|\Psi^{-1}\right| - \frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)} - 2Y^{(i)}\Psi^{-1}B\mathbb{E}(Z'^{(i)}|Y^{(i)}) + tr\left(B'\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right)\right] \quad (5.3)$$

$tr(M)$ is the trace of matrix M. Considering the joint density of Y and Z, we get

$$\mathbb{E}(Z^{(i)}|Y^{(i)}) = Y^{(i)}\Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} \quad (5.4)$$

$$\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)}) = (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \quad (5.5)$$

To find ML estimates of B and Ψ, we maximize (5.3) and find:

$$\hat{B} = \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \quad (5.6)$$

$$\hat{\Psi} = S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B' \quad (5.7)$$

The EMFA algorithm is described hereafter, based on expressions (5.4) and (5.5) for the E-step and (5.6) and (5.7) for the M-step:

1. **Initialization**: rough estimation $\hat{B}_0$ and $\hat{\Psi}_0$.

2. **Iterations**

   - **E-step**: Expectation of the log-likelihood

$$\mathbb{E}(Z^{(i)}|Y^{(i)}) = Y^{(i)}\Psi_0^{-1}B_0(\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1}$$
$$\mathbb{E}(Z^{(i)}Z'^{(i)}|Y^{(i)}) = (\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)})$$

   - **M-step**: Estimation of the ML estimators for B and Ψ

$$B_1 = \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1}$$
$$\Psi_1 = diag\left[S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B_1'\right]$$

3. **Stop**: The E- and M-steps are alternated repeatedly until convergence, which may be determined by using a suitable stopping rule like for example, $tr(\Psi_0 - \Psi_1) < \varepsilon$, $\varepsilon > 0$.

The details of up-dating equations are available in the APPENDIX

Degree of freedom of the FA model We can write a model as $\hat{Y} = PY$ where P represents the smoother matrix. In the linear model theory, the number of degree of freedom corresponds to the trace of the matrix $2P - PP'$. Therefore, the degrees of freedom associated to the residuals are defined as $ddl_r = n - tr(2P - PP')$. These definitions can be extended to the non parametric case, where P is no more independent of Y distribution: $P = P(Y)$ (see Hastie and Tibshirani [1990]).

In the case of Factor Analysis, the model is defined in (3.1). Predictions can be written as $\hat{Y} = ZB'$. Let's call $G = [\mathbb{I}_Q + B'\Psi^{-1}B]^{-1}$.

$$\begin{aligned} \hat{Y} &= ZB' \\ &= \left(Z[nG + Z'Z]^{-1}Z'\right)Y \qquad \text{see (5.6)} \\ &= HY \end{aligned} \tag{5.8}$$

Therefore, the degrees of freedom for the residuals of the Factor Analysis model are:

$$df_r = n - tr(2H - HH') \tag{5.9}$$

5.3. Validation of parameters estimation in high-dimension

Properties of the estimates yielded by the EM algorithm are studied in details in many books and papers, in the classical setting for data dimension. As huge datasets of multidimensional observations are commonplace in real-data analysis, we can wonder whether these properties still hold while considering high-dimensional settings. In this section, we will focus on the estimation of B and Ψ. The estimates brought by the EM algorithm are convergent, but may be biased in the case of finite samples. For small samples, the bias of ML estimators can be substantial.

Exemple 5. We consider simulated datasets to study the impact of high-dimension on FA model parameters estimation. The simulated datasets are characterized by 10 different schemes of dependence structure of $Q = 5$ factors, from a low to a high level of correlation, as described in example 3. Each dataset is composed of $m = 500$ variables and $n = 10, 50, 500$ observations. These 3 configurations for sample sizes allow to compare estimations in small samples and in an asymptotic setting. Each time, 1000 datasets are simulated. Considering the number of factors as known, the EMFA algorithm is computed to get B and Ψ ML estimates.

Estimation of loadings (B) B is determined up to an orthogonal rotation. Estimation is then assessed thanks to the RV coefficient [Escouffier, 1973]. This coefficient evaluates the relationship between the variables of two datasets X and Y, observed on the same individuals. It takes values between 0 (each variable of X is uncorrelated to each variable of Y) and 1 (the configurations of the individuals induced by X and Y are homothetic). Between theses two extreme bounds, the value of an RV coefficient can be tested for its significance [Josse et al., 2008]. This test is performed on the estimations obtained in the simulation study described in example 5. The main result is that the

higher the dependence among tests is, the more accurate the loading estimation is. This results is linked to the fact that for a low dependence level, the number of factors should be less than 5, and most often 0 or 1 (see next SECTION for more details). In this case, it may be more appropriate to use methods designed for the independent case.

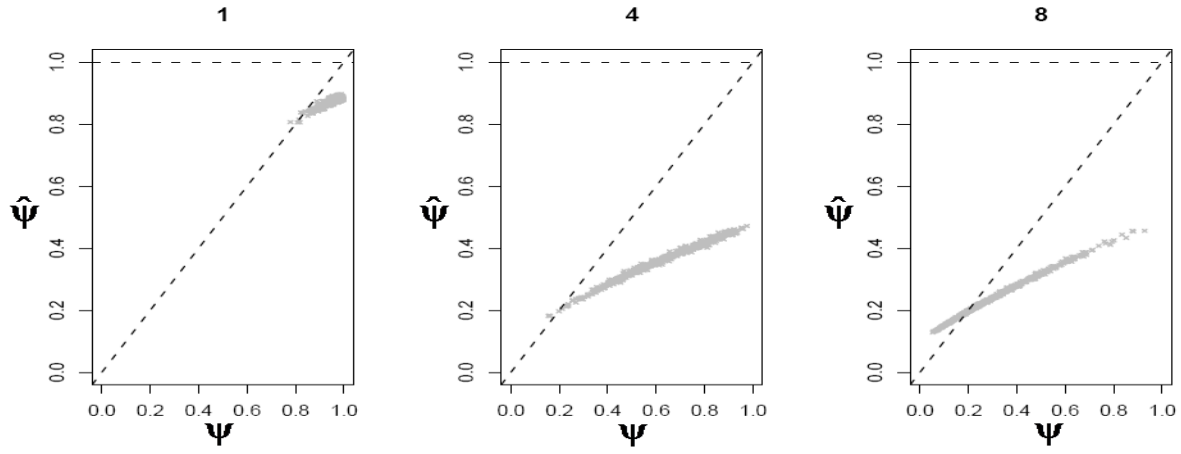
Estimation of specific variances (Ψ) FIGURE 5.1 shows the Ψ estimation with respect to the true Ψ, along with different sample sizes and different dependence structures, from the simulation study of EXAMPLE 5. For small sample sizes (see FIGURES 5.1(a) and 5.1(b)), there is a bias in Ψ estimation: high "uniquenesses" tend to be under-estimated, in particular when the level of dependence increases. As the sample size grows, the bias disappears, which confirms that the ML estimate provided by the EM algorithm converges (see FIGURE 5.1(c)).

Our proposal is to correct the bias in Ψ estimation in the high dimensional framework thanks to the residual degrees of freedom as given in (5.9). Ψ estimation is then the one obtained by the EMFA algorithm, multipied by a coeffifient $cz = \frac{n-1}{df_r}$. We apply this correction to the simulation study presented previously. The results for the different scenarios (levels 1,4,et 8), and the different sample sizes ($n = 10, 50, 500$), are given on FIGURE 5.2. The c_Z correction reduces the bias in Ψ estimation in the case of small samples.

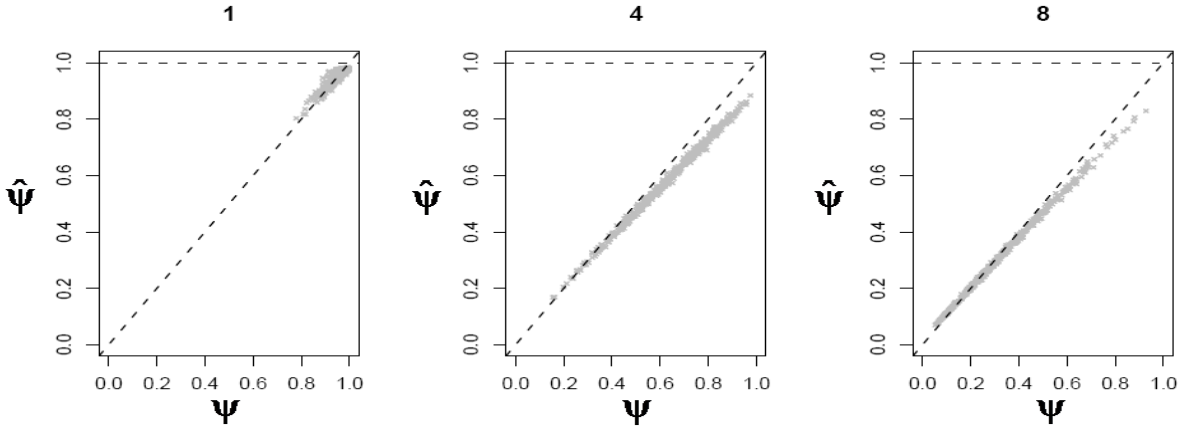
5.4. Number of factors

In Factor Analysis, the first step consists in estimating the number of factors Q to be considered in the model. This step is the most crucial in conducting Factor Analysis. Indeed, we must balance between parsimony, that is to say to consider few parameters in the model, and accuracy in representing the correlation structure. Choosing the "right" number of factors requires to be able to identify major and minor factors. Under-estimation of Q leads to loss of information by ignoring a factor or recombining one with another. The loadings for measured variables are therefore biased and interpretation based on them would not be much reliable, the true structure of the data being concealed. Over-estimation is commonly considered as less severe. But considering too many factors underlines minors factors, with only a few high loadings so that interpretation is difficult and such model is unlikely to be robust for replication. Therefore, considering both too few and too much factors has significant consequences in the reduction of information, affecting parameters estimation and data interpretation. Because of all these reasons, the number of factors issue leads to plenty of methods proposed in the literature, with more or less subjective decision rules.

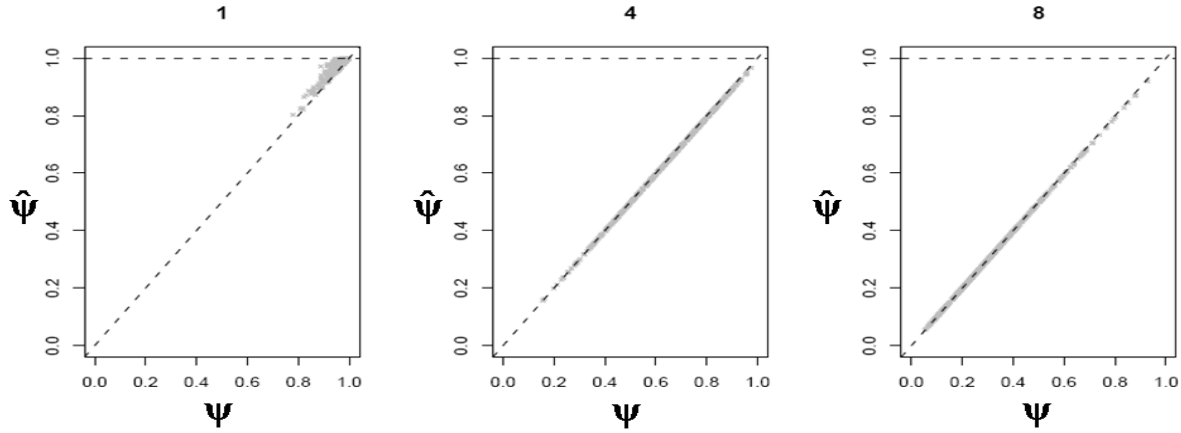
Exiting methods The most famous criteria are the ones proposed by Kaiser [1960] ($K1$: eigenvalues greater than 1) and Cattell [1966] (scree-test: examination of eigen values pattern for discontinuities). The ease of implementation and theorical basis of both K1 and scree test rules make them widely used in practice, as shown by a lot of reviews on the use of factor retention methods in psychology or marketing where FA is commonplace, through the study of articles published in major journals of these research fields (Stewart [1981], Ford et al. [1986], Fabrigar et al. [1999] or



(a) sample size $n = 10$

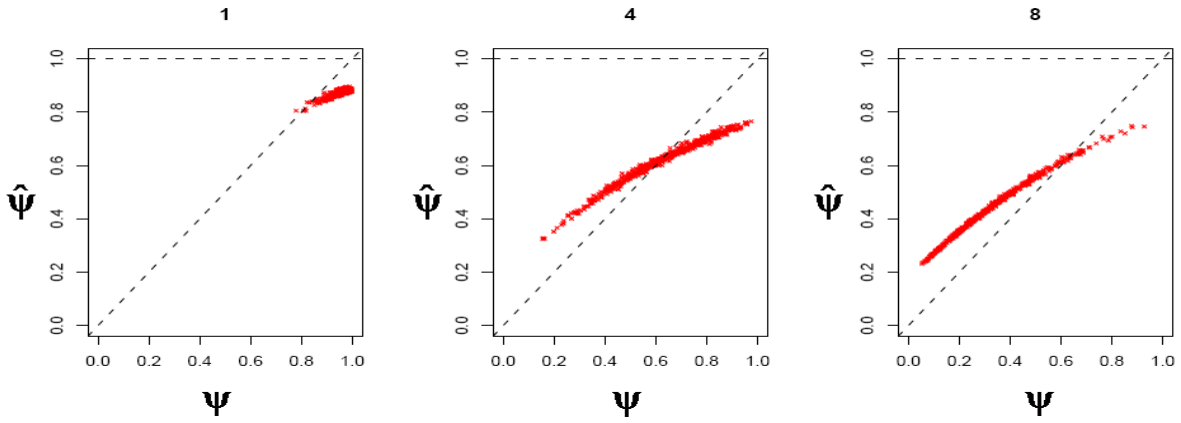


(b) sample size$n = 50$

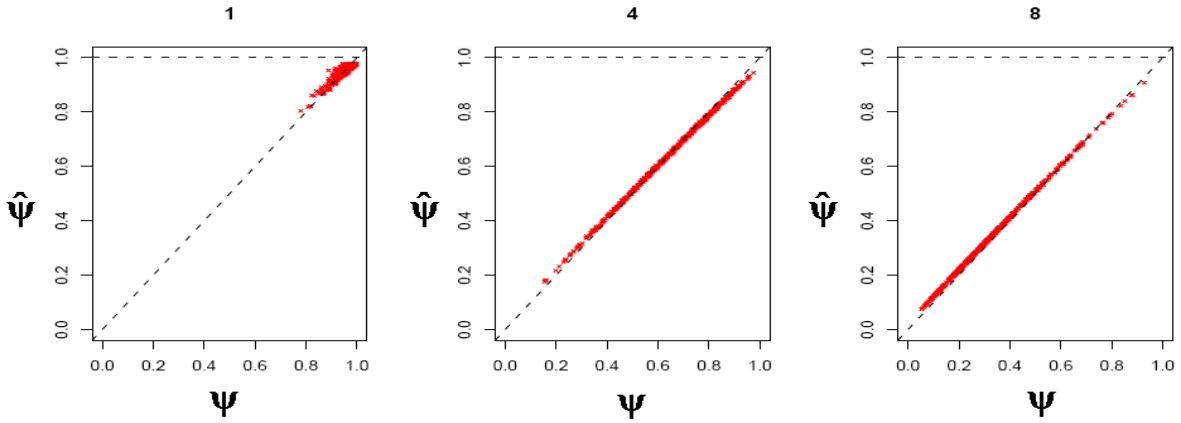


(c) sample size $n = 500$

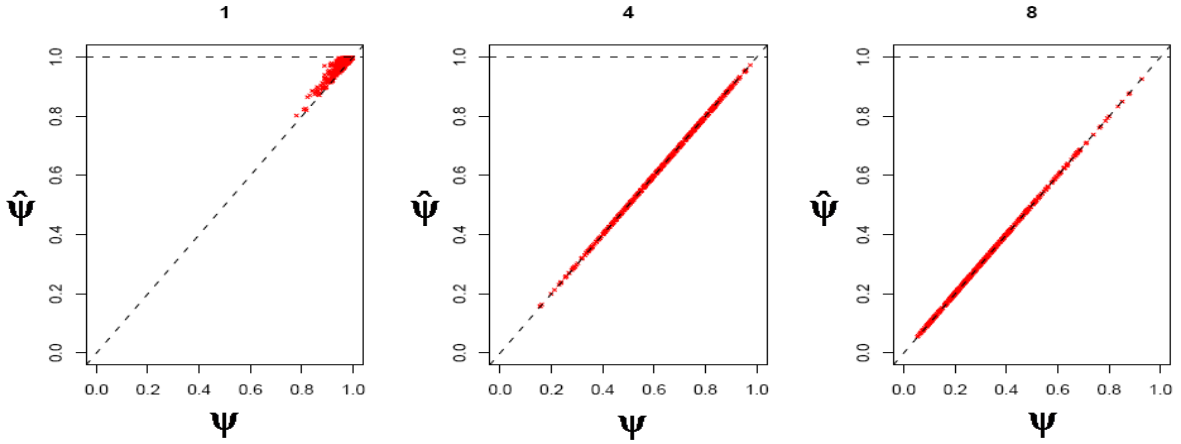
Figure 5.1: Ψ estimation considering 3 different dependence structures: low (1), intermediate (4) and high (8) - 1000 simulated datasets for each scenario - means along with the 1000 replicates are represented on the plots



(a) sample size $n = 10$



(b) sample size $n = 50$



(c) sample size $n = 500$

Figure 5.2: Ψ estimation with residual degrees of freedom correction, considering 3 different dependence structures: low (1), intermediate (4) and high (8) - 1000 simulated datasets for each scenario - means along with the 1000 replicates are represented on the plots

more recently Norris and Lecavalier [2009]). Other methods are proposed in the literature. Comparing eigen-values of the real dataset and those of same-sized datasets of simulated random variables is the principle of Parallel Analysis, suggested by Montanelli and Humphrey [1976] to take into account in the factor number determination the variation of eigen-values due to sampling. Other criteria are sometimes considered such as Very Simple Structure [Revelle and Rocklin, 1979], based on a goodness-of-fit index, or the Minimum Average Partial test [Velicer, 1976], based on the partial correlation matrix.

However, all of these criteria often do not lead to the same, or even similar, results. Therefore, there is no consensus on which one, among the huge number of available criteria, is more appropriate to use. Many authors have conducted studies to compare the performance and accuracy of the previous criteria (see for example Ford et al. [1986] or Velicer et al. [2000]). The general conclusion is that, despite their popularity, $K1$ or scree test criteria are not much accurate, especially when the sample size is small. $K1$ tend to over-estimate Q as it yields a lower bound for the rank of the correlation matrix, and then an upper bound for Q [Horn, 1965]. Scree test suffers from subjectivity, as breaks are less likely evident in the case of small samples and when the correlation structure is not strong. For many authors (Hayton et al. [2004], Velicer et al. [2000]), PA should be preferred for factor retention as the most accurate method, and only slightly over estimates Q in case of error. A major problem in high dimensional data with methods such as PA is that they require the factorisation of huge covariance matrix ($m \times m$) so computation may not be possible from a practical point of view. Considering the ML estimation of the parameters, we could use χ^2-based methods, such as a goodness-of-fit criterion to test for the number of factors to retain. But this is too sensitive to the sample size to be considered in a high-dimensional setting.

A criterion based on the variance inflation of V_t We propose a new criterion to determine the number of factors to retain, which matches with high-dimension constraints. Factor Analysis aims at modeling the covariance (or correlation) structure. Let's consider the following model defined previously in (3.1) and assuming q common factors for each variable k: $Y_k = \mu_k + Zb_k'^{(q)} + \varepsilon_k^{(q)}$.

If the whole correlation structure is well modeled by the Factor Analysis model, then the residual correlation should be zero. Our method is based on this proposition. Indeed, $\text{Cov}(\varepsilon_k; \varepsilon_{k'}) = \sigma_k \sigma_{k'} \rho_{kk'} - b_k b'_{k'}$ and $\mathbb{V}(\varepsilon_k) = \Psi_k$ so we can define the residual correlation, assuming a q-common factors model, by:

$$\rho_{kk'}^{(q)} = \frac{\sigma_k \sigma_{k'} \rho_{kk'} - b_k^{(q)} b_{k'}'^{(q)}}{\sqrt{\Psi_k \Psi_{k'}}} \tag{5.10}$$

Beforehand, let's recall the definition a U-shaped criterion called $D^{kk'}(t)$ that ranges from 0 in $\rho = 0$ to 1 in $\rho = -1$ and $\rho = 1$ (see DEFINITION 3.2.1 and FIGURE 3.7). Considering the $D^{kk'}(t)$ criterion for each pair of variables $\{Y_k, Y_{k'}\}, k \neq k' \in [1; m]$, the proposed method to determine the number of factors is to choose Q satisfying:

PROPOSITION 5.4.1. *If $\rho_{kk'}^{(q)}$) is the residual correlation between Y_k and $Y_{k'}$ as in* (5.10)*, let's define Q as:*

$$Q = \underset{q\in[0;q_{max}]}{\operatorname{argmin}} \frac{1}{m(m-1)} \sum_{k\neq k'\in[1;m]} D^{kk'}(t)^{(q)}) \tag{5.11}$$

Q is the number of factors that minimises the mean of $D^{kk'}(t)$ criterion over all the pairs of variables. In the multiple testing context, this criterion calculated over variables in $\mathcal{M}_0$ allows to extract the number of factors that minimises the variance of V_t. According to the study led in the previous chapters, the consequence is that procedures are stabilized. In practice, this criterion is successively estimated using the residual matrix obtained with an increasing number of factors and the retained number of factors is obtained when the variance inflation is minimised.

Simulation study This estimation procedure for the number of factors is now implemented in the 10 scenarios of simulated data introduced in EXAMPLE 5, in which the true number of factors is $Q = 5$. FIGURE 5.3 reproduces barplots of the distribution of $\hat{Q}$. It clearly shows that when the proportion of common variance is small, the estimated number of factors is relevantly lower than Q and when the factor structure dominates the specific part, $\hat{Q}$ provides a precise estimation of Q.

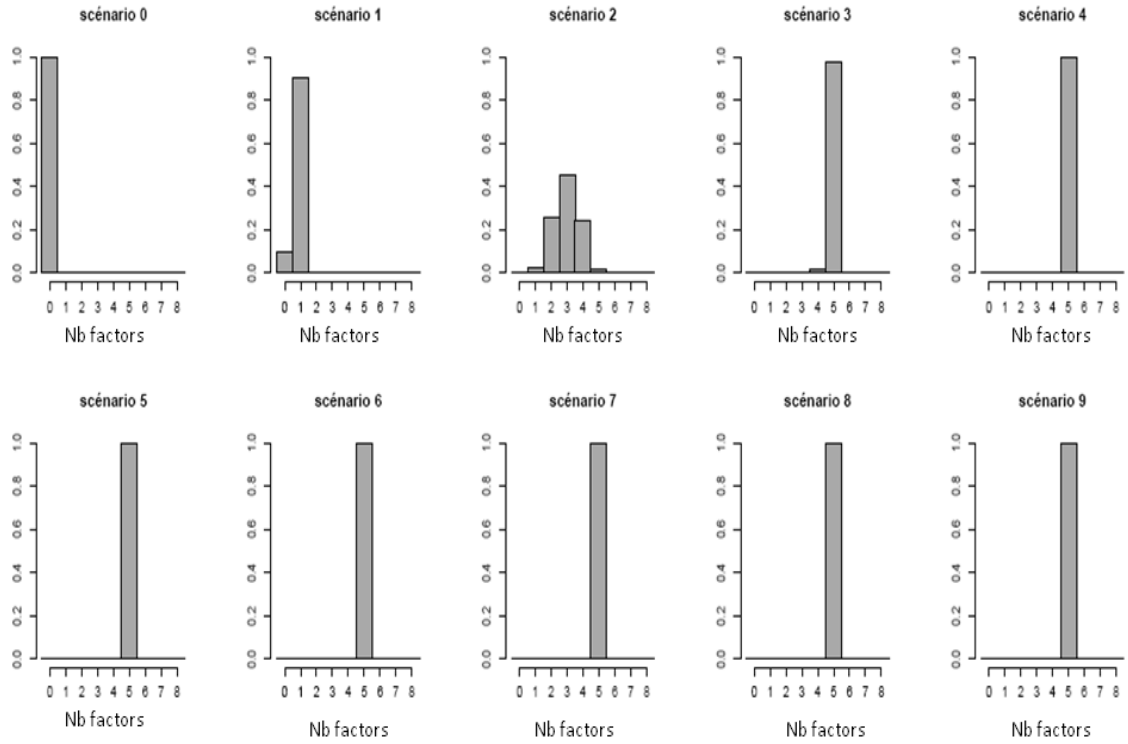


Figure 5.3: Distributions of the estimated number of factors along with the dependence level. From scenarios 4 to 9, $\hat{Q}$ turns out to be constant and equal to 5.

In practice, implementing this method requires to calculate the $D^{kk'}(t)$ criterion for all pairs of variables, namely $m \times (m-1)$. As the correlation matrix is symmetric, it actually requires to consider $m \times (m-1)/2$ pairs. In the multiple testing framework, this is reduced to $m_0 \times (m_0-1)/2$ pairs. In any case, the number of pairs is very huge, and can reach several thousands. As also suggested by Owen [2005], we propose to consider a range of η values, uniformly distributed on $[0;1]$. Then, we count the number of times each value appears in the correlation matrix, considering an approximation of correlations by taking their absolute value and rounding them at the specified significant decimal figure:

$$\sum_{k \neq k' \in \mathcal{M}} D^{kk'}(t) \approx \sum_{j=1}^{\eta} n_j D(t; \rho_j) \tag{5.12}$$

where n_j represents the number of variables pairs for which the rounded correlation is equal to ρ_j. If the correlation distribution is symmetric, which is the case in most of applications, then the approximation given in (5.12) is accurate and leads to a sharp increase in computation time.

In the `R` package called `FAMT`, the minimised criterion is based on the following approximation:

$$\frac{1}{m(m-1)} \sum D^{kk'}(t) \approx \int_{[0;1]} D^{kk'}(t) f(\rho) d\rho$$

where $f(\rho)$ is the correlations density function. It is estimated thanks to sampling in observed correlations.

Conclusion

Several methods are available to estimate the parameters of a latent variables model. We here consider a Maximum Likelihood Factor Analysis using an EM algorithm to deal with high dimension.

Determining the optimal number of factors is a crucial step of Factor Analysis. We propose a criterion allowing to define the model that fit best the covariance structure. By minimising this criterion, the inflation of variance of false-positives is also minimised. Using this criterion is therefore of great interest for the FAMT procedure introduced in CHAPTER 4.

In case of strong dependence structure among tests, the model is easier to fit and estimation by Factor Analysis is more accurate, even for small samples. The consequence when using Factor Analysis for multiple testing is that it may be more appropriate to use procedures derived in the independent case when the dependence structure is very low. Considering our criterion to determine the number of factors leads to this strategy.

The FAMT procedure is now assessed in CHAPTER 6 thanks to an application to gene expressions data and presented as a tutorial of the `R` package called `FAMT`.

CHAPTER 6

FACTOR ANALYSIS FOR MULTIPLE TESTING (FAMT): AN R PACKAGE FOR LARGE-SCALE SIGNIFICANCE TESTING UNDER DEPENDENCE

Abstract The `R` package `FAMT` (Factor Analysis for Multiple Testing) provides a powerful method for large-scale significance testing under dependence. It is especially designed to select differentially expressed genes in microarray data when the correlation structure among gene expressions is strong. Indeed, this method reduces the negative impact of dependence on the multiple testing procedures by modeling the common information shared by all the variables using a factor analysis (FA) structure. New test statistics for general linear contrasts are deduced, taking advantage of the common factor structure to reduce correlation and consequently the variance of error rates. Thus, the FAMT method shows improvements with respect to most usual methods regarding the Non Discovery Rate (NDR) and the control of the False Discovery Rate (FDR). The steps of this procedure, each of them corresponding to `R` functions, are illustrated in this paper by two microarray data analyses. We first present how to import the gene expression data, the covariates and the genes annotations. The second step includes the choice of the optimal number of factors and the FA model fitting, and provides a list of selected genes according to a preset FDR control level. Finally, diagnostic plots are provided to help the user interpret the factors using available external information on either genes or arrays.

Keywords: factor analysis, multiple testing, dependence, false discovery rate, non discovery rate, `R`

Sommaire

6.1. Introduction

Independence among tests is assumed in the settings of most multiple testing procedures. However, for instance in the case of microarray data, the dependence between gene expressions is often strong, due to complex biological regulatory relationships. It has been proved that this dependence has a negative impact on the multiple testing procedures, particularly on the variance of the number of false positive genes, thus on the control of the False Discovery Proportion (Efron [2007], Kim and Van de Wiel [2008], Friguet et al. [2008]). The FAMT procedure deals with this problem by modeling the common information shared by all the variables using a factor analysis structure. New test statistics for general linear contrasts are deduced, taking advantage of the common factor structure to reduce correlation and consequently the variance of error rates. The details of this method are given in Friguet et al. [2008].

The present paper aims at presenting the statistical handling of multiple testing dependence as proposed in the `R` package `FAMT`. The crucial steps of the analysis correspond to core functions: `as.FAMTdata` to import the data and create a single `R` list from multi-sourced datasets, `modelFAMT` to estimate the dependence kernel and adjust the data from heterogeneity components and `defacto` to relate the heterogeneity components to external information if provided. Moreover, additional functions are proposed to summarize the results (`summaryFAMT`) and to optimize the procedure by modifying the default choices implemented in `modelFAMT`, such as the estimation procedure for the proportion of true null hypotheses (`pi0FAMT`) or the optimal number of factors (`nbfactors`).

The FAMT procedure is applied on two microarray datasets, which both describe chicken hepatic transcriptome profiles, and are provided by the INRA Animal Genetics department in Rennes, France. The first microarray data analysis studies the relationships between hepatic gene expression and abdominal fatness (Blum et al. [2010], LeMignon et al. [2009]). The normalized microarray dataset is available in the `FAMT` package, and is used here to describe the method step by step. The second microarray data analysis focuses on the feeding-to-fasting transition in chicken liver by Désert et al. [2008].

The first dataset is a relevant example of a situation where the classical multiple testing method fails to detect differentially expressed genes, due to the high amount of dependence among the gene expressions (see Blum et al. [2010]). Indeed, Figure 6.1 represents the empirical distribution of the F-tests p-values: note that there are less small p-values than expected under the hypothetical situation where all genes would be under the null hypothesis. In situations of highly dependent data, Friguet et al. [2008] show that the empirical distribution of p-values corresponding to true null hypotheses can markedly depart from the uniform distribution. In this situation, it is therefore recommended

to take into account the dependence to improve the data analysis. On the contrary, in our second example, the distribution of the p-values is apparently more common (see Figure 6.8). Yet, we show hereafter that the FAMT method can still be used to give more insight into the multiple testing procedure and increase its overall power.

6.2. Data manipulation

In microarray data analysis, the selection of differentially expressed genes involves at least two datasets with different dimensions. First, the `expression` dataset, which directly results from microarray experiments and usually a normalization procedure, has much more variables (gene expressions) than individuals (arrays). For convenience, it is stored as a m genes $\times$ n microarrays table. In the following illustative example, this dataset concerns hepatic transcriptome profiles for $m = 9893$ genes of $n = 43$ half sib male chickens selected for their variability on abdominal fatness (Af). The data comes from the Animal Genetics department (INRA-Rennes), and it was generated to map quantitative trait loci (QTL) for abdominal fatness in chickens. Animals, marker genotyping and transcriptome data acquisition and normalization are described in LeMignon et al. [2009].

The `covariates` dataset gives information about the experimental conditions: the identifier of each row (arrays), as used in the column names of `expression`, is provided, together with the value of the main explanatory variable in the testing issue (Af in the present example) and possibly of other covariates. This dataset is optional: if not provided, the procedure aims at testing the significance of the mean expressions.

Finally, the `annotations` dataset provides additional information about the response variables of the multiple testing procedure to be used to describe the results: for example, the functional characterization of each gene extracted from the Gene Ontology can be useful to directly connect the list of differentially expressed genes to biological processes. In the present example, some additional variables are also used to locate the spots on the microarray (block, row, column). One column of this dataset must be named `ID` and gives the variable (gene) identifier that will be used in the final output of the procedure. This dataset is optional: if not provided, a basic `annotations` dataset is created with row indices as variable identifiers.

The first step of the FAMT method uses the `as.FAMTdata` function to create a single `R` list containing the multi-sourced datasets. To avoid violations of the correspondence between the columns of the `expression` dataset and the rows of the `covariates` dataset, this function also checks that one column in `covariates`, which number is given by the argument `idcovar`, gives the individual (array) identifier, such as stored in the column names of `expression`. Some simple tests for the compatibility of the datasets' dimensions are also performed: the number of columns of `expression` must correspond to the number of rows of `covariates` and furthermore, `expression` and `annotations` must have the same number of rows. The subclass of the output is named `FAMTdata` and the belonging to this class is required by the other functions of the package.

In our example, three datasets are provided:

- `expression` contains 9893 gene expressions or tests, and 43 individuals.
- `covariates` with variables `AfClass` (the abdominal fatness class, with 3 levels: F (Fat), L (Lean), NC (Intermediate)), `ArrayName` (identifying the arrays), `Mere` (the dam of the offsprings, a factor with 8 levels), `Lot` (the hatch, a factor with 4 levels), `Pds9s` (the body weight, a numeric vector), and `Af` (the abdominal fatness, a numeric vector). `Af` is the experimental condition of main interest in this example.
- `annotations` with variables `ID` (the gene identification), `Name` (the gene functional category), `Block`, `Column` and `Row` (location on the microarray), `Length` (the oligonucleotide size).

The following code creates the `FAMTdata` object called `chicken`.

```
R> chicken = as.FAMTdata(expression, covariates, annotations, idcovar=2)

$‘Rows with missing values‘
integer(0)

$‘Columns with missing values‘
integer(0)
```

By default, `idcovar=1`. Here we use `idcovar=2` because the array identification is given in the second column of `covariates`.

Besides, this step checks for missing data, since the FAMT method cannot be applied with incomplete observations. The `as.FAMTdata` function gives the indices of the rows and columns of `expression` with missing values. If needed, using `na.action=TRUE`, the missing values are imputed by nearest neighbor averaging (function `impute.knn` of the package `impute`, Hastie et al. [2009]). Here, the `expression` component of `chicken` has no missing data.

Some classical componentwise summaries can be obtained on a `FAMTdata` object with the `summaryFAMT` function. The function provides:

- for `expression`: the number of tests which corresponds to the number of rows, the sample size which is the number of columns.
- for `covariates` and `annotations`: classical summaries as returned by the generic function `summary` in package `base`.

The code to perform the summary of a `FAMTdata` object is:

```
R> summaryFAMT(chicken)

$expression
$expression$‘Number of tests‘
```

```
[1] 9893

$expression$`Sample size`
[1] 43

$covariates
 AfClass  ArrayName        Mere    Lot         Pds9s            Af
 F :18   F10    : 1   GMB05555:10   L2:16   Min.   :1994   Min.   :-25.5397
 L :19   F11    : 1   GMB05625: 7   L3:11   1st Qu.:2284   1st Qu.: -8.0042
 NC: 6   F12    : 1   GMB05562: 5   L4: 8   Median :2371   Median :  2.7166
         F13    : 1   GMB05599: 5   L5: 8   Mean   :2370   Mean   :  0.2365
         F14    : 1   GMB05554: 4           3rd Qu.:2474   3rd Qu.:  8.6037
         F15    : 1   GMB05589: 4           Max.   :2618   Max.   : 18.1024
         (Other):37   (Other) : 8

$annotations
         ID            Name               Block          Column
 RIGG00001:   1   Length:9893        23     : 237   9      : 502
 RIGG00002:   1   Class :character   25     : 237   6      : 501
 RIGG00003:   1   Mode  :character   9      : 232   12     : 499
 RIGG00005:   1                     17     : 231   18     : 499
 RIGG00006:   1                     29     : 230   14     : 494
 RIGG00007:   1                     43     : 229   16     : 491
 (Other)  :9887                     (Other):8497   (Other):6907
      Row           Length
 20     : 500   Min.   :60.00
 15     : 498   1st Qu.:70.00
 16     : 493   Median :70.00
 8      : 483   Mean   :69.57
 17     : 481   3rd Qu.:70.00
 9      : 468   Max.   :75.00
 (Other):6970
```

This step is especially useful to check the class of variables in `covariates` and `annotations`. Here, note that all the variables in `annotations` except `ID` are intended to be used as explanatory variables in the description of the latent factors after model fitting. Therefore, character information such as the functional characterization of the genes in microarray data has to be stored as a `character` variable, and not as a `factor` with a large number of levels.

6.3. Multiple testing

This section is dedicated to the use of the `FAMT` method as a classical multiple testing procedure controlling the FDR, without any modeling for the dependence structure across the variables.

6.3.1 Multiple F-tests for general linear hypotheses

The scope of models for the relationship between the responses and the explanatory variable(s) of interest is restricted to linear models. Let $\mathbf{Y} = (Y^{(1)}, \ldots, Y^{(m)})'$ be the m-vector of response variables and $\mathbf{x} = (x^{(1)}, \ldots, x^{(p)})'$ the $p-$vector of explanatory variables. It is assumed that:

$$Y^{(k)} \quad = \quad \beta_0^{(k)} + \mathbf{x}'\beta^{(k)} + \varepsilon^{(k)}, \tag{6.1}$$

where $\varepsilon = (\varepsilon^{(1)}, \ldots, \varepsilon^{(m)})'$ is a normally distributed $m-$vector with mean $\mathbf{0}$ and variance-covariance $\mathbf{\Sigma}$.

The individual tests are the usual Fisher tests for the marginal effect of one or more explanatory variables of interest among x, considering the other ones as covariates. In most of the cases, only one explanatory variable x is included in the model and the aim is then to test the significance of the relationship between each variable and x. However, more complex situations also occur, where the effect of this explanatory variable shall be examined after adjustment from other effects, which have been accounted of in the experimental design. Note also that, if no `covariates` dataset is provided, then model (6.1) is the null model and the significance of the mean of each variable is tested.

The thresholding procedure applied on the p-values of the F-tests to control the FDR at a given level α is the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg [1995]). The cut-off on the p-values under which the hypotheses are rejected is derived from the increasingly ordered p-values $p_{(k)}$ as follows: $t_\alpha = p_{(k^*)}$ with $k^* = \arg\max_k \{m\pi_0 p_{(k)}/k \leq \alpha\}$, provided the proportion of true null hypotheses π_0 is known. Many multiple testing procedures assume that the fraction of non-null hypotheses among the tests is negligible regarding the large number of tests ($\pi_0 \approx 1$). For example, with the Benjamini-Hochberg procedure, approximating π_0 by 1 leads to a FDR control at level $\pi_0\alpha$ instead of α. Generally, plugging-in an estimate of π_0 into the expression of the FDR corrects for this control level and results in a less conservative procedure (see Benjamini and Hochberg [1995], Black [2004] and Storey [2002] for details).

Two methods are proposed in the `FAMT` package to estimate π_0: the first one is based on a non-parametric estimate of the density function of the p-values by a convex curve using Langaas et al. [2005]'s approach and another one uses the smoothing spline approach by Storey and Tibshirani [2003].

In the following, the use of the `modelFAMT` function is illustrated using the `chicken` dataset.

6.3.2 Results

In the `chicken` example, the aim is to test the significance of the relationship between each gene expression and the abdominal fatness variable (6th column of `covariates`), taking into account the

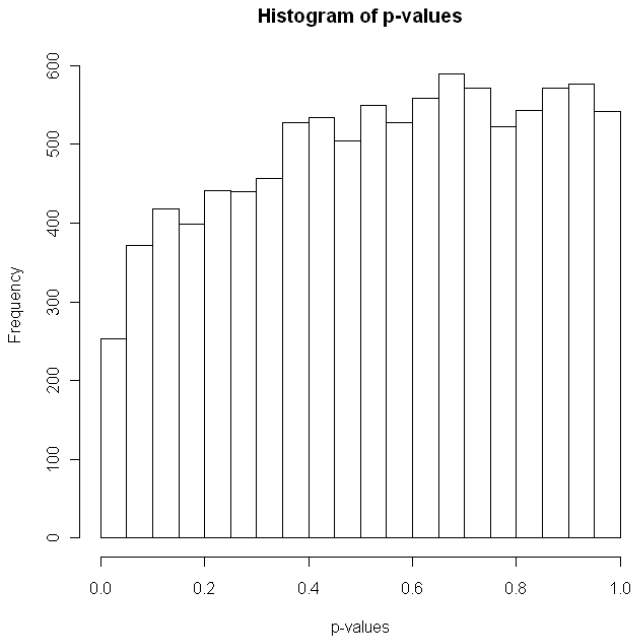


Figure 6.1: Histogram of the raw p-values for the `chicken` dataset

effect of the dam (3rd column of `covariates`). The Fisher test statistics and the corresponding p-values are obtained using the `modelFAMT` function with arguments `x=c(3,6)` to give the column numbers of the explanatory variables in the `covariates` component of `chicken` and `test=6` to give the column number of the explanatory variable of interest. The following code also uses argument `nbf=0`, which means no common factors in the model for the conditional variance Σ, to get the raw F-tests.

```
R> chicken.raw = modelFAMT(chicken, x=c(3,6), test=6, nbf=0)
R> hist(chicken.raw$pval, main="Histogram of p-values", xlab="p-values")
```

Figure 6.1 displays the histogram of the raw p-values, as produced by the command lines above. The shape of the histogram clearly shows an abnormal under-representation of the p-values in the neighborhood of 0. Indeed, if all the gene expressions were all truly under the null hypothesis, the p-values should be uniformly distributed on $[0, 1]$ and the proportion of observed p-values under 0.05 should be close to 0.05, provided the gene expressions are independent. This marked departure of the empirical distribution of p-values from the density function of a uniform distribution has been recently considered by some authors as the impact of a high amount of dependence among tests (see Efron [2007], Leek and Storey [2007] and Friguet et al. [2008]).

The `modelFAMT` function creates a `R` list with subclass `FAMTmodel`. This subclass is required for the main input of the other functions in the package. Thus, the `summaryFAMT` function can be applied to a `FAMTmodel` object to get the list of positive tests for a control of the FDR at a preset level α (the default level is `alpha=0.15`). Moreover, some useful information about the positive responses is

provided, using the argument `info` which gives the names of columns in the `annotations` component of `chicken`. Here, the columns named `ID` and `Name` give the gene identifier and the functional annotation of the significant genes.

```
R> summaryFAMT(chicken.raw, alpha=0.05, info=c("ID","Name"))
$nbreject
  alpha Raw analysis FA analysis
1  0.05            0           0

$DE
integer(0)

$pi0
[1] 1
```

The `nbreject` component in the output of `summaryFAMT` is a table providing the number of positive tests using the raw multiple testing procedure and the factor analytic approach, for possibly different values of the FDR control level α. In this special use of `summaryFAMT` with `nbf=0`, the columns named `Raw analysis` and `FA analysis` gives the same result since they are equivalent. The result shows no positive genes for a FDR control at level 0.05. The `DE` component of the output also provides the additional information on the responses specified in the argument `info`.

Note that, if not explicitly provided as an input of the function `summaryFAMT` using argument `pi0`, the proportion π_0 of true null hypotheses is estimated from the histogram of p-values, by the method proposed by Storey and Tibshirani [2003], and returned in the `pi0` component of the output. Here, $\hat{\pi}_0 = 1$ is a direct consequence of the abnormal shape of the histogram of p-values as displayed in Figure 6.1. Indeed, the dependence across genes induces a bias in the estimation of π_0. Analysis of this dependence among gene expressions is addressed in section 5 but a first possible biological explanation is that all the chickens in this experiment are half sib males, which is to say genetically very similar.

This example is a typical situation where the dependence among genes must be taken into account to have a chance to reveal significant relationships between the hepatic transcriptome profile and the quantity of abdominal fatness.

6.4. Multiple testing dependence using FAMT

6.4.1 Method

The details of the method are described in Friguet et al. [2008]. The main innovation with respect to most classical methods consists in capturing the components of dependence between variables

into latent factors and integrating these latent structure in the calculation of the test statistics. It is indeed assumed that the conditional covariance matrix Σ of the responses, given the explanatory variables is represented by a factor analysis model:

$$\mathbf{\Sigma} \quad = \quad \mathbf{\Psi} + \mathbf{B}\mathbf{B}', \tag{6.2}$$

where $\mathbf{\Psi}$ is a diagonal $m \times m$ matrix of uniquenesses ψ_k^2 and $\mathbf{B}$ is a $m \times q$ matrix of factor loadings. In the above decomposition, the diagonal elements ψ_k^2 in $\mathbf{\Psi}$ are also referred to as the specific variances of the responses. Therefore, $\mathbf{B}'\mathbf{B}$ appears as the shared variance in the common factor structure.

The factor analysis representation of the covariance is equivalent to the following mixed effects regression modeling of the data: for $k = 1, \ldots, m$,

$$Y^{(k)} \quad = \quad \beta_0^{(k)} + \mathbf{x}'\beta^{(k)} + \mathbf{b}_k' Z + \varepsilon^{(k)}, \tag{6.3}$$

where $\mathbf{b}_k$ is the kth row of B, $\mathbf{Z} = (Z^{(1)}, \ldots, Z^{(q)})$ are latent factors supposed to concentrate in a small dimension space the common information in the m responses, $\mathbf{Z}$ is normally distributed with expectation $\mathbf{0}$ and variance $\mathbf{I}_q$ and $\varepsilon = (\varepsilon^{(1)}, \ldots, \varepsilon^{(m)})'$ is a normally distributed $m-$vector, independent of $\mathbf{Z}$, with mean $\mathbf{0}$ and variance-covariance $\mathbf{\Psi}$. Therefore, factor analysis can be viewed as simultaneous mixed effects regression models sharing common covariance components.

An EM algorithm inspired from Rubin and Thayer [1982] is used to estimate $\mathbf{\Psi}$, B and $\mathbf{Z}$ (see Friguet et al. [2008] for details). Since this algorithm only implies inversions of $q \times q$ matrices, fitting the FA model on high-dimensional datasets is computationally much less cumbersome than more usual algorithms such as Principal Factoring used in psychometrics. As recommended in Friguet et al. [2008], the number of factors is chosen according to an *ad-hoc* procedure which consists in minimizing the variance of the number of false discoveries. Once the factor model is estimated, so called factor-adjusted test statistics are derived as F-tests calculated on the adjusted response variables $Y^{(k)} - \mathbf{b}_k' Z$ obtained by subtracting the dependence kernel from the data. Friguet et al. [2008] show that the resulting tests statistics are asymptotically uncorrelated, which improves the overall power of the multiple testing procedure.

6.4.2 Results of the FAMT analysis

Optimal number of factors for the FA model fitting The `modelFAMT` function implements the whole `FAMT` procedure with default options for the estimation of π_0 and the number of factors. As mentioned in the previous section, the method proposed by Storey and Tibshirani [2003] is implemented to estimate π_0.

Concerning the number of factors, the dependence in the residual correlation matrix resulting from the k-factor analysis model fitting induces an inflation of the variance of the number of false positives. This variance has a negative impact on the actual control of the false discovery proportion. Hence, as explained by Friguet et al. [2008], the number of factors considered in the model is chosen to reduce this variance. In order to avoid the overestimation of the number of factors, the function is implemented in such a way that the optimal number of factors corresponds to the largest number of

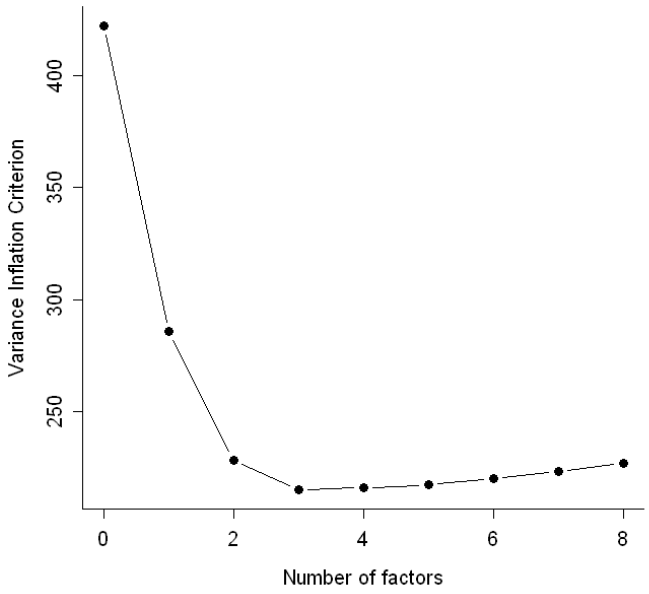


Figure 6.2: Variance inflation criterion for the determination of the optimal number of factors

factors for which the decrease of the variance inflation criterion is lower than 5 % of the previous value (see the Cattell scree test criterion, Cattell [1966]). Nevertheless, the optimal number of factors can also be specified by the `nbf` argument in the `modelFAMT` function (see the second illustrative example of this paper). Once the optimal number of factors is chosen, the model parameters are estimated using an EM algorithm. Factor-adjusted tests statistics are derived, as well as the corresponding p-values.

The testing issue is the same as in the previous section.

```
R> modelfinal=modelFAMT(chicken, x=c(3,6), test=6)
R> modelfinal$nbf
[1] 3
```

A side effect of the `modelFAMT` function is to produce a diagnostic plot, displaying the values of the variance inflation criterion along with the number of factors. Figure 6.2 shows that the optimal number of factors obtained by the `modelFAMT` function, which is `modelfinal$nbf=3`, also corresponds in this case to the minimum value of the variance inflation criterion. The model parameters are estimated with this choice of a 3-factor structure and π_0 is estimated using the method by Storey and Tibshirani [2003] applied on the factor-adjusted p-values.

The number of positive tests is provided for each level of FDR control chosen by the user (in our example below, the levels are defined by the argument `alpha=seq(0, 0.3, 0.05)`). The list of positive genes (`DE` component) is given for the highest `alpha`.

```
R> summaryFAMT(modelfinal, alpha=seq(0, 0.3, 0.05))
```

```
$nbreject
  alpha Raw analysis FA analysis
1  0.00            0           0
2  0.05            0           0
3  0.10            0           2
4  0.15            0           6
5  0.20            0           6
6  0.25            0           8
7  0.30            0          11

$DE
            ID
6722 RIGG05436
3885 RIGG04393
1119 RIGG15056
3484 RIGG05478
463  RIGG09893
124  RIGG12578
9859 RIGG03755
9840 RIGG04507
3925 RIGG10355
4968 RIGG05365
3855 RIGG13434

6722                                                                Same gene X54200
3885               Weakly similar to CAE03429 (CAE03429) OSJNBa0032F06.12 protein
1119                                                             ENSGALT00000015290.1
3484    Weakly similar to Q8IUG4 (Q8IUG4) Rho GTPase activating protein (Fragment)
463                                                              ENSGALT00000000452.1
124                                                              ENSGALT00000008042.1
9859                                                              Contig Hit 348847.1
9840 Weakly similar to Q8AWZ8 (Q8AWZ8) Voltage-gated potassium channel subunit MiR
3925                           Transforming protein p54/c-ets-1. [Source:SWISSPROT
4968                                                            Genome Hit Contig7.437
3855                  Troponin T fast skeletal muscle isoforms. [Source:SWISSPROT


$pi0
[1] 0.9738531
```

With a FDR control at level 0.15, there is no differentially expressed gene with the raw analysis,

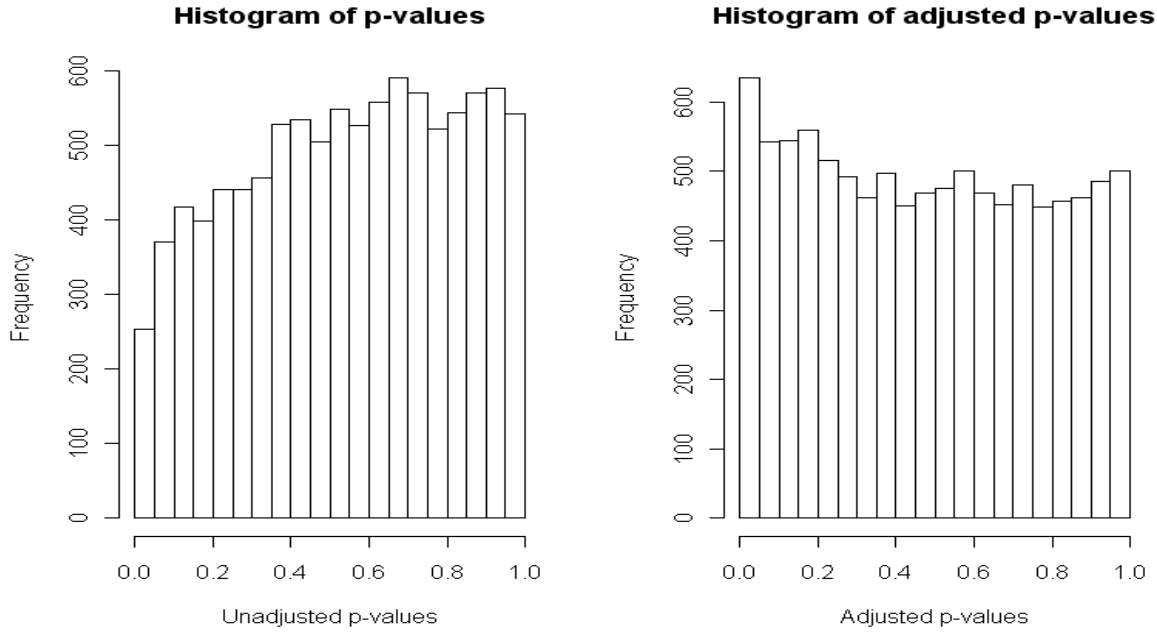


Figure 6.3: Histograms of raw p-values (left panel plot) and factor-adjusted p-values (right panel plot)

whereas 6 genes are differentially expressed with the `FAMT` analysis based on factor-adjusted tests statistics. In order to figure out the differences between both analyses, Figure 6.3 compares the empirical distributions of the raw and the factor-adjusted p-values.

```
R> par(mfrow=c(1,2))
R> hist(modelfinal$pval, main="Histogram of p-values", xlab="Unadjusted p-values")
R> hist(modelfinal$adjpval, main="Histogram of adjusted p-values",
+ xlab="Adjusted p-values")
```

Factor-adjustment restores independence between tests statistics, which results in a correction of the distribution of the p-values from the concave shape observed on the left panel plot of Figure 6.3. Indeed, it seems that a large amount of p-values are uniformly distributed and a few small p-values shall correspond to significant genes.

Note that the user can choose to focus on two aspects of the multiple testing procedure, which are the choice of the optimal number of factors with the `nbfactors` function and the estimation of π_0 with the `pi0FAMT` function.

```
R> nbfactors(chicken, x=c(3,6), test=6, diagnostic.plot=TRUE)
```

This function gives the optimal number of factors as obtained from the `modelFAMT` function and produces the same plot as shown in Figure 6.2.

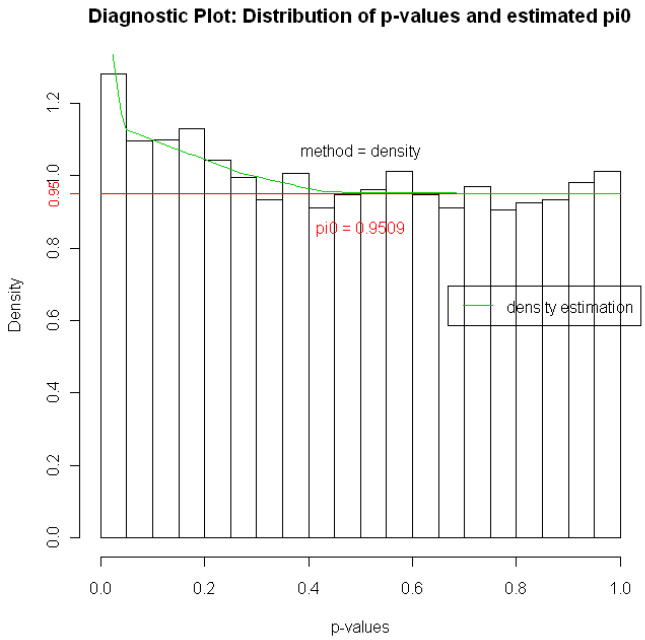


Figure 6.4: Estimation of the proportion of true null hypotheses using a non-parametric estimate of the density function of the p-values proposed by Langaas et al. [2005].

The `pi0FAMT` function provides 2 algorithms to estimate π_0. The `density` method is based on Langaas et al. [2005]'s approach where the density function f of the p-values distribution is estimated assuming f is a convex function: the estimation of π_0 is then $f(p = 1)$. The `smoother` method uses the smoothing spline approach proposed by Storey and Tibshirani [2003]. In most situations, these two methods give similar results but the `smoother` method is numerically less time-consuming.

The following code uses the `density` method to estimate π_0 and produces a histogram of the p-values (Figure 6.4), on which the convex estimation of f is represented.

```
R> pi0FAMT(modelfinal, method="density", diagnostic.plot=TRUE)
```

The estimated value of π_0 is 0.95, which is slightly less than with the `smoother` method ($\hat{\pi}_0 = 0.97$).

6.5. Interpretation of the common factors

The `defacto` function helps the user to give more insight on the common factors using some available external information on either response variables or individuals (see Blum et al. [2010]). The use of this function requires a `FAMTmodel` as returned from the function `modelFAMT` and one or more explanatory variable in `covariates`. The external variables available to describe the responses in `annotations` must be categorical variables. As the factors are designed to be independent from the explanatory variables (the abdominal fatness and the dam in our example), they shall be described

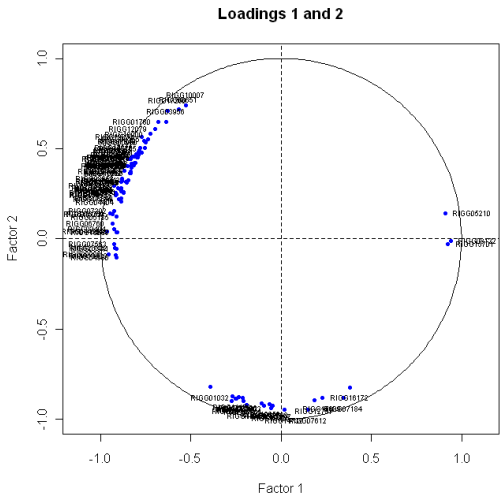


Figure 6.5: Loadings circle plot of the genes in the `chicken` example

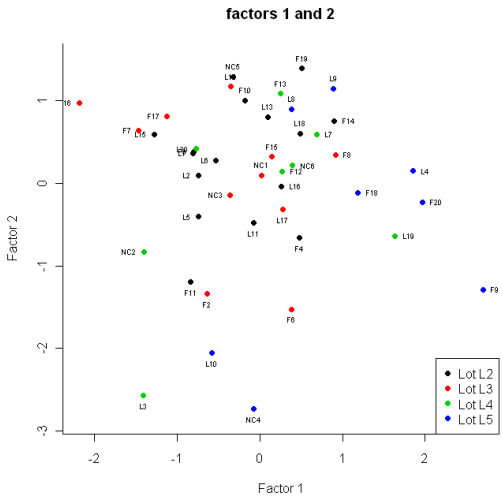


Figure 6.6: Score plot of the microarrays in the `chicken` example

according to the other covariates. In our example, the argument `select.covar=4:5` gives the 2 column numbers in `covariates` picking two external variables: `Lot` and `Pds9s`, which are respectively the hatch and the body weight of the chickens. Similarly, the argument `select.annot=3:6` picks 4 external variables in `annotations`: `Block`, `Column`, `Row` and `Length`.

As for many implementations of PCA-like methods, two plots are provided to summarize the relationships between the latent factors extracted from the data and the external variables. First, if there are at least 2 common factors in the FA model, the `defacto` function provides a loadings circle plot displaying the largest loadings along with two factors which numbers are given by the argument `axes` (the default option is `axes = c(1, 2)`). Points are automatically labelled by their identifier as given in `annotations` (see Figure 6.5). Moreover, the score plot displays the coordinates of the individuals along the two factors, with different colors according to the levels of the factors selected in `covariates` (their hatch in the `chicken` example, see Figure 6.6).

```
R> chicken.defacto = defacto(modelfinal, axes=1:2, select.covar=4:5,
+ select.annot=3:6, cex=0.6)
```

In addition to these plots, F-tests are provided for the significance of the linear relationship between each component of the external information and each factor. The corresponding p-values are given in the `covariates` and `annotations` components of the `defacto` function.

```
R> chicken.defacto$covariates

                  Lot      Pds9s
Factor 1 0.006437319 0.27847793
Factor 2 0.258859549 0.00124608
```

```
Factor 3 0.271648846 0.96813322

R> chicken.defacto$annotations

                  Block    Column        Row
Loadings 1 8.148075e-25 0.2368072 0.21767892
Loadings 2 3.328477e-19 0.9323152 0.01889079
Loadings 3 0.000000e+00 0.1030426 0.68201616
```

Here, `Factor 1` is clearly affected by a hatch effect, and `Factor 2` by a body weight effect. Thus, part of the expression heterogeneity is probably due to these marked biological effects, which are independent of the abdominal fatness, the explanatory variable of main interest in this study.

Moreover, some second-order technological biases turn out to have an impact on the correlation structure of the gene expressions, since the location of the spots on the microarray (`Block`, `Column`, and `Row` here) appears as significantly related to the loadings. According to Qiu et al. [2005], such kinds of effects on the correlation between gene expressions may be induced by the normalization procedure itself. The effect of `Block` is captured by all the factors, and the effect of `Row` by factor 2.

6.6. Second illustrative example

The Animal Genetics Department (INRA-Rennes) studies the transcriptome profiling of the feeding-to-fasting transition in chicken liver. Désert et al. [2008] show that numerous genes are altered by starvation in chickens, and the study suggests a global repression of cellular activity in response to this stressor. In this section, the related gene expression data are used to illustrate that the FAMT method is still useful to increase the power of the multiple testing procedure in a case where a large proportion of genes are significant.

From the 20460 oligos present in the microarray data, 13057 aligning with a unique coding region of the 2.1 Washington University assembly of the chicken sequence genome, were chosen for statistical analyses. The dataset was finally restricted to 7419 genes (out of 13057) presenting a human ortholog with a HUGO (Human Gene Ontology) symbol allowing to recover functional annotations from those databases. 18 microarrays were analyzed: 6 corresponding to fed chickens, 5 to 16-hour fasted animals and 7 to 48-hour fasted animals. We calculate the p-values of the classical Fisher tests. The left panel plot of Figure 6.8 shows that a large number of genes have small p-values, which means that many genes are involved in the fasting process.

Figure 6.7, resulting from the `modelFAMT` function, shows that the variance inflation criterion is minimum for 5 factors. Yet, the `modelFAMT` function proposes `nbf=1` as optimal number of factors, using the Catell scree test criterion (see the previous section). In this case, the plot appears as a useful tool to modify, if necessary, the default number of factors resulting from `modelFAMT`. We finally choose to fit the factor analysis model with 5 factors.

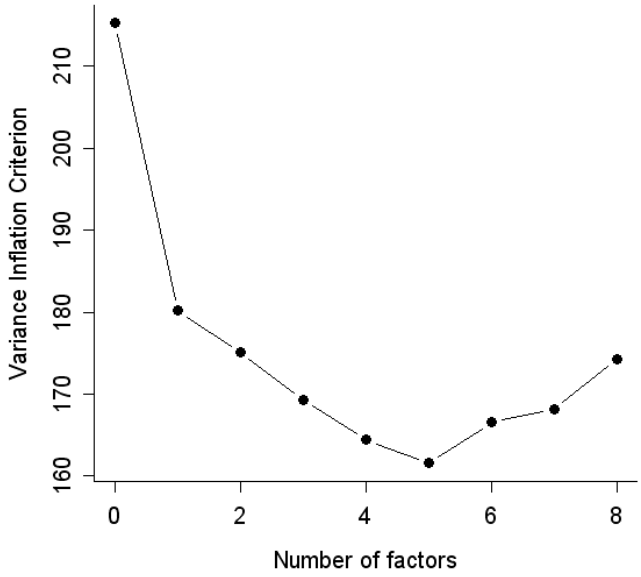


Figure 6.7: Variance inflation criterion

The following code fits the FA model with 5 factors and extracts the numbers of rejected genes for the given FDR control levels.

```
R> model=modelFAMT(Poulets, x=2, nbf=5)
R> rejections =summaryFAMT(model, alpha=seq(0.001,0.01,0.001))$nbreject
R> plot(rejections[,1], rejections[,3], type="l", lty=1, ylab="Number
+ of significant genes", xlab="False Discovery Rate control level",
+   ylim=c(0,7000))
R> lines(rejections[,1], rejections[,2], type="l", lty=2)
R> legend("topleft", c("FAMT","Classical method"), lty=c(1,2))
```

The number of significant genes for various FDR control levels are plotted in Figure 6.9. For a same level of FDR control, more genes are considered as differentially expressed with the FAMT method than using the raw p-values. This illustrates that the FAMT procedure improves the power of the multiple testing procedure since, for a same FDR control level, more genes are significant.

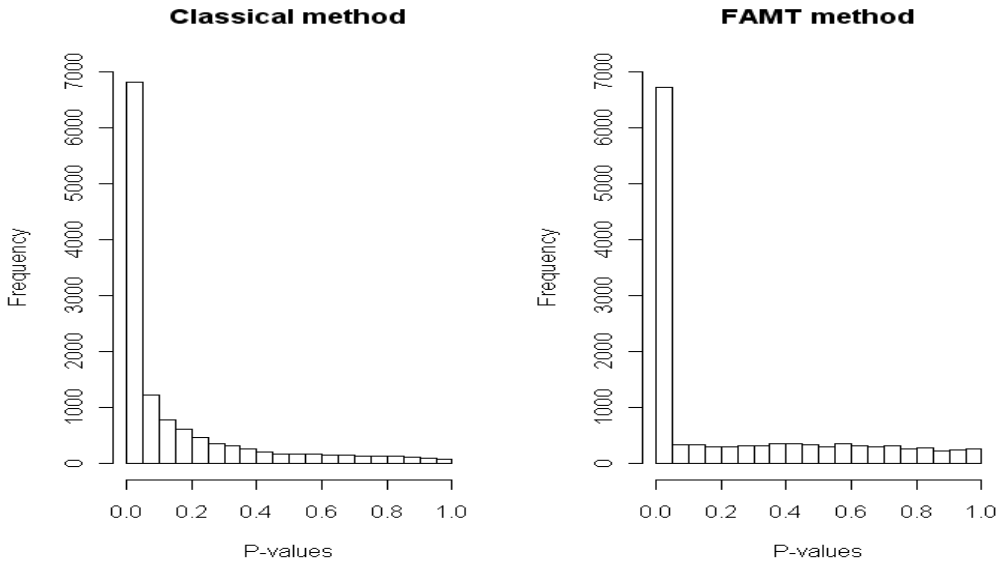


Figure 6.8: Raw p-values and factor-adjusted p-values

6.7. Conclusion

The R package FAMT provides a powerful method for large-scale significance testing under dependence. It is essentially based on a factor modeling of the conditional covariance structure of the response variables. As these factors capture the dependence, there can be used to restore independence among tests, which results in a gain in terms of control of the false discovery proportion and on the overall power of the multiple testing procedure.

The main functions of the package are described in the paper, and are illustrated using a gene expression dataset available in the package. The package offers also tools to help the user describe and interpret the factors using some external informations on either genes or arrays. The functions of the package, their arguments and values, are detailed in the help files. The website http://famt.free.fr sums up the FAMT package and gives news about eventual updates.

Forthcoming versions of the package should include currently studied procedures aiming at inferring on the gene regulatory network using a Gaussian Graphical Model. Excel add-ins should also be included in the next update in order to help non-R users to analyse microarray data using FAMT.

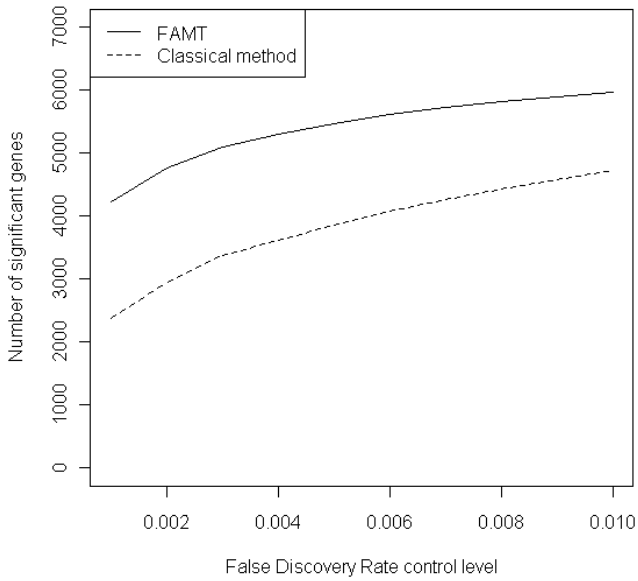


Figure 6.9: Number of significant genes with the raw method and FAMT

CONCLUSION

This thesis deals with studying the impact of dependence in large-scale multiple testing procedures. Motivated by issues raised by the analysis of gene expressions data, the aim is to propose a statistical tool to take into account data heterogeneity in simultaneous hypotheses testing for high-dimensional data.

Heterogeneity may arise from technical or environnemental factors that have not been observed or have not been controled by the experimental design. The proposed method consists in identifying the linear space generated by a set of latent variables that models the heterogeneous structure, catching the common variability shared by all the response variables. The suggested model is related to a Factor Analysis model [Mardia et al., 1979].

The mentioned latent factors are identified as sources of variability which is ignored by the model expectancy and therefore assimilated to heterogeneity. Leek and Storey [2008] also refer to spatial dependence, such as in medical imaging. Finally, the considered modeling of dependence is quite general.

Most of the existing multiple testing procedures rely on the analysis of the empirical process of p-values associated to the individual tests, under the assumption of independence. A thresholding rule takes into account the multiplicity of the tests. There are two main issues in carrying out multiple testing procedures: estimating π_0, the proportion of true null hypotheses, and controlling error rates. The impact of dependence on the stability of these procedures is one of the prime results of this work. Indeed, dependence induces variability that interferes in particular with p-values distribution under the true null hypothesis. The main impact is a sharp deviation from the theorical null distribution when the level of common variability between variables is high. Consequently, the variability of false-positives increases and the estimation of π_0 is biased. More precisely, the variance of the number of false-positives and the variance of π_0 both include a term which explicitly depends on the correlation between the response variables. Dependence has therefore repercussion on the estimation of error rates, leading to high instability in multiple testing procedures.

A procedure is defined from the factor adjusted variables as the data are independent conditionally on the latent structure. Dependence is actually adressed at the level of the original data, integrated

in the model used to calculate the tests statistics. Consequently, the present method illustrates the fact that the well-known individual optimality of the tests indebted to the Neyman-Pearson theory does not imply the global optimality of a multiple testing procedure in situations of dependence between the variables.

As data are independent conditionally on the factors, this framework allows to extend to general dependence the results on error rates control initially gained under independence. Thus, the proposed framework leads to less correlation among tests and shows large improvements of power and stability of simultaneous inference.

Besides, in a high-dimensional setting, representing the covariance matrix with a Factor Analysis structure, through a linear space of moderate dimensions, can be compared to other approaches such as shrinkage methods [Schäfer and Strimmer, 2005]. It turns out to be interesting for the estimation of the partial correlation matrix.

Beyond the study of the model itself, many points concerning the effective implementation of the factor-adjusted multiple testing procedure are addressed in this thesis, including the model parameters estimation thanks to an EM algorithm, the estimation of the proportion of true null hypotheses and the choice of the number of factors. The EMFA algorithm provides accurate estimates of variance parameters in a high-dimensional setting (not asymptotic). We propose a criterion allowing to define the model that fits best the covariance structure, minimising the inflation of variance of false-positives. However, some issues are still to be explored, such as the preliminary estimation of $\mathcal{M}_0$ involved in the calculation of the scores. This last issue is probably very similar to problems encountered by Efron [2007] and by Leek and Storey [2007].

Finally, the method has been applied to microarray data and great improvements for biological interpretation of differential analysis in gene expressions data have been highlighted [Blum et al., 2010]. Representing direct interactions between genes with gene regulatory networks is a continuation of gene expressions differential analysis. Graphical models, where a node corresponds to a gene and an edge corresponds to a biological deppendence, turn out to be interesting tools for modelling multivariate dependence patterns. The estimation of the partial correlations matrix is involved. On-going work on the use of a Factor Analysis model to estimate this matrix gives promising first results (42èmes Journées de la SFdS, 2010). They are also confirmed more analytically [Fan et al., 2008, Ambroise et al., 2009] so that such approach is encouraged.

The `R` package called `FAMT` [Causeur et al., 2010] implements the proposed procedure to make it widely at the users' disposal, mainly in favor of genomics applications. It is available on the `R` project website (`http://www.r-project.org/`) and a website is dedicated to the package (`http://famt.free.fr/`).

APPENDIX

EMFA algorithm

This appendix presents the reasonning behind the up-dating equations in the EM algorithm for Maximum Likelihood Factor Analysis (SECTION 5.2).

In the following, $Y^{(i)}$ is the $1 \times m$-vector corresponding to observation i in matrix Y and $Z^{(i)}$ is the $1 \times Q$-vector corresponding to observation i in matrix Z.

1. **Log-likelihood of the model**

$$\begin{aligned}
\mathcal{L}(B,\Psi) &= \sum_{i=1}^{n} ln\left\{(2\pi)^{-m/2}\left|\Psi\right|^{-1/2} exp\left[-\frac{1}{2}(Y^{(i)}-Z^{(i)}B')\Psi^{-1}(Y^{(i)}-Z^{(i)}B')'\right]\right\} \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[(Y^{(i)}-Z^{(i)}B')\Psi^{-1}(Y^{(i)}-Z^{(i)}B')'\right] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\Big[Y^{(i)}\Psi^{-1}Y'^{(i)}-Z^{(i)}B'\Psi^{-1}Y'^{(i)} \\
&- Y^{(i)}\Psi^{-1}BZ'^{(i)}+Z^{(i)}B'\Psi^{-1}BZ'^{(i)}\Big] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}BZ'^{(i)}+Z^{(i)}B'\Psi^{-1}BZ'^{(i)}\right] \\
&= -\frac{nm}{2}ln(2\pi)+\frac{n}{2}ln\left|\Psi^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\left[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}BZ'^{(i)}+tr\left(B'\Psi^{-1}BZ'^{(i)}Z^{(i)}\right)\right] \quad (4)
\end{aligned}$$

Considering: $Z^{(i)}B'\Psi^{-1}Y'^{(i)} = (Z^{(i)}B'\Psi^{-1}Y'^{(i)})' = Y^{(i)}\Psi^{-1}BZ'^{(i)}$ and $x'Ax = tr(Axx')$ with $A = B'\Psi^{-1}B$ et $x = Z'^{(i)}$.

$$\begin{aligned}
\mathbb{E}(\mathcal{L}|Y) = cst+\frac{n}{2}ln\left|\Psi^{-1}\right| &- \frac{1}{2}\sum_{i=1}^{n}\Big[Y^{(i)}\Psi^{-1}Y'^{(i)}-2Y^{(i)}\Psi^{-1}B\mathbb{E}(Z'^{(i)}|Y^{(i)}) \\
&+ tr\left(B'\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right)\Big] \quad (5)
\end{aligned}$$

2. **Maximisation of** (.5)

- Maximisation wrt B: $\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial B} = 0$

$$\begin{aligned}
\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial B} &= \frac{1}{2}\sum_{i=1}^{n}\left[-2\Psi^{-1}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})' - \Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)}) - \Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'\right] = 0 \\
&\Leftrightarrow \sum_{i=1}^{n}\left[\Psi^{-1}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\right] = \sum_{i=1}^{n}\left[\Psi^{-1}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right] \\
&\Leftrightarrow \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\right] = \sum_{i=1}^{n}\left[B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right] \\
&\Leftrightarrow B = \sum_{i=1}^{n}\left[Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})\right]\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1} \qquad (.6)
\end{aligned}$$

Using $\frac{\partial A'XB}{\partial X} = AB'$ et $\frac{\partial tr(X'AXB)}{\partial X} = AXB + A'XB'$.

- Maximisation wrt Ψ^{-1}: $\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial \Psi^{-1}} = 0$

$$\begin{aligned}
\frac{\partial \mathbb{E}(\mathcal{L}|Y)}{\partial \Psi^{-1}} &= \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n}\left[Y'^{(i)}Y^{(i)} - 2Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' + B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B'\right] \\
&= \frac{n}{2}\Psi - \frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' - \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' = 0 \\
&\Leftrightarrow \Psi = \frac{2}{n}\left[\frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' - \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right]
\end{aligned}$$

Using (.6) to "simplify" this equation:

$$\begin{aligned}
\Psi &= \frac{2}{n}\left[\frac{1}{2}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{2}\sum_{i=1}^{n}B\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B' - \sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} + \frac{1}{n}\left\{\left[\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'\left[\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})\right]^{-1}\right]\sum_{i=1}^{n}\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})'B'\right. \\
&\quad \left. - 2\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B'\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}Y^{(i)} - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z'^{(i)}|Y^{(i)})'B' \\
&= S - \frac{1}{n}\sum_{i=1}^{n}Y'^{(i)}\mathbb{E}(Z^{(i)}|Y^{(i)})B' \qquad (.7)
\end{aligned}$$

Ψ is estrimated considering the diagonal in (.7).

3. **Calcul of** $\mathbb{E}(Z'^{(i)}|Y^{(i)})$ **and** $\mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)})$**.** The density of $Z|Y$, denoted $f(z|y)$ is:

$$\begin{aligned}
f(z|y) &= \frac{f(z,y)}{f(y)} \\
&= \frac{(2\pi)^{-(m+Q)/2}|C|^{-1/2}exp\left\{-\frac{1}{2}(y,z)C^{-1}(y,z)'\right\}}{(2\pi)^{-m/2}|\Sigma|^{-1/2}exp\left\{-\frac{1}{2}y\Sigma^{-1}y'\right\}} \\
&= (2\pi)^{-Q/2}\frac{|C|^{-1/2}}{|\Sigma|^{-1/2}}exp\left\{-\frac{1}{2}\left[(y,z)C^{-1}(y,z)' - y\Sigma^{-1}y'\right]\right\} \\
&= (2\pi)^{-Q/2}\frac{|C|^{-1/2}}{|\Sigma|^{-1/2}}exp\left\{-\frac{1}{2}\theta\right\} \qquad (.8)
\end{aligned}$$

Where $C = \mathbb{V}(Y,Z) = \begin{bmatrix} \Sigma & B \\ B' & \mathbb{I}_Q \end{bmatrix}$.

$$C^{-1} = \begin{bmatrix} (C^{-1})_{11} & (C^{-1})_{12} \\ (C^{-1})_{21} & (C^{-1})_{22} \end{bmatrix} = \begin{bmatrix} \Psi^{-1} & \Sigma^{-1}B'(B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1} \\ (B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1}B\Sigma^{-1} & \mathbb{I}_Q + B'\Psi^{-1}B \end{bmatrix} \quad (.9)$$

Considering the expression of θ in the exponential term of (.8).

$$\begin{aligned} \theta &= y(C^{-1})_{11}y' + y(C^{-1})_{12}z' + z(C^{-1})_{21}y' + z(C^{-1})_{22}z' - y\Sigma^{-1}y' \\ &= y[(C^{-1})_{11} - \Sigma^{-1}]y' + 2y(C^{-1})_{12}z' + z(C^{-1})_{22}z' \end{aligned}$$

Indeed, $(C^{-1})_{12} = (C^{-1})'_{21}$. Moreover, considering $(C^{-1})_{11} - \Sigma^{-1} = \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}$:

$$\begin{aligned} \theta &= y[\Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}]y' + 2y[\Sigma^{-1}B'(B'\Sigma^{-1}B - \mathbb{I}_Q)^{-1}]z' + z[\mathbb{I}_Q + B'\Psi^{-1}B]z' \\ &= [z - y\Sigma^{-1}B][\mathbb{I}_Q + B'\Psi^{-1}B][z' - B'\Sigma^{-1}y'] \\ &= [z - y\Sigma^{-1}B][\mathbb{I}_Q + B'\Psi^{-1}B][z - y\Sigma^{-1}B]' \quad (.10) \end{aligned}$$

As $[B'\Sigma^{-1}B - \mathbb{I}_Q]^{-1} = -[\mathbb{I}_Q + B'\Psi^{-1}B]$. Therefore, considering $f(z|y)$ as a mixture of gaussian distribution, with mean $y\Sigma^{-1}B$ et de variance $G = [\mathbb{I}_Q + B'\Psi^{-1}B]^{-1}$.

Consequently, $\mathbb{E}(Z^{(i)}|Y^{(i)})$ is derived:

$$\begin{aligned} \mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Sigma^{-1}B = Y^{(i)}[\Psi^{-1} - \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}]B \\ &= Y^{(i)}[\Psi^{-1}B - \Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}B] \\ &= Y^{(i)}\Psi^{-1}B[\mathbb{I}_Q - (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}B'\Psi^{-1}B] \\ &= Y^{(i)}\Psi^{-1}B[\mathbb{I}_Q - (\mathbb{I}_Q + B'\Psi^{-1}B)(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} + (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1}] \\ &= Y^{(i)}\Psi^{-1}B(\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} \quad (.11) \end{aligned}$$

And then $\mathbb{E}(Z^{(i)}|Y^{(i)})$ is derived:

$$\begin{aligned} \mathbb{E}(Z'^{(i)}Z^{(i)}|Y^{(i)}) &= \mathbb{V}(Z^{(i)}|Y^{(i)}) + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \\ &= [(C^{-1})_{22}]^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \\ &= (\mathbb{I}_Q + B'\Psi^{-1}B)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \quad (.12) \end{aligned}$$

The EMFA algorithm is then described as the following:

1. **Initialization**: rough estimation $\hat{B}_0$ and $\hat{\Psi}_0$.
2. **Iterations**
   - **E-step**: Expectation of the log-likelihood

$$\begin{aligned} \mathbb{E}(Z^{(i)}|Y^{(i)}) &= Y^{(i)}\Psi_0^{-1}B_0(\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} \\ \mathbb{E}(Z^{(i)}Z'^{(i)}|Y^{(i)}) &= (\mathbb{I}_Q + B_0'\Psi_0^{-1}B_0)^{-1} + \mathbb{E}(Z^{(i)}|Y^{(i)})'\mathbb{E}(Z^{(i)}|Y^{(i)}) \end{aligned}$$

- **M-step**: Estimation of the ML estimators for B and Ψ

$$\begin{aligned} B_1 &= \sum_{i=1}^{n} \left[Y'^{(i)} \mathbb{E}(Z^{(i)}|Y^{(i)}) \right] \left[\sum_{i=1}^{n} \mathbb{E}(Z'^{(i)} Z^{(i)}|Y^{(i)}) \right]^{-1} \\ \Psi_1 &= diag \left[S - \frac{1}{n} \sum_{i=1}^{n} Y'^{(i)} \mathbb{E}(Z^{(i)}|Y^{(i)}) B_1' \right] \end{aligned}$$

3. **Stop**: The E- and M-steps are alternated repeatedly until convergence, which may be determined by using a suitable stopping rule like for example, $tr(\Psi_0 - \Psi_1) < \varepsilon$, $\varepsilon > 0$.

BIBLIOGRAPHY

C. Ambroise, J. Chiquet, and C. Matias. Inferring sparse gaussian graphical models with latent structure. *Electronic Journal of Statistics*, 3:205–238, 2009.

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

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

Y. Benjamini, A. Krieger, and D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93:491–507, 2006.

M. A. Black. A note on the adaptative control of false discovery rates. *Journal of the Royal Statistical Society. Series B*, 66:297–304, 2004.

G. Blanchard and E. Roquain. Two simple sufficient conditions for FDR control. *Electronic journal of Statistics*, 2:963–992, 2008.

Y. Blum, G. LeMignon, S. Lagarrigue, and D. Causeur. A factor model to analyze heterogeneity in gene expression. *BMC bioinformatics*, 11:368, 2010.

C. E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore si Scienze Economiche e Comerciali di Firenze*, pages 3–62, 1936.

R. B. Cattell. The scree test for the number of factors. *Multivariate Behavioural Research*, 1:245–276, 1966.

D. Causeur, C. Friguet, M. Houée, and M. Kloareg. Factor analysis for multiple testing (famt): an r package for large-scale significance testing under dependence. *Journal of Statistical Software, submitted*, 2010.

Alan Dabney, John D. Storey, and with assistance from Gregory R. Warnes. *qvalue: Q-value estimation for false discovery rate control*, 2009. URL `http://CRAN.R-project.org/package=qvalue`. R package version 1.20.0.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 34:1–38, 1977.

C. Désert, M. J. Duclos, P. Blavy, F. Lecerf, F. Moreews, C. Klopp, M. Aubry, F. Herault, P. Le Roy, C. Berri, M. Douaire, C. Diot, and S. Lagarrigue. Transcriptome profiling of the feeding-to-fasting transition in chicken liver. *BMC Genomics*, 9:611, 2008.

S. Dudoit and M.J. VanDerLaan. *Multiple testing procedures with application to genomics*. 2008.

S. Dudoit, J. Fridlyand, and T.P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

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

B. Efron. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99:96–104, 2004.

B. Efron. Correlation and large-scale simultaneous testing. *Journal of the American Statistical Association*, 102:93–103, 2007.

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

Y. Escouffier. Le traitement des variables vectorielles. *Biometrics*, 29:751–760, 1973.

L.R. Fabrigar, R. MacCallum, D.T. Wegener, and E.J. Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4:272–299, 1999.

J. Fan, Y. Fan, and J. Lv. High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, 147:186–197, 2008.

J.K. Ford, R. MacCallum, and M. Tait. The application of exploratory factor analysis in applied psychology: a critical review and analysis. *Personnel Psychology*, 39:291–314, 1986.

C. Friguet, M. Kloareg, and D. Causeur. A factor model approach to multiple testing under dependence. *submitted*, 2008.

C. Genovese and L. Wasserman. Operarting characteristics and extensions of the false discovery rate procedure. *Journal of the Royal Statistical Society. Series B*, 64:499–517, 2002.

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

A. Gordon, G. Glazko, X. Qiu, and A. Yakovlev. Control of the mean number of false discoveries, bonferroni, and stability of multiple testing. *Annals of Applied Statististics*, 1:179–190, 2007.

M. Guedj, G. Nuel, S. Robin, and A. Celisse. *kerfdr: semi-parametric kernel-based approach to local* FDR *estimations*, 2007. URL `http://stat.genopole.cnrs.fr/sg/software/kerfdr`. R package version 1.0.1.

T.J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. 1990.

Trevor Hastie, Robert Tibshirani, Balasubramanian Narasimhan, and Gilbert Chu. *impute: Imputation for microarray data*, 2009. URL `http://CRAN.R-project.org/package=impute`. R package version 1.20.0.

J.C. Hayton, D.G. Allen, and V. Scarpello. Factor retention in exploratory factor analysis: a tutorial on parallel analysis. *Organisational research Methods*, 7:191–205, 2004.

S. Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6:65–70, 1979.

J.L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 32: 179–185, 1965.

J.C. Hsu. The factor analytic approach to simultaneous inference in the general linear model. *Journal of computational and graphical statistics*, 1:151–168, 1992.

J. Josse, J. Pagès, and F. Husson. Testing the significance of the RV coefficient. *Computational Statististics and Data Analysis*, 53:82–91, 2008.

H. F. Kaiser. The application of electronic computers to factor analysis. *Educationnal and Psychological Measurment*, 20:141–151, 1960.

C. Kendziorski, H. Newton, M.and Lan, and M. Gould. On parametric empirical bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine*, 22: 3899–3914, 2003.

K. I. Kim and M. Van de Wiel. Effects of dependence in high-dimensional multiple testing problems. *BMC Bioinformatics*, 9, 2008.

R. Kustra, R. Shioda, and M. Zhu. A factor analysis model for functional genomics. *BMC Bioinformatics*, 7, 2006.

M. Langaas, B. H. Lindqvist, and E. Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society. Series B*, 67: 555–572, 2005.

J. T. Leek and J. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3(9):e161, 2007.

J. T. Leek and J. Storey. A general framework for multiple testing dependence. *Proceedings of the National Academy of Sciences*, 105:18718–18723, 2008.

G. LeMignon, C. Désert, F. Pite, S. Leroux, O. Demeure, G. Guernec, B. Abasht, M. Douaire, P. LeRoy, and S. Lagarrigue. Using transcriptome profiling to characterize qtl regions on chicken chromosome 5. *BMC Genomics*, pages 10–575, 2009.

I. Lönnstedt and T.P. Speed. Replicated microarray data. *Statistica Sinica*, 12:31–46, 2002.

R. MacCallum, K. Widaman, S. Zhang, and S. Hong. Sample size in factor analysis. *Psychological Methods*, 4:84–99, 2008.

K. V. Mardia, J. T. Kent, and J. M. Bibby. *Multivariate Analysis*. 1979.

G.J. McLachlan, T. Krishnan, and S.K. NG. The EM algorithm. http://www.econstor.eu/bitstream/10419/22198/1/24_tk_gm_skn.pdf, 2004. Humboldt-Universitat Berlin, Center for Applied Statistics and Economics.

X.L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: a general framework. *Boimetrika*, 80:267–278, 1993.

R. G. Montanelli and L. G. Humphrey. Latent roots of ranrom data correlatoin matrices with squared multiple correlations on the diagonal: a monte-carlo study. *Psychometrica*, 41:341–348, 1976.

R.M. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and oter variants. In Michael I. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer Academic, 1998.

M. Norris and L. Lecavalier. Evaluating the use of exploratory factor analysis in developmental disability psychological research. *Journal of Autism and Developpement Disorders*, 2009.

A.B. Owen. Variance of the number of false discoveries. *Journal of the Royal Statistical Society. Series B*, 67:411–426, 2005.

K. Pollard, Y. Ge, S. Taylor, and S. Dudoit. *multtest: Resampling-based multiple hypothesis testing*. R package version 1.23.3.

I. Pournara and L. Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC Bioinformatics*, 8:61, 2007.

K. Preacher and R. MacCallum. Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics*, 32:153–161, 2002.

X. Qiu, A. Brooks, L. Klebanov, and A. Yakovlev. The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics*, 6:120, 2005.

W. Revelle and T. Rocklin. Very simple structure: an alternative procedure for estimating the opimal number of interprétable factors. *Multivariate Behavioural Research*, 14:403–414, 1979.

S. Robin, A. Bar-Hen, J-J. Daudin, and L. Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation". *Computational Statistics & Data Analysis*, 51:5483 – 5493, 2007.

D. B. Rubin and D. T. Thayer. EM algorithms for ML factor analysis. *Psychometrika*, 47:69–76, 1982.

D. Salsburg. The lady tasting tea: How statistics revolutionized science in the twentieth century. *ISBN 0-8050-7134-2*, 2002.

S. Sarkar. Two-stage stepup procedures controlling fdr. *Journal of Statistical Planning and Inference*, 138:1072–1084, 2008.

J. Schäfer and K. Strimmer. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4:32, 2005.

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

J. Shaffer. Multiple hypotheses testing: a review. *Annual review of psychology*, 46:561–584, 1995.

Z. Sidak. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62:626–633, 1967.

B. W. Silverman. *Density estimation for statistics and data analysis*. 1986.

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

G. Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3:article 3, 2004.

C. Spearman. General intelligence, objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

D.W. Stewart. The application and misapplication of factor analysis in marketing research. *Journal of Marketing Research*, 18:51–62, 1981.

J. Storey and R Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100:9440–9445, 2003.

J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society. Series B*, 64:479–498, 2002.

J. D. Storey. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035, 2003.

J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society. Series B*, 66:187–205, 2004.

J. D. Storey, J.Y. Dai, and J. T. Leek. The optimal discovery procedure for large-scale significance testing, with application to comparative microarray experiments. *Biostatistics*, 8:414–432, 2007.

W. F. Velicer. Determining the number of components from the matrix of partial correlations. *Psychometrika*, 41:321–327, 1976.

W. F. Velicer, C. A. Eaton, and J. L. Fava. Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. *in R. D. Goffin and E. Helmes, eds.*, pages 41–71, 2000.

C.G.C. Wei and M.A. Tanner. A monte carlo implementaion of the EM algorithm and the poor man's data augmentation algorithm. *Journal of the American Statistical Association*, 85:699–704, 1990.

P. H. Westfall and S. S. Young. *Resampling-based multiple testing: examples and methods for p-value adjustment*. 1993.

H. Yifan, H. Xu, V. Calian, and J.C. Hsu. To permute or not to permute. *Bioinformatics*, 22: 2244–2248, 2006.

LIST OF ARTICLES AND COMMUNICATIONS

Articles are available in the APPENDIX *of the French document. Abstracts of communications are available on my website:* `http://friguetchloe.wordpress.com`.

Articles in international journals

- **C. Friguet**, M. Kloareg & D. Causeur (2009) - A factor model approach to multiple testing under dependence, *Journal of the American Statistical Association* - 104:488,1406-1415
  DOI: 10.1198/jasa.2009.tm08332
- D. Causeur, M. Kloareg. & C. Friguet (2009) - Control of the FWER in Multiple Testing Under Dependence, *Communications in Statistics - Theory and Methods* - 38:16,2733-2747
  DOI: 10.1080/03610910902936281
- **C. Friguet** & D. Causeur (2010) - Estimation of the proportion of true null hypotheses in high-dimensional data under dependence, *submitted, 05/2010*
- D. Causeur, C. Friguet, M. Houée & M. Kloareg - Factor Analysis for Multiple Testing (FAMT): an R package for large-scale significance testing under dependence, *submitted, 06/2010*

Communications

2008 • **C. Friguet**, M. Kloareg & D. Causeur - Accounting for a factor structure in high-dimensional data to improve multiple testing procedures
6th workshop Statistical methods for post-genomic data
Rennes, 2008

- **C. Friguet**, M. Kloareg & D. Causeur - Multiple tests for high-throughput data assuming a factor modeling of dependence
  40èmes Journées de Statistiques Société Française de Statistiques/Société Statistique Canada
  Ottawa, Canada, 2008
- D. Causeur, M. Kloareg & C. Friguet - Impact of dependence on the stability of model selection in supervised classification for high-throughput data
  International Indian Statistical Association (IISA) Conference "Frontiers of Probability and Statistical Science"
  Storrs, Connecticut, USA, 2008
- **C. Friguet** - La Statistique: un outil pour comprendre les données génomiques
  Doctoriales® de Bretagne
  Brest, 2008 (exposé + poster)
- **C. Friguet** - Approche conditionnelle des tests multiples pour données biologiques à haut-débit
  7ème Journée Jeunes Chercheurs en Biométrie (Société Française de Biométrie)
  INSERM, Villejuif, 2008

2009

- **C. Friguet**, M. Kloareg & D. Causeur - Factor Analysis for Multiple Testing: A general approach for differential analysis of genome-scale dependent data
  Workshop Statistical advances in Genome-scale Data Analysis
  Ascona, Suisse, 2009 (poster)
- M. Kloareg, C. Friguet & D. Causeur - Factor Analysis for Multiple Testing (FAMT) : an R package for simultaneous tests under dependence in high-dimensional data
  UseR!2009, conférence des utilisateurs de R
  Agrocampus OUEST, Rennes, 2009
- **C.Friguet** & D. Causeur - Estimation conditionnelle de la proportion d'hypothèses nulles en grande dimension
  41èmes Journées de Statistiques, Société Française de Statistiques
  Bordeaux, 2009
- **C. Friguet** & D. Causeur - Estimation of the proportion of null p-values among dependent tests
  Workshop on Simulation
  St Petersbourg, Russia, 2009
- M. Kloareg, C. Friguet, Y. Blum & D. Causeur - Factor Analysis for Multiple Testing: large scale significance testing under dependance
  EMBL
  Heidelberg, Allemagne, 2009

2010
- **C. Friguet** - Impact de la dépendance en analyse différentielle en grande dimension
  Séminaire du groupe de travail en biostatistique
  Institut Elie Cartan, Nancy, 2010
- Y. Blum, C. Friguet, S. Lagarrigue & D. Causeur - Inférence sur réseaux géniques par Analyse en Facteurs
  42èmes Journées de Statistiques, Société Française de Statistiques
  Marseille, 2010