Sorbonne Université
LPSM

Doctoral School **École Doctorale Sciences Mathématiques de Paris Centre**

University Department **Laboratoire de Probabilités, Statistique et Modélisation**

Thesis defended by **Iqraa MEAH**

Defended on **November 30, 2023**

In order to become Doctor from Sorbonne Université

Academic Field **Applied mathematics**

Speciality **Statistics**

Controlling False Discovery Proportion in Structured Data Sets

Thesis supervised by Sebastian Döhler
Etienne Roquain

Committee members

Referees	Christophe Ambroise	Professor at Université d'Evry	
	Jelle Goeman	Professor at Leiden University	
Examiners	Stephane Robin	Professor at Sorbonne Université	Committee President
	Sylvain Arlot	Professor at Université Paris-Saclay	
	Magalie Fromont	Professor at Université Rennes 2	
	Antje Jahn	Professor at Hochschule Darmstadt	
Supervisors	Sebastian Döhler	Professor at Hochschule Darmstadt	
	Etienne Roquain	Associate Professor at Sorbonne Université	

This thesis has been prepared at the following research units.

Laboratoire de Probabilités, Statistique et Modélisation

Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

☎ +33 1 57 27 93 16
Web Site `https://www.lpsm.paris/`

Department of mathematics

Hochshule Darmstadt
University of Applied Sciences
Haardtring 100
64295 Darmstadt
Germany

☎ +49.6151.16-02
Web Site `https://h-da.de/en/`

Controlling False Discovery Proportion in Structured Data Sets

Abstract

The present work proposes new methodologies for controlling the False Discovery Proportion (FDP) while accommodating different types of data structures arising from the underlying scientific context. Since the seminal work of Benjamini and Hochberg (1995) (BH) introducing the FDP, multiple testing procedures have found widespread applications across diverse domains. The BH procedure has facilitated the identification of significant variables within large data sets, providing insights to scientific questions in fields such as biology, medicine, or marketing research, by ensuring guarantees on the proportion of false discoveries. However, the BH procedure has several limitations, among which e.g. the fact that it is most effective for uniform p-values under the null; it is developed within a batch framework requiring simultaneous availability of all p-values; the false discoveries control guarantee is only in expectation. These limitations can lead to a range of unfavorable outcomes – spanning from reduced interpretability, loss of statistical power, to potential inflation of the Type I error rate – particularly in contexts where we perceive the data as possessing inherent "structure." This work aims to push back those limits by providing new procedures and methodologies that adapt to settings where p-values can be discrete, online, preordered, or weighted. This ultimately gives the practitioner more effective tools for identifying significant variables in structured data sets as we illustrate in various numerical experiments.

Keywords: multiple testing, discrete p-values, online p-values, weighted p-values, preordered p-values, (m)FDR control, FDP confidence bounds, plug-in FDR control

Résumé

Ce travail propose de nouvelles méthodologies pour contrôler la proportion de fausses découvertes (FDP) tout en prenant en compte différentes types de structures de données résultant du contexte scientifique sous-jacent. Depuis le travail fondamental de Benjamini and Hochberg (1995) (BH) introduisant le FDP, les procédures de tests multiples ont trouvé une application dans de nombreux domaines. La procédure de BH a facilité l'identification de variables significatives dans de grands ensembles de données, permettant de répondre à des questions scientifiques dans des domaines tels que la biologie, la médecine ou le marketing, tout en fournissant des garanties sur la proportion de fausses découvertes. Toutefois, la procédure de BH présente plusieurs limites : elle est plus efficace pour des p-valeurs uniformes sous l'hypothèse nulle ; elle est développée dans un cadre *offline* nécessitant la connaissance simultanée de toutes les p-valeurs ; la garantie de contrôle des fausses découvertes est en espérance. Ces limitations peuvent entraîner une perte de puissance, une réduction de l'interprétabilité, voire même une inflation de l'erreur de Type I dans différents contextes où les données sont considérées comme "structurées". Ce travail vise à combler ces lacunes en fournissant de nouvelles procédures et méthodologies qui s'adaptent à des contextes structurels où les p-valeurs peuvent être discrètes, en ligne, pré-ordonnées ou pondérées. Cela donne, in fine, au praticien des outils plus efficaces pour identifier les variables significatives dans un ensemble de données structurées, comme nous l'illustrons dans diverses expériences numériques.

Mots clés : tests multiples, p-valeurs discrètes, p-valeurs en ligne, p-valeurs pondérées, p-valeurs ordonnées, contrôle du (m)FDR, bornes de confiance pour le FDP, contrôle du plug-in FDR

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Contents

Chapter 1

Introduction

Outline of the current chapter

In this chapter, we start by motivating our work with concrete applications and models for which multiple decisions should be made. In Section 1.2, we formalize the Multiple Testing (MT) setting and present a review of the existing research in the field. Then, in Section 1.3, we outline the particular settings of interest in our work pertaining to certain p-value structures and discuss the challenges associated with them. Finally, we provide an overview of the subsequent chapters, and summarize their contents in Section 1.4.

1.1 Multiplicity in scientific research

1.1.1 Application examples

In many applications, scientists face complex problems for which insights are gained by answering multiple questions of the same nature. We briefly describe several examples of such applications

to exemplify the notion of multiplicity in scientific research endeavors.

Clinical trials Clinical trials are research studies that assess a medical, surgical, or behavioral intervention on a given group of people. These trials allow researchers to determine if a new form of treatment or prevention (e.g., a new drug, diet, or medical device) is safe and effective. For instance, to market a new drug, pharmaceutical companies investigate what different side effects the new drug is associated with. Since many possible side effects are considered, multiple assessments need to be provided, for more detailed examples, see e.g. Chapter 1 of Dmitrienko et al. (2009). Such associations are also investigated for already marketed drugs with pharmacovigilance systems. These systems collect and monitor spontaneous reports of suspected adverse events for a number of marketed medicines. As a case in point, Chavant et al. (2011) study the association between drugs and amnesia.

Molecular biology In molecular biology, modern high-throughput technologies have allowed scientists to collect precise genotypic information at different scales containing tens or hundreds of thousands of measurements (see e.g. Uffelmann et al. (2021) for a description of the data collection process). These high-resolution datasets paved the way for a better understanding of associations between the genome and biological traits (e.g., diseases or phenotypic expressions). In this context, a massive amount of variables representing genetic information are investigated for associations with a specific trait. Nowadays, a number of such datasets are publicly available by consortiums of scientists. This is the case e.g. for the *International Mice Phenotyping Consortium* (IMPC) which aims at understanding the genotype effect on the phenotype of mice through gene knock-out studies, see Muñoz-Fuentes et al. (2018) for more details.

Neuroscience Brain mapping helps to associate regions of the brain with cognitive function or disorders to allow neuroscientists gain a better understanding of the brain and its diseases. More precisely, the goal is to identify regions of the brain that are active when a person performs a certain task or when a person's senses are stimulated. For this, functional Magnetic Resonance Imaging (fMRI) technologies provide images of the blood flow in the brain when the task or stimuli of interest is performed. fMRI data are made up of 50 to 400 thousands of 3D pixels called voxels. For specific regions of interest, the magnitude of these voxels are compared altogether with reference measures coding for the inactivity of the corresponding brain regions. Examples of such data are made available by Gorgolewski et al. (2015) and for recent studies on brain mapping using fMRI see e.g. Varoquaux et al. (2018) or Nowinski (2021).

Astrophysics Detecting exoplanets, stars or ultra faint galaxies involves comparing measures of candidate sources with reference benchmarks. For instance, when working with astrophysical images, measures can refer to pixel magnitudes in multi-wavelength. In such contexts, one goal is to detect celestial objects of interest that could stand out by local highlights. However, these highlights can also be caused by instrumental noisy artifacts so the detection of new objects can be erroneous. Instances of such astrophysical data are provided by the multi-unit spectroscopic explorer (MUSE), see Bacon et al. (2010) for more details. See Dumusque et al. (2012); Mary et al. (2020) for detailed studies on exoplanet and ultra-faint galaxy detection.

In all these applications, a common objective is to identify a set of items (drugs, genes, brain regions, galaxies) that seem relevant to tackle the underlying scientific problem. While the general aim is to identify precisely that set of items, sub-tasks may involve evaluating the cardinal of that set (number of interesting items) or quantifying the number of false discoveries

(items declared as interesting while they are not) within a selected subset. These tasks can be undertaken using appropriate statistical models whose definitions are based on the nature of the scientific problem.

1.1.2 Statistical modeling and P-values

Statistical testing is a classical tool used by the scientific community to address the aforementioned types of questions. The distinguishable feature of the applications of Section 1.1.1 is that a potentially large number $m > 1$ of null hypotheses are considered simultaneously. This task is referred to as *Multiple Testing* (MT) and can appear in a range of statistical models related to different underlying scientific questions. We present two statistical models that can instantiate an MT task. As we will solely focus on p-value based testing in this work, for each of these models we briefly present how the p-values are generated.

Two sample tests Consider a variable Y coding for a given outcome (e.g. medication side effects, or phenotypic trait), and another variable X coding for a potentially correlated cause (e.g. new drug, or gene allele). Investigating the association between Y and X can be formulated as a statistical testing problem, where the null hypothesis declares no association between the variables. A classical approach to carry out this test is with two-sample testing where observations of Y are collected from two different random and independent populations that differ in terms of the value of X. In general, the specific test used (e.g., χ^2-test, Fisher's Exact Test (FET), Student's t-test) depends on the data structure (categorical or continuous) and the underlying distributional assumption. For binary counts, FET is suitable and we present it in more detail as it will also appear in further discussions.

FETs are used for data summarized in a 2×2 contingency table when investigating two categorical variables (see Table 1.1). The rows in the matrix represent two different groups where the potential correlated cause X is either observed or not. For example, in medical studies, individuals from the control group are given a placebo ($X = 0$), while individuals from the case group are given the new drug ($X = 1$). The columns in the matrix indicate the observed counts of the outcome variable Y within each respective group. For instance, $Y = 1$ codes the occurrence of a side effect, while $Y = 0$ indicates that no side effect was observed.If there is no association and the columns and rows totals are treated as fixed, the distribution of the entries – n_{11}, n_{12}, n_{21}, n_{22}– can be described according to a hypergeometric distribution. Then, an exact (non-asymptotic) p-value is computed by summing probabilities corresponding to tables, having the same margins, which are at least as extreme under the null as the one observed.

	$Y = 1$	$Y = 0$	Total
$X = 0$	n_{11}	n_{12}	$n_{1\cdot}$
$X = 1$	n_{21}	n_{22}	$n_{2\cdot}$
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	n

Table 1.1: Example of 2×2 contingency table for association study

As motivated in Section 1.1.1, association studies can be designed to investigate several numbers of variables $X_1, \ldots, X_m$ for potential association with one or different outcomes Y's, consequently providing multiple 2×2 contingency tables, and thus multiple p-values, one for each variable $X_i, 1 \leq i \leq m$, we want to test. This is typically the case for the aforementioned IMPC dataset and for pharmacovigilance data, see respectively Karp et al. (2017) and Ahmed et al. (2010) for results of particular studies using FETs.

Gaussian linear model Consider $Y \in \mathbb{R}^n$ a response variable that one wants to predict from explanatory variables $X = (X_1, \dots, X_m) \in \mathbb{R}^{n \times m}$. A standard textbook model is the Gaussian linear model written as

$$Y = X\beta^* + \epsilon,$$

where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ is the noise, with a possibly unknown variance $\sigma^2 > 0$, and $\beta^* \in \mathbb{R}^m$ is the unknown parameter of interest. The gaussian linear model is one of the vanilla statistical models used in many domains including genomics, see e.g. Wu and Ye (2006); Yang et al. (2017b). In this context, one classical goal is to identify the non-zero entries of β. Indeed, a non-null entry β_i^* of β^* indicates that variable X_i is active in the model i.e. that there is an interaction between Y and X_i, making the latter appropriate for the regression. Identifying the non null entries of β^* can be performed via MT: for each variable X_i, $j \in \{1, \dots, m\}$ test the null hypothesis $H_{0,i} : \beta_i = 0$ against the alternative $H_{1,i} : \beta_i \neq 0$.

In high dimensional statistics, where $m \approx n$ or $m \gg n$ this task is more difficult so that sparsity is classically assumed, implying that β^* contains only a few non-zero entries, which means that only a few variables in $X = (X_1, \dots, X_m)$ are in fact necessary to predict Y. While both the Ordinary Least Squares (OLS) and LASSO estimators of Tibshirani (1996) can generate test statistics to carry out the MT task, they have specific limitations. The OLS model is not suitable for high-dimensional settings. The LASSO estimator is better under the sparsity assumption but does not provide distributional information for the test statistic under the null hypothesis, see Giraud (2021) for more details. To counteract these limitations, Barber and Candès (2015) propose an approach that involves augmenting the model with so-called "knockoffs" of the covariates $X = (X_1, \dots, X_m)$ and performing the LASSO on this augmented model. Artificial binary p-values are then constructed to indicate the sign of the test statistic built by comparing the estimated values of β^* for the true variables with those for the knockoff variables.

MT tasks also arise in other contexts. For instance, reliable classification can be performed using conformal p-values generated empirically using non-conformity scores with respect to a null reference set, see e.g. Bates et al. (2023) or Marandon et al. (2022) and references therein for more details. MT can also occur in non-parametric density estimation to gain point-wise knowledge about the density function, see e.g. Blanchard et al. (2014).

1.2 Multiple testing and false discovery control

MT has been developed to assess multiple statistical inferences simultaneously. These methods avoid an excessive number of false conclusions, corresponding to false rejections of the null hypothesis called false discoveries/rejections. To formally discuss the challenges induced by multiplicity we first introduce the general setting.

1.2.1 Setting

Denote the observations by X which is a random variable (r.v) generated by an unknown distribution P that belongs to a set $\mathcal{P}$ of possible distributions. Consider m null hypotheses for P, denoted $H_{0,i}$, $1 \leq i \leq m$. Let the corresponding set of true null hypotheses denoted by $\mathcal{H}_0(P) = \mathcal{H}_0 = \{1 \leq i \leq m \, : \, H_{0,i} \text{ is satisfied by } P\}$, and denote by $m_0(P) = m_0 = |\mathcal{H}_0(P)|$ the number of true nulls. Denote $\mathcal{H}_1(P) = \mathcal{H}_1$ the complement set of $\mathcal{H}_0$ in $\{1, \dots, m\}$ containing the false nulls (also referred to as alternatives or signal). We assume that there exists a set of p-values defined as a set of r.v : $\{p_i(X), 1 \leq i \leq m\}$, with each $p_i = p_i(X) \in [0, 1]$ summarizing

the evidence against the corresponding null hypothesis $H_{0,i}$. The smaller the p-value p_i is, the more evidence we have to support the rejection of the null hypothesis $H_{0,i}$.

In p-value based testing, the decision to reject (or not) null hypotheses is made by comparing p-values with thresholds often called critical values. In single testing, the critical value corresponds to the prespecified testing level $\alpha \in (0,1)$. To perform a valid test based on a p-value $p = p(X) \in [0,1]$, the latter needs to be super-uniform under the null, meaning stochastically larger than the uniform distribution under the null, which can be formally stated as

$$\mathbf{P}_{X\sim P}(p(X) \leq u) \leq u, \text{ for all } u \in [0,1], \text{ when } P \text{ satisfies } H_0 \tag{SuperUnif}$$

Super-uniformity is a core defining property for the validity of the test: for a nominal level $\alpha \in (0,1)$ and a rejection decision taken by $\mathbf{1}_{p\leq\alpha}$, verifying (SuperUnif) allows to bound the probability of falsely rejecting the null hypothesis, i.e. the Type I error probability, by α.

In MT, using the same level α to test the m hypotheses would lead to a considerable number of false discoveries even though the individual Type I error probabilities of each test would be bounded by α. This fact is concealed until all the decisions are assessed jointly. As a matter of fact, assuming all hypotheses to be true nulls (complete null setting) with identically uniform p-values yields on average a number of false rejections equal to $\mathbf{E}\left[\sum_{j=1}^{m} \mathbf{1}_{p_j\leq\alpha}\right] = m\alpha$ which can be extremely large. This puts forward the need to correct the decision.

We define a MT procedure as a function $\mathcal{R} : [0,1]^m \to \mathcal{P}(\{1,\dots,m\})$ that, provided the m p-values, returns a subset $R \subset \{1,\dots,m\}$ containing indices for the rejected null hypotheses. Sometimes, a procedure can be identified to the set of critical values $\{\alpha_i\}_{1\leq i\leq m}$ it designs, corresponding to individual Type I error levels tailored for each hypothesis. The goal is to reject as many as false nulls possible while controlling the amount of false discoveries. This is similar to the Neyman-Pearson paradigm in single testing: MT procedures first focus on controlling the number of false discoveries over all the performed tests, using an error criterion accounting for a global Type I error. While maintaining this global Type I risk bounded by a pre-specified level $\alpha \in (0,1)$, the procedure aims at rejecting correctly as much as possible, i.e. to be as powerful as possible.

1.2.2 Family-Wise Error Rate (FWER) control

Perhaps the earliest criterion introduced, that we can trace back to the work of Tukey (see Benjamini and Braun (2002)) is the Family-Wise Error Rate (FWER). It is defined as the probability of making at least one false rejection among all the decisions:

$$\text{FWER}(R) = \mathbf{P}(|R \cap \mathcal{H}_0| \geq 1), \tag{1.1}$$

where we recall that R denotes the rejection set returned by the MT procedure. By controlling the FWER at level α we control the occurrence of any error with high probability: we are $(1-\alpha)\%$ confident that there is no false discovery within the rejection set R. Thus, FWER has a clear interpretation that supports its use. For instance, it is a criterion of interest in medical research where the Food and Drug Agency strictly regulates research findings before allowing a new drug to be marketed, see Dmitrienko et al. (2009) for more details.

Nevertheless, in other contexts, FWER could be too stringent since by preventing any false discovery with high probability it also limits at the same time the total number of discoveries. This fact can be clearly illustrated with the most popular FWER controlling procedure, the so-called Bonferroni procedure which tests each individual hypothesis at level α/m to return $R = \{i : p_i \leq \frac{\alpha}{m}\}$. In words, the Bonferroni correction divides the global level α by the number

of tested hypotheses m. Thus the testing levels shrink fast when m is large which prevents the procedure from being powerful. Other more powerful FWER controlling procedures were proposed by Šidák (1967), and Holm (1979) but the field took a new turn with the innovative work of Benjamini and Hochberg (1995).

1.2.3 False Discovery Rate (FDR) control

The work of Benjamini and Hochberg (1995) introduces a new way of accounting for the number of errors that scales with the number of tests performed. For any rejection set R, the False Discovery Proportion (FDP) is defined as the ratio between the number of false discoveries within R and the number of discoveries/rejections |R|:

$$\mathrm{FDP}(R) = \frac{|R \cap \mathcal{H}_0|}{1 \vee |R|}. \tag{1.2}$$

Since the rejection set R is random, so is FDP(R), ant it cannot be directly used as a criterion. Thus Benjamini and Hochberg (1995) consider controlling the expectation of the FDP – called the False Discovery Rate (FDR) – at level $\alpha \in (0, 1)$ and propose an MT procedure to do so assuming independent p-values. The BH procedure works by ordering the p-values in ascending order $p_{(1)}, \ldots, p_{(m)}$ and set $\hat{k} = \sup\{k \in \{1, \ldots, m\} : \widehat{\mathrm{FDP}}_k \leq \alpha\}$, where $\widehat{\mathrm{FDP}}_k = \frac{p_{(k)} m}{k}$. Then, the $\hat{k}$ first smallest p-values are rejected, i.e. $R_{BH} = \{i \in \{1, \ldots, m\} : p_{(i)} \leq p_{(\hat{k})}\}$ (and the procedure rejects nothing if the set is empty). An equivalent description of $\hat{k}$ can be given as follows: the ordered p-values $p_{(1)}, \ldots, p_{(m)}$ are compared with the corresponding critical values $\alpha_k = \frac{\alpha k}{m}, 1 \leq k \leq m$ until $p_{(k)} \leq \alpha_k$ happens for the last time, which corresponds to the index $\hat{k}$.

The work of Benjamini and Hochberg (1995) has gained significant recognition since its introduction, accumulating around 100 000 citations today on Google Scholar. It enjoys popularity not only within the field of statistics itself but also across various scientific domains such as biology, medicine, and economics, as shows Figure 1.1.

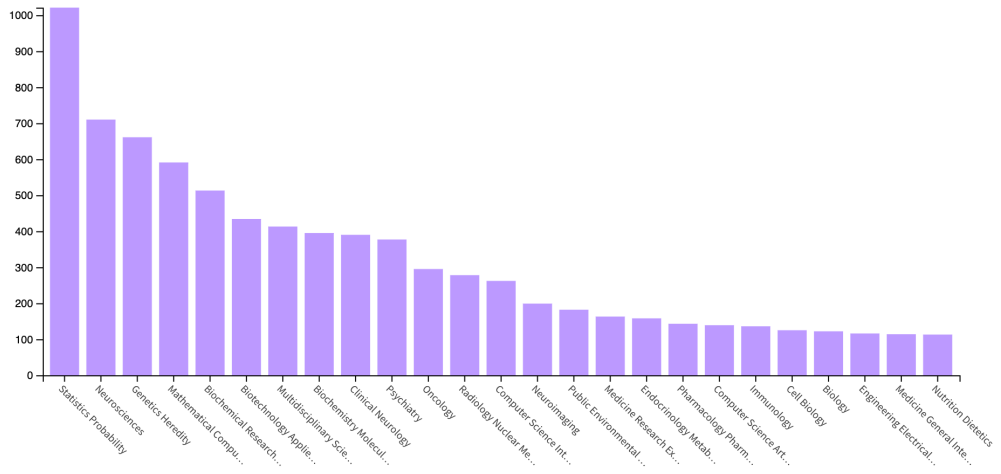

Figure 1.1: Number of papers per research field where the terms "False discovery rate control" appear from 1995 to 2023 according to the *Web of Science*.

Expanding on the work of Benjamini and Hochberg (1995), subsequent efforts have been made in the literature to refine the procedure. A first axis is on extending the dependency assumptions under which the BH procedure controls the FDR. Indeed, the seminal work analyses the procedure under the assumption of independent p-values which is far from being realistic in scientific applications. Then, Benjamini and Yekutieli (2001) proved that BH is still conservative when the p-value family verifies a specific type of positive dependence. Beyond that, for arbitrary dependence, they propose the Benjamini-Yekutieli (BY) procedure which corresponds to the BH procedure run for the shrunk level $\alpha/(1 + \frac{1}{2} + \cdots + \frac{1}{m})$. This procedure, although being free of the independence assumption is not popular as the number of rejections can be worse than the one of Bonferroni. Providing results accounting for the dependence structure is still interesting up to now. Recently, assuming a specific type of weak negative dependence between the p-values, Chi et al. (2022) propose a sharper upper bound than the one of BY.

Another avenue of improvement is dedicated to the power enhancement of the BH procedure to allow for more discoveries. A simple and effective way to do this is to adapt the procedure to the quantity of signal by incorporating an estimator of the proportion of nulls, see Section 1.4.3 for more details and references. Another way to enhance the power is by using weighted p-values which incorporate additional information about the practitioner's confidence in the null hypotheses, see Section 1.3.3 for more details.

Beyond its direct practical relevance, the work of Benjamini and Hochberg (1995) has proven to be valuable in deriving tools for non-parametric MT. Indeed, the FDP and the mathematical tools associated with it are helpful in formulating procedures that are applicable in contexts where p-values are unavailable, see e.g. the works of Sun and Tony Cai (2009) or Barber and Candès (2015). Furthermore, it is also helpful to build connections with the field of machine learning for reliable prediction, see the recent work of Marandon et al. (2022) and references therein.

Overall, the work of Benjamini and Hochberg (1995) introduced a novel criterion that is more liberal allowing the number of errors to grow with the number of rejections, see (1.2). The liberal aspect highly contributed to the popularity of the criterion as it resonates with the flexibility sought after by researchers in certain domains.

1.2.4 False Discovery Proportion (FDP) stochastic control

A stochastic control of the FDP means controlling the FDP of a rejection set below a certain level with high probability. Regarding this guarantee, two distinctions can be made in the literature. One line of research aims at building a rejection set R that has an FDP below a pre-specified level $\alpha \in (0, 1)$ with probability $1 - \delta$, where $\delta \in (0, 1)$ is a pre-specified coverage parameter. This task is referred to as "FDP/FDX control". Another strand of the literature focuses on deriving confidence bounds on the FDP valid for a given family of rejection sets R.

For FDP control, the goal is to control the False Discovery Exceedance (FDX) defined as

$$\mathrm{FDX}(R) = \mathbf{P}(\mathrm{FDP}(R) \leq \alpha) \geq 1 - \delta,$$

where R denotes the rejection set returned by the procedure and $\alpha, \delta \in (0, 1)$ are pre-specified. In words, FDX control involves controlling the $(1 - \delta)$-quantile of the FDP at a pre-specified level $\alpha \in (0, 1)$. To this end, novel procedures are built like the one proposed by Lehmann and Romano (2005) and Guo and Romano (2007) later improved by Döhler and Roquain (2020) for discrete settings (a setting described in more detail in Section 1.3.1). Since stochastic control

is more involved than control in expectation, procedures controlling the FDX have usually less power than the BH procedure but in return one can gain a broader insight into the FDP distribution. FDX control is relevant in practice when scientists want to profit from the liberal aspect of the FDP but at the same time have stronger guarantees than the FDR, see e.g. the work of Tan et al. (2019).

For FDP bounds, one wishes for tools to evaluate the quality of a specific rejection set family. Confidence bounds can be tailored for a given set Π of rejection sets. Formally, the goal is to provide a confidence envelope, i.e. a function $\overline{\mathrm{FDP}}$ valued in $(0,1)$ that takes a subset R and returns an upper bound of $\mathrm{FDP}(R)$ that verifies

$$\mathbf{P}(\text{for any } R \in \Pi, \mathrm{FDP}(R) \leq \overline{\mathrm{FDP}}(R)) \geq 1-\delta, \tag{1.3}$$

for a pre-specified $\delta \in (0,1)$. For instance, Π can be the path of Top-k rejection subsets i.e. $\Pi = \{R_k, 1 \leq k \leq m\}$ where $R_k = \{i : p_{(i)} \leq p_{(k)}\}$, or the set of all possible rejection sets, i.e. $\Pi = \mathcal{P}(\{1,\dots,m\})$.

These guarantees are qualified as *posthoc* since bounds holding uniformly over a family of rejection set allows for a valid analysis even by choosing R *after looking at the data*. For instance, scientists can initially analyze the Top-10 p-values, and based on those results, decide to expand the analysis to the Top-15 p-values if the findings are not conclusive. Alternatively, scientist may desire to analyze selection set, not necessarily designed using the p-values but using auxilliary statistics, which is allowed whenever (1.4.1) holds with $\Pi = \mathcal{P}(\{1,\dots,m\})$. Posthoc guarantees are very desirable in practice e.g. in biology where scientists can be confronted with an overwhelming choice of genes to test. In this context, exploring different selection sets is helpful to narrow down the research scope, allowing scientists to focus on a reduced set of genes for further investigation in confirmatory research. This viewpoint was popularized by the innovative work of Goeman and Solari (2011) who shed light on the importance of providing posthoc guarantees to better align with the way scientists work.

The literature on exploratory MT is well established with notables works of Meinshausen and Bühlmann (2005), Meinshausen (2006), Genovese and Wasserman (2006), Goeman and Solari (2011) or more recently Blanchard et al. (2020), among others.

1.3 P-value structures

In this section, we present different p-value structures that are of interest throughout our work. These structures arise from different underlying data-generating contexts and require design of appropriate procedures to enhance power or to avoid power loss. To better understand the specificities or the challenges related to each structure of interest we first define a "canonical structure", roughly corresponding to the baseline setting described in Section 1.2.1.

Canonical structure Our canonical structure refers to the classical setting of Benjamini and Hochberg (1995) where p-values are available all together as a batch and have uniform marginal distributions under the null. The batch availability is advantageous because all the relevant information is accessible simultaneously. Also, marginally uniformly distributed p-values is the best case scenario to avoid over-conservativeness as we further explain in the following section.

1.3.1 Super-uniform p-values

As mentioned in Section 1.2.1, super-uniformity is a core property for the validity of statistical tests. When the p-value is not uniform, i.e. when the inequality in (SuperUnif) is strict, the

p-value is more conservative than needed. Indeed, when (SuperUnif) is verified, we know that the probability of making a Type I error for the corresponding hypothesis is actually less or equal to α_i:

$$\underbrace{\mathbf{P}_{X\sim P}(p_i(X) \leq \alpha_i)}_{=\text{ Type I error probability}} = \tilde{\alpha}_i \leq \alpha_i, \text{ for all } \alpha_i \in (0,1), \text{ and } P \in \mathcal{P} \text{ with } i \in \mathcal{H}_0.$$

When the effective level $\tilde{\alpha}_i$ is strictly smaller than the nominal level α_i, it means that the testing level α_i is not exhausted. Consequently, the individual power is in some sense capped as we are in reality testing at a smaller level than what is allowed by the procedure. One solution is to enlarge the nominal level to effectively attain the Type I error probability, which requires information on the distribution of the p-values under the null. Thus, in this setting we assume that the underlying model $\mathcal{P}$ contains known functions F_i, $1 \leq i \leq m$ such that:

$$\mathbf{P}_{X\sim P}\left(p_i \leq u\right) \leq F_i(u) \leq u, \text{ for all } u \in [0,1], \text{ and } P \in \mathcal{P} \text{ with } i \in \mathcal{H}_0. \tag{1.4}$$

While deriving such F_i can be difficult in classical contexts (e.g. when testing composite nulls for Gaussian means where H_0: $\mu \leq 0$, with μ denoting the mean), explicit null bounds F_i are available in other specific contexts that we further describe in what follows.

Discrete p-values Discrete test statistics appear whenever dealing with categorial variables describing counts. One instance of such tests is FET described in Section 1.1.2, producing a discrete p-value with a finite support on $[0,1]$. Other instances of discrete tests include the Poisson test for parametric settings or Mann-Whitney and Wilcoxon tests for non-parametric settings, see e.g. Hirji (2005) or Rousson (2013) for detailed examples.
In this context, each p-value p_i has a finite support $\mathcal{S}_i$ that is known and independent of P and satisfies (1.4) with F_i taken as the right continuous step function that jumps at each point of $\mathcal{S}_i$ and $F_i(u) = u$ only when $u \in \mathcal{S}_i$, as illustrated in Figure 1.2. The knowledge of these upper bounds F_i can be used to compensate the power loss caused by the super-uniformity, see e.g. Döhler et al. (2018) or Döhler and Roquain (2020) for some recent works in this setting.

Self-imposed super-uniformity with constrained weighting We describe the classical setting of p-value weighting in Section 1.3.3 and present here a version where the weights are constrained to be less than 1. Assume that the base p-values $p_i, 1 \leq i \leq m$ are uniformly distributed, and denote $r_i > 0, 1 \leq i \leq m$ the raw, i.e. non processed, weights. Denote $w_i, 1 \leq i \leq m$ the corresponding processed weights such that $w_i \in (0,1), 1 \leq i \leq m$. The preprocessing constrains the weights to be less than 1 and consequently enforces super-uniformity of the p-values. Indeed, each weighted p-value is defined as $\tilde{p}_i = \frac{p_i}{w_i}$ so that the c.d.f under the null is

$$\mathbf{P}_{X\sim P}\left(\tilde{p}_i \leq u\right) = \mathbf{P}_{X\sim P}\left(p_i \leq uw_i\right) = uw_i =: F_i(u) \leq u, \text{ for all } u \in [0,1], 1 \leq i \leq m, \tag{1.5}$$

where the last inequality stands because $w_i \in (0,1)$. See Figure 1.3 for an illustration of the weighted c.d.f under the null.
In classical p-value weighting, the weights w_i are only constrained to sum to m, so that weighted p-values can be smaller or larger than their original p-values. More precisely, when well chosen, large weights are associated with confident alternative hypotheses so that a power enhancement can be expected from the weighting scheme. By contrast, here the weighted p-values are always larger (i.e. less significant) than the original p-values. This scheme seems counter-intuitive

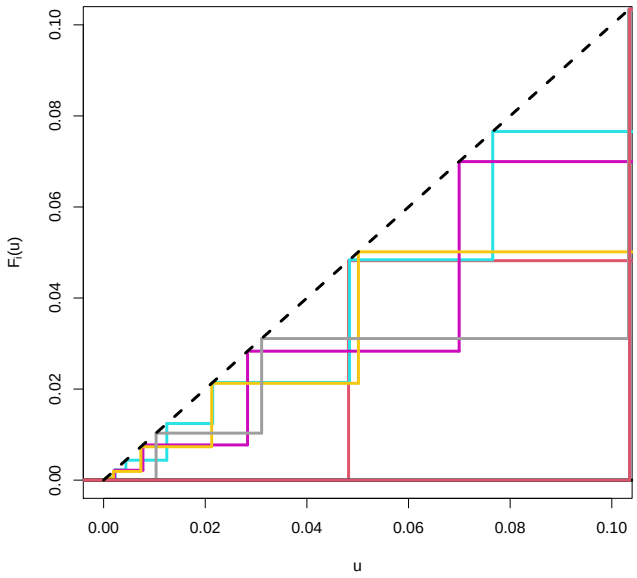

Figure 1.2: Examples of F_i functions in the discrete setting (colored solid lines) and standard Uniform c.d.f (black dashed line).

but this enforced conservativeness can help to better manage the allocation of the overall error level α across all hypotheses in some contexts. For instance, this proves to be beneficial in situations where the classical p-value weighting strategy is unfeasible, as in the online setting (see Chapter 2).

1.3.2 Online p-values

This structure pertains to the temporal availability of the p-values. In the online setting, null hypotheses are formulated sequentially over time for an unknown duration. Thus, p-values are only available one by one, with an overall number of tests m unknown beforehand and potentially infinite. This scenario models modern applications such as perpetual clinical trials where an ever-growing number of new treatments are continuously emerging and tested against a baseline control treatment as described by James et al. (2008). Similarly, tech companies also perform A/B testing dealing with thousands of hypotheses formulated one by one over time to look e.g. for the best layout of a website, see e.g. Berman et al. (2018) for more details. To perform MT in an online setting, one approach could be to run the BH procedure over and over again as each new p-value is made available. In that case, for a prespecified level $\alpha \in (0, 1)$, one would first run a single test for the first p-value p_1, then once p_2 is available, rerun BH at level α for p_1 and p_2, then once p_3 is available, rerun BH at level α for p_1, p_2 and p_3, and so on. However, performing this naive online procedure could break the FDR control and lead to conflicting decisions over time: at one time the BH procedure can reject a null hypothesis while being unable to reject it at subsequent times. The unstable nature of the decisions can be very costly – as some rejections might involve follow-up pricy investigations – or greatly perturb the overall understanding as some non-rejections can lead to state novel hypotheses. This concern highlights the need to

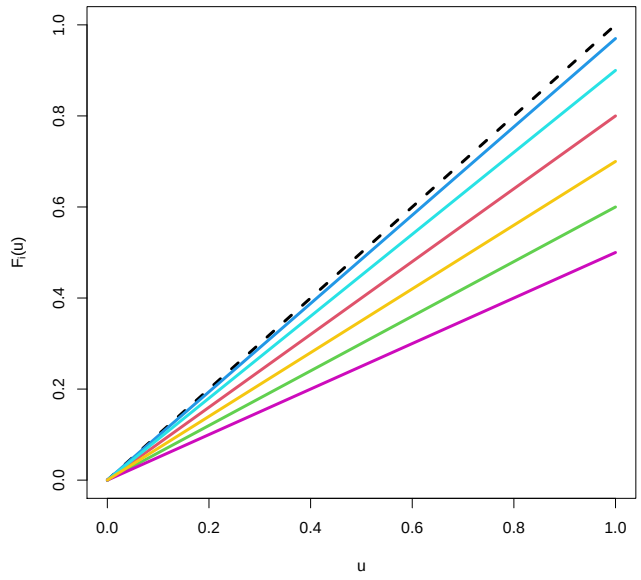

Figure 1.3: Examples of F_i functions in the constrained weighting setting (colored solid lines) and standard Uniform c.d.f (black dashed line).

make irrevocable decisions using procedures specifically designed for the online setting. Several recent works propose such procedures, see Robertson et al. (2022) for an exhaustive review on the topic listing several online error rates and procedures to control them.

1.3.3 Covariates

In practice, a natural aspiration is to integrate valuable insights from the underlying domain, or past experiments into the current decision. In this setting, the challenge is to efficiently incorporate this knowledge without corrupting the control of false discoveries. This knowledge can be incorporated in different ways, and we describe below the cases where it is incorporated in the ordering of the p-values and in their magnitude.

Preordered p-values In this setting, an ordering of the null hypotheses is pre-specified, thus providing a batch of preordered p-values. The preorder can be caused by the underlying mathematical model, as is the case when using the knockoff approach of Barber and Candès (2015) for variable selection in the Gaussian linear model. Alternatively, the preorder can also come from the domain knowledge. For instance, in molecular biology, biologists often have some prior knowledge or intuition about which genes are more likely to be associated with the biological trait of interest.

The preordering aims at putting hypotheses that are the most favorable to be rejected first. However, this information might not be translated in the magnitude of the p-values so running a BH procedure – which involves ordering the p-values in ascending order – could ruin the intended preordering. Specific procedures for the preordered setting have been proposed by Li and Barber

(2017), and more recently by Lei and Fithian (2016) with a procedure more robust to the quality of the ordering.

Weighted p-values Incorporating weights $w_i > 0, 1 \leq i \leq m$, coding the practitioner's confidence in the rejection can help to reject false nulls more efficiently. The testing decision is taken on the weighted p-value $\frac{p_i}{w_i}$ instead of the (raw) p-value p_i. For confident nulls we expect small weights to enlarge the p-values making them more conservative to avoid rejecting them. Conversely, for confident alternatives we expect large weights to shrink the p-values making them more inclined to rejection. In this context, two questions arise at first sight: *(i) How to incorporate the weights without compromising the overall Type I error control? (ii) How to obtain these weights ?*

To address (i), Holm (1979) introduces a weighted Bonferroni procedure that controls the FWER. Later, Benjamini and Hochberg (1997) provides a different approach by integrating weighting into the error criterion for FDR control instead of integrating it into the p-values. The current line of research on weighted MT follows the work of Genovese et al. (2006) who investigates p-value weighting for FDR control and provides a sufficient condition on the weights, which requires the sum of weights to be less than m for maintaining FDR control.

To address concern (ii) there are two main strands. First, weights can come from the underlying domain knowledge. Indeed, useful external knowledge is often available in molecular biology, as mentioned previously, or in neuroscience. Alternatively, another strand of the literature on p-value weighting focuses on deriving weights using the available data. To give a brief review we can cite the work of Roeder and Wasserman (2009) who derive optimal weights in the Gaussian t-testing setting or the work of Roquain and Van de Wiel (2008) who derive oracle optimal weights assuming that the alternative c.d.f of p-values are known. More recently, Ignatiadis et al. (2016) and Durand (2019) propose to derive weights in the context of grouped hypothesis testing.

In a more comprehensive overview of these settings, a differentiation can be drawn based on how the structure under consideration relates to multiplicity. Indeed, some structures are specific to multiplicity whereas others are not. For instance, the discrete structure already appears at the single testing level while the online or the preordered structures only exist in the context of MT.

All the structures can influence the quality of the decisions taken in terms of power. In some cases (e.g., discrete, online), the structure imposes additional constraints on the procedure and the aim is to try to recover the unconstraint power as much as possible. In other situations, like when dealing with covariates, the structure is beneficial and provides information on the underlying true distribution P. In these cases, the aim is to improve the power of classical procedures by incorporating the underlying structure.

In addition, the power of a procedure can also be influenced by a quantitative aspect of the p-value family. Indeed, the prevalence of true nulls limits the power, i.e. it makes detection of alternatives very difficult. By constrast, when the proportion of true nulls $\pi_0 = \frac{m_0}{m}$ is small, the power can be enhanced. These two scenarios correspond respectively to a sparse and dense settings and are also investigated in our work.

1.4 Contributions

In this work we explore combinations of the aforementioned settings with three different MT goals: control the online mFDR (Chapter 2), provide valid FDP envelopes (Chapter 3), and control the plug-in FDR (Chapter 4). Broadly speaking, in each scenario the aim is to use the

structural knowledge of the p-values to improve state-of-the-art procedures or tools for the MT goal of interest.

P-value structure	Chapter 2: Online mFDR	Chapter 3: FDP envelope	Chapter 4: plug-in FDR
Canonical		✓	✓
Discrete	✓		✓
Online	✓	✓	
Preordered		✓	
Weighted	✓		

Table 1.2: Visual summary of the structures studied in each chapter of the manuscript.

1.4.1 Super-uniformity reward

In Chapter 2 we address the two following questions: *(i) How to improve the efficiency of current Online Multiple Testing (OMT) procedures when dealing with discrete p-values ? (ii) How to incorporate external knowledge in OMT procedures using p-value weighting?*

Roughly speaking, an online procedure works as a budgeting scheme that sequentially allocates an amount $\alpha_t \in (0,1)$ of the desired global level $\alpha \in (0,1)$, to test any new hypothesis that comes up at time $t \geq 1$. As with classical MT procedures, the allocation scheme is designed with the aim of rejecting most of the false nulls while controlling an online error criterion that monitors the number of false discoveries at any time point. In our work, we focus on the online FWER and mFDR formally defined as follows:

$$\mathrm{FWER}(\mathcal{A}, P) := \sup_{T \geq 1} \{\mathrm{FWER}(T, \mathcal{A}, P)\},\ \mathrm{FWER}(T, \mathcal{A}, P) := \mathbf{P}_{X \sim P}\Big(|\mathcal{H}_0 \cap \mathcal{R}(T)| \geq 1\Big);$$

$$\mathrm{mFDR}(\mathcal{A}, P) := \sup_{T \geq 1} \{\mathrm{mFDR}(t, \mathcal{A}, P)\},\ \mathrm{mFDR}(T, \mathcal{A}, P) := \frac{\mathbf{E}_{X \sim P}\left(|\mathcal{H}_0 \cap \mathcal{R}(T)|\right)}{\mathbf{E}_{X \sim P}\left(1 \vee |\mathcal{R}(T)|\right)},$$

where $\mathcal{A} = \{\alpha_t, t \geq 1\}$ denotes the procedure (identified to the sequence of critical values), and $\mathcal{R}(T) = \{t \in \{1, \ldots, T\} \,:\, p_t(X) \leq \alpha_t\}$ denotes the set of rejection times of the procedure $\mathcal{A}$, up to time T.

In recent years, the literature on OMT has grown substantially to introduce performant procedures, see Robertson et al. (2022) for further details on the methodology and an extensive review of the existing literature. The existing procedures are most efficient when dealing with uniform p-values under the null, however some (significant) parts of the individual budgets $\{\alpha_t\}_{t \geq 1}$ can be wasted when dealing with discrete p-values due to super-uniformity as described in Section 1.3.1. While discreteness can appear in plenty of scenarios (see Section 1.3.1), no solution has been proposed up to now to deal with the super-uniformity issue in the OMT setting.

To deal with discrete p-values, classical approaches in the offline literature attempt to address the conservativeness of discrete p-values with randomization techniques, see e.g. Habiger (2015) (see also Section 4.5.1 of Chapter 4 for more details and references). While these methods can be theoretically effective, they also introduce reproducibility and interpretability issues which is a major drawback in practice. Recent approaches, e.g. proposed by Döhler et al. (2018), introduce methods that use known null bounds on the p-values c.d.f to correct the loss of power due to super-uniformity. However, these offline solutions are not readily applicable to the online setting. As a result, the need for an online procedure tailored for discrete p-values is still to be addressed.

To address this need, we introduce "rewarded" versions of some classical online procedures that compensate for the power loss caused by the super-uniformity of discrete p-values. These rewarded procedures have two components corresponding to the initial procedure combined with an additional quantity called the super-uniformity reward. The super-uniformity reward represents the gap between the nominal and the effective level that we are able to calculate using the distributional knowledge on the p-value c.d.f. More formally, it is defined as

$$\rho_t = \rho_t(\alpha_t, F_t) := \alpha_t - F_t(\alpha_t), \quad t \geq 1. \tag{1.6}$$

where $\alpha_t \in (0,1)$ is the nominal testing level at time t, and F_t is the known c.d.f of p-value p_t under the null. Since (1.4) holds under super-uniformity, we know that the effective level – i.e. the level attained in reality – is equal to $F_t(\alpha_t)$ rather than the intended α_t. Thus, ρ_t can be interpreted as the unused amount of testing level at time t. This quantity is added in subsequent testing times as a reward to help uplift the next critical values to improve the power. This first part of the work addresses question (i).

To address question (ii), we show that our rewarding method allows to seamlessly handle p-value weighting in the online framework. In the offline setting, "raw" weights $r_i \geq 0$ need to be rescaled to average to 1 to maintain FDR control. However, this sufficient condition cannot be satisfied in the online context because the weights are only available one by one across time. Ramdas et al. (2017) present sufficient criteria for weighting procedures controlling the (m)FDR and also discuss the technical challenges associated with weighted online multiple testing. We propose another approach that relies on the super-uniformity reward. Our method works by constraining the weights to be less than 1, as described in Section 1.3.1, which imposes super-uniformity so that our method described to address point (i) can be applied.

1.4.2 Consistent FDP bounds

Chapter 3 focuses on the question: *How to provide sharp FDP confidence bounds that are tailored for rejections sets of FDR controlling procedures?*

As mentioned in Section 1.2.3, probabilistic guarantees on the FDP are appealing for several reasons: they are stronger than guarantees on the expectation of the FDP, they provide distributional information on the FDP, and they can give the practitioner freedom to choose the selection set to analyze. In this context, Katsevich and Ramdas (2020) propose new confidence envelopes, i.e. a sequence of confidence bounds $(\overline{\mathrm{FDP}}_k, k \geq 1)$ tailored for rejection sets following a path $\Pi = (R_k, k \geq 1)$ in three different p-value settings taking into account the canonical (Section 1.3.1), online (Section 1.3.2), and pre-ordered (Section 1.3.3) structures. These bounds are valid uniformly over the path $\Pi = (R_k, k \geq 1)$, i.e. they verify

$$\mathbf{P}(\forall k \geq 1, \mathrm{FDP}(R_k) \leq \overline{\mathrm{FDP}}_k) \geq 1 - \delta,$$

where $\mathrm{FDP}(R_k) = \frac{|R_k \cap \mathcal{H}_0|}{|R_k| \vee 1}$ is the FDP of the set R_k, and $\delta \in (0,1)$ is a pre-specified coverage parameter.

In each of the three settings, the path Π is a family of rejection sets built to contain an FDR controlling procedure's output. For instance, in the canonical setting, the rejection sets are built following the Top-k path with $R_k = \{i : p_{(i)} \leq p_{(k)}\}$, which includes BH rejection set when $k = \hat{k}$ (defined in Section 1.2.3). Katsevich and Ramdas (2020) show that their new bounds improve classical bounds so-called Simes derived by Robbins (1954) or DKW derived by Massart (1990) – in various regimes. The work of Katsevich and Ramdas (2020) bridges the gap between FDR and FDP by proposing FDP bounds for FDR controlling procedures. However, it appears that their bounds are tailored for the context where the number of rejections is small. As the number

of rejections grows, their FDP bounds can deviate significantly from the actual FDP value for certain rejection sets. As a case in point, in the canonical setting, considering the BH rejection set denoted as $R_{\hat{k}}$, existing bounds yield $\overline{\mathrm{FDP}}_{\hat{k}} = \alpha.c$ where $c > 1$, see Section 3.2.1 for more details. Since c is an incompressible constant even when m tends to infinity, the bound can be far above α which is suboptimal. Indeed, concentration arguments imply that the FDP should closely align with its expected value, namely the FDR (equal to $\alpha\pi_0$ for BH). Thus, these bounds can be described as not "consistent".

In Chapter 3, we formalize the aforementioned consistency notion and we develop new confidence envelopes that are consistent, for each of the three p-value settings of interest mentioned above. We further analyze the consistency of our novel and previous bounds in sparse settings where the amount of signal is weak (π_0 close to 1).

1.4.3 Unifying class of null proportion estimators

In Chapter 4, we focus on the question: *How to adjust existing π_0 estimators to discrete p-values while maintaining plug-in FDR control?*

Plug-in FDR control refers to the FDR control for the adaptive BH procedure where an estimator of π_0, the proportion of null hypotheses, is used to adjust the critical values to allow for more rejections. More specifically, when plugging in the estimator, the BH thresholds are $\alpha_k = \frac{\alpha k}{\hat{\pi}_0} \geq \frac{\alpha k}{m}, : 1 \leq i \leq m$, where the r.h.s corresponds to the base thresholds of the BH procedure (see Section 1.2.3), $\alpha \in (0, 1)$ is the desired control level for the FDR, and $\hat{\pi}_0$ is the estimated proportion of null hypotheses. The additional source of randomness present in the estimation needs to be accounted to avoid breaking the FDR control. Thus, estimators need to verify some conditions to be valid for plug-in FDR control. In our work, we focus on the sufficient criterion introduced in the works of Benjamini et al. (2006) and Blanchard and Roquain (2009), which focus on bounding the inverse moment of the estimator.

In the past, adaptivity to the number of true null hypotheses has been extensively studied, with various works proposing estimation methods and investigating corresponding plug-in FDR control. To give a rough timeline we can mention the work of Storey et al. (2004), who were the first to propose an estimation method for π_0 coupled with FDR plug-in control thus marking a notable milestone in the area. Then, Pounds and Cheng (2006) introduced another type of estimation method that does not require any parameter tuning and that has a version tailored for the discrete setting. However, their estimator lacks theoretical guarantees regarding plug-in FDR control making it less appealing for practitioners.

The current line of work has two main limitations. First, some of the estimators lack theoretical guarantees for plug-in FDR control. Second, most of the proposed estimators have been derived assuming the canonical setting described in Section 1.3, making them less efficient when applied to other settings, like the one of discrete p-value described in Section 1.3.1. Along with the work of Pounds and Cheng (2006), the works of Chen et al. (2018) and Biswas and Chattopadhyay (2020) address the discreteness issue for estimation purposes but they either lack plug-in FDR control or are not better than other classical estimators e.g. like Storey et al. (2004).

In Chapter 4, we aim to bridge these gaps by introducing a new class of estimators that not only encompasses previously proposed estimators but also allows for defining new estimators. This class of estimators comes with mathematical guarantees for plug-in FDR control and offers the flexibility to incorporate adjustments to the p-value distribution without compromising the plug-in control. These adjustments can sometimes significantly improve the performance of the estimators in the discrete setting without requiring sophisticated parameter tuning. The theoretical guarantees are established based on convex ordering arguments, which in essence,

provide moment ordering of random variables.

Outline of the manuscript Each chapter is independent and self-contained therefore the notations may vary from chapter to chapter. For each chapter, the proofs are given in the associated appendices. We give the status quo of the chapters below.

- Chapter 2 is a joint work with Sebastian Döhler (Hochschule Darmstadt) and Etienne Roquain (Sorbonne Université). In revision for the Electronic Journal of Statistics.
- Chapter 3 is a joint work with Gilles Blanchard (Université Paris Saclay) and Etienne Roquain. It has been submitted for publication.
- Chapter 4 is a joint work with Sebastian Döhler. It has been submitted for publication.

Chapter 2

Online multiple testing with super-uniformity reward

Outline of the current chapter

Valid online inference is an important problem in contemporary multiple testing research, to which various solutions have been proposed recently. It is well-known that these existing methods can suffer from a significant loss of power if the null p-values are conservative. In this work, we extend the previously introduced methodology to obtain more powerful procedures for the case of super-uniformly distributed p-values. These types of p-values arise in important settings, e.g. when discrete hypothesis tests are performed or when the p-values are weighted. To this end, we introduce the method of super-uniformity reward (SUR) that incorporates information about the individual null cumulative distribution functions. Our approach yields several new 'rewarded' procedures that offer uniform power improvements over known procedures and come with mathematical guarantees for controlling online error criteria based either on the family-wise error rate (FWER) or the marginal false discovery rate (mFDR). We illustrate the benefit of super-uniform rewarding in real-data analyses and simulation studies. While discrete tests serve as our leading example, we also show how our method can be applied to weighted p-values.

javanmard2018online

2.1 Introduction

2.1.1 Background

Multiple testing is a well-established statistical paradigm for the analysis of complex and large-scale data sets, in which each hypothesis typically corresponds to a scientific question. In the classical situation, the set of hypotheses should be pre-specified before running the statistical inference. However, in contrast to the former 'offline' setting, in many contemporary applications questions arise sequentially. A first instance of such sequential application is when testing a *single* null hypothesis repeatedly as new data are collected, as for continuous monitoring of A/B tests in the information technology industry or marketing research, see Kohavi et al. (2013); Johari et al. (2019) and references therein, or Howard et al. (2021) for recent developments. A second situation is when the null hypotheses are (potentially) different and arise in a continuous stream, and accordingly decisions have to be made one at a time and prior to the termination of the stream. This is generally referred to as the *online multiple testing* (OMT) framework and is the focus of this paper, see, e.g., Lark (2017); Robertson et al. (2019); Kohavi et al. (2020) for application examples. This second situation also occurs in combination with the first one to form a 'doubly-sequential' experiment (Ramdas, 2019).

2.1.2 Existing literature on online multiple testing

The literature aiming at control of various error rates in OMT has grown rapidly in the last few years. As a starting point, the family-wise error rate (FWER) is the probability of making at least one error in the past discoveries, and a typical aim is to control it at each time of the stream (for a formal definition of this and other error rates, see Section 2.2.2). Since controlling FWER at a given level α is a strong constraint, it requires employing a procedure that is conservative, thus generally leading to few discoveries. The typical strategy is to distribute over time the

initial *wealth* α, e.g., testing the i-th test at level $\alpha\gamma_i$ for a sequence $\{\gamma_i\}_{i\geq 1}$ summing to 1. This approach is generally referred to as α-*spending* in the literature (Foster and Stine, 2008).

A less stringent criterion is the false discovery rate (FDR), which corresponds to the expected proportion of false discoveries. This versatile criterion allows many more discoveries than the FWER and has known a huge success in offline multiple testing literature since its introduction by Benjamini and Hochberg (1995), both from a theoretical and practical point of view. In their seminal work on OMT, Foster and Stine (2008) extended the FDR in an online setting by considering the expected proportion of errors among the past discoveries (actually, considering rather the marginal FDR, denoted below by mFDR, which is defined as the ratio of the expectations, rather than the expectation of the ratio). The novel strategy in Foster and Stine (2008), which is called α-*investing*, is based on the idea that an mFDR controlling procedure is allowed to recover some α-wealth after each rejection, which slows down the natural decrease of the individual test levels. In subsequent papers, many further improvements of this method have been proposed: first, the α-investing rule has been generalized by Aharoni and Rosset (2014), while maintaining marginal FDR control. Later, Javanmard and Montanari (2018) establish the (non-marginal) FDR control of these rules, including the LORD (Levels based On Recent Discovery) procedure. Then, a uniform improvement of LORD, called LORD++, has been proposed by Ramdas et al. (2017), that maintains FDR/mFDR control while extending the theory in several directions (weighting, penalties, decaying memory).

Extensions to other specific frameworks have been proposed, including rules that allow asynchronous online testing (Zrnic et al., 2021), maintain privacy (Zhang et al., 2020), and accommodate a high-dimensional regression model (Johnson et al., 2020). Other online error criteria have also been explored, with false discovery exceedance (Javanmard and Montanari, 2018; Xu and Ramdas, 2021), post hoc false discovery proportion bounds (Katsevich and Ramdas, 2020), or confidence intervals with false coverage rate control (Weinstein and Ramdas, 2020).

Since the online framework is more constrained than the offline framework, the employed procedures are generally less powerful in that context. Hence, another important branch of the literature aims at proposing improved rules that gain more discoveries: first, following the classical 'adaptive' offline strategy, procedures can be made less conservative by implicitly estimating the amount of true null hypotheses, see the SAFFRON procedure for FDR and the adaptive-spending procedure for FWER. Second, under an assumption on the null distribution, increasing the number of discoveries is possible by 'discarding' tests with a too large p-value (Ramdas et al., 2018; Tian and Ramdas, 2021, 2019).

A power enhancement can also be obtained by combining online procedures with other methods. A natural idea is to use more sophisticated individual tests in the first place, e.g., based on multi-armed bandits (Yang et al., 2017a), or so-called 'always valid p-values', see Johari et al. (2019) and references therein. Another idea is to combine offline procedures to form 'mini-batch' rules, see Zrnic et al. (2020). Further improvements are also possible by incorporating contextual information as done by Chen and Kasiviswanathan (2020) or using local FDR-like approach, see Gang et al. (2020). Lastly, performance boundaries have been derived by Chen and Arias-Castro (2021).

2.1.3 Super-uniformity

This paper consider OMT in the setting of super-uniformly distributed p-values (defined in detail in Section 4.2). Super-uniformity may originate from various sources. The first main example we have in mind, and which has been extensively investigated in the statistical literature, is super-uniformity arising from discrete p-values (described in detail in Section 2.5). Additionally, we show that super-uniformity can also be used in a more indirect way as a device for dealing

with online p-value weighting. In the offline setting, this is a powerful and extensively studied approach, which has, however, in the online case, received little attention so far (described in detail in Section 2.6).

Discrete tests often originate when the tests are based on counts or contingency tables, for example:

- in clinical studies, the efficiency or safety of drugs are compared by counting patients who survive a certain period after being treated, or who experience a certain type of adverse drug reaction;
- in biology, the genotype effect on the phenotype can be tested by knocking out genes sequentially in time.

The latter case is met for instance with the data from the International Mouse Phenotyping Consortium (IMPC, see Muñoz-Fuentes et al. (2018)), which contains many categorical variables, and thus are described with counts and contingency tables. While this data set is frequently used (see e.g., Tian and Ramdas, 2021; Xu and Ramdas, 2021; Karp et al., 2017), the classical OMT procedures do not exploit the discrete nature of the tests, and it turns out that much more powerful procedures can be developed, see Section 2.5.3.

In the literature, different solutions have been proposed for dealing with the conservatism of discrete tests, the most straightforward one being randomization (see Habiger (2015) and references therein). While this approach possesses attractive theoretical properties, randomization is usually unacceptable in practice (Lehmann and Romano, 2022). An active research area explores this phenomenon in the offline multiple testing setting, with the seminal works of Tarone (1990); Westfall and Wolfinger (1997); Gilbert (2005) and the subsequent studies of Heyse (2011); Heller and Gur (2011); Dickhaus et al. (2012); Habiger (2015); Chen et al. (2015); Döhler (2016); Chen et al. (2018); Döhler et al. (2018); Durand et al. (2019), see also references therein. The present work shows that such an improvement is also possible in the online setting, as far as FWER or mFDR control is concerned.

Error rate	Procedure	Critical values	Results
FWER	OB	$\alpha_T^{\mathrm{OB}} = \alpha\gamma_T$	Tian and Ramdas (2021)
	AOB	$\alpha_T^{\mathrm{AOB}} = \alpha(1-\lambda)\gamma_{\mathcal{T}(T)}$	Tian and Ramdas (2021)
mFDR	LORD	$\alpha_T^{\mathrm{LORD}} = W_0\gamma_T + (\alpha - W_0)\gamma_{T-\tau_1} + \alpha\sum_{j\geq 2}\gamma_{T-\tau_j}$	Javanmard and Montanari (2018) and Ramdas et al. (2017)
	ALORD	$\alpha_T^{\mathrm{ALORD}} = (1-\lambda)\cdot\Big(W_0\gamma_{\mathcal{T}_0(T)} + (\alpha - W_0)\gamma_{\mathcal{T}_1(T)} + \alpha\sum_{j\geq 2}\gamma_{\mathcal{T}_j(T)}\Big)$	Ramdas et al. (2018) (slightly improved)

Table 2.1: Overview of the critical values of the base procedures for some choice of level $\alpha \in (0,1)$, adaptivity parameter $\lambda \in [0,1)$, initial wealth $W_0 \in (0,\alpha)$, and spending sequence $(\gamma_j)_{j\geq 1}$. The quantities $\mathcal{T}(\cdot)$, τ_j, $\mathcal{T}_j(\cdot)$ are given by (2.16), (B.14), (2.26), respectively.

Finally, weighting p-values is a well-established and popular approach for improving the performance of offline multiple testing procedures. It can be traced back to Holm (1979) and has been further developed, in, e.g., Genovese et al. (2006); Wasserman and Roeder (2006);

Rubin et al. (2006); Blanchard and Roquain (2008); Roeder and Wasserman (2009); Hu et al. (2010); Zhao and Zhang (2014); Ignatiadis et al. (2016); Durand (2019); Ramdas et al. (2019) with weights that can be driven for instance by sample size, groups, or more generally by some covariates. By approaching the problem from the perspective of super-uniformity, our general method also allows seamless and flexible integration of such weighting schemes in an online context.

Error rate	Procedure	Critical values	Results
FWER	ρOB	$\alpha_T^{\rho\mathrm{OB}} = \alpha_T^{\mathrm{OB}} + \sum_{t=1}^{T-1} \gamma'_{T-t}\rho_t$	Theorem 2.3.1
	ρAOB	$\alpha_T^{\rho\mathrm{AOB}} = \alpha_T^{\mathrm{AOB}} + \sum_{\substack{1\le t\le T-1 \\ p_t>\lambda}} \gamma'_{T-t}\rho_t + \varepsilon_{T-1}$	Theorem 2.3.2
mFDR	ρLORD	$\alpha_T^{\rho\mathrm{LORD}} = \alpha_T^{\mathrm{LORD}} + \sum_{t=1}^{T-1} \gamma'_{T-t}\rho_t$	Theorem 2.4.1
	ρALORD	$\alpha_T^{\rho\mathrm{ALORD}} = \alpha_T^{\mathrm{ALORD}} + \sum_{\substack{1\le t\le T-1 \\ p_t>\lambda}} \gamma'_{T-t}\rho_t + \varepsilon_{T-1}$	Theorem 2.4.2

Table 2.2: Overview of the critical values of the rewarded procedures denoted as the corresponding base procedures, with an additional symbol "ρ" in the name. Here, $\alpha_T^{\mathrm{OB}}, \alpha_T^{\mathrm{AOB}}, \alpha_T^{\mathrm{LORD}}, \alpha_T^{\mathrm{ALORD}}$ are the base procedures from Table 2.1 (with the adaptivity parameter λ defined there), ρ_t is the super-uniformity reward at time t given by (2.8), γ' is the SURE spending sequence defined in Section 2.2.4 and $\varepsilon_T = \mathbf{1}\{p_T < \lambda\}(\alpha_T - \alpha_T^0)$ is an additional adaptivity reward, for either $(\alpha_T^0, \alpha_T) = (\alpha_T^{\mathrm{AOB}}, \alpha_T^{\rho\mathrm{AOB}})$, or $(\alpha_T^0, \alpha_T) = (\alpha_T^{\mathrm{ALORD}}, \alpha_T^{\rho\mathrm{ALORD}})$, depending on the case.

2.1.4 Contributions of the paper

In this paper, we propose uniform improvements of the classical base procedures listed in Table 2.1, and prove control of the corresponding error rates. A distinguishing feature of our work is that we assume that a (non-trivial) upper bound for the null cumulative distribution function's (c.d.f.), called the *null bounding family*, is *known* (see Section 4.2). By combining this information with base procedures, we construct more efficient OMT procedures (see Table 2.2). The key quantity involved in this construction can be interpreted as a reward (more details will be provided in Section 2.2.3) induced by the super-uniformity of the null bounding family. Therefore, we use the acronym SUR (Super-Uniform-Reward) to refer to these new procedures. When we use the uniform null bounding family (i.e., in the classical framework), our SUR procedures reduce to their base counterparts. Our main contributions are as follows:

- We propose two new SUR procedures for online FWER control in Section 2.3: the first one (ρOB) uniformly improves upon the Online Bonferroni procedure (OB), while the second (ρAOB) uniformly improves upon the adaptive spending procedure of Tian and Ramdas (2021) (AOB).
- We propose two new SUR procedures for online mFDR control in Section 2.4: the first one (ρLORD) uniformly improves upon the LORD++ procedures of Javanmard and Montanari (2018); Ramdas et al. (2017) (LORD), while the second one (ρALORD) uniformly improves upon the SAFFRON procedure of Ramdas et al. (2018) (ALORD).

- We present a general and simple way of constructing SUR procedures for any base procedure satisfying some mild conditions, see Section 2.3.4 for FWER and Section 2.4.4 for mFDR. This allows us to obtain concise proofs for all our results, which are deferred to the supplement, see Section A.1.

- Application to discrete data: we evaluate the performances of the new SUR procedures on discrete data, with simulated experiments (Section 2.5.2) and for a classical real data set (Section 2.5.3), where each hypothesis is tested using a (discrete) Fisher exact test. The gain in power is shown to be substantial.

- Application to p-value weighting: our new SUR procedures can be used to derive weighted online FWER and mFDR controlling procedures. The p-value weighting is carried out by rescaling in a certain way the 'raw' weights so that the weighted p-value distributions become super-uniform and our methodology can be applied. The new online procedures are shown to outperform existing ones both on simulated and real data (Section 2.6).

For easier readability of the paper, a succinct overview of our work is presented in Tables 2.1 and 2.2. It lists the base and SUR procedures and provides links to definitions and results for error rate control. All our numerical experiments (simulations and application) are reproducible from the code provided in the repository `https://github.com/iqm15/SUREOMT`.

2.1.5 Relation to adaptive discarding

As Tian and Ramdas (2019) pointed out, online multiple testing procedures frequently suffer from significant power loss if the null p-values are too conservative. In Tian and Ramdas (2021) (FWER control) and Tian and Ramdas (2019) (mFDR control), the authors propose adaptive discarding (ADDIS) approaches as improved methods. In particular, an idea is to use a discarding rule, that avoids testing a null when the corresponding p-value exceeds a given threshold. For the particular type of super-uniformity induced by discrete tests, we show that the discarding rule is less efficient than the SUR method, at least in the settings of Sections 2.5.2 and 2.5.3.

2.2 Preliminaries

2.2.1 Setting, procedure and assumptions

Let $X = (X_t, t \in \{1, 2, \ldots\})$ be a process composed of random variables. We denote the distribution of X by P, which is assumed to belong to some distribution set $\mathcal{P}$. We consider an online testing problem where, at each time $t \geq 1$, the user only observes variable X_t and should test a new null hypothesis H_t, which corresponds to some subset of $\mathcal{P}$, typically defined from the distribution of X_t. We let $\mathcal{H}_0 = \mathcal{H}_0(P) = \{t \geq 1 \, : \, H_t \text{ is satisfied by } P\}$ the set of (unknown) times where the corresponding null hypothesis is true. Throughout the manuscript, we focus on decisions based upon p-values. Hence, we suppose that at each time t, we have at hand a p-value $p_t = p_t(X) \in [0, 1]$ (typically depending only on X_t although this is not necessary) for testing H_t, and we consider online multiple testing procedures based on p-value thresholding. This means that each null H_t is rejected whenever $p_t(X) \leq \alpha_t$, where $\alpha_t \in [0, \infty)$ is a nonnegative threshold, called *a critical value*, that is allowed to depend on the past decisions. More precisely, we denote $R_t = \mathbf{1}\{p_t(X) \leq \alpha_t\}$, $C_t = \mathbf{1}\{p_t(X) \geq \lambda\}$ for all $t \geq 1$ and assume that each α_t is measurable with respect to the σ-field $\mathcal{F}_{t-1} = \sigma(R_1, \ldots, R_{t-1}, C_1, \ldots, C_{t-1})$. Here, $\lambda \in [0, 1]$ is a parameter that is used for designing adaptive procedures. The particular non-adaptive case is obtained by setting $\lambda = 0$, in which case $\mathcal{F}_{t-1} = \sigma(R_1, \ldots, R_{t-1})$.

In the literature, this property is referred to as predictability, see Ramdas et al. (2017). Throughout the manuscript, an online multiple testing procedure is identified with a family $\mathcal{A} = \{\alpha_t, t \geq 1\}$ of such predictable critical values. Let us now state the assumptions used in what follows. First, recall the classical *super-uniformity* assumption:

$$\mathbf{P}_{X\sim P}(p_t(X) \leq u) \leq u \text{ for all } u \in [0,1], \text{ and } P \in \mathcal{P} \text{ with } t \in \mathcal{H}_0, \tag{2.1}$$

which means that each test rejecting $H_{0,t}$ when $p_t(X)$ is smaller than or equal to u is of level u. Here, we typically consider a setting where these tests may have a more stringent level. Formally, at each time t, there is a *known* null function $F_t : [0,1] \to [0,1]$ satisfying

$$\mathbf{P}_{X\sim P}\left(p_t(X) \leq u\right) \leq F_t(u) \leq u, \text{ for all } u \in [0,1], \text{ and } P \in \mathcal{P} \text{ with } t \in \mathcal{H}_0. \tag{2.2}$$

Note that we will sometimes also consider $F_t(u)$ for $u \geq 1$, in which it is to be understood as $F_t(u \wedge 1)$. The family $\mathcal{F} = \{F_t, t \geq 1\}$ will be referred to as the *null bounding family*. Note that (2.2) reduces to (2.1) when choosing $F_t(u) = u$ for all u, but encompasses other cases by choosing differently the null bounding family. Typically, for discrete tests, it is well-known that $F_t(u)$ can be (much) smaller than u, see Example 2.2.1 for more details. Second, another important assumption is the *online independence* within the p-value process:

$$p_t(X) \text{ is independent of the past decisions } \mathcal{F}_{t-1} \text{ for all } t \in \mathcal{H}_0 \text{ and } P \in \mathcal{P}. \tag{2.3}$$

For instance, Assumption (2.3) holds in the case where $p_t(X)$ only depends on X_t and the variables in $(X_t, t \geq 1)$ are all mutually independent, which means that the data are collected independently at each time.

Remark 2.2.1 *In this manuscript, results are often based on assumptions* (2.2) *and* (2.3)*. In all these results, these two assumptions can be replaced by the weaker condition*

$$\mathbf{P}_{X\sim P}(p_t(X) \leq u \,|\, \mathcal{F}_{t-1}) \leq F_t(u) \leq u \quad \textit{a.s. for all } u \in [0,1], \textit{ for all } t \in \mathcal{H}_0 \textit{ and } P \in \mathcal{P}. \tag{2.4}$$

When choosing the null bounding family $F_t(u) = u$ for all u, the latter condition is sometimes referred to as SuperCoAD (super-uniformity conditionally on all discoveries), see Ramdas et al. (2017).

Throughout the paper, we investigate the two following prototypical examples of super-uniformity.

Example 2.2.1 *Our leading example is the case where a discrete test statistic is used for inference in each individual test. Typical instances include tests for analyzing counts represented by contingency tables, such as Fisher's exact test, see Section 2.5.2. In discrete testing, each p-value $p_t(X)$ has its own support $\mathcal{S}_t$ (known and not depending on P), that is a finite set (or, in full generality, a countable set with 0 as the only possible accumulation point). A null bounding family satisfying* (2.2) *can easily be derived by considering F_t, the right-continuous step function that jumps at each point of $\mathcal{S}_t$, see Figure 2.2 below. Note that the support $\mathcal{S}_t$ depends on t so that discrete testing also induces heterogeneity over time.*

Example 2.2.2 *Our secondary example is p-value weighting, where we start from continuous p-values (uniform under the null), which are weighted using external a priori information in order to increase power, see Section 2.6.*

2.2.2 Error rates and power

Let us define the criteria that we use to measure the quality of a given procedure $\mathcal{A} = \{\alpha_t, t \geq 1\}$. For each $T \geq 1$, let $\mathcal{R}(T) = \{t \in \{1, \dots, T\} \,:\, p_t(X) \leq \alpha_t\}$ denote the set of rejection times of the procedure $\mathcal{A}$, up to time T. We consider the two following classical online criteria for type I error rates:

$$\mathrm{FWER}(\mathcal{A}, P) := \sup_{T \geq 1} \{\mathrm{FWER}(T, \mathcal{A}, P)\},\ \mathrm{FWER}(T, \mathcal{A}, P) := \mathbf{P}_{X \sim P}\Big(|\mathcal{H}_0 \cap \mathcal{R}(T)| \geq 1\Big); \quad (2.5)$$

$$\mathrm{mFDR}(\mathcal{A}, P) := \sup_{T \geq 1} \{\mathrm{mFDR}(T, \mathcal{A}, P)\},\ \mathrm{mFDR}(T, \mathcal{A}, P) := \frac{\mathbf{E}_{X \sim P}\left(|\mathcal{H}_0 \cap \mathcal{R}(T)|\right)}{\mathbf{E}_{X \sim P}\left(1 \vee |\mathcal{R}(T)|\right)}, \quad (2.6)$$

with the convention $0/0 = 0$. In words, when controlling the online FWER at level α, one has the guarantee that, at each fixed time T, the probability of making at least one false discovery before time T is below α. Since FWER control does not tolerate any false discovery (with high probability), it is generally considered a stringent criterion. By contrast, when controlling the online mFDR, at each time T, the expected number of false discoveries before time T can be non-zero, but in an amount controlled by the expected number of discoveries. While online FWER has been investigated in Tian and Ramdas (2021), online mFDR control is generally less conservative (that is, allows more discoveries), and is widely used in an online context, see Foster and Stine (2008); Ramdas et al. (2017, 2018). The false discovery rate (FDR) is close to the mFDR: it is defined by using the expectation of the ratio, instead of the ratio of the expectations as in (2.6). Controlling the FDR generally requires more assumptions, while mFDR is particularly useful in an online context (we refer the reader to Section 1.1 of Zrnic et al. (2021) for more discussions on this). For a given error rate, we aim at deriving procedures that maximize power. For any procedure $\mathcal{A}$, we define the power as the expected proportion of signal the procedure can detect, that is,

$$\mathrm{Power}(T, \mathcal{A}, P) := \frac{\mathbf{E}_{X \sim P}\left(|\mathcal{H}_1 \cap \mathcal{R}(T)|\right)}{1 \vee |\mathcal{H}_1|}, \quad (2.7)$$

where $\mathcal{H}_1$ is the set of times of false nulls, that is, the complement of $\mathcal{H}_0$ in $\{1, 2, \dots\}$.

While this power notion will be used in our numerical experiments to compare procedures, our theoretical results will use a stricter comparison criterion. For two procedures $\mathcal{A} = \{\alpha_t, t \geq 1\}$ and $\mathcal{A}' = \{\alpha'_t, t \geq 1\}$, we say that $\mathcal{A}'$ *uniformly dominates* $\mathcal{A}$ when $\alpha'_t \geq \alpha_t$ for all $t \geq 1$ (almost surely). This implies that, almost surely, $\mathcal{A}'$ makes more discoveries than $\mathcal{A}$, in the sense that the set of discoveries of $\mathcal{A}$ is contained in the one of $\mathcal{A}'$, that is, $\mathcal{R}(T) \subset \mathcal{R}'(T)$ for all $T \geq 1$ (a.s.). In particular, this implies the same domination for the true discovery sets and thus in particular $\mathrm{Power}(T, \mathcal{A}, P) \leq \mathrm{Power}(T, \mathcal{A}', P)$ for all $T \geq 1$. With this terminology, we can restate the aim of this work as follows: construct valid OMT procedures that uniformly dominate their base procedures by incorporating the null bounding family F_t given in (2.2).

Remark 2.2.2 *There is no consensus regarding the most adequate definition of power in online testing literature. The concept of uniform domination that we use in this paper is much stronger than, e.g., the asymptotic power considered by Javanmard and Montanari (2018). It may, however, not be particularly appropriate if the base procedure $\mathcal{A}$ is chosen poorly. Since the base procedures given in Table 2.1 are standard in our settting, the domination criterion seems to be reasonable.*

2.2.3 Wealth and super-uniformity reward

In the Generalized Alpha-Investing (GAI) paradigm (see Xu and Ramdas (2021) and the references given therein), the nominal level α, at which one wants to control the type I error rate, can be seen as an overall error budget – or *wealth* – that may be spent on testing hypotheses in the course of an online experiment. For a given OMT procedure $\mathcal{A}$, it is possible to define a suitable *wealth function* $W(T) = W(T, \mathcal{A}, P)$, such that $W(T)$ represents the wealth available at time T for further testing. As a case in point, Xu and Ramdas (2021) define the (nominal) wealth function for the online Bonferroni procedure by $W^{\text{nom}}(T) = \alpha - \sum_{t=1}^{T} \alpha\gamma_t$. Generalizing this expression for arbitrary null distributions we obtain the 'true' or 'effective' wealth $W^{\text{eff}}(T) = \alpha - \sum_{t=1}^{T} F_t(\alpha\gamma_t)$, where F_t is a null-bounding function. In the super-uniform setting, assumption (2.2) implies $W^{\text{nom}}(T) \leq W^{\text{eff}}(T)$, and as the two orange curves in Figure 2.1 illustrate, the discrepancy can be quite large.

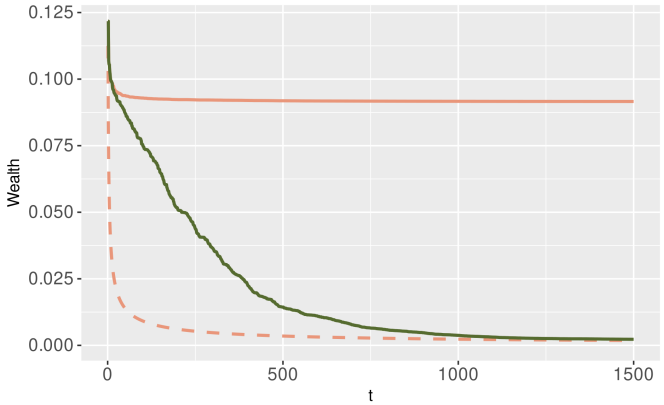

Figure 2.1: Nominal wealth for OB (dashed orange curve), effective wealth for OB (solid orange curve) and effective wealth for ρOB (solid green curve) for the male mice from the IMPC data (see Section 2.5.3 for more details).

However, while the user thinks the procedure is spending the budget over time according to the nominal wealth given by the dashed orange curve, in reality, the procedure is under-utilizing wealth, as the solid orange true wealth curve indicates. This unnecessarily austere spending behaviour makes the online Bonferroni procedure sub-optimal. In addition, this phenomenon extends to the other procedures and error rates listed in Table 2.1 as well. Our proposed solution incorporates super-uniformity so that its wealth function behaves more like the targeted nominal wealth, as depicted by the green curve in Figure 2.1.

For incorporating super-uniformity, we introduce the *super-uniformity reward* (SUR), a key quantity in our work. For any procedure $\mathcal{A} = \{\alpha_t, t \geq 1\}$ and null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$, the super-uniformity reward ρ_t at time t is defined by

$$\rho_t = \rho_t(\alpha_t, F_t) := \alpha_t - F_t(\alpha_t), \quad t \geq 1. \tag{2.8}$$

Note that (2.2) always implies $\rho_t \geq 0$ for all $t \geq 1$. In the case of discrete testing (Example 2.2.1), we have $F_t(\alpha_t) = 0$ when α_t is below the infimum of the support $\mathcal{S}_t$. This produces the maximum possible super-uniformity reward at time t, that is, $\rho_t = \alpha_t$. Conversely, when $\alpha_t \in \mathcal{S}_t$, we have $F_t(\alpha_t) = \alpha_t$ and we have no super-uniformity reward at time t, that is, $\rho_t = 0$. In general, we have $\rho_t \in [0, \alpha_t]$, its actual value depending on the discreteness of the test (that is on the steps of F_t) and of the value of α_t. The super-uniformity reward is illustrated in Figure 2.2 for a single

distribution F_t and value α_t. Mathematically, ρ_t is simply the difference between the nominal

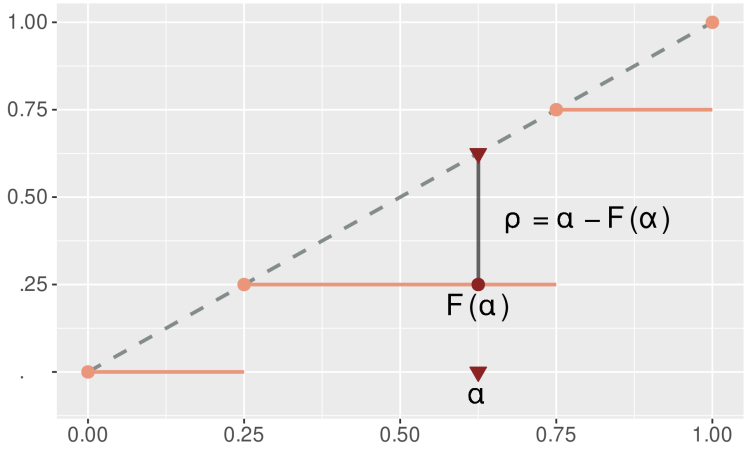

Figure 2.2: Super-uniformity reward ρ_t at time t (length of the vertical line) as defined by (2.8) for a given function F_t (orange step function) and a critical value α_t (triangle). The dashed line is the identity function $x \in [0,1] \mapsto x$.

significance level α_t and the truly achieved significance level $F_t(\alpha_t)$. In terms of wealth, ρ_t can be interpreted as the fraction of nominal significance level which the OMT procedure was unable to 'spend' due to super-uniformity. Intuitively, it seems clear that this amount can be put aside and be re-allocated to the subsequent tests to increase the future critical values $(\alpha_T, T \geq t+1)$. In Sections 2.3 and 2.4, we show in detail how this can be done without sacrificing type I error control.

2.2.4 Spending sequences

As Table 2.1 displays, the base procedures we use are parametrized by a sequence $\gamma = (\gamma_t)_{t\geq 1}$ of non-negative values, such that $\sum_{t\geq 1} \gamma_t \leq 1$, which we refer to as the *spending sequence*. The spending sequence controls the rate at which the wealth is spent in the course of the online experiment (for instance, see (2.10) for the online Bonferroni procedure). However, finding suitable spending sequences is not trivial: there is a trade-off between saving wealth for large values of T and the ability to make discoveries in the not-too-distant future. Typical choices for γ in the literature are:

- $\gamma_t \propto t^{-q}$ for all t for some $q > 1$, see Tian and Ramdas (2021);
- $\gamma_t \propto (t+1)^{-1} \log^{-q}(t+1)$ for all t, for some $q > 1$, see Tian and Ramdas (2021);
- $\gamma_t \propto \frac{\log((t+1)\vee 2)}{(t+1)\exp(\sqrt{\log(t+1)})}$, see Javanmard and Montanari (2018).

Throughout the paper, we choose $\gamma_t \propto t^{-q}$ with $q = 1.6$, as suggested by previous literature. In the base procedures listed in Table 2.1, there are two potential sources of wealth: the initial wealth invested at $T = 0$, and the *rejection reward* that can be earned by rejections for investing procedures (i.e., mFDR controlling procedures). When one can use super-uniformity reward as described in Section 2.2.3, an additional source of wealth comes into play. Indeed, our approach is to use an additional *SUR spending sequence* γ' to smoothly incorporate all the rewards collected

up to time T to compute the new critical value α_T. This SUR spending sequence could be chosen for instance from one of the smoothing sequences listed above. Here, we focus on the following choice:

$$\gamma'_t = \gamma'_t(h) = \mathbf{1}\{t \leq h\}/h, \quad t \geq 1, \tag{2.9}$$

where $h \geq 1$ is a suitably chosen integer. Since this leads to procedures that spread rewards uniformly over a finite horizon of length h, we refer to (2.9) – by analogy with non-parametric density estimation – as a *rectangular kernel* with *bandwidth* h. Finally, another idea introduced by Ramdas et al. (2018); Tian and Ramdas (2021) in order to slow down the natural decay in the α_t sequence is to consider $\gamma_{\mathcal{T}(t)}$ where $\mathcal{T}(t)$ is a slowed down clock, see (2.16) and (2.26) below. As we will see in Section 2.3.3 and Section 2.4.3, this technique can also be combined with a suitable super-uniformity reward.

2.3 Online FWER control

In this section, we aim at finding procedures $\mathcal{A}$ such that $\mathrm{FWER}(\mathcal{A}, P) \leq \alpha$ for some targeted level $\alpha \in (0, 1)$. We begin with a simple application of our approach to improve the online Bonferroni procedure with a 'greedy' super-uniformity reward, and then turn to a smoother spending of the super-uniformity reward (Theorem 2.3.1). This approach is then applied in combination with the adaptive online procedure introduced by Tian and Ramdas (2021) (Theorem 2.3.2). Finally, a general result is provided (Theorem 2.3.3) that allows to reward any procedure controlling the online FWER in some specific way. This allows unifying all results obtained in this section while further extending the scope of our methodology.

2.3.1 Warming-up: online Bonferroni procedure and a first greedy reward

For any given spending sequence sequence $\gamma = (\gamma_t)_{t\geq 1}$, a well-known online FWER controlling procedure is the online Bonferroni procedure, $\mathcal{A}^{\mathrm{OB}} = \mathcal{A}^{\mathrm{OB}}(\alpha, \gamma) := \{\alpha_t^{\mathrm{OB}}, t \geq 1\}$, defined by

$$\alpha_T^{\mathrm{OB}} := \alpha\gamma_T, \quad T \geq 1. \tag{2.10}$$

It is also called Alpha-Spending rule (Foster and Stine, 2008) in the context of online FWER control, see Tian and Ramdas (2021). It is straightforward to check that $\mathcal{A}^{\mathrm{OB}}$ controls the FWER under the classical super-uniformity condition (2.1): by the Markov inequality, for all $T \geq 1$,

$$\mathrm{FWER}(T, \mathcal{A}^{\mathrm{OB}}, P) \leq \mathbf{E}_{X\sim P}\Big(\sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0, p_t \leq \alpha\gamma_t\}\Big) \tag{2.11}$$

$$\leq \sum_{t\in\mathcal{H}_0} \mathbf{P}_{X\sim P}(p_t \leq \alpha\gamma_t) \leq \sum_{t\in\mathcal{H}_0} \alpha\gamma_t \leq \alpha. \tag{2.12}$$

Let us now present the rationale behind our approach in this simple case. Assume more generally that we have at hand a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying (2.2). The above reasoning leads to the following valid bound for any procedure $\mathcal{A} = \{\alpha_t, t \geq 1\}$ (with deterministic α_t):

$$\mathrm{FWER}(T, \mathcal{A}, P) \leq \sum_{t=1}^{T} F_t(\alpha_t) \leq \alpha_T + \sum_{t=1}^{T-1} F_t(\alpha_t) = \alpha \sum_{t=1}^{T} \gamma_t \leq \alpha, \tag{2.13}$$

by choosing $\alpha_T = \sum_{t=1}^{T} \alpha\gamma_t - \sum_{t=1}^{T-1} F_t(\alpha_t)$. The latter is a recursive relation that allows to define a new procedure $\mathcal{A} = \{\alpha_t, t \geq 1\}$ controlling the FWER. Since $\alpha_1 = \alpha\gamma_1$ and for $T \geq 2$, $\alpha_T - \alpha_{T-1} = \alpha\gamma_T - F_{T-1}(\alpha_{T-1})$, this leads to the simple rule

$$\alpha_T = \alpha\gamma_T + \rho_{T-1}, \quad T \geq 1, \tag{2.14}$$

where $\rho_{T-1} = \alpha_{T-1} - F_{T-1}(\alpha_{T-1})$ is the super-uniformity reward (2.8) at time $T-1$ (with the convention $\rho_0 = 0$). In addition, from (2.2), we have $\rho_{T-1} \geq 0$, and the critical values (2.14) uniformly dominate the online Bonferroni critical values (2.10) (the obtained critical values are in particular nonnegative, thus defining a valid OMT procedure). The approach behind critical values (2.14) is said here to be 'greedy', because it spends the complete super-uniformity reward ρ_{T-1} obtained at step $T-1$ for increasing the next critical value α_T.

2.3.2 Smoothing out the super-uniformity reward

The greedy policy described in the previous section is not always appropriate when time is considered on a potentially large period, because the sequence of critical values might fall too abruptly. Instead, we can smooth this effect over time, by distributing the reward collected at time $T-1$ over all times following T. To formalize this idea, we introduce a *SUR spending sequence* (see also Section 2.2.4), which is defined as a non-negative sequence $\gamma' = (\gamma'_t)_{t\geq 1}$ such that $\sum_{t\geq 1} \gamma'_t \leq 1$. While this definition is mathematically the same as the definition of a spending sequence, the role of the SUR spending sequence is different, so we use a different name for it.

Definition 2.3.1 *For any spending sequence γ and any SUR spending sequence γ', the online Bonferroni procedure with super-uniformity reward, denoted by $\mathcal{A}^{\rho OB} = \{\alpha_t^{\rho OB}, t \geq 1\}$, is defined by the recursion*

$$\alpha_T^{\rho OB} = \alpha\gamma_T + \sum_{t=1}^{T-1} \gamma'_{T-t}\rho_t, \quad T \geq 1, \tag{2.15}$$

where $\rho_t = \alpha_t^{\rho OB} - F_t(\alpha_t^{\rho OB})$ denotes the super-uniformity reward at time t for that procedure.

Note that taking $\gamma' = (1, 0, \dots 0)$ recovers the 'greedy' critical values (2.14). For the rectangular kernel SUR spending sequence given by (2.9), we have $\sum_{t=1}^{T-1} \gamma'_{T-t}\rho_t = h^{-1} \sum_{t=1\vee(T-h)}^{T-1} \rho_t$, which we interpret as a uniform spending of the SUR reward over the last h time points. As shown in Figure 2.3, the corresponding sequence of critical values (green line) is more 'stable' than the one using the greedy approach (blue line), allowing for some additional discoveries (on this simulated data). The following result provides FWER control of the new rewarded critical values (2.15), for a general SUR spending sequence.

Theorem 2.3.1 *Consider the setting of Section 4.2, where a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying (2.2) is at hand. For any spending sequence γ and any SUR spending sequence γ', consider the online Bonferroni procedure $\mathcal{A}^{OB} = \{\alpha_t^{OB}, t \geq 1\}$ (2.10), and the online Bonferroni with super-uniformity rewards $\mathcal{A}^{\rho OB} = \{\alpha_t^{\rho OB}, t \geq 1\}$ (2.15). Then we have $\mathrm{FWER}(\mathcal{A}^{\rho OB}, P) \leq \alpha$ for all $P \in \mathcal{P}$, while $\mathcal{A}^{\rho OB}$ uniformly dominates $\mathcal{A}^{OB}$.*

This result will be a consequence of a more general result, see Section 2.3.4.

2.3.3 Rewarded Adaptive Online Bonferroni

It is apparent from (2.11)-(2.12) that there is some looseness when upper-bounding $\sum_{t\in\mathcal{H}_0} \gamma_t$ by $\sum_{t\geq 1} \gamma_t$ which may lead to unnecessarily conservative procedures. We may attempt to avoid

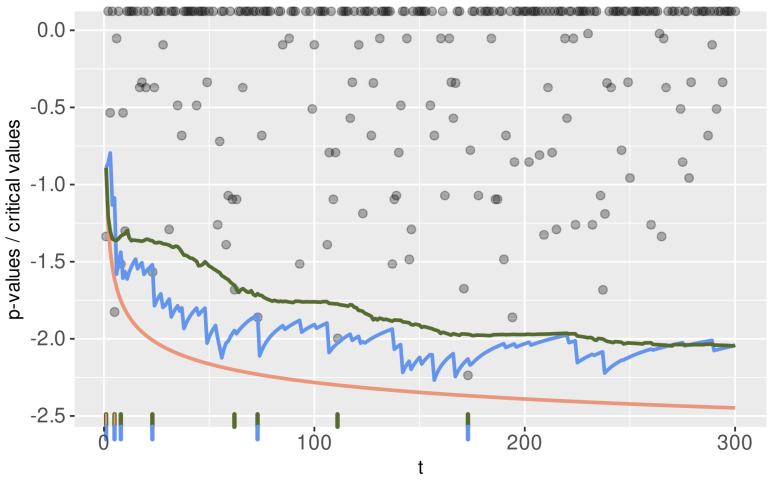

Figure 2.3: Sequences of critical values for Bonferroni procedures with different rewards over time $1 \leq t \leq T = 300$ (simulated data): base Bonferroni critical values (2.10) (orange line), rewarded with the greedy approach (2.14) (blue line), and with the rectangular kernel SUR spending sequence (2.15) ($h = 100$, green line). The rug plots display the time of discoveries for each procedure with the corresponding color. The Y-axis has been transformed by $y \mapsto -\log(-\log(y))$. The grey dots denote the p-value sequence (those equal to 1 are displayed at the top of the picture). The spending sequence is $\gamma_t \propto t^{-1.6}$.

this loss in efficiency by considering a spending sequence γ satisfying the condition $\sum_{t \in \mathcal{H}_0} \gamma_t \leq 1$ which is more liberal than $\sum_{t \geq 1} \gamma_t \leq 1$. In words, this means that the index t in the sequence $\{\gamma_t, t \geq 1\}$ should only be incremented when we are testing an hypothesis H_t with $t \in \mathcal{H}_0$. Since $\mathcal{H}_0$ is unknown, such a modification cannot be implemented directly in the γ sequence. Nevertheless, an approach proposed by Tian and Ramdas (2021) works by replacing the unknown set $\mathcal{H}_0$ by an estimate $\{1\} \cup \{t \geq 2 : p_{t-1} \geq \lambda\}$ for some parameter $\lambda \in (0, 1)$, and to correct the introduced error in the thresholds α_t to maintain the FWER control. More formally, we follow Tian and Ramdas (2021) by introducing the re-indexation functional $\mathcal{T} : \{1, \ldots\} \to \{1, \ldots\}$ defined by

$$\mathcal{T}(T) = 1 + \sum_{t=2}^{T} \mathbf{1}\{p_{t-1} \geq \lambda\}, \quad T \geq 1. \tag{2.16}$$

Since a large p-value is more likely to be linked to a true null, $\mathcal{T}(T)$ is used to account for the number of true nulls before time T (note that this estimate is nevertheless biased). From an intuitive point of view, $\mathcal{T}(T)$ slows down the time by only incrementing time when the preceding p-value is large enough. This idea leads to the adaptive online Bonferroni procedure introduced by Tian and Ramdas (2021) (called there 'Adaptive spending' [1]), with spending sequence γ and *adaptivity parameter* $\lambda \in [0, 1)$, denoted here by $\mathcal{A}^{\text{AOB}} = \{\alpha_t^{\text{AOB}}, t \geq 1\}$, and given by

$$\alpha_T^{\text{AOB}} = \alpha(1 - \lambda)\gamma_{\mathcal{T}(T)}, \quad T \geq 1. \tag{2.17}$$

[1] The so-called 'discarding' part of the method proposed by Tian and Ramdas (2021) cannot be implemented in our setting because the F_t are not convex, as discussed in Section 2.1.5.

It recovers the standard online Bonferroni procedure when $\lambda = 0$ (because $\mathcal{T}(T) = T$ for $T \geq 1$ in that case), but leads to different thresholds when $\lambda > 0$. Comparing $\mathcal{A}^{\text{AOB}}$ to $\mathcal{A}^{\text{OB}}$, no procedure uniformly dominates the other. An improvement of $\mathcal{A}^{\text{AOB}}$ over $\mathcal{A}^{\text{OB}}$ is expected to hold when there are many false null hypotheses in the data, and increasingly so if the signal occurs early in the time sequence, see the numerical experiments in Section 2.5.2. In addition, note that the critical value α_T^{AOB} depends on the data $X_1, \ldots, X_{T-1}$ and thus is random. As a result, the adaptive approach requires additional distributional assumptions compared with the online Bonferroni procedure. In Tian and Ramdas (2021), $\mathcal{A}^{\text{AOB}}$ is proved to control the FWER under (2.1) and (2.3) (actually under the slightly more general condition (2.4) with F_t equal to identity). Let us now use this approach in combination with the super-uniformity reward.

Definition 2.3.2 *For any spending sequence γ, any SUR spending sequence γ', and $\lambda \in [0, 1)$, the adaptive online Bonferroni procedure with super-uniformity reward, denoted by $\mathcal{A}^{\rho AOB} = \{\alpha_t^{\rho AOB}, t \geq 1\}$, is defined by*

$$\alpha_T^{\rho AOB} = \alpha(1-\lambda)\gamma_{\mathcal{T}(T)} + \sum_{\substack{1 \leq t \leq T-1 \\ p_t \geq \lambda}} \gamma'_{T-t}\rho_t + \varepsilon_{T-1}, \quad T \geq 1, \tag{2.18}$$

where $\rho_t = \alpha_t^{\rho AOB} - F_t(\alpha_t^{\rho AOB})$ denotes the super-uniformity reward a time t, and $\varepsilon_{T-1} = \mathbf{1}\{p_{T-1} < \lambda\}(\alpha_{T-1} - \alpha(1-\lambda)\gamma_{\mathcal{T}(T-1)})$ is an additional 'adaptive' reward (convention $\varepsilon_0 = 0$).

This class of procedures reduces to the class of procedures (2.15) introduced in the previous section by setting $\lambda = 0$. However, when $\lambda > 0$ the class is different since the term $\alpha(1-\lambda)\gamma_{\mathcal{T}(T)}$, which comes from α_T^{AOB}, makes the threshold random. Also, the super-uniformity reward is only collected at time $t \leq T-1$ where $p_t \geq \lambda$. The latter is well expected from the motivation of the adaptive approach described above: when $p_t < \lambda$, no testing is performed so no reward could be obtained from ρ_t. Nevertheless, note that the additional term ε_{T-1} allows to collect some reward at time $T-1$ in the case where $p_{T-1} < \lambda$. Since this term only appears in critical values of adaptive procedures, we call it the 'adaptive' reward. It is linked to the super-uniformity reward in that no adaptive reward can be obtained if no super-uniformity reward has been collected in the past. The following result shows that this approach is valid from the FWER control perspective.

Theorem 2.3.2 *Consider the setting of Section 4.2 where a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying (2.2) is at hand. For any spending sequence γ, any SUR spending sequence γ' and $\lambda \in [0, 1)$, consider the adaptive online Bonferroni procedure $\mathcal{A}^{AOB} = \{\alpha_t^{AOB}, t \geq 1\}$ (2.17) and the adaptive online Bonferroni with super-uniformity rewards $\mathcal{A}^{\rho AOB} = \{\alpha_t^{\rho AOB}, t \geq 1\}$ (2.18). Then, assuming that the model $\mathcal{P}$ is such that (2.3) holds, we have $\text{FWER}(\mathcal{A}^{\rho AOB}, P) \leq \alpha$ for all $P \in \mathcal{P}$, while $\mathcal{A}^{\rho AOB}$ uniformly dominates $\mathcal{A}^{AOB}$.*

Theorem 2.3.2 relies on a more general result (Theorem 2.3.3 below). Note that, contrary to Theorem 2.3.1, Theorem 2.3.2 needs an independence assumption. This was already the case without the super-uniformity reward since this is due to the adaptive methodology that makes the critical values random. If this independence assumption holds, we show in Section 2.5.2 that $\mathcal{A}^{\rho\text{AOB}}$ can indeed improve $\mathcal{A}^{\rho\text{OB}}$, while it always improves the procedure $\mathcal{A}^{\text{AOB}}$ of Tian and Ramdas (2021) (as guaranteed by the above theorem).

2.3.4 Rewarded version for base FWER controlling procedures

In this section we present a general result stating that any procedure ensuring online FWER control (in a specific way) can be rewarded using super-uniformity while maintaining the FWER control.

Theorem 2.3.3 *Assuming that* (2.2) *holds, consider any procedure* $\mathcal{A}^0 = (\alpha_t^0, t \geq 1)$ *satisfying almost surely, for some* $\lambda \in [0, 1)$ *and for all* $T \geq 1$,

$$\alpha_T^0 + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \alpha_t^0 \leq (1 - \lambda)\alpha. \tag{2.19}$$

Then the following holds:

(i) $\mathcal{A}^0$ *controls the online FWER, that is,* $\mathrm{FWER}(\mathcal{A}^0, P) \leq \alpha$ *for all* $P \in \mathcal{P}$*, either if the* α_T^0 *are deterministic for all* $T \geq 1$*, or if* (2.3) *holds;*

(ii) for any SUR spending sequence $\gamma' = (\gamma'_t, t \geq 1)$*, the procedure* $\mathcal{A} = (\alpha_t, t \geq 1)$*, corresponding to the rewarded* $\mathcal{A}^0$*, and defined by*

$$\alpha_T = \alpha_T^0 + \sum_{\substack{1 \leq t \leq T-1 \\ p_t \geq \lambda}} \gamma'_{T-t}(\alpha_t - F_t(\alpha_t)) + \mathbf{1}\{p_{T-1} < \lambda\}(\alpha_{T-1} - \alpha_{T-1}^0), \quad T \geq 1, \tag{2.20}$$

controls the online FWER, that is, $\mathrm{FWER}(\mathcal{A}, P) \leq \alpha$ *for all* $P \in \mathcal{P}$*, either if the* α_T *are deterministic for all* $T \geq 1$*, or if* (2.3) *holds.*

Theorem 2.3.3 is proved in Section A.1.1. Condition (2.19) is essentially the same as Condition (20) derived in Tian and Ramdas (2021). It is satisfied by the online Bonferroni procedure ($\mathcal{A}^0 = \mathcal{A}^{\mathrm{OB}}$), and the online adaptive Bonferroni procedure ($\mathcal{A}^0 = \mathcal{A}^{\mathrm{AOB}}$). While this is obvious for $\mathcal{A}^{\mathrm{OB}}$, the case of $\mathcal{A}^{\mathrm{AOB}}$ requires to carefully check how the functional $\mathcal{T}(\cdot)$ (2.16) slows down the time, which is done in Lemma A.1.3. Statement (i) of Theorem 2.3.3 thus proves the online FWER control for these procedures. Statement (ii) of Theorem 2.3.3 is our main contribution and reduces to Theorems 2.3.1 and 2.3.2, when choosing $\mathcal{A}^0 = \mathcal{A}^{\mathrm{OB}}$ and $\mathcal{A}^0 = \mathcal{A}^{\mathrm{AOB}}$, respectively. This recovers the rewarded procedures $\mathcal{A}^{\rho\mathrm{OB}}$ and $\mathcal{A}^{\rho\mathrm{AOB}}$ discussed in the previous sections: compare (2.20) to (2.15) (with $\lambda = 0$), and (2.20) to (2.18). Nevertheless, other choices for $\mathcal{A}^0$ satisfying (2.19) are possible. According to our general result, any such choice is compatible with our reward methodology.

2.4 Online mFDR control

In this section, we aim at finding procedures $\mathcal{A}$ such that $\mathrm{mFDR}(\mathcal{A}, P) \leq \alpha$ for some targeted level $\alpha \in (0, 1)$. We follow the same route as for the FWER: we start with an application of the super-uniformity reward to the classical LORD++ procedure (Ramdas et al., 2017, called just LORD hereafter for short), and then turn to adaptive counterparts. Finally, we propose a general result encompassing all these cases. In this section, we follow the notation of Ramdas et al. (2017) for online mFDR control. For any procedure $\mathcal{A} = \{\alpha_t, t \geq 1\}$ and realization of the p-value process, let us denote

$$R(T) = \sum_{t=1}^{T} \mathbf{1}\{p_t(X) \leq \alpha_t\} \tag{2.21}$$

the number of rejections of the procedure up to time T, and

$$\tau_j = \min\{t \geq 1 \,:\, R(t) \geq j\} \quad (\tau_j = +\infty \text{ if the set is empty}), \tag{2.22}$$

the first time that the procedure makes j rejections, for any $j \geq 1$.

2.4.1 Warming up: LORD procedure and a first greedy reward

While a sufficient condition for online FWER control is $\sum_{t\geq 1}\alpha_t \leq \alpha$ (see the previous section and in particular (2.19)), the mFDR control is ensured when $\sum_{t\geq 1}\alpha_t \leq \alpha(1 \vee R(T))$, as proved in Theorem 2 of Ramdas et al. (2017) (applicable, e.g., under assumptions (2.2) and (2.3)). Consequently, for each rejection we earn back wealth α with which we are allowed to increase α_t; typically by starting a new online Bonferroni critical value process. This idea is referred to as α-investing in the literature, see Foster and Stine (2008); Aharoni and Rosset (2014); Javanmard and Montanari (2018). This idea leads to the LORD (Levels based On Recent Discovery) procedure (Javanmard and Montanari, 2018), with the improvement given by Ramdas et al. (2017):

$$\alpha_T^{\mathrm{LORD}} = W_0\gamma_T + (\alpha - W_0)\gamma_{T-\tau_1} + \alpha\sum_{j\geq 2}\gamma_{T-\tau_j}, \quad T \geq 1, \tag{2.23}$$

where by convention $\gamma_t = 0$ at any time $t \leq 0$ and where γ is an arbitrary spending sequence. Note that the test level at time T splits the initial α-wealth between the cases where $R(T) = 0$ and $R(T) = 1$, because the bound is equal to $\alpha(1 \vee R(T)) = \alpha$ in both cases so the first rejection does not provide an extra room for false discoveries. The resulting additional parameter $W_0 \in (0, \alpha)$ balances the initial α-wealth between these two cases to maintain the mFDR control. The procedure $\mathcal{A}^{\mathrm{LORD}} = \{\alpha_t^{\mathrm{LORD}}, t \geq 1\}$ controls the mFDR under (2.1) and (2.3), because $\sum_{t\geq 1}\alpha_t^{\mathrm{LORD}} \leq \alpha(1\vee R(T))$ (see Section A.1.2 for a proof). Now, let us consider our more general framework where we have at hand a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying (2.2). In that case, we can prove that a sufficient condition on the critical values for mFDR control is that, almost surely,

$$\sum_{t=1}^{T} F_t(\alpha_t) \leq \alpha_T + \sum_{t=1}^{T-1} F_t(\alpha_t) \leq \alpha(1 \vee R(T)),$$

see the general condition (2.30) below. This can be achieved by choosing

$$\alpha_T = \sum_{t=1}^{T}\alpha_t^{\mathrm{LORD}} - \sum_{t=1}^{T-1} F_t(\alpha_t), \quad T \geq 1.$$

This leads to the thresholds

$$\alpha_T = \alpha_T^{\mathrm{LORD}} + \rho_{T-1}, \quad T \geq 1, \tag{2.24}$$

where $\rho_{T-1} = \alpha_{T-1} - F_{T-1}(\alpha_{T-1})$ is the super-uniformity reward (2.8) at time $T-1$ (with the convention $\rho_0 = 0$). Since $\rho_t \geq 0$ for all t by (2.2), this procedure uniformly dominates the procedure $\mathcal{A}^{\mathrm{LORD}}$. Furthermore, depending on the magnitude of the super-uniformity reward, this new procedure is potentially much more powerful.

2.4.2 Smoothing out the super-uniformity reward

As discussed for FWER control (see Section 2.3.2), the preliminary procedure (2.24) spends immediately at time T all of the super-uniformity reward collected at time $T-1$. However, it is more advantageous to redistribute this reward over subsequent times $T, T+1, \ldots$, by using a SUR spending sequence $\gamma' = (\gamma'_t)_{t\geq 1}$. This gives rise to the following more general class of online procedures.

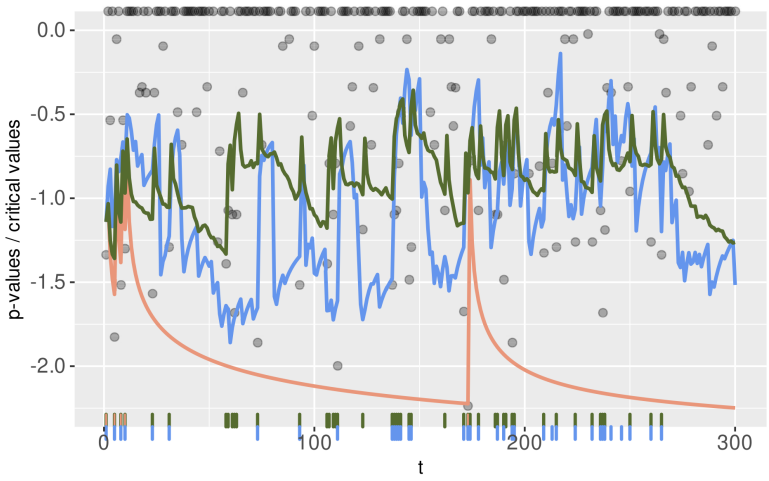

Figure 2.4: Sequences of critical values of LORD procedure with different rewards over time $1 \leq t \leq T = 300$ (simulated data): base LORD critical values (B.13)(orange line), rewarded with the greedy approach (2.24) (blue line), and with the rectangular kernel SUR spending sequence (2.25) ($h = 10$, green line). The rug plots display the time of discoveries for each procedure with the corresponding color. The y-axis has been transformed by $y \mapsto -\log(-\log(y))$. The grey dots denote the p-value sequence (those equal to 1 are displayed at the top of the picture). The spending sequence is $\gamma_t \propto t^{-1.6}$.

Definition 2.4.1 *For a spending sequence γ and a SUR spending sequence γ', the LORD procedure with super-uniformity reward, denoted by $\mathcal{A}^{\rho LORD} = \{\alpha_t^{\rho LORD}, t \geq 1\}$, is defined by the recursion*

$$\alpha_T^{\rho LORD} = \alpha_T^{LORD} + \sum_{t=1}^{T-1} \gamma'_{T-t} \rho_t \quad T \geq 1, \tag{2.25}$$

where α_T^{LORD} is given by (B.13) *and $\rho_t = \alpha_t^{\rho LORD} - F_t(\alpha_t^{\rho LORD})$ denotes the super-uniformity reward at time t.*

Figure 2.4 displays the critical values of the LORD procedure, and of those rewarded with the greedy SUR spending sequence $\gamma' = (1, 0, \dots)$ or rewarded with the rectangular kernel SUR spending sequence (2.25) ($h = 10$). First, the reward given by the α-investing, which is possible for mFDR control, is visible at each discovery for which all critical value curves 'jump'. Second, the effect of the super-uniformity reward is visible between these jumps, and the kernel sequence is able to better smooth the critical value sequence. As a result, the corresponding procedure is likely to make more discoveries (as it is the case on the simulated data presented in Figure 2.4). The following result establishes the mFDR control of this new class of rewarded procedures.

Theorem 2.4.1 *Consider the setting of Section 4.2 where a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying* (2.2) *is at hand. For any spending sequence γ and any SUR spending sequence γ', consider the LORD procedure $\mathcal{A}^{LORD} = \{\alpha_t^{LORD}, t \geq 1\}$* (B.13) *and the LORD procedure with super-uniformity rewards $\mathcal{A}^{\rho LORD} = \{\alpha_t^{\rho LORD}, t \geq 1\}$* (2.25)*. Then, assuming that the model $\mathcal{P}$ is such that* (2.3) *holds, we have* $\mathrm{mFDR}(\mathcal{A}^{\rho LORD}, P) \leq \alpha$ *for all $P \in \mathcal{P}$ while $\mathcal{A}^{\rho LORD}$ uniformly dominates $\mathcal{A}^{LORD}$.*

This theorem is proved in Section A.1.2, as a corollary of a more general result (Theorem 2.4.3 below). As shown in the numerical experiments (Section 2.5.2), the improvement of $\mathcal{A}^{\rho \mathrm{LORD}}$ with

respect to $\mathcal{A}^{\text{LORD}}$ can be substantial.

Remark 2.4.1 *$\mathcal{A}^{\rho LORD}$ can be also expressed by using the paradigm of generalized α investing (GAI) rules, as introduced in Foster and Stine (2008); Aharoni and Rosset (2014); Ramdas et al. (2017), see Section A.3.3.*

2.4.3 Rewarded Adaptive LORD

In this section, we apply the re-indexation trick of the γ sequence presented in Section 2.3.3 to improve the performance of the procedures $\mathcal{A}^{\text{LORD}}$ and $\mathcal{A}^{\rho\text{LORD}}$. For this, we follow essentially the reasoning used by Ramdas et al. (2018) for deriving the SAFFRON procedure, with a slight modification, as explained below. To start, let us define, for some parameter $\lambda \in [0, 1)$,

$$\mathcal{T}_j(T) = \begin{cases} 1 + \sum_{t=\tau_j+2}^{T} \mathbf{1}\{p_{t-1} \geq \lambda\} & \text{if } T \geq \tau_j + 1 \\ 0 & \text{if } T \leq \tau_j \end{cases}, \quad j \geq 1, \tag{2.26}$$

with $\mathcal{T}_0(T) = \mathcal{T}(T)$ given by (2.16) by convention. From an intuitive point of view, $\mathcal{T}_j(T)$ is like a 'stopwatch' starting after τ_j and suspended at each time t for which $p_{t-1} < \lambda$. Hence, having $p_t < \lambda$ allows to delay the natural dissipation of α-wealth due to online testing. Then, the SAFFRON procedure (Ramdas et al., 2018) is defined by the threshold

$$\alpha_T = \min\left(\lambda, (1-\lambda)\left(W_0\gamma_{\mathcal{T}_0(T)} + (\alpha - W_0)\gamma_{\mathcal{T}_1(T)} + \alpha\sum_{j\geq 2}\gamma_{\mathcal{T}_j(T)}\right)\right). \tag{2.27}$$

This procedure controls the mFDR under (2.1) and (2.3) as proved by Ramdas et al. (2018). However, examining the proof in Ramdas et al. (2018), it turns out that the capping with λ is not necessary. The capping prevents the critical values from exceeding λ, thus avoiding to get $p_t \geq \lambda$ when $p_t \leq \alpha_t$. However, to our knowledge, the latter does not play any role in the mFDR control, and we work with the (uniformly dominating) procedure

$$\alpha_T^{\text{ALORD}} = (1-\lambda)\left(W_0\gamma_{\mathcal{T}_0(T)} + (\alpha - W_0)\gamma_{\mathcal{T}_1(T)} + \alpha\sum_{j\geq 2}\gamma_{\mathcal{T}_j(T)}\right). \tag{2.28}$$

With the capping (2.27), an mFDR control is provided in Theorem 1 in Ramdas et al. (2018). For our version (2.28), the mFDR control follows as a special case of Theorem 2.4.2 below with $F_t(u) = u$ for all t, u. Also note that $\mathcal{A}^{\text{ALORD}}$ reduces to $\mathcal{A}^{\text{LORD}}$ (B.13) when $\lambda = 0$, because $\mathcal{T}_j(T) = 0 \vee (T - \tau_j)$ in that case. Now, we generalize this method to our present framework.

Definition 2.4.2 *For a spending sequences γ, a SUR spending sequence γ' and $\lambda \in [0, 1)$, the adaptive LORD procedure with super-uniformity reward denoted by $\mathcal{A}^{\rho ALORD} = \{\alpha_t^{\rho ALORD}, t \geq 1\}$, is defined by*

$$\alpha_T^{\rho ALORD} = \alpha_T^{ALORD} + \sum_{\substack{1\leq t\leq T-1 \\ p_t\geq\lambda}} \gamma'_{T-t}\rho_t + \varepsilon_{T-1}, \quad T \geq 1, \tag{2.29}$$

where α_T^{ALORD} is defined by (2.28), $\rho_t = \alpha_t^{\rho ALORD} - F_t(\alpha_t^{\rho ALORD})$ denotes the super-uniformity reward a time t and $\varepsilon_{T-1} = \mathbf{1}\{p_{T-1} < \lambda\}(\alpha_{T-1}^{\rho ALORD} - \alpha_{T-1}^{ALORD})$ is an additional 'adaptive' reward at time $T-1$ (convention $\varepsilon_0 = 0$).

Note that $\mathcal{A}^{\rho\text{ALORD}}$ reduces to $\mathcal{A}^{\rho\text{LORD}}$ (2.25) when $\lambda = 0$, and to $\mathcal{A}^{\text{ALORD}}$ when $F_t(u) = u$ for all u, t. The following result shows that this class of procedures controls the mFDR.

Theorem 2.4.2 *Consider the setting of Section 4.2 where a null bounding family* $\mathcal{F} = \{F_t, t \geq 1\}$ *satisfying* (2.2) *is at hand. For any spending sequence* γ *and any SUR spending sequence* γ'*, consider the adaptive LORD procedure* $\mathcal{A}^{ALORD} = \{\alpha_t^{ALORD}, t \geq 1\}$ (2.28)*, and the adaptive LORD procedure with super-uniformity rewards* $\mathcal{A}^{\rho ALORD} = \{\alpha_t^{\rho ALORD}, t \geq 1\}$ (2.29)*. Then, assuming that the model* $\mathcal{P}$ *is such that* (2.3) *holds, we have* $\mathrm{mFDR}(\mathcal{A}^{\rho ALORD}, P) \leq \alpha$ *for all* $P \in \mathcal{P}$ *while* $\mathcal{A}^{\rho ALORD}$ *uniformly dominates* $\mathcal{A}^{ALORD}$ *and thus also the SAFFRON procedure of Ramdas et al. (2018).*

Theorem 2.4.2 follows from Theorem 2.4.3 below. Let us underline that $\mathcal{A}^{\rho\mathrm{ALORD}}$ both incorporates α-investing and super-uniformity reward. Thus, it is expected to be the most powerful among the procedures considered in the present paper. This is supported both by the numerical experiments of Section 2.5.2 and the real data analysis in Section 2.5.3.

Remark 2.4.2 *Note that the critical values of ALORD and* ρ*-ALORD can exceed* 1 *(e.g., when all p-values are zero). Since the rejection decision is the same for a critical value larger than* 1 *or equal to* 1*, this may appear at first sight as wasted wealth. While this is indeed the case for ALORD, we emphasize that this is not the case for* ρ*-ALORD, because the super-uniformity reward allows to reuse the exceeding amount of wealth engaged in* $\alpha_t^{\rho ALORD}$*; namely* $\rho_t = \alpha_t^{\rho ALORD} - 1$ *when* $\alpha_t^{\rho ALORD} \geq 1$*.*

2.4.4 Rewarded version for base mFDR controlling procedures

The following result establishes that any base online mFDR controlling procedure (of a specific type) can be rewarded with super-uniformity.

Theorem 2.4.3 *Assuming that both* (2.2) *and* (2.3) *hold, consider any procedure* $\mathcal{A}^0 = (\alpha_t^0, t \geq 1)$ *satisfying almost surely, for some* $\lambda \in [0, 1)$ *and for all* $T \geq 1$*,*

$$\alpha_T^0 + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \alpha_t^0 \leq (1 - \lambda)\alpha \, (1 \vee R(T)), \tag{2.30}$$

where $R(T)$ *denotes the number of rejections up to time* T *for this procedure, see* (2.21)*. Then the following holds*

(i) $\mathcal{A}^0$ *controls the online mFDR, that is,* $\mathrm{mFDR}(\mathcal{A}^0, P) \leq \alpha$ *for all* $P \in \mathcal{P}$*;*

(ii) for any SUR spending sequence $\gamma' = (\gamma'_t, t \geq 1)$*, the procedure* $\mathcal{A} = (\alpha_t, t \geq 1)$*, corresponding to the rewarded* $\mathcal{A}^0$*, and defined by* (2.20)*, controls the online mFDR, that is,* $\mathrm{mFDR}(\mathcal{A}, P) \leq \alpha$ *for all* $P \in \mathcal{P}$*.*

Theorem 2.4.3 is proved in Section A.1.2. Condition (2.30) is essentially the same as the condition found in Theorem 1 of Ramdas et al. (2018). Our main contribution is thus in statement (ii), showing that the super-uniformity reward can be used with any base procedure $\mathcal{A}^0$ satisfying (2.30). Since the latter condition holds for the LORD procedure $\mathcal{A}^0 = \mathcal{A}^{\mathrm{LORD}}$, and the adaptive LORD procedure $\mathcal{A}^0 = \mathcal{A}^{\mathrm{ALORD}}$ (see Lemma A.1.3), Theorem 2.4.3 entails Theorem 2.4.1 and Theorem 2.4.2, respectively. Finally, let us emphasize the similarity between Theorem 2.3.3 (FWER) and Theorem 2.4.3 (mFDR). Strikingly, the reward takes exactly the same form (2.20), which makes the range of improvement comparable for these two criteria.

2.5 SUR procedures for discrete tests

In this section, we study the performances of our newly derived SUR procedures in discrete online multiple testing problems for simulated and real data. We defer some of the numerical results to Appendix A.4.

2.5.1 Considered procedures

The considered procedures are the base (non-rewarded) procedures $\mathcal{A}^{\text{OB}}$ (2.10), $\mathcal{A}^{\text{AOB}}$ (2.17), $\mathcal{A}^{\text{LORD}}$ (B.13), and $\mathcal{A}^{\text{ALORD}}$ (2.28), and their rewarded counterparts $\mathcal{A}^{\rho\text{OB}}$ (2.18), $\mathcal{A}^{\rho\text{AOB}}$ (2.18), $\mathcal{A}^{\rho\text{LORD}}$ (2.29), and $\mathcal{A}^{\rho\text{ALORD}}$ (2.29), respectively. As mentioned in Section 2.1.5, we also consider the ADDIS-spending and ADDIS procedures (see Tian and Ramdas, 2021, 2019) although the type I error rate control is not guaranteed for these two procedures, in our (discrete) setting. The parameters of the OMT procedures are set to $\alpha = 0.2$, $W_0 = \alpha/2$ and $\lambda = 0.5$. For ADDIS and ADDIS-spending, we use the default values $W_0 = \frac{\alpha\lambda\tau}{2}$, with $\lambda = 0.25$ and $\tau = 0.5$ (the latter being the discarding parameter, see Tian and Ramdas, 2021, 2019). Following Tian and Ramdas (2019), we set $\gamma_t \propto t^{-1.6}$ with a normalizing constant chosen such that $\sum_{t=1}^{+\infty} \gamma_t = 1$. For the SUR spending sequence $(\gamma'_t)_{t\geq 1}$ we use a rectangular kernel with bandwidth h, as defined by (2.9), with $h = 100$ for FWER and $h = 10$ for mFDR. We discuss different choices for tuning parameters in the SUR procedures (adaptivity parameter λ and the rectangular kernel bandwidth h) in Appendices A.4.4 and A.4.5.

2.5.2 Application to simulated data

Simulation setting

We simulate m experiments in which the goal is to detect differences between two groups by counting the number of successes/failures in each group. More specifically, we follow Gilbert (2005), Heller and Gur (2011) and Döhler et al. (2018) by simulating a two-sample problem in which a vector of m independent binary responses is observed for N subjects in both groups. The goal is to test the m null hypotheses H_{0i}: '$p_{1i} = p_{2i}$', $i = 1, ..., m$ in an online fashion, where p_{1i} and p_{2i} are the success probabilities for the i^{th} binary response in group A and B respectively. Thus, for each hypothesis i, the data can be summarized by a 2×2 contingency table, and we use (two-sided) Fisher's exact test for testing H_{0i}. The m hypotheses are split in three groups of size m_1, m_2, and m_3 such that $m = m_1 + m_2 + m_3$. Then, the binary responses are generated as i.i.d Bernoulli of probability 0.01 ($\mathcal{B}(0.01)$) at m_1 positions for both groups, i.i.d $\mathcal{B}(0.10)$ at m_2 positions for both groups, and i.i.d $\mathcal{B}(0.10)$ at m_3 positions for one group and i.i.d $\mathcal{B}(p_3)$ at m_3 positions for the other group. Thus, the null hypotheses are true for $m_1 + m_2$ positions (set $\mathcal{H}_0$), while the null hypotheses are false for m_3 positions (set $\mathcal{H}_1$). Therefore, we interpret p_3 as the strength of the signal while $\pi_A = \frac{m_3}{m}$, corresponds to the proportion of signal. Also, m_1 and m_2 are both taken equal to $\frac{m-m_3}{2}$. In these experiments, we fix $m = 500$, and vary each one of the parameters $\mathcal{H}_1$ (Section 2.5.2), π_A (Section 2.5.2), N (Section A.4.1), p_3 (Section A.4.2) while keeping the others fixed. The default values are $\pi_A = 0.3$, $N = 25$, $p_3 = 0.4$ and $\mathcal{H}_1 \subset \{1, \dots, m\}$ chosen randomly for each simulation run. We estimate the different criteria (FWER (2.5), mFDR (2.6), power (2.7)) using empirical mean over 10 000 independent simulation trials.

Position of signal

We start by studying how the position of the signal can affect the performances of the procedures (it is well-known to be critical, see Foster and Stine, 2008; Ramdas et al., 2017). We investigate

different positioning schemes in which the signal can be clustered at the beginning of the stream, or at the end, or clustered between the two, as described in the caption of Figure 2.5. Consistently

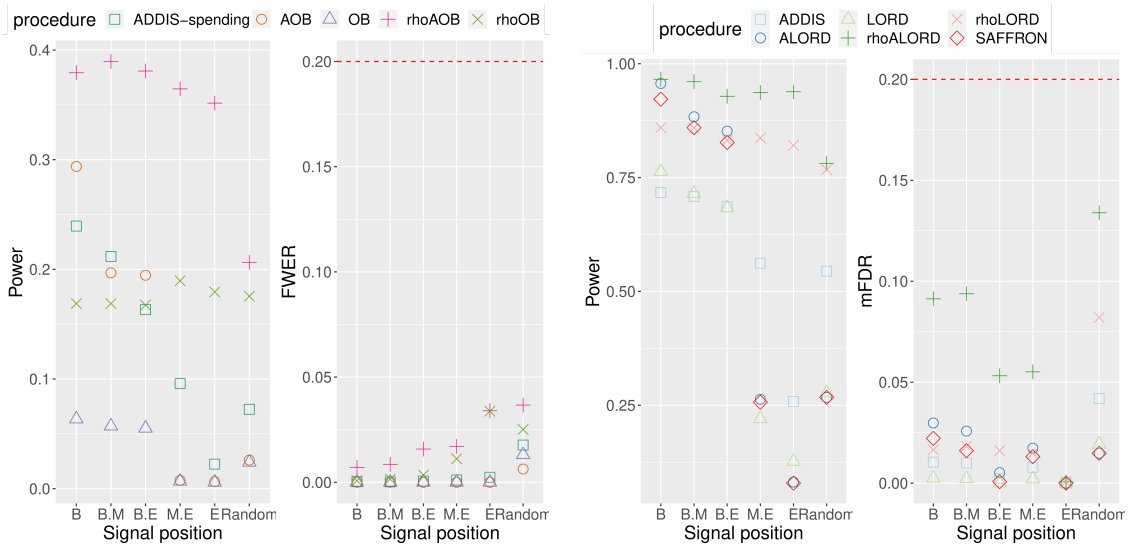

Figure 2.5: Power and type I error rates of the different considered OMT procedures versus positions of the signal: at the beginning (B), the end (E), half at the beginning and half in the middle of the stream (BM), half at the beginning and half at the end of the stream (BE), half in the middle and half at the end of the stream (ME), and taken uniformly at random (Random).

with our theoretical results, Figure 2.5 shows that all procedures control the type I error rate at level $\alpha = 0.2$. In terms of power, we can see that the rewarded procedures have greater power than the associated base procedures. More specifically, $\mathcal{A}^{\rho\text{ALORD}}$ uniformly dominates the other procedures for mFDR control and $\mathcal{A}^{\rho\text{AOB}}$ for FWER control. The gain in power is most noticeable when the signal is not localized at the beginning of the stream (i.e. positions ME, E, and Random) for which the online testing problem is more difficult. These first results indicate that the rewarded procedures may protect against 'α-death'.

Proportion of signal

Figure 2.6 displays the results for π_A varying in $\{0.1, \dots, 1\}$. It shows that the aforementioned superiority of the rewarded procedures holds in this whole range. Also note that the SUR reward can affect the monotonicity of the power curves: while most curves are increasing with π_A, the power of the rewarded procedure $\mathcal{A}^{\rho\text{OB}}$ decreases. An explanation could be that when π_A increases, the marginal counts increase, and thus the degree of discreteness decreases providing a smaller super-uniformity reward. However, using adaptivity seems to compensate for this effect, thus providing better results.

Finally, let us mention that the additional numerical results in Section A.4 provide qualitatively similar conclusions for all other explored parameter configurations: the SUR procedures $\mathcal{A}^{\rho\text{AOB}}$ and $\mathcal{A}^{\rho\text{ALORD}}$ always improve, often substantially, the existing OMT procedures.

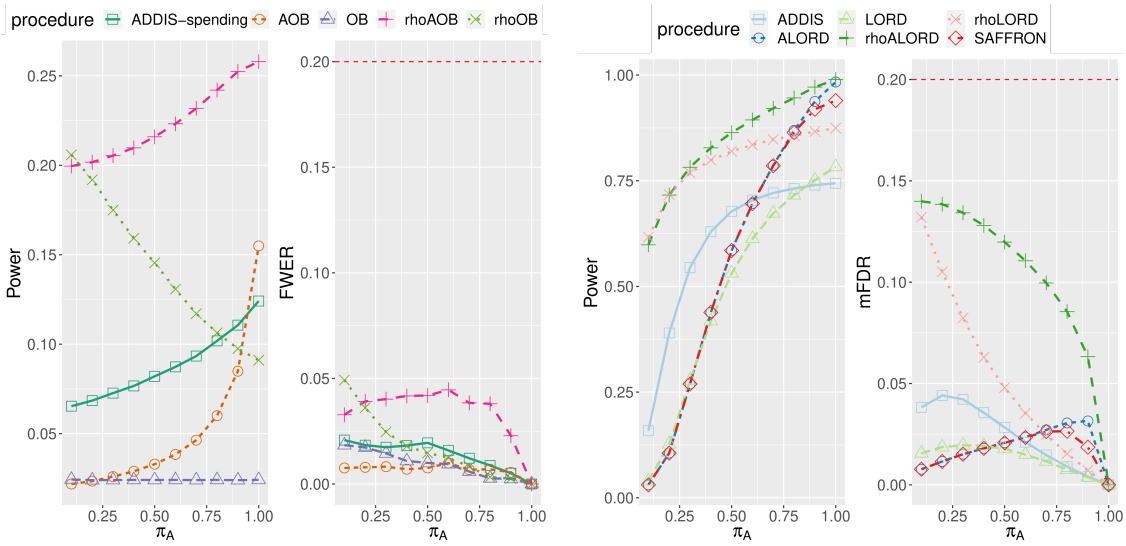

Figure 2.6: Power and type I error rates of the considered procedures for $\pi_A \in \{0.1, 0.2, \ldots, 0.9, 1\}$.

Table 2.3: Number of discoveries for FWER controlling OMT procedures (left) and mFDR controlling OMT procedures (right). These numbers are obtained by running the procedures on the first 30 000 genes for male (second row) and female (third row) mice in the IMPC data.

Procedures	OB	ρOB	AOB	ρAOB	LORD	ρLORD	ALORD	ρALORD
# discoveries (male)	229	377	281	697	882	972	972	1041
# discoveries (female)	267	481	764	811	839	946	966	1046

2.5.3 Application to IMPC data

In this section we analyse data from the International Mouse Phenotyping Consortium (IMPC), which coordinates studies on the genotype influence on mouse phenotype. More precisely, scientists test the hypotheses that the knock-out of certain genes will not change certain phenotypic traits (e.g., the coat or eye color). Since the data set is constantly evolving as new genes are studied for new phenotypic traits of interest, online multiple testing is a natural approach for analysing such data, see also Tian and Ramdas (2021); Xu and Ramdas (2021). We use the data set provided by Karp et al. (2017) which includes, for each studied gene, the count of normal and abnormal phenotype for female and male mice (separately), thus providing two by two contingency tables, which can be analysed using Fisher exact tests. In this section, we investigate the genotype effect on the phenotype separately for male and female. The data set originally contains nearly 270 000 genes studies, but we focus on the first 30 000 genes for simplicity. We set the global level α to 0.2 and 0.05, respectively for FWER and mFDR procedures. For the procedure parameters, we follow the choice made in Section 2.5.1. Table 2.3 presents the number of discoveries for the FWER controlling procedures OB, AOB, ρOB, ρAOB (left) and for the mFDR controlling procedures LORD, ALORD, ρLORD, ρALORD (right). The results show that ignoring the discreteness of the tests causes the scientist to miss (potentially many) discoveries. Hence, using the SUR methods helps to reduce this risk.

Figure 2.7 (FWER procedures) and Figure 2.8 (mFDR procedures) illustrate in more detail how the super-uniformity reward leads to more discoveries, in the case of male mice (similar findings hold for the female mice for which the corresponding figures can be found in Section A.5.2). First, note that the smallest p-values occur at the beginning of the stream (see Figure A.9 in Section A.5.1), so that we limit the visual analysis to the first 1500 p-values for clarity of exposition. For the ρOB procedure, the benefit of incorporating the super-uniformity reward is visible in the left panel of Figure 2.7. As expected from Figure 2.3, applying a rectangular kernel to these rewards yields a smooth curve. For the ρAOB procedure, presented in the right panel of Figure 2.7, the improvement is even stronger, but the resulting critical value curve is less smooth. This is due to the 'adaptive' reward, that is, the ε_{T-1}-component of our improvement, recall (2.18). More precisely, an explanation of this 'saw-tooth' shape is that during a period with p-values smaller than λ, we have $\alpha_T^{\rho\mathrm{AOB}} - \alpha_T^{\mathrm{AOB}} \geq \alpha_{T-1}^{\rho\mathrm{AOB}} - \alpha_{T-1}^{\mathrm{AOB}}$ so the gain increases. Also, if this period lasts for a while (as for $500 \lesssim t \lesssim 1240$ here), the ρ-part of the reward vanishes and we end up with a constant gain $\alpha_T^{\rho\mathrm{AOB}} - \alpha_T^{\mathrm{AOB}} \approx \alpha_{T-1}^{\rho\mathrm{AOB}} - \alpha_{T-1}^{\mathrm{AOB}}$, explaining the flat part of the curve, until the next $p_T \geq \lambda$ occurs. After this point, we switch from the ε-regime back to the ρ-regime, i.e., $\alpha_{T+1}^{\rho\mathrm{AOB}} = \alpha_{T+1}^{\mathrm{AOB}} + \gamma_1' \rho_T$. Since typically $\gamma_1' \rho_T \ll \alpha_{T-1}^{\rho\mathrm{AOB}} - \alpha_{T-1}^{\mathrm{AOB}}$, this causes the downward jump in the green curve. For the mFDR procedures presented in Figure 2.8, there is an additional 'rejection' reward as described in Section 2.4. Note that this makes some critical values exceed 1 (both for ALORD and ρALORD), which thus cannot be displayed in the Y-axis scale considered in that figure. However, these values are still used in ρALORD algorithm to compute the future critical values (see Remark 2.4.2). The obtained results are qualitatively similar to the FWER setting: our proposed reward makes the green curves run above the orange ones, uniformly over the considered time, hence inducing significantly more discoveries.

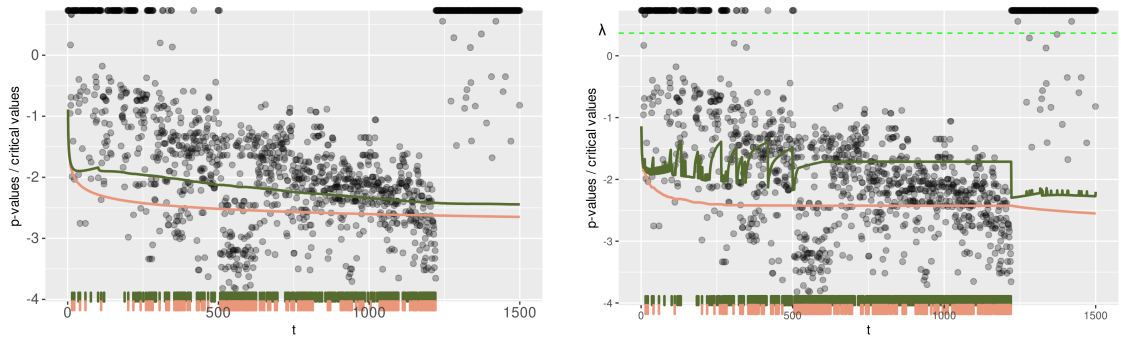

Figure 2.7: Applying online FWER controlling procedures to the male mice IMPC data set. Left panel: p-values and critical values for OB (orange curve) and ρOB (green curve). Right panel: AOB (orange curve) and ρAOB (green curve). Representation similar to Figure 2.3 (Y-axis transformed by $y \mapsto -\log(-\log(y))$; p-values equal to 1 displayed at the top of the picture).

2.6 SUR procedures for weighted p-values

In this section, we show how our SUR approach can be easily used to construct valid online p-value weighting procedures.

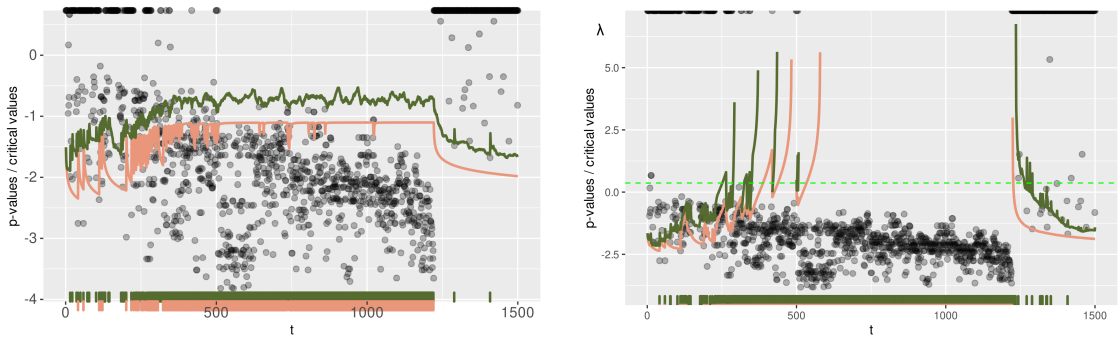

Figure 2.8: Applying online mFDR controlling procedures to the male mice IMPC data set. Left panel: p-values and critical values for LORD (orange curve) and ρLORD (green curve). Right panel: ALORD (orange curve) and ρALORD (green curve). Representation similar to Figure 2.4 (Y-axis transformed by $y \mapsto -\log(-\log(y))$; p-values equal to 1 are displayed at the top of the picture).

2.6.1 Setting and benchmark procedure

Consider a standard continuous online multiple testing setting where each p-value is super-uniformly distributed under the null, that is, (2.1) holds. Assume in addition that, at each time t, the p-value p_t is associated with a quantity $r_t \geq 0$, called the *raw weight* (as opposed to the *rescaled weight* defined further on), which is assumed to be measurable w.r.t. $\mathcal{F}_{t-1}$. The magnitude of r_t is interpreted as the level of belief in a potential true discovery at time t: a large weight indicates a strong belief that the corresponding null hypothesis is false. Throughout the section, the weights r_t are assumed to be available a priori and we will not discuss how to derive them (for this task, we refer to Wasserman and Roeder (2006); Rubin et al. (2006); Roeder and Wasserman (2009); Hu et al. (2010); Zhao and Zhang (2014); Ignatiadis et al. (2016); Chen and Kasiviswanathan (2020) among others).

While p-value weighting is a classical tool for improving the performance of multiple testing methods in the offline setting (see references in Section 2.1.3), the incorporation of weights has received little attention in the online case. The only relevant work to our knowledge is Ramdas et al. (2017) (Section 5 therein), which presents sufficient criteria for weighting procedures controlling the (m)FDR based on so-called GAI++ procedures and also discusses the technical challenges associated with weighted online multiple testing. An explicit algorithm which satisfies these criteria is used in Ramdas et al. (2017)[2], which is detailed in Appendix A.3.2 for completeness. This method, which will be our benchmark procedure, works by weighting the p-values and adjusting for this weighting in the rejection reward.

2.6.2 New weighting approach

The main idea of our new approach is as follows: consider weighted p-values $\tilde{p}_t = p_t/w_t$ for some rescaled weight $w_t \in [0,1]$ which gives rise to the null bounding family $\mathcal{F} = \{F_t : u \in [0,1] \mapsto uw_t, t \geq 1\}$. Since the weights are constrained to take their values in $[0,1]$, the functions of $\mathcal{F}$ are super-uniform, that is, (2.2) holds. Hence, one can apply our SUR approach with respect to that family $\mathcal{F}$.

[2]An implementation of this procedure can be found on the website `https://github.com/fanny-yang/OnlineFDRCode`

Table 2.4: Number of discoveries for weighted controlling OMT procedures for the 'airway' data set , with the weights taken from Ignatiadis et al. (2016).

Procedures	OB	wOB (new)	AOB	wAOB (new)	LORD	wGAI$_1$	wGAI$_2$	wLORD (new)
# discoveries	1092	1195	1188	1273	3550	1308	3631	4445

More specifically, our approach takes into account the null bounding family $\mathcal{F}$ in a simple two-step process, which proceeds as follows: for each time t,

1. enforce super-uniformity by computing the *rescaled weight* $w_t = \xi_t(r_t|r_1, \ldots, r_{t-1})$, $t \geq 1$, for some given *rescaling function* ξ_t valued in $[0, 1]$ (see below for more details and an explicit choice);

2. apply any one of the SUR methods from Section 2.3 or Section 2.4, depending on whether FWER or mFDR control is desired.

We denote these new procedures by wX, where X stands for the name of the base procedure (either OB (2.10), AOB (2.17), LORD (B.13) or ALORD (2.28)). These procedures all come with the corresponding FWER or mFDR control (by additionally assuming (2.3) if needed). In particular, to the best of our knowledge, this also provides the first method for weighted online FWER control.

At first sight, these SUR weighting approaches may seem to be ineffective due to the conservatism induced by the rescaling step. However, this is countered in the second step by using SUR procedures that provide larger values α_t, due to the super-uniform rewards accumulated in the past. The hope is that these two effects balance out in such a way as to favor rejection of hypotheses associated with larger values of (raw) weights.

Finally, let us mention that a simple choice for ξ_t is given by $\xi_t(x|r_1, \ldots, r_{t-1}) = \hat{F}_{t-1}(x)\mathbf{1}\{x > 0\}$, where $\hat{F}_{t-1}(x) = (t-1)^{-1} \sum_{i=1}^{t-1} \mathbf{1}\{r_i \leq x\}$ is the empirical c.d.f. of the sample $r_1, \ldots, r_{t-1}$ (and by convention $\hat{F}_0(x) = 1$). This particular choice is easy to compute in a sequential manner, and it satisfies the following intuitive and desirable properties: $\xi_t(x) \in [0, 1]$ (ensures super-uniformity of $\mathcal{F}$), $\xi_t(x)$ is nondecreasing in x (a larger raw weight leads to a larger rescaled weight), $\xi_t(0) = 0$ (raw zero weights rescaled to zero), $\xi_t(\lambda r_t|\lambda r_1, \ldots, \lambda r_{t-1}) = \xi_t(r_t|r_1, \ldots, r_{t-1})$ for all $\lambda > 0$ (scale invariance) and if all raw weights are equal then all rescaled weights are equal to 1.

2.6.3 Analysis of RNA-Seq data

We revisit an analysis of the RNA-Seq data set 'airway' using results from the Independent Hypothesis Weighting (IHW) approach (for details, see Ignatiadis et al. (2016) and the vignette accompanying its software implementation). While the original data was not collected in an online fashion, we use it here nevertheless to provide a proof of concept for weighted SUR procedures. The 'airway' data set contains data from 64102 genes and the corresponding (offline) weights are taken from the output of the ihw function from the bioconductor package 'IHW'. These 'raw' weights are then transformed into rescaled weights by using the function ξ_t described in the previous section. For the procedure parameters, we use the same choices as for the analysis of the IMPC data, see Section 2.5.3.

Table 2.4 (left part) presents the result for the FWER controlling procedures OB, AOB (non-weighted), and wOB, wAOB (SUR weighted approaches). It is clear that incorporating the weights leads to more rejections, which corroborates the fact that the weights coming from Ignatiadis et al. (2016) are indeed informative.

As for mFDR control, the (non-weighted) LORD is compared to our weighted version wLORD in Table 2.4 (right part). As additional competitors, we also added the weighted GAI++ procedure proposed in Ramdas et al. (2017) (see Section A.3.2 for a detailed description), that we use either with the raw weights (denoted by $w\text{GAI}_1$) or with the rescaled weights (denoted by $w\text{GAI}_2$). As one can see, the effect of rescaling the weights is highly beneficial, and the new wLORD proposal is the one that incorporates these weights in the most efficient way.

2.7 Discussion

2.7.1 Conclusion

Existing OMT procedures often suffer from a lack of power due to conservativeness of the p-values. This occurs typically for discrete test statistics, which is a common situation in data sets where testing is based upon counts. To fill the gap, we introduced new SUR versions of some existing classical procedures, that 'reward' the base procedures by spending more efficiently the α-wealth according to known bounds on the null cumulative distribution functions. We showed that our new SUR procedures provide rigorous control of online error criteria (FWER or mFDR) under classical assumptions while offering a systematic power enhancement. When using discrete Fisher exact test statistics, the improvement is substantial, both for simulated and real data.

In addition, even in the standard case of uniformly distributed p-values, our approach allowed us to derive new weighted procedures that incorporate external covariates. This provides improvements w.r.t. existing online weighting strategies.

2.7.2 Another viewpoint

In the discrete setting, let us consider the following constrained spending problem: at each step t, choose the critical value α_t to be in the support S_t (including 0) so that the following contraint holds

$$\sum_{t\geq 1} \alpha_t \leq \alpha. \tag{2.31}$$

It solves the super-uniformity problem, because $F_t(\alpha_t) = \alpha_t$ for all t, while it controls the online FWER. This general principle, that we refer to as 'constrained spending strategies', can be implemented in many ways.

Markedly, the *SUR approach is a way to achieve this*, by additionally following some reference critical values — here the online Bonferroni critical values α_t^{OB} (2.10). Indeed, the rejection decision $p_t \leq \alpha_t^{\rho\text{OB}}$ and $p_t \leq \alpha_t = F_t(\alpha_t^{\rho\text{OB}})$ are almost surely identical and we have calibrated $\alpha_t^{\rho\text{OB}}$ such that (2.31) holds, see (2.13). In other words, even if our critical values are not constrained to be in the support initially, the effective critical values $\alpha_t = F_t(\alpha_t^{\rho\text{OB}})$ that are actually used in the decision rule will automatically belong to the support. Thus, our approach can be equivalently seen as a way of implementing the constrained spending strategies delineated above.

Obviously, there are other ways to implement the constrained spending strategy. One instance is the delayed spending (DS) approach, that we describe in detail in Appendix A.2.

2.7.3 Future directions

While our results address several issues, they also raise new questions. First, the bandwidth of the kernel-based SUR spending sequence γ' given by (2.9) has been chosen in a loose way

here, but tuning the bandwidth is certainly interesting from a power enhancement perspective (see Section A.4.5). Also, in applications, the user would possibly like to select the bandwidth in a data dependent fashion without losing control over type I error rate. These two issues are interesting extensions for future developments. Second, while our work focuses on marginal FDR, it would be desirable to build rewarded OMT procedures that control the (non-marginal) FDR. However, usual proofs rely on a monotonicity property of the critical value sequence (Ramdas et al., 2017) that is difficult to satisfy here, because the super-uniformity reward naturally varies over time. Hence, deriving rewarded FDR controlling procedures is a challenging issue that is left for future investigations. Third, most of our results rely on an independence assumption, see (2.3). While this can be considered as a mild restriction in an online framework, relaxing it or incorporating a known dependence structure in OMT is an interesting avenue.

Chapter 3

False discovery proportion envelopes with consistency

Outline of the current chapter

We provide new false discovery proportion (FDP) confidence envelopes in several multiple testing settings relevant for modern high dimensional-data methods. We revisit the scenarios considered in the recent work of Katsevich and Ramdas (2020) (top-k, preordered — including knockoffs —, online) with a particular emphasis on obtaining FDP bounds that have both non-asymptotical coverage and asymptotical consistency, i.e. converge below the desired level α when applied to a classical α-level false discovery rate (FDR) controlling procedure. This way, we derive new bounds that provide improvements over existing ones, both theoretically and practically, and are suitable for situations where at least a moderate number of rejections is expected These improvements are illustrated with numerical experiments and real data examples. In particular, the improvement is significant in the knockoffs setting, which shows the impact of the method for a practical use. As side results, we introduce a new confidence envelope for the empirical cumulative distribution function of i.i.d. uniform variables and we provide new power results in sparse cases, both being of independent interest.

3.1 Introduction

3.1.1 Background

Multiple inference is a crucial issue in many modern, high dimensional, and massive data sets, for which a large number of variables are considered and many questions naturally emerge, either simultaneously or sequentially. Recent statistical inference has thus turned to designing methods that guard against false discoveries and selection effect, see Cui et al. (2021); Robertson et al. (2022) for recent reviews on that topic. A key quantity is typically the false discovery proportion (FDP), that is, the proportion of false discoveries within the selection (Benjamini and Hochberg, 1995).

Among classical methods, finding confidence bounds on the FDP that are valid after a user data-driven selection ('post hoc' FDP bounds), has retained attention since the seminal works of Genovese and Wasserman (2004, 2006); Goeman and Solari (2011). The strategy followed by these works is to build confidence bounds *valid uniformly over all selection subsets*, which *de facto* provides a bound valid for any data-driven selection subset. A number of such FDP bounds have been proposed since, either based on a 'closed testing' paradigm (Hemerik et al., 2019; Goeman et al., 2019, 2021; Vesely et al., 2021), a 'reference family' (Blanchard et al., 2020; Durand et al., 2020), or a specific prior distribution in a Bayesian framework (Perrot-Dockès et al., 2021). It should also be noted that methods providing bounds valid uniformly over *some particular* selection subsets can also be used to provide bounds valid on *any* subsets by using an 'interpolation' technique, see, e.g., Blanchard et al. (2020). This is the case for instance for bounds based upon an empirical distribution function confidence band, as investigated by Meinshausen and Bühlmann (2005); Meinshausen (2006); Meinshausen and Rice (2006); Dümbgen and Wellner (2023). Loosely, we will refer to such (potentially partial) FDP bounds as *FDP confidence envelopes* in the sequel.

Recently, finding FDP confidence envelopes has been extended to different contexts of interest in Katsevich and Ramdas (2020) (KR below for short), including knockoffs (Barber and Candès, 2015; Candès et al., 2018) and online multiple testing (Aharoni and Rosset, 2014). For this, their bounds are tailored on particular nested 'paths', and employ accurate martingale techniques. In addition, Li et al. (2022) have recently investigated specifically the case of the knockoffs setting by using a 'joint' k-FWER error rate control (see also Genovese and Wasserman, 2006; Meinshausen, 2006; Blanchard et al., 2020), possibly in combination with closed testing.

3.1.2 New insight: consistency

The main point of this paper is to look at FDP confidence envelopes towards the angle of a particular property that we call *consistency*. First recall that the false discovery rate (FDR) is the expectation of the FDP, which is a type I error rate measure with increasing popularity since the seminal work of Benjamini and Hochberg (1995). Informally, an FDP confidence envelope is consistent, if its particular value on an FDR-controlling selection set is close to (or below) the corresponding nominal value, at least asymptotically. This property is important for several reasons:

- FDR controlling procedures are particular selection sets that are widely used in practice. Hence, it is very useful to provide an accurate FDP bound for these particular rejection sets. This is the case for instance for the commonly used Benjamini-Hochberg (BH) procedure at a level α — or even for a data dependent choice of the level $\hat{\alpha}$ — for which the FDP bound should be close to α (or $\hat{\alpha}$), at least in 'favorable' cases;

- a zoo of FDP confidence envelopes have been proposed in previous literature, and we see the consistency as a principled way to discard some of them while putting the emphasis on others;

- searching for consistency can also lead to new bounds that are accurate for a moderate sample size.

It turns out that most of the existing bounds, while being accurate in certain regimes, are not consistent. In particular, this is the case for those of Katsevich and Ramdas (2020), because of a constant factor (larger than 1) in front of the FDP estimate. The present paper proposes to fill this gap by proposing new envelopes that are consistent. In a nutshell, we replace the constant in front of the FDP estimate by a function that tends to 1 in a particular asymptotical regime.

Since we evoke consistency, it is worth emphasizing that the envelopes developed in this work have coverage holding in a *non-asymptotical* sense. Here, consistency means that on top of this strong non-asymptotical guarantee, the bound satisfies an additional sharpness condition in an asymptotical sense and for some scenarios of interest, including sparse ones.

3.1.3 Settings

Following Katsevich and Ramdas (2020), we consider the three following multiple testing settings for which a 'path' means a (possibly random) nested sequence of candidate rejection sets:

- Top-k: the classical multiple testing setting where the user tests a finite number m of null hypotheses and observes simultaneously a family of corresponding p-values. This is the framework of the seminal paper of Benjamini and Hochberg (1995) and of the majority of the follow-up papers. In that case, the path is composed of the hypotheses corresponding to the top-k most significant p-values (i.e. ranked in increasing order), for varying k.

- Pre-ordered: we observe p-values for a finite set of cardinal m of null hypotheses, which are *a priori* arranged according to some ordering. In that setting, the signal (if any) is primarily carried by the ordering: alternatives are expected to be more likely to have a small rank. Correspondingly the path in that case is obtained by p-value thresholding (for fixed threshold) of the first k hypotheses w.r.t. that order, for varying k. A typical instance is the knockoffs setting (Barber and Candès, 2015; Candès et al., 2018), where the null hypotheses come from a high-dimensional linear regression model and one wants to test whether each of the m variables is associated with the response. The ordering is

data-dependent and comes from an ancillary statistic independent of the tests themselves, so that one can argue conditionally and consider the ordering (and path) as fixed.

- Online: the null hypotheses come sequentially, and there is a corresponding potentially infinite stream of p-values. An irrevocable decision (reject or not) has to be taken in turn for each new hypothesis, depending on past observations only. The path is naturally defined according to the set of rejections until time t, for varying t.

Let us introduce notation that encompasses the three settings mentioned above: the set of hypotheses is denoted by $\mathcal{H}$ (potentially infinite), the set of null hypotheses $\mathcal{H}_0$ is an unknown subset of $\mathcal{H}$, and a path $\Pi = (R_k, k \geq 1)$ (with convention $R_0 = \emptyset$) is an ordered sequence of nested subsets of $\mathcal{H}$ that depends only on the observations. A confidence envelope is a sequence $(\overline{\mathrm{FDP}}_k, k \geq 1)$ (with convention $\overline{\mathrm{FDP}}_0 = 0$) of random variables valued in $[0, 1]$, depending only on the observations, such that, for some pre-specified level δ, we have

$$\mathbf{P}(\forall k \geq 1, \mathrm{FDP}(R_k) \leq \overline{\mathrm{FDP}}_k) \geq 1 - \delta, \tag{3.1}$$

where $\mathrm{FDP}(R_k) = \frac{|R_k \cap \mathcal{H}_0|}{|R_k| \vee 1}$ is the FDP of the set R_k. In (3.1), the guarantee is uniform in k, which means that it corresponds to confidence bounds valid uniformly over the subsets of the path. Also, distribution $\mathbf{P}$ is relative to the p-value model, which will be specified further on and depends on the considered framework.

Remark 3.1.1 (Interpolation) *On can notice here that any FDP confidence envelope of the type* (3.1) *can also lead to a post hoc FDP bound valid uniformly for all $R \subset \mathcal{H}$: specifically, by using the interpolation method (see, e.g., Blanchard et al., 2020; Goeman et al., 2021; Li et al., 2022), if* (3.1) *holds then the relation also holds with the sharper bound* $(\widetilde{\mathrm{FDP}}_k, k \geq 1)$ *given by*

$$\widetilde{\mathrm{FDP}}_k = \frac{\min_{k' \leq k} \{|R_k \cap (R_{k'})^c| + |R_{k'}| \overline{\mathrm{FDP}}_{k'}\}}{|R_k| \vee 1}, \tag{3.2}$$

due to the fact that the number of false positives in R_k is always bounded by the number of false positives in $R_{k'} \subset R_k$ plus the number of elements of $R_k \cap (R_{k'})^c$.

Particular subsets of $\Pi = (R_k, k \geq 1)$ that are of interest are those controlling the FDR. Given a nominal level α, a 'reference' procedure chooses a data-dependent $\hat{k}_\alpha$ such that $\mathbb{E}\left[\mathrm{FDP}(R_{\hat{k}_\alpha})\right] \leq \alpha$. Depending on the setting, we consider different reference procedures:

- Top-k setting: the reference FDR controlling procedure is the Benjamini-Hochberg (BH) step-up procedure, see Benjamini and Hochberg (1995);
- Pre-ordered setting: the reference procedure is the Lei-Fithian (LF) adaptive Selective sequential step-up procedure, see Lei and Fithian (2016) (itself being a generalization of the procedure of Li and Barber, 2017);
- Online setting: the reference procedure is the (LORD) procedure, see Javanmard and Montanari (2018) and more precisely the improved version of Ramdas et al. (2017).

As announced, for all these procedures, the *expectation* of the FDP (that is, the FDR) is guaranteed to be below α. On the other hand, it is commonly the case that in an appropriate asymptotic setting, the FDP concentrates around its expectation, see, e.g., Genovese and Wasserman (2004); Neuvial (2008, 2013). Therefore, an adequate confidence bound on the FDP should asymptotically converge to (or below) α when applied to a reference procedure. Furthermore, we emphasize

once more that we aim at a bound which is valid non-asymptotically, and uniformly over the choice of α (or equivalently k) to account for possible 'data snooping' from the user (that is, $\alpha = \hat{\alpha}$ is possibly depending on the data).

Let us now make the definition of consistency more precise.

Definition 3.1.1 (Consistency for top-k and pre-ordered settings) *Let $\delta \in (0,1)$ be fixed. For each $m \geq 1$, let $\mathbf{P}^{(m)}$ be a multiple testing model over the hypotheses set $\mathcal{H} = \{1, \dots, m\}$, $\Pi = (R_k, k \geq 1)$ be a possibly random path of nested subsets of $\mathcal{H}$, and $(\overline{\mathrm{FDP}}_k, k \geq 1)$ a confidence envelope at level $1-\delta$ over that path, i.e. satisfying (3.1) (for $\mathbf{P} = \mathbf{P}^{(m)}$). For any $\alpha \in (0,1)$, let $\hat{k}_\alpha$ be an FDR controlling procedure at level α, i.e. satisfying $\mathbf{E}^{(m)}\left[\mathrm{FDP}(R_{\hat{k}_\alpha})\right] \leq \alpha$. Then the confidence envelope is said to be consistent for the sequence $(\mathbf{P}^{(m)}, m \geq 1)$ and for the FDR controlling procedure $R_{\hat{k}_\alpha} \in \Pi$ at a level α in a range $[\alpha_0, 1) \subset (0,1)$, if for all $\epsilon > 0$,*

$$\lim_{m \to \infty} \mathbf{P}^{(m)} \left(\sup_{\alpha \in [\alpha_0, 1)} \left\{ \overline{\mathrm{FDP}}_{\hat{k}_\alpha} - \alpha \right\} \geq \epsilon \right) = 0. \tag{3.3}$$

In the above definition, $\mathbf{P}^{(m)}$ stands for a multiple testing model with m hypotheses that is to be specified. We will be interested in standard model sequences that represent relevant practical situations, in particular sparse cases where a vanishing proportion of null hypotheses are false when m tends to infinity. This definition applies for the two first considered settings (top-k and pre-ordered). Note that due to (3.1), we have

$$\mathbf{P}(\forall \alpha \in (0,1), \mathrm{FDP}(R_{\hat{k}_\alpha}) \leq \overline{\mathrm{FDP}}_{\hat{k}_\alpha}) \geq 1 - \delta. \tag{3.4}$$

Hence, (3.3) comes as an additional asymptotical accuracy guarantee to the non-asymptotical coverage property (3.4). Moreover, the uniformity in α in (3.4)-(3.3) allows for choosing α in a post hoc manner, while maintaining the false discovery control and without paying too much in accuracy, that is, for any data-dependent choice of $\hat{\alpha}$, $\mathrm{FDP}(R_{\hat{k}_{\hat{\alpha}}}) \leq \overline{\mathrm{FDP}}_{\hat{k}_{\hat{\alpha}}}$ with probability at least $1-\delta$, with $\overline{\mathrm{FDP}}_{\hat{k}_{\hat{\alpha}}} \lesssim \hat{\alpha}(1 + o(1))$ in 'good' cases.

In the third setting, an online FDR controlling procedure provides in itself a sequence $(R_k, k \geq 1)$ and not a single set $R_{\hat{k}_\alpha}$. As a consequence, a confidence envelope $(\overline{\mathrm{FDP}}_k, k \geq 1)$ is defined specifically for each procedure $(R_k, k \geq 1)$. Hence, the definition should be slightly adapted:

Definition 3.1.2 (Consistency for online setting) *Let $\delta \in (0,1)$ be fixed and $\mathbf{P}$ be an online multiple testing model over the infinite hypothesis set $\mathcal{H} = \{1, 2, \dots\}$. Let $(R_k, k \geq 1)$ be an (online) FDR controlling procedure at level α, i.e. such that $\sup_{k \geq 1} \mathbb{E}\left[\mathrm{FDP}(R_k)\right] \leq \alpha$, and $(\overline{\mathrm{FDP}}_k, k \geq 1)$ be a corresponding confidence envelope at level $1-\delta$, i.e., satisfying (3.1). Then $(\overline{\mathrm{FDP}}_k, k \geq 1)$ is said to be consistent for the model $\mathbf{P}$ if for all $\epsilon > 0$,*

$$\lim_{k \to \infty} \mathbf{P}\left(\overline{\mathrm{FDP}}_k - \alpha \geq \epsilon\right) = 0. \tag{3.5}$$

Note that both in (3.1) and (3.5) no uniformity w.r.t. the level α is imposed in the online setting.

3.1.4 Contributions

Our findings are as follows:

- In each of the considered settings (top-k, pre-ordered, online), we provide new (non-asymptotical) FDP confidence envelopes that are consistent under some mild conditions, in-

cluding sparse configurations, see Proposition 3.2.2 (top-k), Proposition 3.3.1 (pre-ordered) and Proposition 3.4.1 (online). Table 3.1 provides a summary of the considered procedures in the different contexts, including the existing and new ones. It is worth noting that in the top-k setting, the envelope based on the DKW inequality (Massart, 1990) is consistent under moderate sparsity assumptions only, while the new envelope based on the Wellner inequality (Shorack and Wellner, 2009) covers all the sparsity range (Proposition 3.2.2).

- As a byproduct, our results provide (non-asymptotical) confidence bounds on the FDP for standard FDR-controlling procedures which are asymptotically sharp (consistency) and for which a data-driven choice of the level α is allowed. In particular, in the top-k setting, this gives a new sharp confidence bound for the achieved FDP of the BH procedure while tuning the level from the same data, see (3.18) below.
- In the top-k setting, we also develop *adaptive* envelopes, for which the proportion of null hypotheses is simultaneously estimated, see Section 3.2.5. This is a novel approach with respect to existing literature and it is shown to improve significantly the bounds on simulations in 'dense' situations, see Section 3.5.
- In the pre-ordered setting, including the 'knockoff' case, we introduce new envelopes, called 'Freedman' and 'KR-U', which are the two first (provably-)consistent confidence bounds in that context to our knowledge. This is an important contribution since the knockoff method is one of the leading methodology in the literature of the last decade. In addition, KR-U is shown to behave suitably, even for moderate sample size, see Section 3.5.
- Our study is based on dedicated tools of independent interest, based on uniform versions of classical deviation inequalities, see Corollary 3.2.1 (Wellner's inequality), Corollary B.3.2 (Freedman's inequality). Both can be seen as a form of 'stitching' together elementary inequalities, see Howard et al. (2021) for recent developments of this principle. The bounds developed here are presented in a self-contained manner.

	Simes	DKW	KR	Wellner (new)	Freedman (new)	KR-U (new)
Top-k	No	Yes	No	Yes		
Pre-ordered			No		Yes	Yes
Online			No		Yes	Yes

Table 3.1: Consistency property (Yes or No) for different envelopes, depending on the considered contexts. 'Consistent' means consistent at least in a particular (reasonable) configuration. Unfilled means undefined in that context.

3.2 Results in the top-k case

3.2.1 Top-k setting

We consider the classical multiple setting where we observe m independent p-values $p_1, \dots, p_m$, testing m null hypotheses $H_1, \dots, H_m$. The set of true nulls is denoted by $\mathcal{H}_0$, which is of cardinal m_0 and we denote $\pi_0 = m_0/m \in (0, 1)$. We assume that the p-values are uniformly distributed under the null, that is, for all $i \in \mathcal{H}_0$, $p_i \sim U(0, 1)$.

We consider here the task of building a $(1 - \delta)$-confidence envelope (3.1) for the top-k path

$$R_k = \{1 \le i \le m \, : \, p_i \le p_{(k)}\}, \quad k = 1, \dots, m. \tag{3.6}$$

A rejection set of particular interest is the BH rejection set, given by $R_{\hat{k}_\alpha}$ where

$$\hat{k}_\alpha = \max\left\{k \in \mathbb{N} \,:\, \widehat{\mathrm{FDP}}_k \leq \alpha\right\},\ \widehat{\mathrm{FDP}}_k = mp_k/k, \tag{3.7}$$

(with the convention $R_0 = \emptyset$).

3.2.2 Existing envelopes

Let us first review the prominent confidence envelopes that have been considered in the literature. Let $U_1, \ldots, U_n$ be $n \geq 1$ i.i.d. uniform random variables. For $\delta \in (0,1)$, each of the following (uniform) inequalities holds with probability at least $1-\delta$:

- Simes (or Robbins, 1954): for all $t \in (0,1)$, $n^{-1}\sum_{i=1}^n \mathbf{1}\{U_i \leq t\} \leq t/\delta$.
- DKW (Massart, 1990): for all $t \in (0,1)$, $n^{-1}\sum_{i=1}^n \mathbf{1}\{U_i \leq t\} \leq t + \sqrt{\log(1/\delta)/2}\, n^{-1/2}$.
- KR (Katsevich and Ramdas, 2020) (for $\delta \leq 0.31$), for all $t \in (0,1)$, $n^{-1}\sum_{i=1}^n \mathbf{1}\{U_i \leq t\} \leq \frac{\log(1/\delta)}{\log(1+\log(1/\delta))}(1/n + t)$.

Taking $(U_1, \ldots, U_n) = (p_i, i \in \mathcal{H}_0)$, $n = m_0$, and $t = p_{(k)}$ in the bounds above gives the following confidence envelopes (in the sense of (3.1)) for the top-k path: for $k \in \{1, \ldots, m\}$,

$$\overline{\mathrm{FDP}}_k^{\mathrm{Simes}} = 1 \wedge \frac{mp_{(k)}}{k\delta}; \tag{3.8}$$

$$\overline{\mathrm{FDP}}_k^{\mathrm{DKW}} = 1 \wedge \left(\frac{mp_{(k)}}{k} + \frac{m^{1/2}\sqrt{0.5\log 1/\delta}}{k}\right); \tag{3.9}$$

$$\overline{\mathrm{FDP}}_k^{\mathrm{KR}} = 1 \wedge \left(\frac{\log(1/\delta)}{\log(1+\log(1/\delta))}\left(\frac{mp_{(k)}}{k} + 1/k\right)\right), \tag{3.10}$$

the last inequality requiring in addition $\delta \leq 0.31$. Please note that we can slightly improve these bounds by taking appropriate integer parts, but we will ignore this detail further on for the sake of simplicity.

3.2.3 New envelope

In addition to the above envelopes, this section presents a new one deduced from a new 'uniform' variation of Wellner's inequality (recalled in Lemma B.4.2). Let us first define the function

$$h(\lambda) = \lambda(\log\lambda - 1) + 1, \quad \lambda > 1. \tag{3.11}$$

Lemma B.4.1 gathers some properties of h, including explicit accurate bounds for h and h^{-1}.

Proposition 3.2.1 (Uniform version of Wellner's inequality) *Let $U_1, \ldots, U_n$ be $n \geq 1$ i.i.d. uniform random variables and $\kappa = \pi^2/6$. For all $\delta \in (0,1)$, we have with probability at least $1-\delta$,*

$$\forall t \in (0,1),\ n^{-1}\sum_{i=1}^n \mathbf{1}\{U_i \leq t\} \leq t\, h^{-1}\left(\frac{\log(\kappa/\delta) + 2\log\left(\lceil \log_2(1/t)\rceil\right)}{ng(t)}\right), \tag{3.12}$$

for $g(t) = 2^{-\lceil \log_2(1/t) \rceil}/(1 - 2^{-\lceil \log_2(1/t) \rceil}) \geq t/2$ *and* $h(\cdot)$ *defined by* (3.11)*. In particular, with probability at least* $1 - \delta$*,*

$$\forall t \in (0,1),\ \ n^{-1} \sum_{i=1}^{n} \mathbf{1}\{U_i \leq t\} \leq t\, h^{-1}\left(\frac{2\log(\kappa/\delta) + 4\log\left(1 + \log_2(1/t)\right)}{nt} \right). \tag{3.13}$$

The proof of Proposition 3.2.1 is given in Section B.2.1. It immediately leads to the following result.

Theorem 3.2.1 *In the top-k setting of Section 3.2.1, the following quantity is a* $(1-\delta)$*-confidence envelope in the sense of* (3.1) *for the top-k path:*

$$\overline{\mathrm{FDP}}_k^{Well} = 1 \wedge \left(\frac{m p_{(k)}}{k}\, h^{-1}\left(\frac{2\log(\kappa/\delta) + 4\log\left(1 + \log_2(1/p_{(k)})\right)}{m p_{(k)}} \right) \right), \tag{3.14}$$

with $\kappa = \pi^2/6$*.*

Proof 3.2.1 *We use* (3.13) *for* $(U_1, \ldots, U_n) = (p_i, i \in \mathcal{H}_0)$*,* $n = m_0$*, and* $t = p_{(k)}$*. We conclude by using* $m_0 \leq m$ *and the monotonicity property of Lemma B.4.1.*

Remark 3.2.1 *Denoting by* $\overline{F}_n(t)$ *the RHS of* (3.13)*, we can easily check*

$$\sup_{t \in ((\log\log n)/n, 1)} \left(\sqrt{n} \frac{\overline{F}_n(t) - t}{\sqrt{t \log\left(1 + \log_2(1/t)\right)}} \right) = O(1),$$

with a constant possibly depending on δ*. The iterated logarithm in the denominator is known from classical asymptotic theory (convergence to a Brownian bridge) to be unimprovable for a uniform bound in the vicinity of* 0*; in this sense the above is a 'finite law of the iterated logarithm (LIL) bound' (Jamieson et al., 2014).*

3.2.4 FDP confidence bounds for BH and consistency

Applying the previous bounds for the particular BH rejection sets $R_{\hat{k}_\alpha}$ (see (3.7)) leads to the following result.

Corollary 3.2.1 *In the top-k setting of Section 3.2.1, for any* $\alpha, \delta \in (0,1)$*, the following quantities are* $(1-\delta)$*-confidence bounds for* $\mathrm{FDP}(R_{\hat{k}_\alpha})$*, the FDP of the BH procedure at level* α*:*

$$\overline{\mathrm{FDP}}_\alpha^{Simes} = 1 \wedge (\alpha/\delta); \tag{3.15}$$

$$\overline{\mathrm{FDP}}_\alpha^{DKW} = 1 \wedge \left(\alpha + \frac{m^{1/2}\sqrt{0.5 \log 1/\delta}}{1 \vee \hat{k}_\alpha} \right); \tag{3.16}$$

$$\overline{\mathrm{FDP}}_\alpha^{KR} = 1 \wedge \left(\frac{\log(1/\delta)}{\log(1 + \log(1/\delta))} \left(\alpha + 1/(1 \vee \hat{k}_\alpha) \right) \right); \tag{3.17}$$

$$\overline{\mathrm{FDP}}_\alpha^{Well} = 1 \wedge \left(\alpha\, h^{-1}\left(\frac{2\log(\kappa/\delta) + 4\log\left(1 + \log_2\left(\frac{m}{\alpha(1 \vee \hat{k}_\alpha)}\right)\right)}{\alpha(1 \vee \hat{k}_\alpha)} \right) \right), \tag{3.18}$$

where $\kappa = \pi^2/6$, $\hat{k}_\alpha$ denotes the number of rejections of the BH procedure (3.7) *at level α, and where the KR bound requires in addition $\delta \leq 0.31$. Moreover, these bounds are also valid uniformly in $\alpha \in (0,1)$, in the sense that*

$$\mathbf{P}(\forall \alpha \in (0,1), \mathrm{FDP}(R_{\hat{k}_\alpha}) \leq \overline{\mathrm{FDP}}_\alpha^{Meth}) \geq 1-\delta, \;\; Meth \in \{Simes, DKW, KR, Well\},$$

and thus also when using a post hoc choice $\alpha = \hat{\alpha}$ of the level.

Proof 3.2.2 *For* (3.18)*, we use* (3.13) *for $(U_1, \ldots, U_n) = (p_i, i \in \mathcal{H}_0)$, $n = m_0$, and $t = \alpha(1 \vee \hat{k}_\alpha)/m$.*

Let us now consider the consistency property (3.3). Among the four above bounds, it is apparent that Simes and KR are never consistent, because of the constant in front of α; namely, for all m,

$$\overline{\mathrm{FDP}}_\alpha^{\mathrm{Simes}} \wedge \overline{\mathrm{FDP}}_\alpha^{\mathrm{KR}} \geq 1 \wedge (c\alpha),$$

for some constant $c > 1$. By contrast, $\overline{\mathrm{FDP}}_\alpha^{\mathrm{DKW}}$ and $\overline{\mathrm{FDP}}_\alpha^{\mathrm{Well}}$ are consistent in the sense of (3.3) in a regime such that $m^{1/2}/\hat{k}_{\alpha_0} = o_P(1)$ and $(\log\log m)/\hat{k}_{\alpha_0} = o_P(1)$, respectively. The latter means that the BH procedure at level α_0 should make enough rejections. This is discussed for a particular setting in the next result.

Proposition 3.2.2 *Let us consider the sequence of sparse one-sided Gaussian location models $(\mathbf{P}_{b,c,\beta}^{(m)}, m \geq 1)$ with fixed parameters $b \in \mathbb{R}$, $c \in (0,1)$ and a sparsity parameter $\beta \in [0,1)$, as defined in Section B.1.1. Then we have for all $\alpha \in (0,1)$,*

$$\overline{\mathrm{FDP}}_\alpha^{DKW} - \alpha \asymp_{\mathbf{P}_{b,c,\beta}^{(m)}} m^{-1/2+\beta};$$
$$\overline{\mathrm{FDP}}_\alpha^{Well} - \alpha \asymp_{\mathbf{P}_{b,c,\beta}^{(m)}} \sqrt{\log\log(m)}\, m^{-1/2+\beta/2},$$

where $u_m \asymp_P v_m$ stands for $u_m = O_P(v_m)$ and $v_m = O_P(u_m)$. In particular, concerning the consistency (3.3) *for the sequence $(\mathbf{P}_{b,c,\beta}^{(m)}, m \geq 1)$ and the BH procedure:*

- *for the DKW envelope* (3.9) *and the corresponding bound* (3.16)*, the consistency* (3.3) *holds when $\beta < 1/2$ but fails for $\beta \geq 1/2$;*

- *for the Wellner envelope* (3.14) *and the corresponding bound* (3.18)*, the consistency* (3.3) *holds for any arbitrary $\beta \in (0,1)$.*

Proof 3.2.3 *By Theorem B.1.1, we have $\hat{k}_\alpha \asymp_{\mathbf{P}_{b,c,\beta}^{(m)}} m^{1-\beta}$. This gives the result (by applying in addition Lemma B.4.1 for the Wellner bound).*

Proposition 3.2.2 shows the superiority of the Wellner bound on the DKW bound for achieving the consistency property on a particular sparse sequence models: while the DKW bound needs a model dense enough ($\beta < 1/2$), the Wellner bound covers the whole sparsity range $\beta \in (0,1)$.

3.2.5 Adaptive envelopes

Let us consider the following upper-bounds for m_0:

$$\hat{m}_0^{\text{Simes}} = m \wedge \inf_{t\in(0,\delta)} \frac{V_t}{1-t/\delta}; \tag{3.19}$$

$$\hat{m}_0^{\text{DKW}} = m \wedge \inf_{t\in(0,1)} \left(\frac{C^{1/2}}{2(1-t)} + \sqrt{\frac{C}{4(1-t)^2} + \frac{V_t}{1-t}} \right)^2; \tag{3.20}$$

$$\hat{m}_0^{\text{KR}} = m \wedge \inf_{t\in(0,1/C')} \frac{C' + V_t}{1 - C't}; \tag{3.21}$$

$$\hat{m}_0^{\text{Well}} = m \wedge \inf_{t\in(0,1)} \left(\sqrt{\frac{tC_t}{2(1-t)^2}} + \sqrt{\frac{C_t}{2(1-t)^2} + \frac{V_t}{1-t}} \right)^2, \tag{3.22}$$

where $V_t = \sum_{i=1}^m \mathbf{1}\{p_i > t\}$, $C = \log(1/\delta)/2$, $C' = \frac{\log(1/\delta)}{\log(1+\log(1/\delta))}$, $C_t = 2\log(\kappa/\delta) + 4\log\left(1+\log_2(1/t)\right)$, $\kappa = \pi^2/6$. Since $V_t/(1-t)$ corresponds to the so-called Storey estimator Storey et al. (2004), these four estimators can all be seen as Storey-type confidence bounds, each including a specific deviation term that takes into account the probability error δ. Note that $\hat{m}_0^{\text{DKW}}$ was already proposed in Durand et al. (2020).

Proposition 3.2.3 *In the top-k setting of Section 3.2.1, the envelopes defined by* (3.8), (3.9), (3.10) *and* (3.14) *with* m *replaced by the corresponding bound* $\hat{m}_0^{Simes}$ (3.19), $\hat{m}_0^{DKW}$ (3.20), $\hat{m}_0^{KR}$ (3.21) *or* $\hat{m}_0^{Well}$ (3.22)*, respectively, are also* $(1-\delta)$*-confidence envelopes in the sense of* (3.1) *for the top-k path.*

We can easily check that these four adaptive envelopes all uniformly improve their own non-adaptive counterpart. The proof of Proposition 3.2.3 is provided in Section B.2.2.

Remark 3.2.2 *In practice, the bounds* $\hat{m}_0^{Simes}$ (3.19), $\hat{m}_0^{DKW}$ (3.20), $\hat{m}_0^{KR}$ (3.21) *or* $\hat{m}_0^{Well}$ (3.22) *can be computed by taking an infimum over* $t = p_{(k)}$, $1 \le k \le m$ *and by replacing* V_t *by* $m-k$.

Applying Proposition 3.2.3 for the BH procedure, this gives rise to the following adaptive confidence bounds.

Corollary 3.2.2 *In the top-k setting of Section 3.2.1, for any* $\alpha, \delta \in (0,1)$*, the following quantities are* $(1-\delta)$*-confidence bounds for the FDP of the BH procedure at level* α*:*

$$\overline{\text{FDP}}_\alpha^{Simes\text{-}adapt} = 1 \wedge \alpha(\hat{m}_0^{Simes}/m)/\delta; \tag{3.23}$$

$$\overline{\text{FDP}}_\alpha^{DKW\text{-}adapt} = 1 \wedge \left(\alpha(\hat{m}_0^{DKW}/m) + \frac{(\hat{m}_0^{DKW})^{1/2}\sqrt{0.5\log 1/\delta}}{1 \vee \hat{k}_\alpha} \right); \tag{3.24}$$

$$\overline{\text{FDP}}_\alpha^{KR\text{-}adapt} = 1 \wedge \left(\frac{\log(1/\delta)}{\log(1+\log(1/\delta))} \left(\alpha(\hat{m}_0^{KR}/m) + 1/(1 \vee \hat{k}_\alpha) \right) \right); \tag{3.25}$$

$$\overline{\text{FDP}}_\alpha^{Well\text{-}adapt} = 1 \wedge \left(\alpha(\hat{m}_0^{Well}/m)\, h^{-1} \left(\frac{2\log(\kappa/\delta) + 4\log\left(1+\log_2\left(\frac{m}{\alpha(1\vee\hat{k}_\alpha)}\right)\right)}{\alpha(1 \vee \hat{k}_\alpha)\hat{m}_0^{Well}/m} \right) \right), \tag{3.26}$$

where $\kappa = \pi^2/6$, $\hat{k}_\alpha$ *denotes the number of rejections of BH procedure* (3.7) *at level* α*, and where the KR-adapt bound requires in addition* $\delta \le 0.31$*. Moreover, these bounds are also valid uniformly in* $\alpha \in (0,1)$ *and thus also when using a post hoc choice* $\alpha = \hat{\alpha}$ *of the level.*

Proof 3.2.4 *For* (3.26)*, we use* (3.13) *for* $(U_1, \dots, U_n) = (p_i, i \in \mathcal{H}_0)$*,* $n = m_0$*,* $t = \alpha(1 \vee \hat{k}_\alpha)/m$*, and the fact that* $m_0 \leq \hat{m}_0^{Well}$ *on the considered event by the proof in Section B.2.2. The other bounds are proved similarly.*

3.2.6 Interpolated bounds

According to Remark 3.1.1, the coverage (3.1) is still valid after the interpolation operation given by (3.2). As a result, the above confidence envelopes can be improved as follows:

$$\widetilde{\mathrm{FDP}}_k^{\mathrm{Simes}} = \min_{k' \leq k} \{k - k' + k' \wedge (mp_{(k')}/\delta)\}/k; \tag{3.27}$$

$$\widetilde{\mathrm{FDP}}_k^{\mathrm{DKW}} = \min_{k' \leq k} \{k - k' + k' \wedge (mp_{(k')} + m^{1/2}\sqrt{0.5 \log 1/\delta})\}/k; \tag{3.28}$$

$$\widetilde{\mathrm{FDP}}_k^{\mathrm{KR}} = \min_{k' \leq k} \left\{ k - k' + k' \wedge \left(\frac{\log(1/\delta)}{\log(1 + \log(1/\delta))} \left(mp_{(k')} + 1 \right) \right) \right\} /k; \tag{3.29}$$

$$\widetilde{\mathrm{FDP}}_k^{\mathrm{Well}} = \min_{k' \leq k} \left\{ k - k' + k' \wedge \left(mp_{(k')}\, h^{-1} \left(\frac{2\log(\kappa/\delta) + 4\log\left(1 + \log_2(1/p_{(k')})\right)}{mp_{(k')}} \right) \right) \right\} /k, \tag{3.30}$$

respectively. When applied to BH rejection set, this also provides new confidence bounds $\widetilde{\mathrm{FDP}}_\alpha^{\mathrm{Simes}}$, $\widetilde{\mathrm{FDP}}_\alpha^{\mathrm{DKW}}$, $\widetilde{\mathrm{FDP}}_\alpha^{\mathrm{KR}}$, $\widetilde{\mathrm{FDP}}_\alpha^{\mathrm{Well}}$, that can further be improved by replacing m by the corresponding estimator $\hat{m}_0$.

3.3 Results in the pre-ordered case

In this section, we build consistent envelopes in the case where the p-values are ordered a priori, which covers the famous 'knockoff' case.

3.3.1 Pre-ordered setting

Let $\pi : \{1, \dots, m\} \to \{1, \dots, m\}$ be some ordering of the p-values that is considered as given and deterministic (possibly coming from independent data). The pre-ordered setting is formally the same as the one of Section 3.2.1, except that the p-value set is explored according to $\pi(1), \pi(2), \dots, \pi(m)$. The rationale behind this is that the alternative null hypotheses $\mathcal{H}_1 = \{1, \dots, m\} \backslash \mathcal{H}_0$ are implicitly expected to be more likely to have a small rank in the ordering π (although this condition is not needed for the future controlling results to hold).

Formally, the considered path is

$$R_k = \{\pi(i) \,:\, 1 \leq i \leq k,\ p_{\pi(i)} \leq s\}, \quad k = 1, \dots, m, \tag{3.31}$$

for some fixed additional threshold $s \in (0, 1]$ (possibly coming from independent data) and can serve to make a selection. The aim is still to find envelopes $(\overline{\mathrm{FDP}}_k)_k$ satisfying (3.1) for this path while being consistent. To set up properly the consistency, we should consider an FDR controlling procedure that is suitable in this setting. For this, we consider the Lei Fithian (LF) adaptive Selective sequential step-up procedure (Lei and Fithian, 2016). The latter is defined by $R_{\hat{k}_\alpha}$ where

$$\hat{k}_\alpha = \max \left\{ k \in \{0, \dots, m\} : \widehat{\mathrm{FDP}}_k \leq \alpha \right\}, \ \widehat{\mathrm{FDP}}_k = \frac{s}{1 - \lambda} \frac{1 + \sum_{i=1}^k \mathbf{1}\{p_{\pi(i)} > \lambda\}}{1 \vee \sum_{i=1}^k \mathbf{1}\{p_{\pi(i)} \leq s\}}, \tag{3.32}$$

where $\lambda \in [0,1)$ is an additional parameter. The 'knockoff' setting of Barber and Candès (2015) can be seen as a particular case of this pre-ordered setting, where the p-values are independent and binary, the ordering is independent of the p-values and $s = \lambda = 1/2$. The LF procedure reduces in that case to the classical Barber and Candès (BC) procedure.

3.3.2 New confidence envelopes

The first envelope is as follows.

Theorem 3.3.1 *Consider the pre-ordered setting of Section 3.3.1 with $s \in (0,1]$. For all $\delta \in (0,1)$, $\lambda \in [0,1)$, the following is a $(1-\delta)$-confidence envelope for the ordered path* (3.31) *in the sense of* (3.1)*:*

$$\overline{\mathrm{FDP}}_k^{Freed} = 1 \wedge \frac{\frac{s}{1-\lambda}\sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} > \lambda\} + \Delta(\nu k)}{\sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} \le s\}},\ k \ge 1, \tag{3.33}$$

where $\Delta(u) = 2\sqrt{\varepsilon_u}\sqrt{(u \vee 1)} + \frac{1}{2}\varepsilon_u$, $\varepsilon_u = \log((1+\kappa)/\delta) + 2\log\left(1 + \log_2(u \vee 1)\right)$, $u > 0$, $\kappa = \pi^2/6$ *and* $\nu = s(1 + \min(s,\lambda)/(1-\lambda))$.

The proof of Theorem 3.3.1 is a direct consequence of a more general result (Theorem B.3.1), itself being a consequence of a uniform version of Freedman's inequality (see Section B.3.2).

The second result comes from the KR envelope (Katsevich and Ramdas, 2020):

$$\overline{\mathrm{FDP}}_k^{\mathrm{KR}} = 1 \wedge \left(\frac{\log(1/\delta)}{a \log(1 + \frac{1-\delta^{B/a}}{B})} \frac{a + \frac{s}{1-\lambda}\sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} > \lambda\}}{1 \vee \sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} \le s\}} \right), \tag{3.34}$$

where $a > 0$ is some parameter, $B = s/(1-\lambda)$ and it is assumed $\lambda \ge s$. While the default choice in KR is $a = 1$, we can build up a new envelope by taking a union bound over $a \in \mathbb{N}\backslash\{0\}$:

Theorem 3.3.2 *Consider the pre-ordered setting of Section 3.3.1 with $s \in (0,1]$. For all $\delta \in (0,1)$ and $\lambda \in [s,1]$, the following is a $(1-\delta)$-confidence envelope for the ordered path* (3.31) *in the sense of* (3.1)*:*

$$\overline{\mathrm{FDP}}_k^{KR\text{-}U} = 1 \wedge \min_{a \in \mathbb{N}\backslash\{0\}} \left\{ \frac{\log(1/\delta_a)}{a \log(1 + \frac{1-\delta_a^{B/a}}{B})} \frac{a + \frac{s}{1-\lambda}\sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} > \lambda\}}{1 \vee \sum_{i=1}^{k} \mathbf{1}\{p_{\pi(i)} \le s\}} \right\},\ k \ge 1, \tag{3.35}$$

for $\delta_a = \delta/(\kappa a^2)$, $a \ge 1$, for $B = s/(1-\lambda)$, $\kappa = \pi^2/6$.

The envelope (3.35) is less explicit than (3.33) but has a better behavior in practice, as we will see in the numerical experiments of Section 3.5.

3.3.3 Confidence bounds for LF and consistency

Recall that the LF procedure (3.32) is the reference FDR-controlling procedure in this setting. Applying the above envelopes for the LF procedure gives the following confidence bounds.

Corollary 3.3.1 *In the pre-ordered setting of Section 3.3.1 with a selection threshold $s \in (0,1]$, for any $\alpha, \delta \in (0,1)$, $\lambda \in [s,1]$ the following quantities are $(1-\delta)$-confidence bounds for the FDP*

of the LF procedure with parameters s, λ at level α:

$$\overline{\mathrm{FDP}}_\alpha^{KR} = 1 \wedge \left(\frac{\log(1/\delta)}{\log(1+\frac{1-\delta^B}{B})} \left(\alpha + 1/(1 \vee \hat{r}_\alpha)\right) \right); \tag{3.36}$$

$$\overline{\mathrm{FDP}}_\alpha^{Freed} = 1 \wedge \left(\alpha + \Delta(\nu \hat{k}_\alpha)/(1 \vee \hat{r}_\alpha) \right) \tag{3.37}$$

$$\overline{\mathrm{FDP}}_\alpha^{KR\text{-}U} = 1 \wedge \min_{1 \le a \le 1 \vee \hat{r}_\alpha} \left\{ \frac{\log(1/\delta_a)}{a \log(1+\frac{1-\delta_a^{B/a}}{B})} \left(\alpha + a/(1 \vee \hat{r}_\alpha)\right) \right\}, \tag{3.38}$$

for $\nu = s(1+s/(1-\lambda))$, $B = s/(1-\lambda)$, $\delta_a = \delta/(\kappa a^2)$, $a \ge 1$, $\kappa = \pi^2/6$, $\Delta(\cdot)$ defined in Theorem 3.3.1 and where $\hat{k}_\alpha$ is as in (3.32) *and $\hat{r}_\alpha = \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_{\pi(i)} \le s\}$ denotes the number of rejections of LF procedure at level α. In addition, these bounds are also valid uniformly in $\alpha \in (0,1)$ in the sense that*

$$\mathbf{P}(\forall \alpha \in (0,1), \mathrm{FDP}(R_{\hat{k}_\alpha}) \le \overline{\mathrm{FDP}}_\alpha^{Meth}) \ge 1-\delta, \quad \textit{for } Meth \in \{KR, Freed, KR\text{-}U\},$$

and thus also when using a post hoc choice $\alpha = \hat{\alpha}$ of the level.

Proof 3.3.1 *This is direct by applying* (3.34) *($a = 1$),* (3.33) *and* (3.35) *to the rejection set $R_{\hat{k}_\alpha}$.*

Let us now study the consistency property (3.3). It is apparent that KR is never consistent: namely, for all $m \ge 1$,

$$\overline{\mathrm{FDP}}_\alpha^{\mathrm{KR}} \ge 1 \wedge c\alpha,$$

for some constant $c > 1$. By contrast, $\overline{\mathrm{FDP}}_\alpha^{\mathrm{Freed}}$ is consistent if $\Delta(\nu m)/\hat{r}_\alpha$ tends to 0 in probability, that is, $(m \log\log m)^{1/2}/\hat{r}_\alpha = o_P(1)$. For $\overline{\mathrm{FDP}}_\alpha^{\mathrm{KR\text{-}U}}$, we always have

$$\overline{\mathrm{FDP}}_\alpha^{\mathrm{KR\text{-}U}} \le \frac{\log(1/\delta_{\hat{a}})}{\hat{a} \log(1+\frac{1-\delta_{\hat{a}}^{B/\hat{a}}}{B})} \left(\alpha + 1/(1 \vee \hat{r}_\alpha)^{1/2}\right)$$

by considering $\hat{a} = \lfloor (1 \vee \hat{r}_\alpha)^{1/2} \rfloor$. By Lemma B.4.3, this provides consistency (3.3) as soon as $1/\hat{r}_\alpha = o_P(1)$. The following proposition gives an example where the latter condition holds in the varying coefficient two-groups (VCT) model of Lei and Fithian (2016), that we generalize to the possible sparse case in Section B.1.2.

Proposition 3.3.1 *Consider the sequence of generalized VCT models $(\mathbf{P}^{(m)}_{\pi,\beta,F_0,F_1}, m \ge 1)$, as defined in Section B.1.2. Assume that the parameters π, β, F_0, F_1 satisfy the assumptions of Theorem B.1.2 given in Appendix B.1.2 (assuming in particular that $\alpha_0 > \underline{\alpha}$ where $\underline{\alpha}$ is defined by* (B.7)*). Then the consistency* (3.3) *holds for the sequence $(\mathbf{P}^{(m)}_{\pi,\beta,F_0,F_1}, m \ge 1)$ and for any LF procedure using $\lambda \ge s$ in either of the two following cases:*

- *for the KR-U envelope* (3.35) *and the corresponding bound* (3.38)*.*
- *for the Freedman envelope* (3.33) *and the corresponding bound* (3.37) *if either $\lambda = s$ or $\beta < 1/2$;*

Proof 3.3.2 *This is a direct consequence of Theorem B.1.2 because $m^{1-\beta}/\hat{r}_\alpha = O_P(1)$ in that context and $\hat{r}_\alpha$ is nondecreasing in α. To see why the Freedman envelope is consistent when $\lambda = s$, we note that in this case $\hat{k}_\alpha = \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_{\pi(i)} \le s\} + \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_{\pi(i)} > \lambda\} \le (1+\alpha s/(1-\lambda))(1 \vee \hat{r}_\alpha)$, hence the quantity $\Delta(\nu \hat{k}_\alpha)/(1 \vee \hat{r}_\alpha)$ is $o_P(1)$ as $1/\hat{r}_\alpha = o_P(1)$.*

We would like to emphasize that the power analysis made in Appendix B.1.2 provides new insights with respect to Lei and Fithian (2016). First, it accommodates the sparse case for which the probability of generating an alternative is tending to zero as m tends to infinity. Second, it introduces a new criticality-type assumption (see (B.7)), which was overlooked in Lei and Fithian (2016), but is necessary to get a non zero power at the limit (even in the dense case). Finally, it allows to deal with binary p-values, which corresponds to the usual 'knockoff' situation.

Remark 3.3.1 *Similarly to Section 3.2.6 in the top-k setting, the bounds KR, Freedman and KR-U can be improved by performing the interpolation operation* (3.2) *in the pre-ordered setting.*

3.4 Results in the online case

3.4.1 Online setting

We consider an infinite stream of p-values $p_1, p_2, \dots$ testing null hypotheses $H_1, H_2, \dots$, respectively. In the online setting, these p-values come one at a time and a decision should be made at each time immediately and irrevocably, possibly on the basis of past decisions.

The decision at time k is to reject H_k if $p_k \leq \alpha_k$ for some critical value α_k only depending on the past decisions. An online procedure is thus defined by a sequence of critical values $\mathcal{A} = (\alpha_k, k \geq 1)$, that is predictable in the following sense

$$\alpha_{k+1} \in \mathcal{G}_k = \sigma(\mathbf{1}\{p_i \leq \alpha_i\}, i \leq k), \;\; k \geq 1.$$

A classical assumption is that each null p-value is super-uniform conditionally on past decisions, that is,

$$\mathbf{P}(p_k \leq x \,|\, \mathcal{G}_k) \leq x, \;\; k \in \mathcal{H}_0, \tag{3.39}$$

where $\mathcal{H}_0 = \{k \geq 1 \mid H_k = 0\}$. Condition (3.39) is for instance satisfied if the p-values are all mutually independent and marginally super-uniform under the null.

For a *fixed* procedure $\mathcal{A}$, we consider the path

$$R_k = \{1 \leq i \leq k \;:\; p_i \leq \alpha_i\}, \;\; k \geq 1. \tag{3.40}$$

We will also denote

$$R(k) = \sum_{i=1}^{k} \mathbf{1}\{p_i \leq \alpha_i\}, \;\; k \geq 1, \tag{3.41}$$

the number of rejections before time k of the considered procedure. A typical procedure controlling the online FDR is the LORD procedure

$$\alpha_k = W_0 \gamma_k + (\alpha - W_0)\gamma_{k-\tau_1} + \alpha \sum_{j \geq 2} \gamma_{k-\tau_j}, \tag{3.42}$$

where $W_0 \in [0, \alpha]$, each τ_j is the first time with j rejections, $(\gamma_k)_k$ is a non-negative ('spending') sequence with $\sum_{k \geq 0} \gamma_k \leq 1$ and $\gamma_k = 0$ for $k < 0$. The latter has been extensively studied in the literature (Foster and Stine, 2008; Aharoni and Rosset, 2014; Javanmard and Montanari, 2018), and further improved by Ramdas et al. (2017). Under independence of the p-values and super-uniformity of the p-values under the null, the LORD procedure controls the online FDR in the sense of

$$\sup_{k \geq 1} \mathbf{E}[\mathrm{FDP}(R_k)] \leq \alpha,$$

see Theorem 2 (b) in Ramdas et al. (2017). Here, we consider the different (and somehow more demanding) task of finding a bound on the realized online FDP, by deriving confidence envelopes (3.1). Note that this will be investigated for any online procedure and not only for LORD, see Section 3.4.2. Also, we will study the consistency of the envelope for any LORD-type procedure in Section 3.4.3.

3.4.2 New confidence envelopes

The first envelope is a consequence of the general result stated in Theorem B.3.1.

Theorem 3.4.1 *In the online setting described in Section 3.4.1, consider any online procedure* $\mathcal{A} = (\alpha_k, k \geq 1)$ *and assume* (3.39). *Then for any* $\delta \in (0,1)$, *the following is a* $(1-\delta)$*-confidence envelope for the path* (3.40) *in the sense of* (3.1)*:*

$$\overline{\mathrm{FDP}}^{Freed}_{\mathcal{A},k} = 1 \wedge \frac{\sum_{i=1}^{k} \alpha_i + \Delta\left(\sum_{i=1}^{k} \alpha_i\right)}{1 \vee R(k)}, \ \ k \geq 1, \tag{3.43}$$

where $R(k)$ *is given by* (3.41), $\Delta(u) = 2\sqrt{\varepsilon_u}\sqrt{u \vee 1} + \frac{1}{2}\varepsilon_u$, $\varepsilon_u = \log((1+\kappa)/\delta) + 2\log\left(1 + \log_2\left(u \vee 1\right)\right)$, $u > 0$ *and* $\kappa = \pi^2/6$.

Proof 3.4.1 *We apply Theorem B.3.1 in the online setting for* $\lambda = 0$ *(and further upper-bounding each term* $\mathbf{1}\{p_{\pi(i)} > 0\}$ *by* 1*),* $\pi(k) = k$, *because* (B.15) *is satisfied by* (3.39).

Next, the envelope of Katsevich and Ramdas (2020) is as follows

$$\overline{\mathrm{FDP}}^{\mathrm{KR}}_{\mathcal{A},k} = 1 \wedge \left(\frac{\log(1/\delta)}{a \log(1 + \log(1/\delta)/a)} \frac{\left(a + \sum_{i=1}^{k} \alpha_i\right)}{1 \vee R(k)} \right), \tag{3.44}$$

for some parameter $a > 0$ to choose. While the default choice in Katsevich and Ramdas (2020) is $a = 1$, applying a union w.r.t. $a \in \mathbb{N}\backslash\{0\}$ provides the following result.

Theorem 3.4.2 *In the online setting described in Section 3.4.1, and for any online procedure* $\mathcal{A} = (\alpha_k, k \geq 1)$, *for any* $\delta \in (0,1)$, *the following is a* $(1-\delta)$*-confidence envelope for the path* (3.40) *in the sense of* (3.1)*:*

$$\overline{\mathrm{FDP}}^{KR\text{-}U}_{\mathcal{A},k} = 1 \wedge \min_{a \in \mathbb{N}\backslash\{0\}} \left\{ \frac{\log(1/\delta_a)}{a \log(1 + \log(1/\delta_a)/a)} \frac{\left(a + \sum_{i=1}^{k} \alpha_i\right)}{1 \vee R(k)} \right\}, \ \ k \geq 1, \tag{3.45}$$

where $R(k)$ *is given by* (3.41), $\delta_a = \delta/(\kappa a^2)$, $a \geq 1$, *for* $\kappa = \pi^2/6$.

Remark 3.4.1 *Note that the guarantee* (3.1) *is not uniform in the procedure* $\mathcal{A}$ *(by contrast with the envelopes in top-k and preordered cases which were uniform in k and thus also in the cut-off procedure).*

3.4.3 Confidence envelope for LORD-type procedures and consistency

We now turn to the special case of online procedures satisfying the following condition:

$$\sum_{i=1}^{k} \alpha_i \leq \alpha(1 \vee R(k)), \quad k \geq 1. \tag{3.46}$$

Classically, this condition is sufficient to control the online FDR (if the p-values are independent and under an additional monotonicity assumption), see Theorem 2 (b) in Ramdas et al. (2017). In particular, it is satisfied by LORD (3.42).

Corollary 3.4.1 *In the online setting described in Section 3.4.1, consider any online procedure $\mathcal{A} = (\alpha_k, k \geq 1)$, satisfying* (3.46) *for some $\alpha \in (0,1)$, and assume* (3.39)*. Then for any $\delta \in (0,1)$, the following quantities are $(1-\delta)$-confidence bounds for the FDP of the procedure: for all $k \geq 1$,*

$$\overline{\mathrm{FDP}}^{KR}_{\alpha,k} = 1 \wedge \left(\frac{\log(1/\delta)}{\log(1+\log(1/\delta))} \left(\alpha + 1/(1 \vee R(k))\right) \right); \tag{3.47}$$

$$\overline{\mathrm{FDP}}^{Freed}_{\alpha,k} = 1 \wedge \left(\alpha + \frac{\Delta\left(\alpha(1 \vee R(k))\right)}{1 \vee R(k)} \right), \ k \geq 1; \tag{3.48}$$

$$\overline{\mathrm{FDP}}^{KR\text{-}U}_{\alpha,k} = 1 \wedge \min_{a \geq 1} \left\{ \frac{\log(1/\delta_a)}{a \log(1+\log(1/\delta_a)/a)} \left(\alpha + a/(1 \vee R(k))\right) \right\}, \tag{3.49}$$

for $\delta_a = \delta/(\kappa a^2)$, $a \geq 1$, $\kappa = \pi^2/6$, $\Delta(\cdot)$ defined in Theorem 3.4.2 and where $R(k)$ is given by (3.41)*.*

Proof 3.4.2 *This is direct by applying* (3.44) *($a = 1$),* (3.48) *and* (3.49) *and by using the inequality* (3.46) *in the corresponding bound.*

Let us now consider these bounds for the LORD procedure (3.42), and study the consistency property (3.5). Clearly, we have $\overline{\mathrm{FDP}}^{\mathrm{KR}}_{\alpha} \geq 1 \wedge (c\alpha)$ for all $k \geq 1$, where $c > 1$ is a constant. Hence, the envelope KR is not consistent. By contrast, it is apparent that both the Freedman envelope and the uniform KR envelope are consistent provided that $1/R(k) = o_P(1)$ as k tends to infinity (consider $a = \sqrt{1 \vee R(k)}$ and use Lemma B.4.3 for the KR-U envelope). This condition is met in classical online models, as the following result shows.

Proposition 3.4.1 *Consider the online one-sided Gaussian mixture model $\mathbf{P}_{\pi_1,F_1}$ of Section B.1.3 and the LORD procedure with $W_0 \in (0,\alpha)$ and a spending sequence $\gamma_k = \frac{1}{k(\log(k))^\gamma}$, $k \geq 1$ for $\gamma > 1$. Then both the Freedman envelope* (3.48) *and the uniform KR envelope* (3.49) *are consistent in the sense of* (3.5) *for the model $\mathbf{P}_{\pi_1,F_1}$.*

Proof 3.4.3 *This is a direct consequence of Theorem B.1.3, which provides that $k^{1/2}/R(k) = O_{\mathbf{P}_{\pi_1,F_1}}(1)$ when k tends to infinity.*

Remark 3.4.2 *Similarly to Section 3.2.6 in the top-k setting, the bounds KR, Freedman and KR-U can be improved by performing the interpolation operation* (3.2) *in the online setting.*

3.5 Numerical experiments

In this section, we illustrate our findings by conducting numerical experiments in each of the considered settings: top-k, pre-ordered and online. Throughout the experiments, the default value for δ is 0.25 and the default number of replications to evaluate each FDP bound is 1000. All our numerical experiments are reproducible from the code provided in the repository https://github.com/iqm15/ConsistentFDP.

3.5.1 Top-k

Here, we consider the top-k setting of Section 3.2.1, for alternative p-values distributed as $F_1(x) = \overline{\Phi}(\overline{\Phi}^{-1}(x) - \mu)$ (one sided Gaussian location model), and for different values of μ and of π_0. To investigate the consistency property, we take m varying in the range $\{10^i, 2 \leq i \leq 6\}$, and we consider the FDP bounds $\overline{\mathrm{FDP}}_\alpha^{\mathrm{Simes}}$ (3.15), $\overline{\mathrm{FDP}}_\alpha^{\mathrm{DKW}}$ (3.16), $\overline{\mathrm{FDP}}_\alpha^{\mathrm{KR}}$ (3.17), $\overline{\mathrm{FDP}}_\alpha^{\mathrm{Well}}$ (3.18) for $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$. We also add for comparison the hybrid bound

$$\overline{\mathrm{FDP}}_{\alpha,\delta}^{\mathrm{Hybrid}} = \min\left(\overline{\mathrm{FDP}}_{\alpha,\delta/2}^{\mathrm{KR}}, \overline{\mathrm{FDP}}_{\alpha,\delta/2}^{\mathrm{Well}}\right),$$

which also provides the correct coverage while being close to the best between the Wellner and KR bounds.

Figure 3.1 displays boxplots of the different FDP bounds in the dense case for which $\pi_0 = 1/2$, $\mu = 1.5$. When m gets large, we clearly see the inconsistency of the bounds Simes, KR and the consistency of the bounds Wellner, Hybrid, DKW, which corroborates the theoretical findings (Proposition 3.2.2). In sparser scenarios, Figure 3.2 shows that the consistency is less obvious for the Wellner and Hybrid bounds and gets violated for the DKW bound when $m_1 \propto m^{0.55}$, as predicted from Proposition 3.2.2 (regime $\beta \geq 1/2$). Overall, the new bounds are expected to be better as the number of rejections gets larger and KR bounds remain better when the number of rejections is expected to be small. The hybrid bound hence might be a good compromise for a practical use.

The adaptive versions of the bounds (Section 3.2.5) are displayed on Figure 3.3. By comparing the left and the right panels, we see that the uniform improvement can be significant, especially for the Wellner and DKW bounds. By contrast, the improvement for KR is slightly worse. This can be explained from Figure 3.4, that evaluates the quality of the different π_0 estimators. DKW, which is close to an optimized Storey-estimator, is the best, followed closely by the Wellner estimator.

Remark 3.5.1 *For clarity, the bounds are displayed without the interpolation improvement* (3.2) *(for top-k and preordered). The figures are reproduced together with the interpolated bounds in Appendix B.5 for completeness. In a nutshell, the interpolation operation improves significantly the bounds mainly when they are not very sharp (typically small m or very sparse scenarios). Hence, while it can be useful in practice, it does not seem particularly relevant to study the consistency phenomenon.*

3.5.2 Pre-ordered

We consider data generated as in the pre-ordered model presented in Section 3.3.1 and more specifically as in the VCT model of Section B.1.2. The trueness/falseness of null hypotheses are generated independently, and the probability of generating an alternative is decreasing with the position $1 \leq k \leq m$, and is given by $\pi(m^{\beta-1}k)$, where $\pi : [0, \infty) \to [0, 1)$ is some function (see below) and $\beta \in [0, 1)$ is the sparsity parameter. Once the process of true/false nulls is given, the p-values are generated according to either:

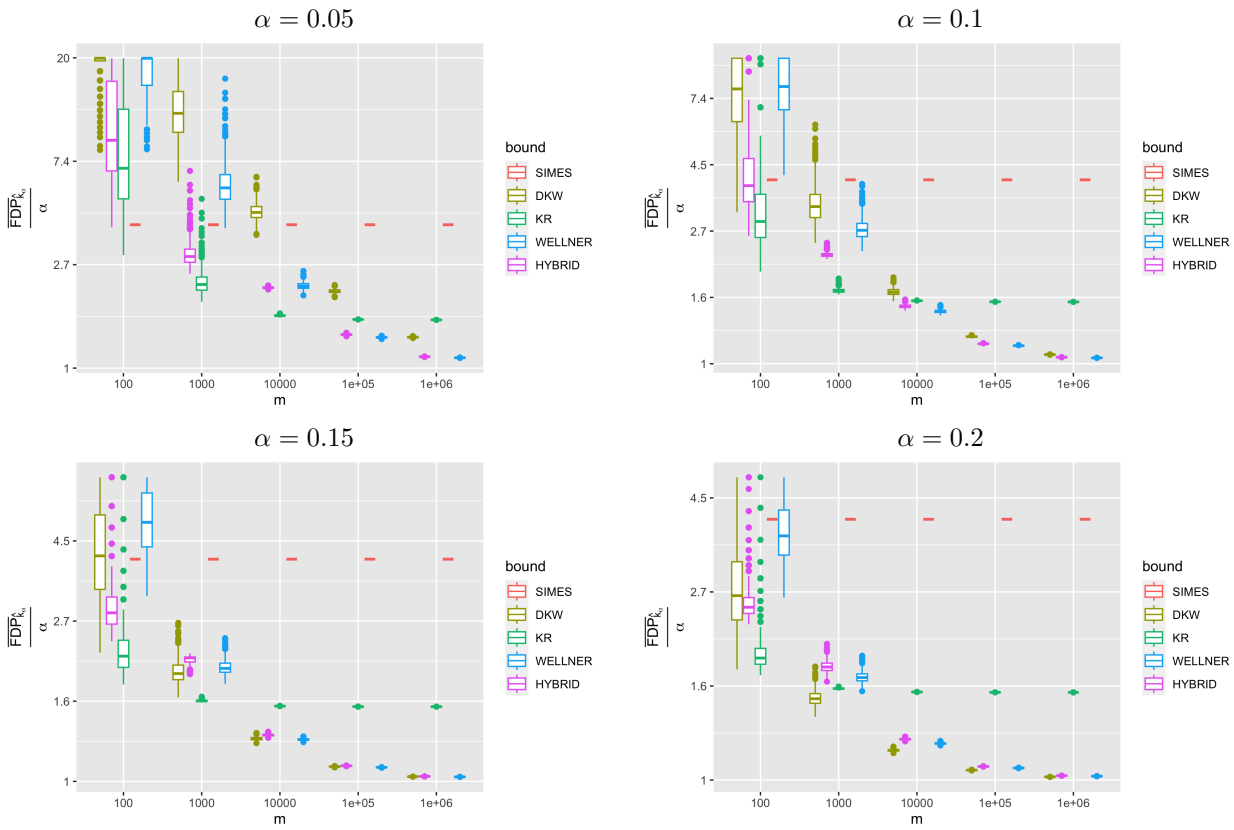

Figure 3.1: Top-k dense case ($\pi_0 = 0.5$, $\mu = 1.5$).

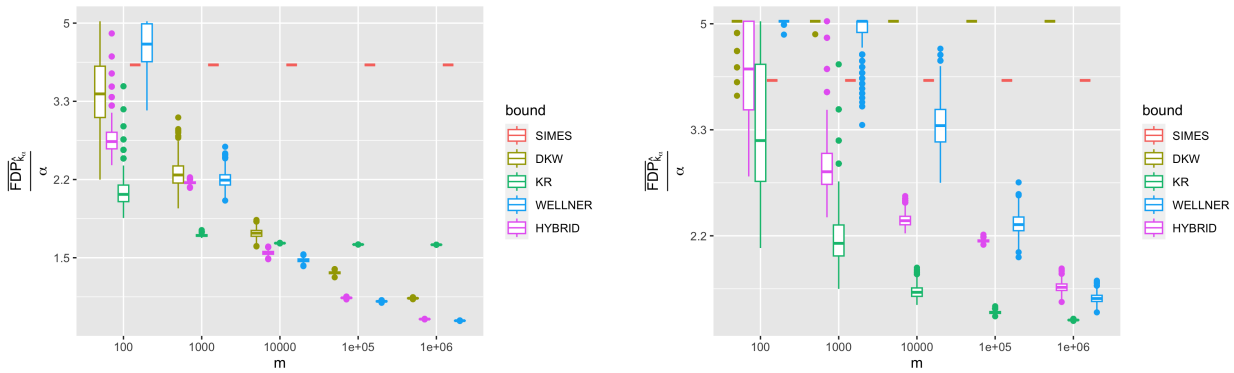

Figure 3.2: Top-k sparse case $\pi_0 = 1 - 0.5m^{-0.25}$, $\mu = \sqrt{2\log(m)}$ (left) $\pi_0 = 1 - 0.5m^{-0.55}$, μ =10 (right), $\alpha = 0.2$.

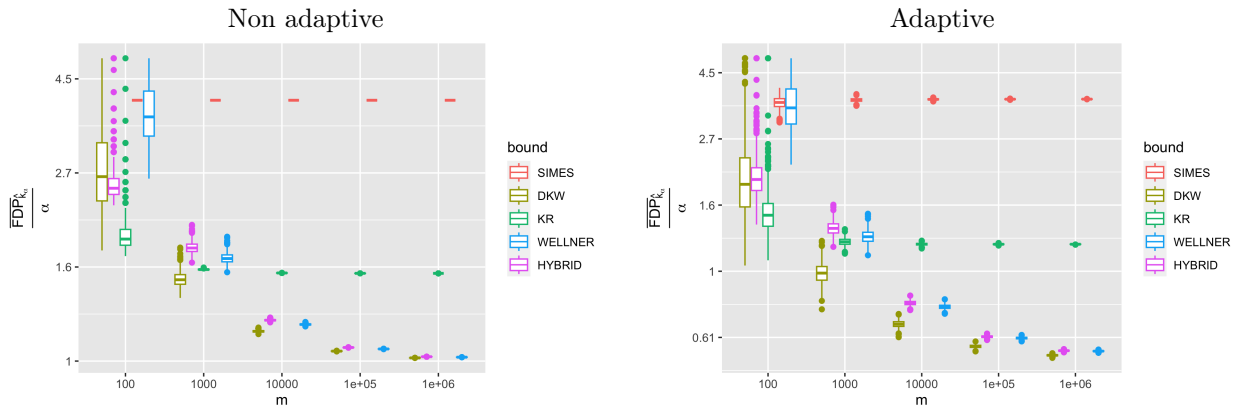

Figure 3.3: Top-k dense case with nonadaptive bounds (left) and adaptive bounds (right) ($\pi_0 = 0.5$, $\alpha = 0.2$).

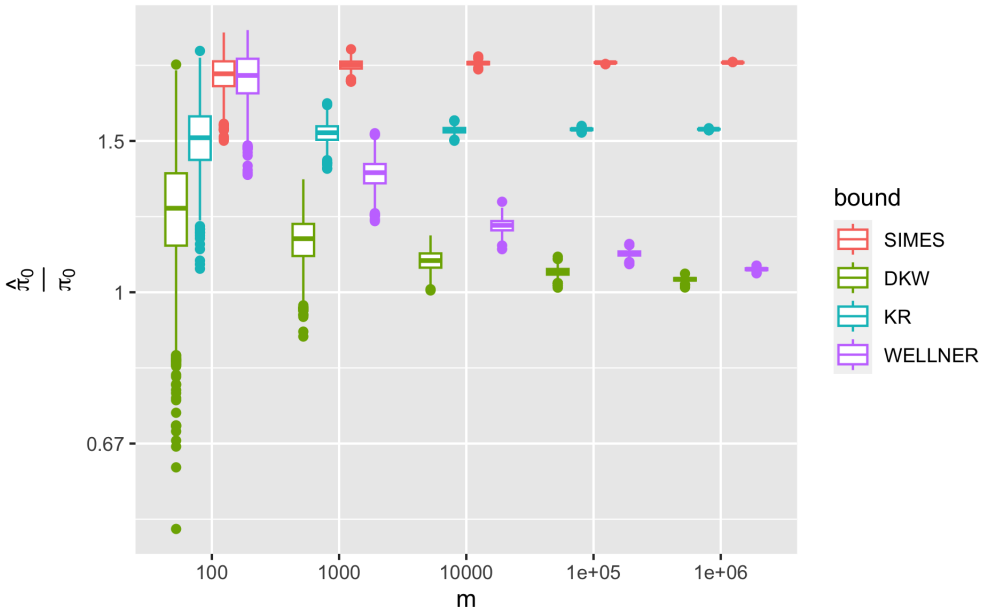

Figure 3.4: Boxplots of the estimators $\hat{\pi}_0$ in the top-k dense case ($\pi_0 = 0.5$, $\alpha = 0.2$).

- LF setting: $\pi(t) = \pi_1 e^{-bt}\frac{b}{1-e^{-b}}$, $t \geq 0$, so that $\Pi(1) = \pi_1$. Here π_1 is equal to 0.4 and b, measuring the quality of the prior ordering, is equal to 2. In addition, the alternative p-values are one-sided Gaussian with $\mu = 1.5$. Note that this is the setting considered in the numerical experiments of Lei and Fithian (2016).

- Knockoff setting: $\pi(t) = 1/2 + (0 \vee 1/2(\frac{z-t}{z-1}))$, $t \geq 0$, with $z > 1$ a parameter that determines how slowly the probability of observing signal deteriorates, taken equal to 30. Then, the binary p-values are as follows: under the null $p_i = 1/2$ or 1 with equal probability. Under the alternative, $p_i = 1/2$ with probability 0.9 and $p_i = 1$ otherwise.

In both settings, the dense (resp. sparse) case refers to the sparsity parameter value $\beta = 0$ (resp. $\beta = 0.25$).

We consider the bounds $\overline{\mathrm{FDP}}_{\alpha}^{\mathrm{KR}}$ (3.36), $\overline{\mathrm{FDP}}_{\alpha}^{\mathrm{Freed}}$ (3.37) and $\overline{\mathrm{FDP}}_{\alpha}^{\mathrm{KR\text{-}U}}$ (3.38) for the LF procedure across different values of $(\lambda, s) \in \{(1/2, 0.1\alpha), (1/2, 1/2)\}$, $m \in \{10^i, 2 \leq i \leq 6\}$, and $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$. The procedure LF with $(\lambda, s) = (1/2, 1/2)$ is referred to as the Barber and Candès (BC) procedure.

Figure 3.5 displays the boxplots of these FDP bounds for the LF procedure with $(\lambda, s) = (1/2, 0.1\alpha)$ in the LF setting with $\beta = 0$ (dense case). It is apparent that KR is not consistent, while the new bounds Freedman and KR-U are. Also, the bound KR-U is overall the best, losing almost nothing w.r.t. KR when the number of rejections is very small (say $m = 100$ and $\alpha = 0.05$ or 0.1) and making a very significant improvement otherwise. Similar conclusions hold for the case of BC procedure, see Figure 3.7. Next, to stick with a very common scenario, we also investigate the sparse situation where the fraction of signal is small in the data, see Figures 3.6 and 3.8. As expected, while the conclusion is qualitatively the same, the rejection number gets smaller so that the consistency is reached for largest values of m (i.e., the convergence is 'slowed down').

3.5.3 Online

We now consider the online case, by applying our method to the real data example coming from the International Mice Phenotyping Consortium (IMPC) (**?**), which is a consortium interested in the genotype effect on the phenotype. This data is collected in an online fashion for each gene of interest and is classically used in online detection works (see Ramdas et al. (2017) and references therein).

Figure 3.9 displays the FDP time-wise envelopes $k \mapsto \overline{\mathrm{FDP}}_{\alpha,k}^{\mathrm{KR}}$ (3.47), $k \mapsto \overline{\mathrm{FDP}}_{\alpha,k}^{\mathrm{Freed}}$ (3.48) and $k \mapsto \overline{\mathrm{FDP}}_{\alpha,k}^{\mathrm{KR\text{-}U}}$ (3.49), for the LORD procedure (3.42) ($W_0 = \alpha/2$ with the spending sequence $\gamma_k = k^{-1.6}$, $k \geq 1$). As we can see, the Freedman and KR-U envelopes both tend to the nominal level α, as opposed to the KR envelope, which is well expected from the consistency theory. In addition, KR-U seems to outperform the Freedman envelope and while KR is (slightly) better than KR-U in the initial segment of the process ($k < 300$), we can see that KR-U gets rapidly more accurate.

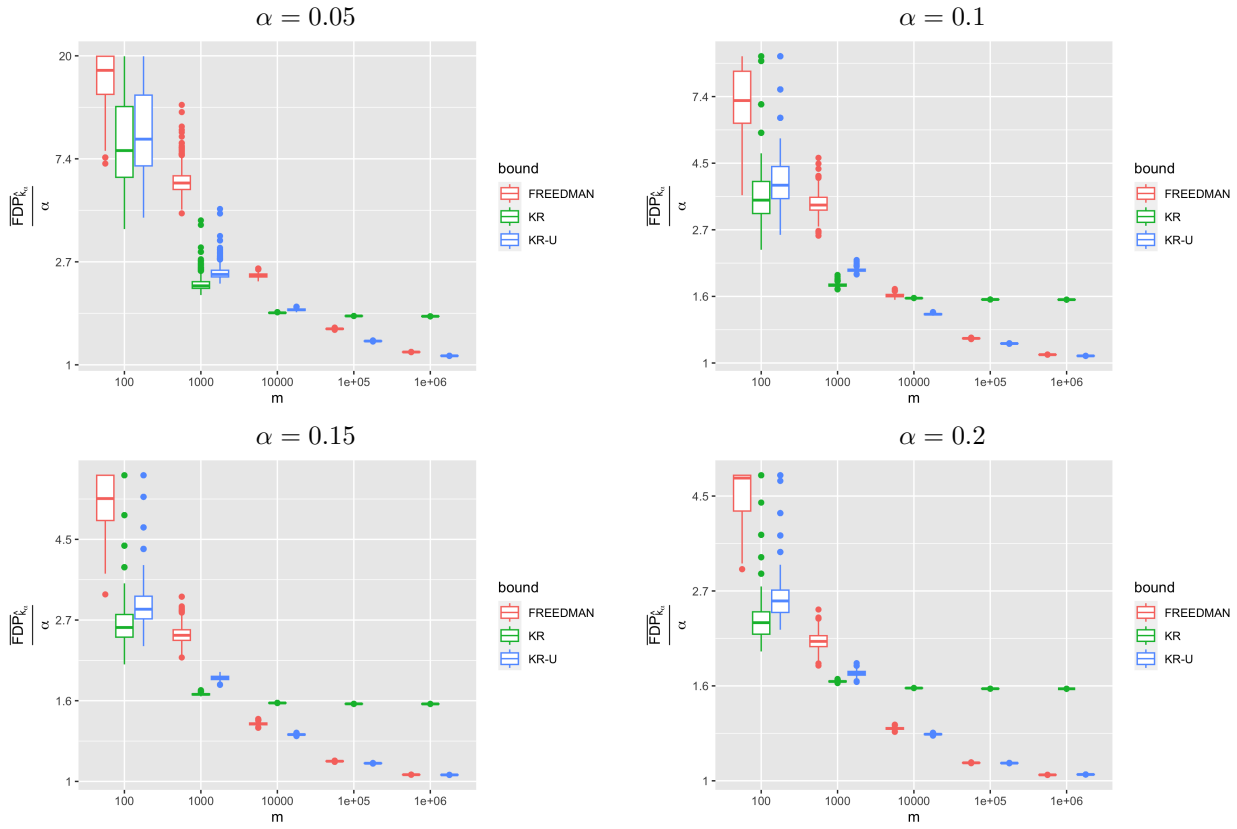

Figure 3.5: Preordered dense ($\beta = 0$) LF setting with LF procedure ($s = 0.1\alpha$, $\lambda = 0.5$).

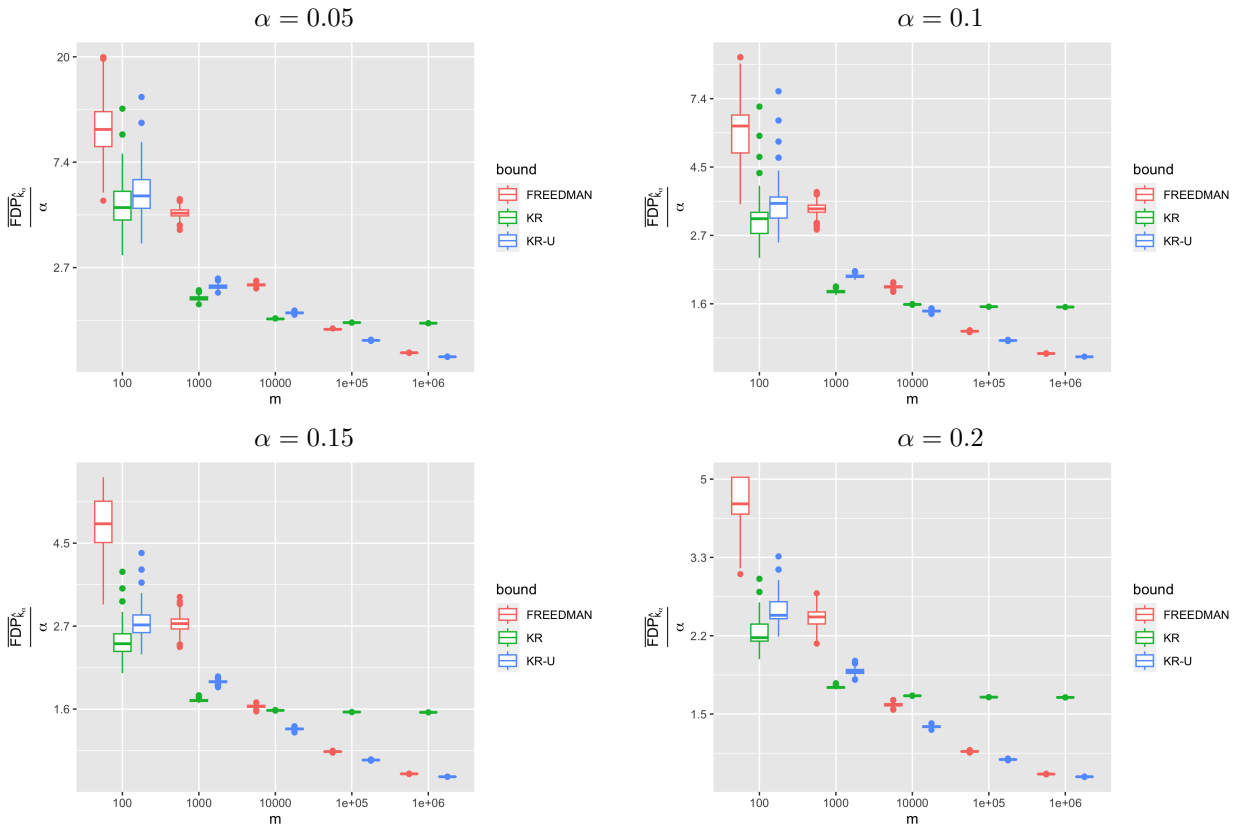

Figure 3.6: Preordered sparse ($\beta = 0.25$) LF setting with LF procedure ($s = 0.1\alpha$, $\lambda = 0.5$).

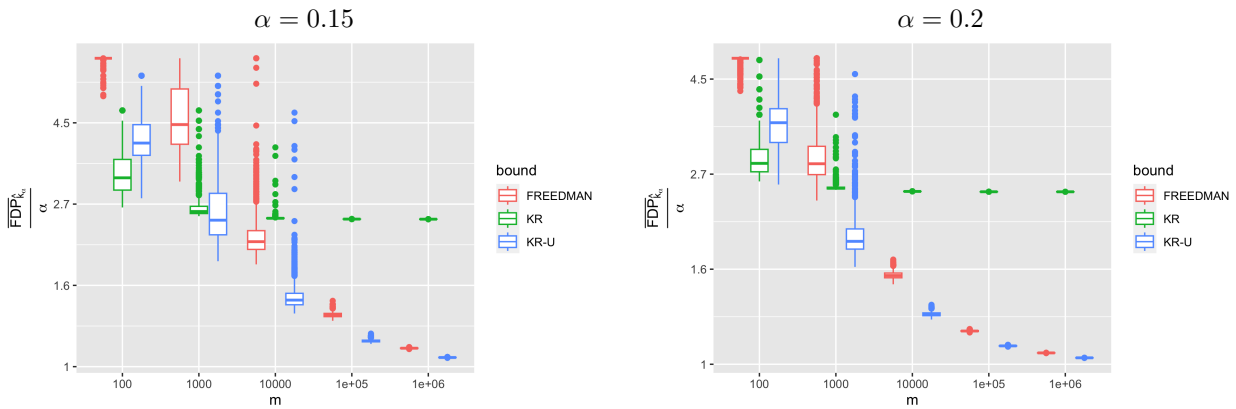

Figure 3.7: Pre-ordered dense ($\beta = 0$) knockoff setting with BC procedure (i.e., LF procedure with $s = \lambda = 0.5$).

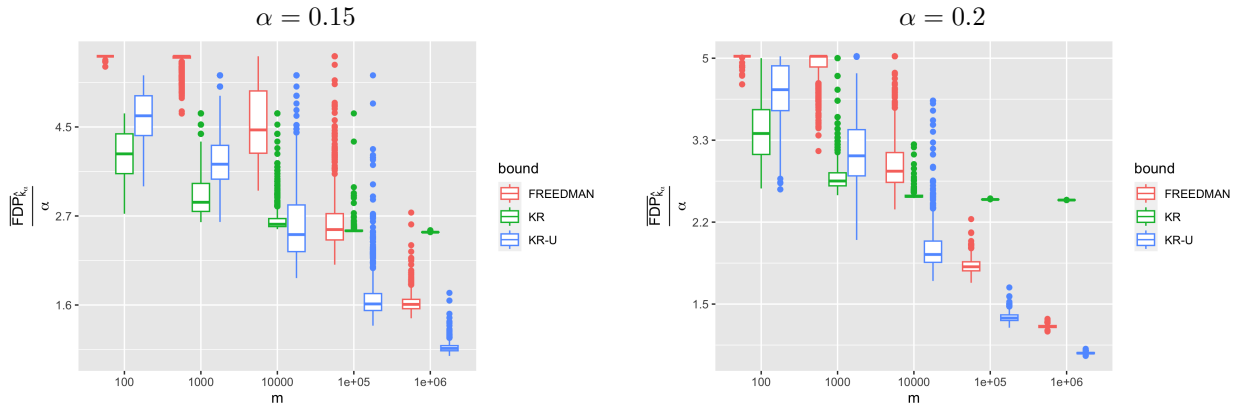

Figure 3.8: Pre-ordered sparse ($\beta = 0.25$) knockoff setting with BC procedure (i.e., LF procedure with $s = \lambda = 0.5$).

3.5.4 Comparison to Li et al. (2022)

In this section, we compare the performances of the KR-U bound with respect to the recent bounds proposed in Li et al. (2022). For this, we reproduce the high dimensional Gaussian linear regression setting of Section 5.1 (a) therein, which generates binary p-values by applying the fixed-X 'sdp' knockoffs and the signed maximum lambda knockoff statistic of Barber and Candès (2015). Doing so, the p-values follow the preordered setting of Section 3.3.1 and thus our bounds are non-asymptotically valid (note however that the p-values do not follow strictly speaking the VCT model of Section B.1.2). To be more specific, the considered Gaussian linear model $Y \sim \mathcal{N}(X\beta, I_n)$ is obtained by first generating X and β as follows: the correlated design matrix X of size $n \times m$ is obtained by drawing $n = 1500$ i.i.d. samples from the multivariate m-dimensional distribution $\mathcal{N}_m(0, \Sigma)$ where $\Sigma_{i,j} = 0.6^{|i-j|}$, $1 \leq i, j \leq m$; the signal vector $\beta \in \mathbb{R}^m$ is obtained by first randomly sampling a subset of $\{1, \dots, m\}$ of size $\lfloor \pi_1 m \rfloor$ for the non-zero entries of β and then by setting all non-zero entries of β equal to $a/\sqrt{n}$ for a given amplitude $a > 0$.

First, in the spirit of Figure 3 in Li et al. (2022), we display in Figure 3.10 the envelope $(\widetilde{\mathrm{FDP}}_k^{\text{KR-U}}, k \geq 1)$ given by the interpolation (3.2) of the envelope $(\overline{\mathrm{FDP}}_k^{\text{KR-U}}, k \geq 1)$ defined by (3.35) (with $s = \lambda = 1/2$), and compare it to those obtained in Li et al. (2022) (namely, KJI A/B/C/D) for $\pi_1 \in \{0.1, 0.5\}$, $a \in \{15, 25\}$. We also set here $\delta = 0.05$ to stick with the choice of Li et al. (2022) (note that this requires to further calibrate the parameters of their method according to this value of δ) and the number of replications is here only taken equal to 10 for computational reasons. Markedly, the KR-U envelope becomes much better than KR and is competitive w.r.t. KJI A/B/C/D, at least when k is moderately large. As expected, the most favorable case for KR-U is when the signal has a large amplitude and is dense.

Second, to stick with the consistency-oriented plots of the previous sections, we also display the corresponding FDP bounds for the BC procedure at level $\alpha \in \{0.15, 0.2\}$ in Figure 3.11. The conclusions are qualitatively similar.

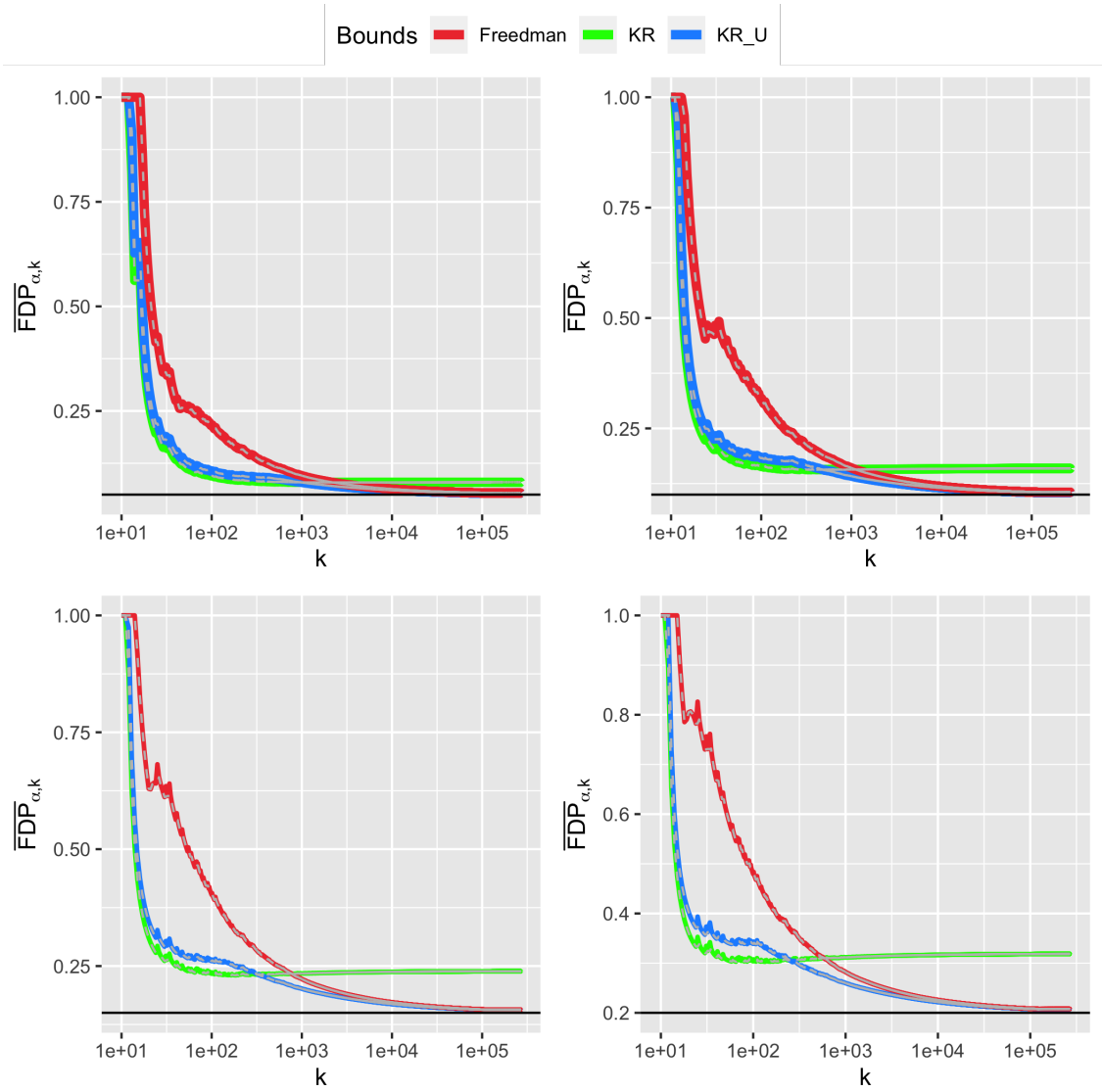

Figure 3.9: Online FDP envelopes for LORD applied on IMPC data for four values of $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$ (horizontal black bars). The interpolated bounds are displayed for each procedure as a gray dashed line.

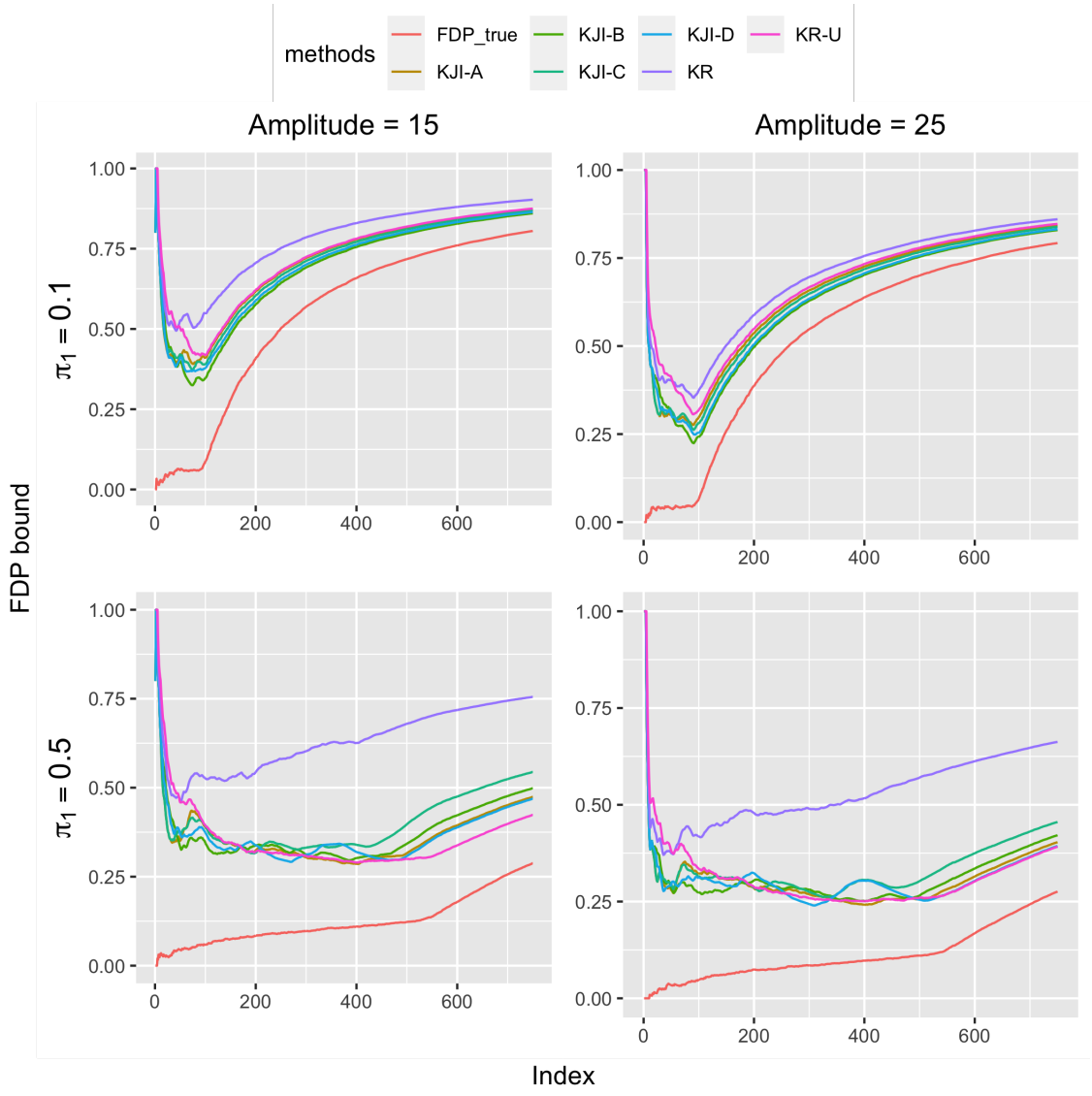

Figure 3.10: Comparing the envelope $\widetilde{\mathrm{FDP}}_k^{\text{KR-U}}$, $k \geq 1$ given by (3.2)-(3.35) ($s = \lambda = 0.5$) to those of Li et al. (2022) in the Gaussian linear regression setting of Section 3.5.4 for $m = 1000$ (see text for more details).

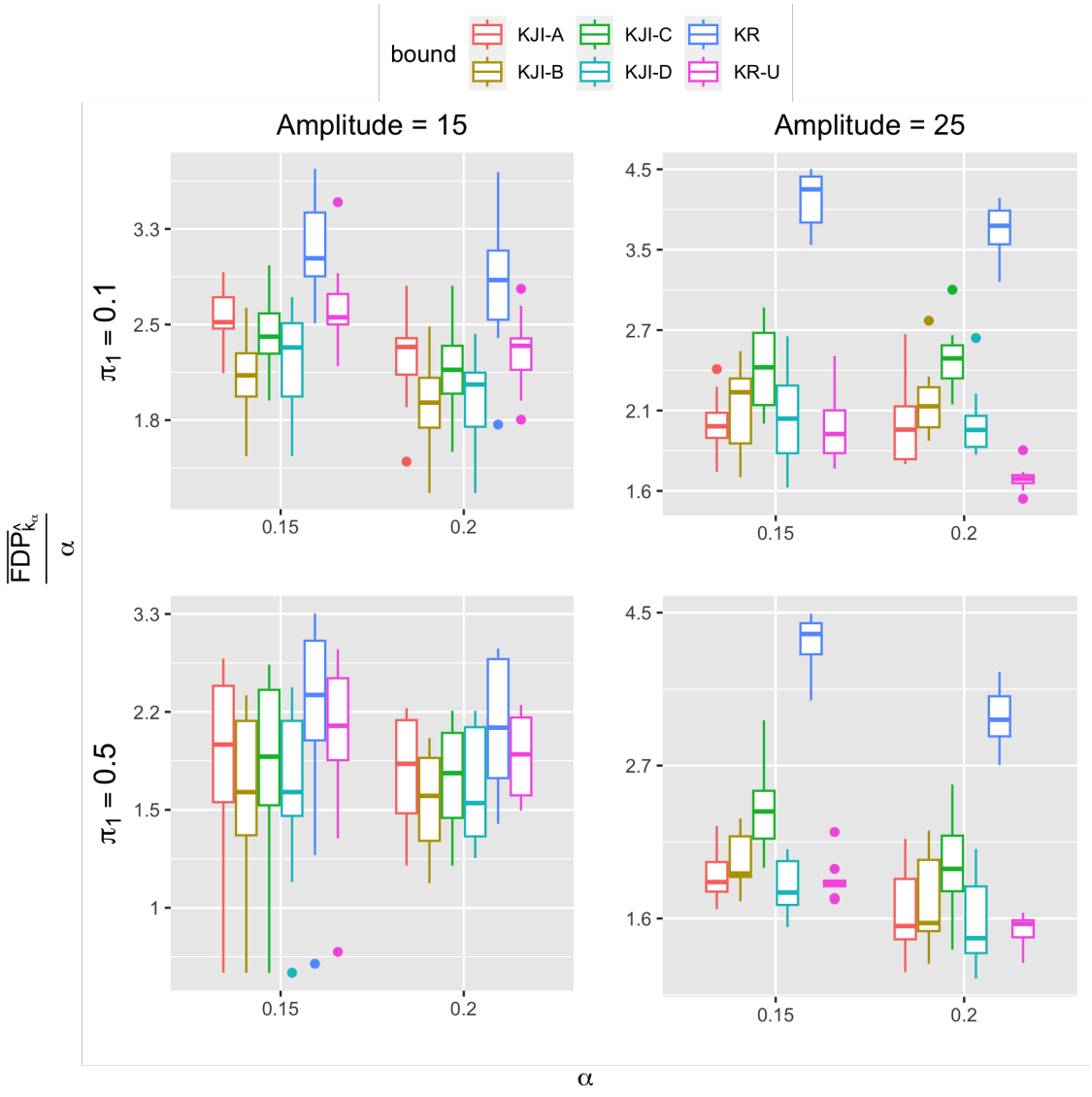

Figure 3.11: Comparing the FDP bound $\widetilde{\mathrm{FDP}}^{\text{KR-U}}_{\hat{k}_\alpha}$ for $\hat{k}_\alpha$ the BC procedure (3.32) ($s = \lambda = 0.5$) to those of Li et al. (2022) with respect to $\alpha \in \{0.15, 0.2\}$ in the Gaussian linear regression setting of Section 3.5.4 for $m = 1000$ (see text for more details).

3.6 Conclusion

The main point if this paper is to provide another point of view on FDP confidence bounds: we introduced a notion of consistency, a desirable asymptotical property which should act as a guiding principle when building such bounds, by ensuring that the bound is sharp enough on particular FDR controlling rejection sets. Doing so, some previous bounds were shown to be inconsistent, as the original KR bounds; while some other known FDP confidence bounds, in particular based on the DKW inequality, are consistent under certain assumptions, we have introduced new ones shown to satisfy this condition under more general conditions (in particular high sparsity). New bounds based on the classical Wellner/Freedman inequalities showed interesting behaviors, however simple modifications of KR bounds Hybrid/KR-U by 'stitching' have been shown to be the most efficient, both asymptotically and for moderate sample size. Overall, this work shows that consistency is a simple and fruitful criterion, and we believe that using it will be beneficial in the future to make wise choices among the rapidly increasing literature on FDP bounds.

Chapter 4

A unified class of null proportion estimators with plug-in FDR control

Outline of the current chapter

Since the work of Storey et al. (2004), it is well-known that the performance of the Benjamini-Hochberg (BH) procedure can be improved by incorporating estimators of the number (or proportion) of null hypotheses, yielding an adaptive BH procedure which still controls FDR. Several such plug-in estimators have been proposed since then, for some of these, like Storey's estimator, plug-in FDR control has been established, while for some others, e.g. the estimator of Pounds and Cheng (2006), some gaps remain to be closed. In this work we introduce a unified class of estimators, which encompasses existing and new estimators and unifies proofs of plug-in FDR control using simple convex ordering arguments. We also show that any convex combination of such estimators once more yields estimators with guaranteed plug-in FDR control. Additionally,

the flexibility of the new class of estimators also allows incorporating distributional informations on the p-values. We illustrate this for the case of discrete tests, where the null distributions of the p-values are typically known. In that setting, we describe two generic approaches for adapting any estimator from the general class to the discrete setting while guaranteeing plug-in FDR control. While the focus of this paper is on presenting the generality and flexibility of the new class of estimators, we also include some analyses on simulated and real data.

4.1 Introduction

4.1.1 Background

When many statistical tests are performed simultaneously, an ubiquituous way to account for the false rejections is the false discovery rate (FDR), that is the expected proportion of errors among the rejections. The seminal Benjamini and Hochberg (1995) procedure (abbreviated in the sequel as BH procedure) works by rejecting $H_{(1)}, \ldots, H_{(\hat{k})}$, where $\hat{k}$ is determined in the following *step-up* manner

$$\hat{k} = \max\left\{\ell \in \{0, \ldots, m\} : p_{(\ell)} \leq \frac{\ell}{m} \cdot \alpha\right\}, \tag{4.1}$$

where $p_{(1)} \leq \ldots \leq p_{(m)}$ denote the ordered p-values, and $H_{(1)}, \ldots, H_{(m)}$ the corresponding null hypotheses and $p_{(0)} := 0$. According to results in Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), this procedure guarantees that FDR $\leq \pi_0 \alpha$ when the p-values are independent or positively dependent, while for arbitrarily dependent p-values the Benjamini and Yekutieli (2001) procedure is available. The simplicity of the BH procedure and its many useful theoretical properties have made it an indispensable tool in modern high dimensional data analysis, see e.g. Benjamini (2010). Much work has gone into analyzing, extending and adapting this procedure to various settings.

In this context, Storey et al. (2004) showed that the *plug-in BH procedure*

$$\hat{k} = \max\left\{\ell \in \{0, \ldots, m\} : p_{(\ell)} \leq \frac{\ell}{\widehat{m}_0} \cdot \alpha\right\}, \tag{4.2}$$

obtained by replacing m in (4.1) by an estimate $\widehat{m_0}$ of m_0 still provides so called *adaptive* or *plug-in* FDR control at level α while allowing for more power. Classical examples of such estimates were proposed by Storey et al. (2004)

$$\widehat{m}_0^{\text{Storey}} = \frac{1 + \sum_{i=1}^{m} \mathbf{1}\{p_i > \lambda\}}{1 - \lambda}, \tag{4.3}$$

where $\lambda \in [0, 1)$ is a tuning parameter, and by Pounds and Cheng (2006)

$$\widehat{m}_0^{\text{PC,2006}} = 1 \wedge \left(2 \sum_{i=1}^{m} p_i\right). \tag{4.4}$$

While the focus of this work is on plug-in FDR control, estimates of m_0 (or equivalently $\pi_0 = m_0/m$) can also be used for FDR estimation purposes (see Storey (2002)). Thus, there is a large body of literature on this topic and numerous methods for establishing plug-in FDR control are available, see e.g. Benjamini et al. (2006); Sarkar (2008); Blanchard and Roquain

(2009); Heesen and Janssen (2016); Ditzhaus and Janssen (2019) and references therein. In this paper, we use a classical condition proposed by Blanchard and Roquain (2009) for establishing plug-in FDR control for a unified class of m_0-estimators, see Section 4.2.2 for more details.

Previous work on plug-in FDR control has focused on continuous test statistics, for which the null p-values are distributed according to the uniform distribution. Considering the abundance of super-uniform p-values in real life applications, the uniformity assumption is often violated which may lead to undesirable conservatism of the m_0-estimators, see Section 4.5 for more details. Super-uniform p-values can be observed when testing composite null hypotheses or when dealing with discrete tests, the latter being the setting of our interest in this work.

Discrete tests often originate when the tests are based on counts or contingency tables: for example in clinical studies, the efficacy or safety of drugs is determined by counting patients who survive a certain period, or experience a certain type of adverse drug reaction after being treated, see e.g. Chavant et al. (2011); and also in biology, where the genotype effect on the phenotype can be analyzed by knocking out genes and counting the number of individuals with a changed phenotype, see e.g. Muñoz-Fuentes et al. (2018). In discrete testing, each p-value is super-uniform and (potentially) has its own support, thus producing heterogeneous p-values. Pounds and Cheng (2006) recognized the need for developing methods tailored to discrete p-values and introduced $\widehat{m}_0^{\mathrm{PC},2006}$ as a simple and robust m_0-estimate in this setting. They did not, however, provide a proof of plug-in FDR control, not even in the uniform setting. Further works adressing the discreteness and heterogeneity include Chen et al. (2018) who introduced a m_0-estimator for discrete p-values based on averaging Storey type estimators for plug-in control. Biswas and Chattopadhyay (2020) pointed out an error in the proof of Chen et al. (2018) and provided a corrected version. However it is unclear whether this estimator actually provides an improvement over the classical (uniform) Storey estimator in practice. Thus, there are gaps to be filled on m_0-estimation both for the uniform and discrete case.

4.1.2 Contributions

In this paper we address some of the gaps and limitations mentioned above by introducing a simple and flexible class of m_0 estimators which has the following properties:

- Plug-in FDR control is guaranteed for all estimators contained in this class under independence of p-values. We give a unified proof using simple convex ordering arguments (for the reader's convenience we restate some definitions and classical results on stochastic and convex ordering in Appendix C.1).
- It provides a simple and flexible generic formulation which is useful for designing new estimators. In particular, we obtain a simple modification of $\widehat{m}_0^{\mathrm{PC},2006}$ with guaranteed plug-in FDR control.
- Additional distributional information on the p-values like heterogeneity and super-uniformity can be incorporated easily into estimators from the class. In particular, the estimators of this class can be used in conjunction with classical discrete p-value transformations like the mid-p transformation.
- Combining several weighted estimators from the class preserves plug-in FDR control.

The paper is organized as follows: Section 4.2 presents the statistical setting and restates a classical sufficient criterion for plug-in FDR control. Section 4.3 introduces the new class of estimators, and presents the main mathematical results on plug-in FDR control, followed by some numerical results in Section 4.4. In Section 4.5 we present approaches for adjusting estimators to

discreteness, and investigate their performance on simulated and real data. The paper concludes with a discussion in Section 4.6. Technical details – including classical results on stochastic and convex ordering– and further analyses are deferred to the Appendices.

4.2 Framework

4.2.1 Distributional assumptions

We use a classical setting for multiple testing encompassing homogeneous and heterogeneous nulls, see e.g. Döhler et al. (2018). We observe X, defined on an abstract probabilistic space, valued in an observation space $(\mathcal{X}, \mathfrak{X})$ and generated by a distribution P that belongs to a set $\mathcal{P}$ of possible distributions. We consider m null hypotheses for P, denoted $H_{0,i}$, $1 \leq i \leq m$, and we denote the corresponding set of true null hypotheses by $\mathcal{H}_0(P) = \{1 \leq i \leq m : H_{0,i} \text{ is satisfied by } P\}$. We also denote by $\mathcal{H}_1(P)$ the complement of $\mathcal{H}_0(P)$ in $\{1, \ldots, m\}$ and by $m_0(P) = m_0 = |\mathcal{H}_0(P)|$ the number of true nulls.
We assume that there exists a set of p-values that is a set of random variables $\{p_i(X), 1 \leq i \leq m\}$, valued in $[0, 1]$. We introduce the following dependence assumptions between the p-values:

All the p-values $\{p_i(X),\ 1 \leq i \leq m\}$ are mutually independent in the model $\mathcal{P}$. (Indep)

The (maximum) null cumulative distribution function (c.d.f) of each p-value is denoted

$$F_i(t) = \sup_{P \in \mathcal{P} \,:\, i \in \mathcal{H}_0(P)} \{\mathbf{P}_{X \sim P}(p_i(X) \leq t)\},\ \ t \in [0, 1],\ \ 1 \leq i \leq m. \tag{4.5}$$

We assume that the set of c.d.f $\mathcal{F} = \{F_i, 1 \leq i \leq m\}$ is *known* and we consider the following possible situations for the functions in $\mathcal{F}$:

For all $i \in \{1, \ldots, m\}$, F_i is continuous on $[0, 1]$ (Cont)

For all $i \in \{1, \ldots, m\}$, there exists some finite set $\mathcal{A}_i \subset [0, 1]$ such that F_i is a step function, right continuous, that jumps only at some points of $\mathcal{A}_i$. (Discrete)

The case (Discrete) typically arises when for all $P \in \mathcal{P}$ and $i \in \{1, \ldots, m\}$, $\mathbf{P}_{X \sim P}(p_i(X) \in \mathcal{A}_i) = 1$ for some given finite sets $\mathcal{A}_i \subset [0, 1]$. Throughout the paper, we will assume that we are either in the case (Cont) or (Discrete) and we denote $\mathcal{A} = \cup_{i=1}^m \mathcal{A}_i$, with by convention $\mathcal{A}_i = [0, 1]$ when (Cont) holds. We will also make use of the following classical assumption:

For all $i \in \{1, \ldots, m\}$, $F_i(t) \leq t$ for all $t \in [0, 1]$. (SuperUnif)

In this paper we will always assume that the p-values are mutually independent ((Indep) holds) and super-uniform under the null ((SuperUnif) holds).

4.2.2 FDR control for plug-in estimates

The following Theorem is a central result for plug-in FDR control, providing a sufficient condition based on bounding the inverse moment of the estimator $\widehat{m}_0$ by the inverse of m_0. Our presentation follows Blanchard and Roquain (2009), similar results can be found in Benjamini et al. (2006); Sarkar (2008); Zeisel et al. (2011).

Theorem 4.2.1 *Let $\widehat{m}_0 = \widehat{m}_0(p_1, \ldots, p_m)$ be a coordinatewise non-decreasing function of the p-values $(p_1, \ldots, p_m)$. Assume that $(p_1, \ldots, p_m)$ are mutually independent* (Indep) *and* (SuperUnif). *For $h \in \mathcal{H}_0$, denote by $p_{0,h}$ the set of p-values where p_h has been replaced by 0. If*

$$\mathbf{E}\left(\frac{1}{\widehat{m}_0(p_{0,h})}\right) \leq \frac{1}{m_0} \tag{IMC}$$

holds for all $h \in \mathcal{H}_0$, then the plug-in BH procedure given by (4.2) *controls FDR at level α.*

Throughout this paper, we will only consider coordinatewise non-decreasing estimators and will always assume that the p-values are mutually independent. Thus, when additionally (SuperUnif) holds true, the inverse moment criterion (IMC) is sufficient for establishing plug-in FDR control in our proofs. We mostly present results in term of the absolute number of null hypotheses m_0, but clearly equivalent statements using the proportion of null hypotheses $\pi_0 = m_0/m$ hold, and in some cases we will present results in terms of π_0 instead of m_0.

4.3 A unified class of plug-in estimators

In this section, we introduce a new class of estimators for m_0 (or equivalently π_0) that mathematically guarantees plug-in FDR control. It is based on sums of suitably transformed p-values, allowing us to recover classical estimators, such as the Storey (4.3) and the PC (slightly modified) (4.4) estimators, and also to define new estimators. We first present a general result for single estimators and then show that plug-in FDR control is also preserved for convex combinations of such estimators. To start, assume that the p-values are transformed by certain functions $g \in \mathcal{G}$, with

$$\mathcal{G} = \{g : [0,1] \to [0,1] : \ g \text{ is non-decreasing and } \mathbf{E}g(U) > 0, \text{ where } U \sim \mathcal{U}[0,1]\}. \tag{4.6}$$

Accordingly we define the class of estimators $\mathcal{F}_0$ as

$$\mathcal{F}_0 = \left\{ \widehat{m}_0 : [0,1]^m \to [0,\infty) |\ \widehat{m}_0(p_1, \ldots, p_m) = \frac{1}{\nu(g)} \left(1 + \sum_{i=1}^{m} g(p_i)\right), g \in \mathcal{G} \right\}, \tag{4.7}$$

where $\nu(g) = \mathbf{E}g(U)$ for any $g \in \mathcal{G}$ with $U \sim \mathcal{U}[0,1]$ (for brevity we sometimes omit the g in ν when there is no ambiguity concerning the function g). The class $\mathcal{F}_0$ contains the classical estimator $\widehat{m}_0^{\text{Storey}}$ (4.3) by taking $g(u) = \mathbf{1}\{u > \lambda\}$ and $\nu = 1 - \lambda$. It also contains a slightly modified version $\widehat{m}_0^{\text{PC,new}}$ of the classical estimator $\widehat{m}_0^{\text{PC,2006}}$ (4.4) obtained from taking $g(u) = u$ with $\nu = 1/2$, i.e.

$$\widehat{m}_0^{\text{PC,new}} = 2 + 2\sum_{i=1}^{m} p_i. \tag{4.8}$$

In Section 4.4 we will introduce some additional estimators and discuss $\widehat{m}_0^{\text{PC,new}}$ in more detail. The rationale behind the definitions of the classes $\mathcal{G}$ and $\mathcal{F}_0$ is two-fold. Requiring that g is non-decreasing ensures that $\widehat{m}_0$ is coordinatewise non-decreasing, allowing us to apply Theorem 4.2.1. The quantity $g(p_i)/\nu$ can be interpreted as the (local) contribution of p_i to the estimate of m_0. If we expect large p-values to provide evidence for null hypotheses, then it seems reasonable to require g to be non-decreasing. Rescaling $g(p_i)$ by $\nu = \mathbf{E}g(U)$ is a simple way of ensuring that

$\sum_{i=1}^m g(p_i)/\nu$ is conservatively biased in the sense that $\mathbf{E}(\sum_{i=1}^m g(p_i)/\nu) \geq m_0$ in any constellation of null and alternative hypotheses. This type of conservativeness may however not be strong enough for plug-in control. As our main result – Proposition 4.3.1 below – shows, simply adding $1/\nu$ as a 'safety margin' to the above estimate is enough for ensuring plug-in FDR control.
In some situations p-values under the null may be heterogeneous, i.e. the p-values may have different distributions under the null, so that using an individual transformation for each p-value may be helpful. To this end, we introduce the following richer and more flexible class of estimators.

$$\mathcal{F} = \Big\{ \widehat{m}_0 : [0,1]^m \to [0,\infty) \;\Big|\; \widehat{m}_0(p_1,\dots,p_m) = \frac{1}{\min(\nu_1,\dots,\nu_m)} + \sum_{i=1}^m \frac{g_i(p_i)}{\nu_i}, \quad \text{with } g_i \in \mathcal{G} \text{ and } \nu_i = \mathbf{E}[g_i(U)],\ U \sim \mathcal{U}[0,1] \text{ for all } i \Big\}. \tag{4.9}$$

We state our main result on plug-in FDR control for this more general class below. Clearly, $\mathcal{F}_0 \subset \mathcal{F}$, so that the results stated for $\mathcal{F}$ also hold for $\mathcal{F}_0$. We now present an upper bound in the convex order for transformed uniform random variables in terms of Bernoulli random variables, which is the main technical tool we use for proving plug-in FDR control for the class $\mathcal{F}$.

Lemma 4.3.1 *For any $g \in \mathcal{G}$ we have $g(U) \leqslant_{cx} \mathbf{Bin}(1,\nu)$ and $\nu = \mathbf{E}g(U)$, where $U \sim \mathcal{U}[0,1]$, and the notation $\leqslant_{cx}$ denotes the convex ordering (see Definition C.1.2).*

Proof 4.3.1 *For $U \sim \mathcal{U}[0,1]$ define $X = g(U)$, so that $\mathbf{E}(X) = \nu$. Let $l_X = \inf_{x\in[0,1]} g(x)$, $u_X = \sup_{x\in[0,1]} g(x)$ be the lower and upper endpoints of the support of X, and define a two-point distribution Y concentrated on $\{l_X, u_X\}$ by $P(Y = l_X) = (u_X - \nu)/(u_X - l_X)$ and $P(Y = u_X) = (\nu - l_X)/(u_X - l_X)$. By Lemma C.1.1 we have $X \leqslant_{cx} Y$.*
Now let $Z \sim \mathbf{Bin}(1,\nu)$ and denote the distribution function of Y and Z by F and G. Clearly, $\mathbf{E}Y = \nu = \mathbf{E}Z$. Since $[l_X, u_X] \subset [0,1]$ and both Y and Z are two-point distributions, the function $G - F$ posesses one crossing point on $[0,1)$. Indeed, for $t \in [0, l_X)$, $\quad F(t) = 0$ while $G(t) = 1 - \nu$ so that $G - F$ is positive, and for $t \in [u_X, 1), F(t) = 1$ while $G(t) = 1 - \nu$ so that $G - F$ is negative. For $t \in [l_X, u_X)$, $G - F$ can be positive or negative depending on ν. Overall, the sign sequence of $G - F$ is $+,-$ so that Lemma C.1.2 implies that $Y \leqslant_{cx} Z$ and the claim follows.

The following proposition is our main result on plug-in FDR control.

Proposition 4.3.1 *Assume that $p_1,\dots,p_m$ are mutually independent and* (SuperUnif) *holds. Then* (IMC) *holds true for any estimator $\widehat{m}_0 \in \mathcal{F}$, where $\mathcal{F}$ is defined by* (4.9)*. In particular, the BH plug-in procedure* (4.2) *using $\widehat{m}_0$ controls FDR at level α.*

Proof 4.3.2 *Since $\widehat{m}_0$ is coordinatewise non-decreasing, it is sufficient to verify* (IMC)*. For any $h \in \mathcal{H}_0$, monotonicity and super-uniformity give us $\widehat{m}_0(p_{0,h}) \geqslant_{st} 1/\nu + S_0$, where $\nu = min_{l\in\mathcal{H}_0\setminus\{h\}}\nu_l$, and $S_0 = \sum_{\ell\in\mathcal{H}_0\setminus\{h\}} g_\ell(U_\ell)/\nu_\ell$ with $(U_\ell)_{\ell\in\mathcal{H}_0}$ i.i.d random variables distributed according to $\mathcal{U}[0,1]$. By Lemma 4.3.1 we have $g_\ell(U_\ell) \leqslant_{cx} \mathbf{Bin}(1,\nu_\ell)$ and Lemma C.1.3 gives $\mathbf{Bin}(1,\nu_i)/\nu_i \leqslant_{cx} \mathbf{Bin}(1,\nu)/\nu$. Since the convex ordering is preserved under convolutions (see Lemma C.1.1) we obtain $\nu S_0 \leqslant_{cx} \mathbf{Bin}(m_0 - 1, \nu)$. Finally, the mapping $x \mapsto \nu/(1+x)$ is convex on $[0,\infty)$ and therefore from the Definition C.1.2 of $\leqslant_{cx}$ we obtain that*

$$\mathbf{E}\left(\frac{1}{\widehat{m}_0(p_{0,h})}\right) \leq \mathbf{E}\left(\frac{1}{\frac{1}{\nu} + S_0}\right) = \mathbf{E}\left(\frac{\nu}{1+\nu S_0}\right) \leq \mathbf{E}\left(\frac{\nu}{1+\mathbf{Bin}(m_0-1,\nu)}\right) \leq \frac{1}{m_0}, \tag{4.10}$$

where the last bound is a well-known result for the inverse moment of Binomial distributions (see e.g. C.1.4 in Appendix) so that (IMC) *is proved. The statement on plug-in FDR control now follows from Theorem 4.2.1.*

By taking $g(u) = \mathbf{1}\{u > \lambda\}$ we have $\nu S_0 \sim \mathbf{Bin}(m_0 - 1, \nu)$ and therefore the second inequality from the right in (4.10) can be replaced by an equality. Thus, it may be tempting to conclude that $\widehat{m}_0^{\text{Storey}}$ is optimal. In the case of a Dirac-Uniform constellation of p-values (see Blanchard and Roquain (2009)) this is indeed true, since $\nu\widehat{m}_0^{\text{Storey}} \sim 1 + \mathbf{Bin}(m_0 - 1, \nu)$ and therefore the left inequality in (4.10) can also be replaced by an equality. In more general settings however, other choices of g may be better, as the results in Section 4.4.1 show.

We highlight that introducing a general class of estimators as (4.9) with Proposition 4.3.1 allows a unified proof of plug-in FDR control for known estimators like $\widehat{m}_0^{\text{Storey}}$ and $\widehat{m}_0^{\text{PC,new}}$ and also for new estimators that we will define in Section 4.4. Additionally, the classes $\mathcal{F}_0$ and $\mathcal{F}$ possess stability properties that make it possible to combine various plug-in estimators while maintaining FDR control.

Proposition 4.3.2 *Let $\widehat{m}_1, \widehat{m}_2 \in \mathcal{F}$, where $\mathcal{F}$ is defined by* (4.9) *and let $\lambda \in [0, 1]$. Then the BH plug-in procedure* (4.2) *using $\widehat{m}_0 = \lambda\widehat{m}_1 + (1 - \lambda)\widehat{m}_2$ controls FDR at level α.*

Proof 4.3.3 *We show that $\widehat{m}_0$ satisfies* (IMC). *Let $\widehat{m}_1, \widehat{m}_2 \in \mathcal{F}$ have the representation*

$$\widehat{m}_1 = \frac{1}{\nu} + \sum_{i=1}^{m} \frac{g_i(p_i)}{\nu_i} \qquad \text{and} \qquad \widehat{m}_2 = \frac{1}{\mu} + \sum_{i=1}^{m} \frac{h_i(p_i)}{\mu_i},$$

where $\nu = min(\nu_1, \ldots, \nu_m)$ and $\mu = min(\mu_1, \ldots, \mu_m)$ so that

$$\widehat{m}_0 = \frac{\lambda}{\nu} + \frac{1 - \lambda}{\mu} + \sum_{i=1}^{m} \frac{\lambda g_i(p_i)}{\nu_i} + \frac{(1 - \lambda)h_i(p_i)}{\mu_i}$$

and define weights $\kappa_i = \frac{\lambda\mu_i}{\lambda\mu_i + (1-\lambda)\nu_i}$ and transformations $f_i = \kappa_i g_i + (1-\kappa_i)h_i$. Clearly, $\kappa_i \in [0, 1]$ and $f_i \in \mathcal{G}$ and we introduce $\epsilon_i = \mathbf{E}(f_i) = \kappa_i\nu_i + (1-\kappa_i)\mu_i$. From the above definitions we obtain with some straightforward algebra

$$\lambda = \frac{\kappa_i\nu_i}{\epsilon_i} \qquad \text{and} \qquad 1 - \lambda = \frac{(1 - \kappa_i)\mu_i}{\epsilon_i} \tag{4.11}$$

which yields

$$\frac{\lambda}{\nu_i} g_i + \frac{(1 - \lambda)}{\mu_i} h_i = \frac{\kappa_i}{\epsilon_i} g_i + \frac{(1 - \kappa_i)}{\epsilon_i} h_i = \frac{f_i}{\epsilon_i}. \tag{4.12}$$

From (4.11) *we have*

$$\frac{\lambda}{\nu} = \max\left(\frac{\lambda}{\nu_1}, \ldots, \frac{\lambda}{\nu_m}\right) = \max\left(\frac{\kappa_1}{\epsilon_1}, \ldots, \frac{\kappa_m}{\epsilon_m}\right) \qquad \text{and}$$
$$\frac{1 - \lambda}{\mu} = \max\left(\frac{1 - \lambda}{\mu_1}, \ldots, \frac{1 - \lambda}{\mu_m}\right) = \max\left(\frac{1 - \kappa_1}{\epsilon_1}, \ldots, \frac{1 - \kappa_m}{\epsilon_m}\right)$$

so that the sub-additivity of the max *function now yields the bound*

$$\frac{\lambda}{\nu} + \frac{1-\lambda}{\mu} \geq \frac{1}{\epsilon}, \tag{4.13}$$

where $\epsilon = \min(\epsilon_1, \ldots, \epsilon_m)$. *Combining* (4.12) *and* (4.13) *now gives us*

$$\widehat{m}_0 = \frac{\lambda}{\nu} + \frac{1-\lambda}{\mu} + \sum_{i=1}^{m} \frac{\lambda g_i(p_i)}{\nu_i} + \frac{(1-\lambda)h_i(p_i)}{\mu_i} \geq \frac{1}{\epsilon} + \sum_{i=1}^{m} \frac{f_i(p_i)}{\epsilon_i} := \widetilde{m}_0$$

with $\widetilde{m}_0 \in \mathcal{F}$ *and from Proposition 4.3.1 we know that* (IMC) *holds true for* $\widetilde{m}_0$ *and therefore also for* $\widehat{m}_0$.

The proof shows that $\mathcal{F}$ is "almost" convex in the sense that whenever equality holds in (4.13) we have $\widehat{m}_0 = \widetilde{m}_0 \in \mathcal{F}$. If $\widehat{m}_1, \widehat{m}_2 \in \mathcal{F}_0$, i.e. each estimator uses only a single transformation function then it is easy to see that equality holds in (4.13) which leads to the following result:

Proposition 4.3.3 *The class of estimators* $\mathcal{F}_0$ *given by* (4.7) *is convex. In particular this implies that for any* $\widehat{m}_1, \widehat{m}_2 \in \mathcal{F}_0$ *and* $\lambda \in [0,1]$ *the BH plug-in procedure* (4.2) *controls FDR at level* α *for the estimator* $\widehat{m}_0 = \lambda\widehat{m}_1 + (1-\lambda)\widehat{m}_2$.

4.4 Homogeneous estimators

In this section we focus on the class of homogeneous estimators $\mathcal{F}_0$ given by (4.7), i.e. on estimators of the form

$$\widehat{m}_0 = \widehat{m}_0(p_1, \ldots, p_m) = \frac{1}{\nu}\left(1 + \sum_{i=1}^{m} g(p_i)\right),$$

where $g \in \mathcal{G}$ and $\nu = \nu(g) = \mathbf{E}g(U) > 0$, with $U \sim \mathcal{U}[0,1]$. As mentioned before, this class includes the classical estimator $\widehat{m}_0^{\text{Storey}}$ (4.3) and the new estimator $\widehat{m}_0^{\text{PC,new}}$ (4.8), and also gives the scientist freedom to define new estimators with proven plug-in FDR control thanks to Proposition 4.3.1 There are many conceivable ways in which this can be done. As an ad hoc example, we define a polynomial estimator of degree $r \geq 0$ and thresholding parameter $\lambda \in [0,1)$ by taking $\widehat{m}_0$ as above in (4.7) with $g(u) = g_{r,\lambda}(u) = u^r \cdot \mathbf{1}\{u > \lambda\}$, so that $\nu = \frac{1-\lambda^{r+1}}{r+1}$. This gives us

$$\widehat{m}_0^{\text{Poly}}(r,\lambda) = \frac{r+1}{1-\lambda^{r+1}} + \frac{r+1}{1-\lambda^{r+1}} \sum_{i=1}^{m} p_i^r \cdot \mathbf{1}\{p_i > \lambda\}. \tag{4.14}$$

It is easily seen that the classical estimators $\widehat{m}_0^{\text{Storey}}$ and $\widehat{m}_0^{\text{PC,new}}$ are particular instances of $\widehat{m}_0^{\text{Poly}}(r,\lambda)$ with $r = 0$ for $\widehat{m}_0^{\text{Storey}}$ and $r = 1$ and $\lambda = 0$ for $\widehat{m}_0^{\text{PC,new}}$. Taking $r = 1$ and $\lambda > 0$ yields a hybrid estimator which combines $\widehat{m}_0^{\text{Storey}}$ and $\widehat{m}_0^{\text{PC,new}}$ which has the potential to combine the strengths of both methods. For all estimators $\widehat{m}_0^{\text{Poly}}(r,\lambda)$ plug-in FDR control follows immediately from Proposition 4.3.1. For illustrational purposes we effectively only use r as a parameter and set the thresholding parameter to the classical value of $\lambda = 1/2$ throughout the paper. These examples are primarily meant to illustrate the freedom and flexibility Proposition 4.3.1 allows for the class $\mathcal{F}_0$ and should not be interpreted as recommendations for optimal choices. These examples are investigated further in the following section.

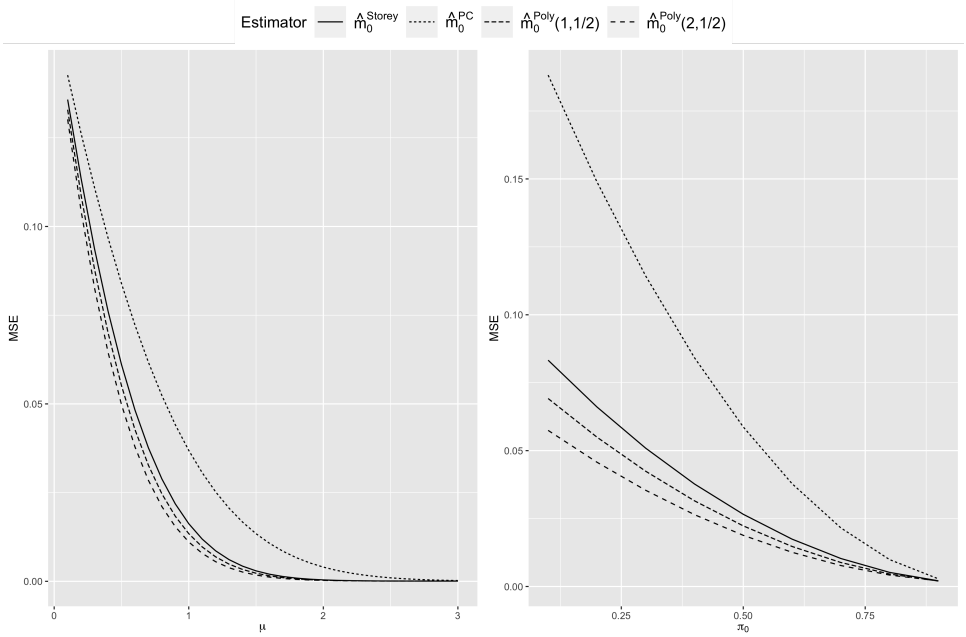

Figure 4.1: MSE against signal strength $\mu \in [0, 3]$ with $\pi_0 = 0.6$(left) and MSE against $\pi_0 \in [0.1, 0.9]$ with $\mu = 1.5$(right)

4.4.1 Numerical results

Here we compare the performance of several estimators from the class $\mathcal{F}_0$ in a Gaussian one-sided testing setting. We assume that we observe $X_1, \ldots, X_m$ independent random variables with $X_i \sim \mathcal{N}(0, 1)$ for $i \in \mathcal{H}_0$ and $X_i \sim \mathcal{N}(\mu, 1)$ for $i \in \mathcal{H}_1$, with $\mu > 0$, and we test $H_{0,i} : \mu = 0$ vs. $H_{1,i} : \mu > 0$. For a given signal strength $\mu > 0$ under the alternative, closed-form expressions for the expectation and variance of $\widehat{m}_0$ are available, see Appendix C.2 for more details. Thus we can numerically compare the mean squared error (MSE) of estimators from $\mathcal{F}_0$.

For this analysis, we fix $m = 10\,000$ and first compare the MSE w.r.t. to the signal strength μ of the alternatives with a fixed proportion of true nulls $\pi_0 = 0.6$. Then, we compare the MSE w.r.t. the proportion of true nulls π_0 with a fixed signal strength $\mu = 1.5$. The considered estimators for this comparison are $\widehat{m}_0^{\text{Storey}}$ (4.3), $\widehat{m}_0^{\text{PC,new}}$ (4.8), $\widehat{m}_0^{\text{Poly}}(1, 1/2)$, and $\widehat{m}_0^{\text{Poly}}(2, 1/2)$ (see (4.14) for both). The MSE is evaluated in terms of π_0 for better readability and displayed in Figure 4.1. The qualitative comparison between the estimators remains consistent across both panels of Figure 4.1: $\widehat{m}_0^{\text{PC,new}}$ has the poorest performance, characterized by the largest MSE, followed by $\widehat{m}_0^{\text{Storey}}$. While the polynomial approach shows some benefits, the improvement is not particularly remarkable except for small to moderate values of π_0. For larger values of π_0 or μ, there are no noticeable differences in performance.

4.4.2 More details on the Pounds and Cheng estimator

While FDR control for $\widehat{m}_0^{\text{Storey}}$ is a classical result following from Theorem 4.2.1 (Blanchard and Roquain (2009); Benjamini et al. (2006)), much less is known about the validity of $\widehat{m}_0^{\text{PC,2006}}$ as a

plug-in estimator. Indeed, Pounds and Cheng (2006) introduced their estimator (4.4) primarily to obtain a robust estimate of FDR. To the best of our knowlegde, the only related result on plug-in FDR control was obtained by Zeisel et al. (2011), who defined the following modified version of $\widehat{m}_0^{\mathrm{PC},2006}$:

$$\widehat{m}_0^{\mathrm{PC,ZZD}} = C(m) \cdot \min\left[m, \max\left(s(m), 2 \cdot \sum_{i=1}^{m} p_i\right)\right], \tag{4.15}$$

where the correction factors $C(m)$ and $s(m)$ are chosen in such a way so that (IMC) holds. However, determining these factors is non-trivial and requires extensive use of numerical integration and approximations methods (see Supplement B in Zeisel et al. (2011) for further details) so that no simple representation of $C(m)$ and $s(m)$ is available (for selected values of m, Table S1 in Zeisel et al. (2011) lists values for the correction factors).

By contrast, our new modification (4.8) is extremely simple and, as we show in Section 4.5, can be adapted easily to e.g. discrete tests, thus confirming a conjecture in Pounds and Cheng (2006). Its validity for plug-in FDR control follows directly from Proposition 4.3.1 and involves no sophisticated asymptotic or numerical approximations. Supplementary material in Appendix C.3 shows that the two versions of the PC estimator behave more or less identically. Nevertheless, we argue in favor of using $\widehat{m}_0^{\mathrm{PC,new}}$ since it is both conceptually and computationally much simpler than $\widehat{m}_0^{\mathrm{PC,ZZD}}$.

4.5 Adjusted estimators for discrete p-values

In this section we assume – additionally to mutual independence of the p-values – that the null distribution functions $F_1, \ldots, F_m$ are known. As a particular application we consider the setting of discrete p-values (see Section 4.1.1 for more detailed references). The classical plug-in estimators, like the Storey (2002) estimator defined in (4.3), were developed for uniformly distributed p-values under the nulls, and can thus suffer of an inflated bias when computed under (SuperUnif) assumption.

To illustrate the problem, we compare the bias of an arbitrary estimator $\widehat{m}_0 \in \mathcal{F}_0$ under the uniform setting with the bias under the super-uniform setting. In the classical uniform case, considering marginally independent p-values $p_i \sim X_0$ for $i \in \mathcal{H}_0$, and $p_i \sim X_1$ for $i \in \mathcal{H}_1$, for some variables X_0, X_1 defined on $[0,1]$, the bias is seen to be

$$\mathrm{Bias}[\widehat{m}_0] = \mathbf{E}[\widehat{m}_0] - m_0 = \frac{1}{\nu}(1 + m_1 \mathbf{E}[g(X_1)]). \tag{4.16}$$

In contrast, under the super-uniform setting, still considering independent p-values under the null and the alternative, the bias is

$$\mathrm{Bias}[\widehat{m}_0] = \frac{1}{\nu}(1 + m_1 \mathbf{E}[g(X_1)]) + \frac{1}{\nu} m_0 (\mathbf{E}[g(X_0)] - \nu). \tag{4.17}$$

Recall that $\nu = \mathbf{E}[g(U)]$ with $U \sim \mathcal{U}[0,1]$, thus under super-uniformity $\mathbf{E}[g(X_0)] \geq \nu$ (see the characterization of the usual stochastic order in Appendix C.1), which shows that an additional source of conservativeness is present in this case. In general, practitioners use classical estimators without worrying about p-values distributions, ingenuously expecting the estimator to perform according to the "uniform" bias (4.16) when in fact it often performs according to the "super-uniform" bias (4.17). This motivates the need for a correction in the estimator that will aim at

deflating (4.17).
Super-uniformity does not solely appear in the discrete setting, it also occurs e.g. when testing composite nulls, however in the discrete setting additional information on the p-values c.d.f (as defined by (4.5)) may be available and leveraged to correct the over-conservativeness. In this section, we present two such approaches that incorporate the available knowledge of F_i – the p-value c.d.f under the null – in the estimators. The standard way of defining p-values for discrete tests leads to distribution functions that satisfy (Discrete) and (SuperUnif). As we later introduce transformed p-values with transformed distribution, for clearer distinguishability the c.d.f associated to these standard discrete p-values are denoted by $F_1^{\mathrm{sd}}, \dots, F_m^{\mathrm{sd}}$ (where the upper-script "sd" denotes "standard discrete").

4.5.1 Transformations of discrete p-values

In order to reduce the individual conservatism of p-values caused by super-uniformity, various transformation of discrete p-values have been proposed, see e.g. Habiger (2015). Perhaps the most popular transformation is the so-called *mid-p-value*. For the realization x of the random variable X, let $p(x)$ be the (realized) standard p-value. Now define the *mid-p-value* (Rubin-Delanchy et al., 2019) $q(x)$ given the observation x as

$$q(x) = p(x) - \frac{1}{2} P_0(p(X) = p(x)), \tag{4.18}$$

where P_0 denotes the distribution of X under the null (for simplicity, we assume that such a unique distribution exists). Transforming the p-value through (4.18) helps to mimic the behavior of a uniform random variable in expectation. Indeed, we always have $\mathbb{E}[q(X)] = 1/2$, see Berry and Armitage (1995) for more details. However, the distribution of the mid-p-value is no longer super-uniform but shrunk toward 0 as displayed in Figure 4.2. In what follows, we denote by $F_1^{\mathrm{mid}}, \dots, F_m^{\mathrm{mid}}$ the distribution functions of the mid-p-values associated with the distribution functions $F_1^{\mathrm{sd}}, \dots, F_m^{\mathrm{sd}}$ of the standard p-values. In Section 4.5.2 we show how the distribution functions of standard discrete or mid-p-values can be used in m_0-estimators introduced in Section 4.3, while preserving plug-in FDR control.

Another transformation to reduce the conservativeness of discrete p-values uses so-called *randomized p-values* which are defined in our context by

$$r(x, u) = p(x) - u \cdot P_0(p(X) = p(x)), \tag{4.19}$$

where u is the realization of a uniform random variable $U \sim \mathcal{U}[0, 1]$, independent of X. Alternatively to the notation $r(x, u)$, we will also use (with a slight abuse) $r(p, u)$, where $p = p(x)$ is the standard p-value obtained from observation x. Randomized p-values and mid-p values are related via the conditional expectation on the observations $q(x) = \mathbf{E}_U[r(x, U)|X = x]$. Randomization leads to an (unconditional) uniform behavior, however at the cost of introducing an additional source of randomness which makes its use controversial for decisions on individual hypotheses, see e.g. Habiger and Pena (2011) for a discussion. We show in Section 4.5.3 that for estimation purposes however, randomized p-values can be beneficial for obtaining an efficient non-randomized estimator.

4.5.2 Adjusting the rescaling constants

The first approach for adjusting estimators to discrete p-values is tailored to estimators from the

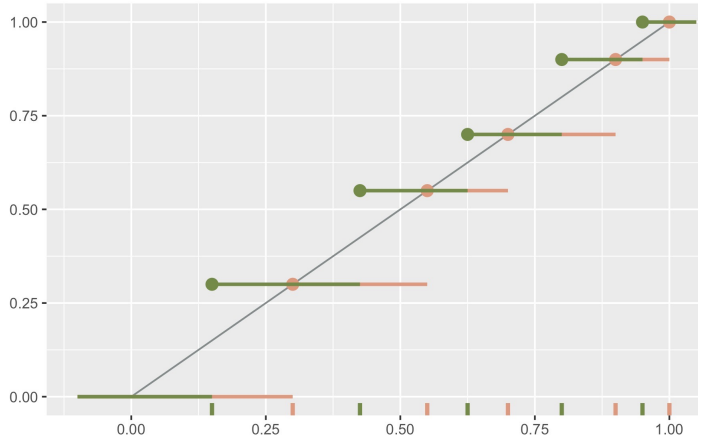

Figure 4.2: Distribution functions of a standard p-value and its associated mid-p-value.The standard p-value has support $\{0.3, 0.55, 0.7, 0.9, 1\}$ (orange rug plot), the c.d.f (orange solid line) is super-uniform, i.e. below the uniform c.d.f (gray line). The support of the associated mid-p-value $\{0.15, 0.465, 0.625, 0.8, 0.95\}$ (green rug plot) is shifted to the left, but the probabilities (given by the jumps in the green c.d.f) remain the same.

class $\mathcal{F}_0$ and adjusts the rescaling constant ν in $\widehat{m}_0$. In fact, this approach is not limited to the discrete setting, and can also be applied for arbitrary p-value distributions.

Proposition 4.5.1 *Assume that $p_1, \ldots, p_m$ are mutually independent and the null distribution functions $F_1, \ldots, F_m$ are known. For any $\widehat{m}_0 \in \mathcal{F}_0$ (4.7) with*

$$\widehat{m}_0(p_1, \ldots, p_m) = \frac{1}{\nu(g)} \left(1 + \sum_{i=1}^{m} g(p_i) \right) \tag{4.20}$$

define the adjusted estimator

$$\widehat{m}_0^{adj}(p_1, \ldots, p_m) = \frac{1}{min(\nu_1^{adj}, \ldots, \nu_m^{adj})} + \sum_{i=1}^{m} \frac{g(p_i)}{\nu_i^{adj}} \tag{4.21}$$

where $\nu_i^{adj} = \mathbf{E}_{p_i \sim F_i}[g(p_i)]$, is the expectation of the transformed p-value taken w.r.t. F_i. Then the BH plug-in procedure (4.2) using $\widehat{m}_0^{adj}$ controls FDR at level α.

Proof 4.5.1 *Without loss of generality we assume that $\nu_i^{adj} > 0$, otherwise the sum and minimum in (4.21) is to be taken over the index set $\{i : \nu_i^{adj} > 0\}$.*

For any $i \in \{1, \ldots, m\}$ define $g_i : [0,1] \to [0,1]$ by $g_i(y) = g \circ F_i^{-1}(y)$ for $y \in (0,1]$, where $F_i^{-1}(y) = \inf\{x \in \mathbb{R} : F_i(x) \geq y\}$ is the generalized inverse of F_i, and set $g_i(0) = g(0)$. Since $g \in \mathcal{G}$ and F_i^{-1} are both nondecreasing, so is g_i. For $i \in \mathcal{H}_0$, with $U \sim \mathcal{U}[0,1]$, we have $p_i \sim F_i^{-1}(U)$ by Proposition 2 in Embrechts and Hofert (2013), so that $g_i(U) \sim g(p_i)$, which implies that $\mathbf{E}[g_i(U)] = \mathbf{E}_{p_i \sim F_i}[g(p_i)] = \nu_i^{adj}$, so that (4.21) belongs to the class $\mathcal{F}$.
Now let $z_1, \ldots z_m$ be independent random variables with $z_i \sim \mathcal{U}[0,1]$ for $i \in \mathcal{H}_0$ and $z_i \sim \delta_0$ for

$i \in \mathcal{H}_1$ *(Dirac-Uniform configuration). Since* $g_i(U) \sim g(p_i)$ *for* $i \in \mathcal{H}_0$*, we have*

$$\begin{aligned}\widetilde{m}_0^{adj}(z_1,\dots,z_m) &= \frac{1}{min(\nu_1^{adj},\dots,\nu_m^{adj})} + \sum_{i\in\mathcal{H}_0} \frac{g_i(z_i)}{\nu_i^{adj}} \\ &\sim \frac{1}{min(\nu_1^{adj},\dots,\nu_m^{adj})} + \sum_{i\in\mathcal{H}_0} \frac{g(p_i)}{\nu_i^{adj}} \leq \widehat{m}_0^{adj}(p_1,\dots,p_m) \quad \textit{(a.s.)}\end{aligned}$$

Since $\widetilde{m}_0^{adj}(z_1,\dots,z_m) \in \mathcal{F}$*, by Proposition 4.3.1* (IMC) *holds for* $\widetilde{m}_0^{adj}(z_1,\dots,z_m)$*, and since* $\widetilde{m}_0^{adj}(z_1,\dots,z_m) \leq \widehat{m}_0^{adj}(p_1,\dots,p_m)$ *(a.s.),* (IMC) *also holds for* $\widehat{m}_0^{adj}(p_1,\dots,p_m)$*.*

Following Proposition 4.5.1, we define the discrete-uniform estimator using (4.21) with standard discrete p-values $p_1,\dots,p_m$ and their distribution functions $F_1^{\mathrm{sd}},\dots,F_m^{\mathrm{sd}}$

$$\widehat{m}_0^{\mathrm{du}}(p_1,\dots,p_m) = \frac{1}{min(\nu_1^{\mathrm{du}},\dots,\nu_m^{\mathrm{du}})} + \sum_{i=1}^{m} \frac{g(p_i)}{\nu_i^{\mathrm{du}}}, \tag{4.22}$$

where $\nu_i^{\mathrm{du}} = \mathbf{E}_{p_i\sim F_i^{\mathrm{sd}}}[g(p_i)]$.

Corollary 4.5.1 *Assume that* $p_1,\dots,p_m$ *are mutually independent and* (SuperUnif) *holds with null distribution functions* $F_1^{sd},\dots,F_m^{sd}$ *that are known. Then the BH plug-in procedure* (4.2) *using* $\widehat{m}_0^{du}$ *as in* (4.22)*controls FDR at level* α*. Moreover* $\widehat{m}_0^{du} \leq \widehat{m}_0$ *(a.s.), where* $\widehat{m}_0$ *is the base non-adjusted estimator* (4.20)*.*

The last statement of the corollary shows that for standard discrete p-values the estimator $\widehat{m}_0^{\mathrm{du}}$ is guaranteed to perform better than $\widehat{m_0}$. This follows from the fact that $\nu_1^{\mathrm{du}},\dots,\nu_m^{\mathrm{du}} \geq \nu$ because g is non-decreasing and (SuperUnif) holds (see Appendix C.1).

For classical estimators, the adjusted rescaling constants can be computed easily, using

- $\nu_i^{\text{du-Storey}} = 1 - F_i^{\mathrm{sd}}(\lambda)$;
- $\nu_i^{\text{du-PC}} = \sum_{x\in\mathcal{A}_i} x \cdot P(p_i = x)$, where $\mathcal{A}_i$ denotes the support of F_i^{sd}.

Similarly to (4.22), we define a mid p-value estimator using (4.21) with mid-p-values $q_1,\dots,q_m$ and their distribution functions $F_1^{\mathrm{mid}},\dots,F_m^{\mathrm{mid}}$

$$\widehat{m}_0^{\mathrm{mid}}(q_1,\dots,q_m) = \frac{1}{min(\nu_1^{\mathrm{mid}},\dots,\nu_m^{\mathrm{mid}})} + \sum_{i=1}^{m} \frac{g(q_i)}{\nu_i^{\mathrm{mid}}} \tag{4.23}$$

where $\nu_i^{\mathrm{mid}} = \mathbf{E}_{q_i\sim F_i^{\mathrm{mid}}}[g(q_i)]$ is the expectation taken under the null using the mid p-value distribution function F_i^{mid}. For $\widehat{m}_0^{\mathrm{Storey}}$ we have $\nu_i^{\text{mid-Storey}} = 1 - F_i^{\mathrm{mid}}(\lambda) \leq 1 - F_i^{\mathrm{sd}}(\lambda) = \nu_i^{\text{du-Storey}}$ and $g(q_i) = \mathbf{1}\{q_i > \lambda\} \leq \mathbf{1}\{p_i > \lambda\} = g(p_i)$ so that $\widehat{m}_0^{\mathrm{mid}}$ can be smaller or larger than $\widehat{m}_0^{\mathrm{du}}$, depending on the specific constellation. In the case of the PC estimator we have $g(x) = x$ and since $\mathbf{E}q_i = 1/2$ for any mid-p-value (see Berry and Armitage (1995)) we have $\nu_1^{\mathrm{mid}} = \ldots = \nu_m^{\mathrm{mid}} = 1/2$ so that in this case the mid-p estimator has a particularly simple representation. Combining this with the fact that $q_i \leq p_i$ $(a.s.)$ gives us the following result.

Corollary 4.5.2 *Assume that* $p_1,\dots,p_m$ *are mutually independent and super-uniform under the null (i.e.* (SuperUnif) *holds), and let* $q_1,\dots,q_m$ *denote the corresponding mid-p-values. Then*

the mid-p estimator of $\widehat{m}_0^{PC,new}$ *is given by*

$$\widehat{m}_0^{mid\text{-}PC}(q_1,\dots,q_m) = 2 + 2\cdot\sum_{i=1}^{m} q_i$$

and the BH plug-in procedure (4.2) *using* $\widehat{m}_0^{mid\text{-}PC}(q_1,\dots,q_m)$ *controls FDR at level* α*. Moreover,* $\widehat{m}_0^{mid\text{-}PC} \leq \widehat{m_0}$ *(a.s.), where* $\widehat{m}_0$ *is the base non-adjusted estimator* (4.20)*.*

This result implies that for the PC estimator with discrete data we can simply use $2+2\cdot\sum_{i=1}^{m} q_i$ instead of the more conservative $2+2\cdot\sum_{i=1}^{m} p_i$ estimator without losing plug-in FDR control. We point out that the mid-p-values are used exclusively for estimating m_0 in the plug-in procedure defined by (4.2) while the (ordered) standard discrete p-values $p_{(k)}$ are used in the final decision step.

4.5.3 A randomization approach

Here we describe an approach related to Dickhaus et al. (2012) who argue for using randomization methods in estimating m_0 on discrete data. For any estimator $\widehat{m}_0$, not necessarily belonging to $\mathcal{F}_0$ define the associated *expected randomized estimator* as

$$\widehat{m}_0^{\text{rand}}(p_1,\dots,p_m) = \left[\mathbf{E}_{(U_1,\dots,U_m)}\left(\frac{1}{\widehat{m}_0(r(p_1,U_1),\dots,r(p_m,U_m))}\right)\right]^{-1} \tag{4.24}$$

where $U_1,\dots,U_m \sim \mathcal{U}[0,1]$ denote i.i.d uniform random variables independent of $(p_1,\dots,p_m)$. Thus, for fixed $(p_1,\dots,p_m)$ this estimator is obtained by taking the expectation over the randomized p-values associated with $(p_1,\dots,p_m)$. In most cases (4.24) is analytically intractable, we therefore use Monte-Carlo approximation of $\widehat{m}_0^{\text{rand}}$ obtained by averaging a large number of simulations of $\widetilde{m}_0(r(p_1,U_1),\dots,r(p_m,U_m))$ (the vector $(U_1,\dots,U_m)$ is simulated many times, while $(p_1,\dots,p_m)$ is kept fixed). Again, this approach comes with guaranteed FDR plug-in control.

Corollary 4.5.3 *Assume that* $p_1,\dots,p_m$ *are mutually independent and super-uniform under the null (i.e.* (SuperUnif) *holds) and let* $\widehat{m}_0$ *satisfy the conditions of Theorem 4.2.1. Then the BH plug-in procedure* (4.2) *using* $\widehat{m}_0^{rand}(p_1,\dots,p_m)$ *defined by* (4.24) *controls FDR at level* α*. Moreover* $\widehat{m}_0^{rand} \leq \widehat{m}_0$ *(a.s.).*

Proof 4.5.2 *The proof uses Theorem 4.2.1. First, we show that* $\widehat{m}_0^{rand}(p_1,\dots,p_m)$ *is coordinatewise non-decreasing. For fixed* $(u_1,\dots,u_m) \in [0,1]^m$ *each (realized) randomized p-value* $r_i = r(p_i,u_i)$ *is non-decreasing in* p_i*. Since* $\widehat{m}_0 \in \mathcal{F}$ *is coordinatewise non-decreasing in* $(p_1,\dots,p_m)$*, the function* $1/\widehat{m}_0(r(\cdot,u_1),\dots,r(\cdot,u_m))$ *is coordinatewise decreasing for all* $(u_1,\dots,u_m) \in [0,1]^m$ *and so is its expectation which implies that* $\widehat{m}_0^{rand}$ *is coordinatewise non-decreasing. To establish* (IMC)*, we denote for* $h \in \mathcal{H}_0$ *by* $r_{0,h}$ *the set of randomized p-values* $(r_1,\dots,r_m)$*, where* r_h *has been replaced by* 0*. By the definition of* $\widehat{m}_0^{rand}$ *we have*

$$\mathbf{E}_{(p_1,\dots,p_m)}\left[\frac{1}{\widehat{m}_0^{rand}(p_{0,h})}\right] = \mathbf{E}_{(p_1,\dots,p_m)}\left[\mathbf{E}_{(U_1,\dots,U_m)}\frac{1}{\widehat{m}_0(r_{0,h})}\right] = \mathbf{E}_{(r_1,\dots,r_m)}\left[\frac{1}{\widehat{m}_0(r_{0,h})}\right].$$

where the second equality follows from the fact that for super-uniform p-value $p_h = 0$*, the associated randomized p-value* $r(p_h,u) = 0$ *(a.s) by Definition* (4.19)*. Since the* $(r_1,\dots,r_m)$ *are mutually independent and uniform under the null and* $\widehat{m}_0 \in \mathcal{F}$*, the bound* (IMC) *for* $\widehat{m}_0(r_{0,h})$ *now follows since* $\widehat{m}_0$ *satisfies the conditions of Theorem 4.2.1. Therefore, the r.h.s. of the last equation*

can be bounded by $1/m_0$ *and plug-in FDR control for* $\widehat{m}_0^{rand}$ *now follows from Theorem 4.2.1. To see that the last statement of the corollary holds true, observe that since* $\widehat{m}_0$ *is coordinatewise non-decreasing and* $r(p_i, U_i) \leq r(p_i, 0) = p_i$ *we have* $\widehat{m}_0(r(p_1, U_1), \ldots, r(p_m, U_m)) \leq \widehat{m}_0(p_1, \ldots, p_m)$ *and therefore the r.h.s. of* (4.24) *is bounded by* $\widehat{m}_0(p_1, \ldots, p_m)$ *(a.s.).*

Dickhaus et al. (2012) argue for using randomized p-values in (essentially) Storey's estimator, i.e. applying $\widehat{m}_0^{\text{Storey}}$ to $(r_1, \ldots, r_m)$ instead of $(p_1, \ldots, p_m)$ which yields a random estimate that should provide a better estimate for m_0. They show that plugging this estimator into the Bonferroni procedure yields asymptotic control of the Familywise Error Rate (FWER) under certain assumptions. They also point out that if fully reproducible results are desired it may be more appropriate to work with the conditional expectation w.r.t. randomization, i.e. using $\mathbf{E}_{(U_1,\ldots,U_m)}\left(\widehat{m}_0^{\text{Storey}}(r_1, \ldots, r_m)\right)$. Corollary 4.5.3 shows that we can obtain similar guarantees w.r.t. to plug-in FDR control in a finite-sample setting for any estimator $\widehat{m}_0$ satisfying the conditions of Theorem 4.2.1 and in particular for $\widehat{m}_0 \in \mathcal{F}_0$ by using conditional expectation w.r.t. randomization. The slightly complicated form of (4.24) is a natural consequence of Theorem 4.2.1, but if the variance of $\widehat{m}_0(r(p_1, U_1), \ldots, r(p_m, U_m))$ w.r.t. $U_1, \ldots, U_m$ is small we have the approximation $\widehat{m}_0^{\text{rand}}(p_1, \ldots, p_m) \approx \mathbf{E}_{(U_1,\ldots,U_m)}\widehat{m}_0(r(p_1, U_1), \ldots, r(p_m, U_m))$.

4.5.4 Simulation results

In this section, we analyze how the discrete adjustments can improve the base estimators on simulated data. More specifically, we follow Döhler et al. (2018) by simulating a two-sample problem in which a vector of $m = 500$ independent binary responses is observed for $N = 25$ subjects in both groups. The goal is to test the m null hypotheses $H_{0,i}$: '$p_{1i} = p_{2i}$', $i = 1, ..., m$ where p_{1i} and p_{2i} are the success probabilities for the i^{th} binary response in group A and B respectively. Thus, for each hypothesis i, the data can be summarized by a 2×2 contingency table, and we use (two-sided) Fisher's exact tests (FETs) for testing H_{0i}. The m hypotheses are split in three groups of size m_1, m_2, and m_3 such that $m = m_1 + m_2 + m_3$. Then, the binary responses are generated as i.i.d Bernoulli of probability 0.01 ($\mathbf{Bin}(1, 0.01)$) at m_1 positions for both groups, i.i.d $\mathbf{Bin}(1, 0.10)$ at m_2 positions for both groups, and i.i.d $\mathbf{Bin}(1, 0.10)$ at m_3 positions for one group and i.i.d $\mathbf{Bin}(1, p_3)$ at m_3 positions for the other group. Thus, null hypotheses are true for $m_1 + m_2$ positions, while they are false for m_3 positions (set $\mathcal{H}_1$). We interpret p_3 as the strength of the signal and set it to 0.4, while $\pi_1 = \frac{m_3}{m}$, corresponds to the proportion of signal. Also, m_1 and m_2 are both taken equal to $\frac{m-m_3}{2}$.

We first compare the base estimators $\widehat{m}_0^{\text{Storey}}$ (4.3), $\widehat{m}_0^{\text{PC,new}}$ (4.8), and $\widehat{m}_0^{\text{Poly}}(2, 1/2)$ (4.14) with their standard discrete rescaled versions. Figure 4.3 displays the estimation results for a grid of true $\pi_0 \in \{0.1, \ldots, 0.9\}$. We can see that incorporating discreteness leads to considerable improvements for all estimators over the entire range of π_0 values. This is particularly relevant for large values of π_0 where the base estimators may lead to a strong deterioration in the power of the plug-in BH procedure. On another note, among the base estimators we can see that $\widehat{m}_0^{\text{Poly}}(2, 1/2)$ performs poorly compared to the results of Section 4.4.1. This seems plausible since a large portion of p-values are equal to 1 in the discrete setting in contrast to the Gaussian setting. For these p-values, the contribution in the $\widehat{m}_0^{\text{Poly}}(2, 1/2)$ estimator is equal to the constant $\nu = \frac{3}{1-1/2^3} = \frac{24}{7}$ which is much larger than the corresponding contribution of $\nu = 2$ in the Storey estimator.

In a second step, we compare the different discrete adjustments $\widehat{m}_0^{\text{du}}$ as in (4.22), $\widehat{m}_0^{\text{mid}}$ as in (4.23), and $\widehat{m}_0^{\text{rand}}$ as in (4.24) in Figure 4.4, where we display the estimation results for three

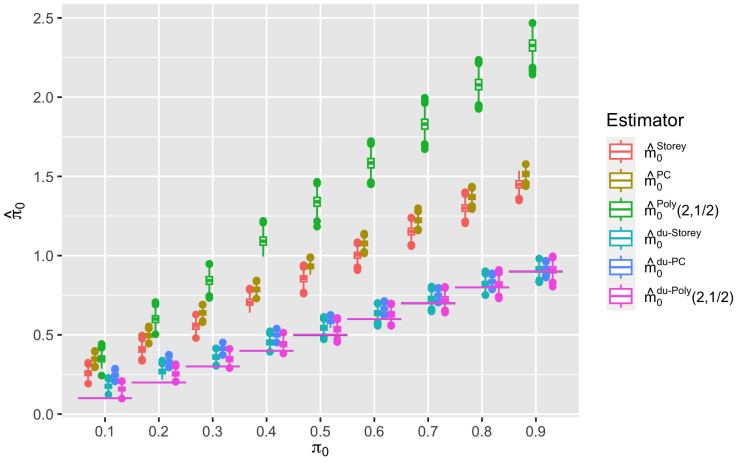

Figure 4.3: Comparison between base estimators $\widehat{m}_0^{\text{Storey}}$, $\widehat{m}_0^{\text{PC,new}}$ and $\widehat{m}_0^{\text{Poly}}(2, 1/2)$ and their standard discrete rescaled versions on simulated data.

values of true $\pi_0 \in \{0.2, 0.5, 0.7\}$. We can see that there are no relevant differences between the different adjustments. Therefore, there is no strong reason to advocate a specific type of adjustment since they yield similar outcomes.

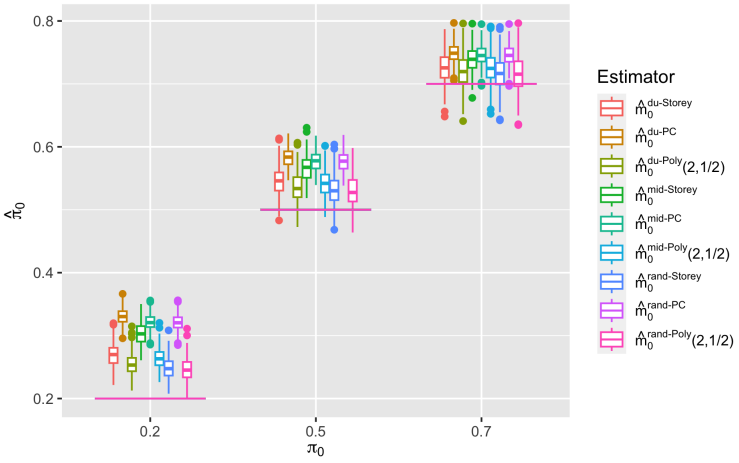

Figure 4.4: Comparison between discrete adjustments $\widehat{m}_0^{\text{du}}$, $\widehat{m}_0^{\text{mid}}$ and $\widehat{m}_0^{\text{rand}}$ on simulated data.

4.5.5 Real data analysis

Finally, we compare the performance of base and discrete estimators on three different datasets. The first dataset consists of data provided by the International Mice Phenotyping Consortium (IMPC) (Karp et al., 2017), which coordinates studies on the genotype influence on mouse phenotype. This dataset includes, for each of the $m = 266952$ studied genes, the counts of normal

and abnormal phenotypes thus providing multiple two by two contingency tables, which can be analysed using FETs. Then we analyze the methylation dataset for cytosines of Arabidopsis in Lister et al. (2008) which is part of the R-package `fdrDiscreteNull` of Chen and Doerge (2015). This dataset contains $m = 3525$ counts for a biological entity under two different biological conditions or treatments also analyzed using FETs. Finally, the third dataset, provided by the Regulatory Agency in the United Kingdom, includes adverse drug reactions due to medicines and healthcare products. It contains the number of reported cases of amnesia as well as the total number of adverse events reported for each of the $m = 2446$ drugs in the database. For more details we refer to Heller and Gur (2011) and to the accompanying R-package `discreteMTP` of Heller et al. (2012), which also contains the data. Heller and Gur (2011) investigate the association between reports of amnesia and suspected drugs by performing for each drug (one-sided) FETs.

From the results in Table 4.1 we can see that taking discreteness into account is always beneficial, regardless of the adjustment used. Depending on the type of discreteness and the amount of signal contained in the data, adjusting for discreteness can provide a great improvement in some cases. Indeed, as the example of the IMPC data shows, base estimators may not be able to recognize the presence of any alternatives. However, the discrete estimators clearly suggest that a considerable amount of alternatives is present.

Table 4.1: π_0-estimates for base estimators and adjusted discrete estimators on three different datasets containing discrete data.

		Dataset		
Adjustment	Estimator	IMPC	Arabidopsis	Pharmacovigilance
standard (none)	Storey	1.26	0.67	1.79
	PC	1.26	0.73	1.79
	Poly$(2, 1/2)$	2.16	0.75	2.97
rescaled (du)	Storey	0.63	0.59	1.05
	PC	0.63	0.64	1.04
	Poly$(2, 1/2)$	0.63	0.57	1.10
rescaled (mid)	Storey	0.63	0.63	1.03
	PC	0.63	0.64	1.05
	Poly$(2, 1/2)$	0.63	0.58	1.11
randomized	Storey	0.63	0.58	1.08
	PC	0.63	0.64	1.06
	Poly$(2, 1/2)$	0.63	0.56	1.14

4.6 Discussion

In this paper we introduce a unified class of m_0-estimators with mathematical guarantees on plug-in FDR control in a classical setting. We also describe some general approaches for adjusting m_0-estimators constructed for continuous p-values to discrete p-values. While we show that these results are useful both from a methodological viewpoint and for practical purposes, there a numerous possibilities for further investigations, some of which we describe now.

While we focus on FDR control in this paper, it is clear, that our new estimators can also be used for FDR estimation, see Storey (2002). For the discrete estimators from Section 4.5 that are uniformly better than their classical counterparts, this implies that the corresponding FDR estimators are uniformly better es well.

In Section 4.4, we describe the performance of various polynomial estimators in a one-sided Gaussian testing framework with the aim of illustrating the flexibility of our Proposition 4.3.1 on plug-in FDR control. The numerical results in Section 4.4.1 show that $\widehat{m}_0^{\mathrm{Poly}}(2, 1/2)$ performs uniformly better in terms of MSE than the other estimators in this specific framework. In a second step, it might be interesting to study whether optimal estimators can be derived, either within the whole class $\mathcal{F}_0$ or perhaps within some sub-class like polynomial estimators. This way, it may eventually be possible to obtain more efficient estimators in practice, or at least give the user some guidance for choosing estimators from the class $\mathcal{F}_0$.

We would also like to point out some connections to the work of Heesen and Janssen (2016), who split the unit interval into an estimation region on which an estimator of m_0 is constructed and a rejection region on which the BH procedure is run. Thus these estimators do not use all available p-values, in contrast to our approach. Heesen and Janssen (2016) derive a general sufficient criterion, similar to Theorem 4.2.1, for finite sample plug-in FDR control which they apply to Storey-type estimators and histogram-type estimators (see MacDonald et al. (2019)). In contrast to our approach, the transformation function g applied to p-values need not be monotone, however it is unclear whether e.g. smooth functions fit into this framework. Their approach also accommodates "dynamization", which allows data-dependent tuning of parameters, see MacDonald et al. (2019). We may wonder if this approach can be used for the estimators of our class, but as this question exceeds the scope of this paper, we leave it for future research.

While we use polynomial estimators as simple examples that include both the classical Storey and Pounds and Cheng estimators, other choices are conceivable. Taking, for instance, certain kernel-type transformation functions would lead to estimators that are advantageous from an asymptotic viewpoint, see Neuvial (2013). We leave this topic for future research.

In Section 4.5 we illustrate how information on the null distribution functions of discrete p-values can be used to obtain more efficient m_0-estimators. This information can be seen as a special case of auxiliary covariates, for which it is well-known that their incorporation into multiple testing procedures e.g. by weighting, can be highly beneficial (see Ignatiadis et al. (2016); Durand (2019)). We would like to mention that our methods for m_0-estimation are not limited to the special case of discrete p-values but should be able to acomodate these types of heterogeneity as well. Thus, they might also be useful in the settings described above for obtaining more efficient m_0-estimators.

While we assume independence of the p-values throughout this paper, it is well-known that dependence may adversely affect the performance of m_0-estimators or may require re-adjustments of tuning parameters like λ in $\widehat{m}_0^{\mathrm{Storey}}$ (see e.g. Blanchard and Roquain (2009)). Thus, it might be interesting to investigate the behaviour of our new estimators under various types of dependency.

Constructing multiple testing procedures for discrete data that provide finite sample plug-in FDR control is challenging. In this paper we make some progress by obtaining improved discrete estimators for m_0. While using these discrete estimators in the plug-in BH procedure

provide more power than using classical estimators (based on uniformly distributed p-values under the null), it is still not ideal because the discreteness is ignored in the rejection stage of the procedure. Döhler et al. (2018) propose discrete variants of the standard (i.e. non plug-in) BH procedure. They also sketch a possible plug-in method based on combining this procedure with estimators of m_0, but caution that it comes without mathematical guarantees. Thus, as MacDonald et al. (2019) pointed out, it still remains an open problem to develop procedures that integrate discreteness of the data in both the estimation of m_0 and the rejection of p-values.

Conclusion and perspectives

The work presented in this manuscript focuses on improving state-of-the-art MT methods in different contexts involving specific types of data structures. In each of the three chapters, we focus on three types of MT methods revolving around online procedures, confidence bounds, and estimation. The improvement of these methods can be described in terms of power gain (Chapters 2 and 4), asymptotic consistency (Chapter 3), and closure of theoretical gaps (Chapter 4). To achieve these improvements we sometimes leveraged information available in the setting of interest or specific mathematical tools: in Chapters 2 and 4 we used the available knowledge on the discrete p-value distribution to compensate the power loss when dealing with discrete p-values. In Chapter 2 this allowed to improve online testing procedures for discrete data and to handle online weighting. This is achieved by incorporating a quantity called the super-uniformity reward, which roughly speaking, represents the amount of wasted testing level due to the super-uniformity of the discrete p-values. In Chapter 3, we used existing and novel concentration inequalities to derive asymptotically consistent FDP bounds. This step further narrows the gap between FDR control and FDP confidence bounds, as consistent bounds offer an asymptotic control at level α of the FDP with high probability. These improvements allow to extend classical MT methods to different contexts that are ubiquitous in real-life applications.

Nevertheless, the current work has also limitations. First, the presented results mostly hold under the assumption of independence between the p-values, which is quite unrealistic in real life. Recently, Wang and Ramdas (2022) introduced the eBH procedure which corresponds to the BH procedure using e-values. An E-value is any statistic for which the expectation is less than 1 under the null hypothesis. For instance, in a parametric setting (where we have a specific value for the tested parameter under the alternative), an e-value can be set as the likelihood ratio. E-values are very appealing since they allows to seamlessly handle dependencies and sequentiality. Indeed, the definition based on the expectation allows to easily verify the FDR control for the eBH under any dependency assumption between the p-values, see Wang and Ramdas (2022) for more details. Additionally, many efforts have been made in the last years to transfer p-value based results into e-value based results, see e.g. the works of Xu et al. (2022) on e-value based FDP bounds or Ignatiadis et al. (2022) on e-value weighting. Further investigations could be undertaken to propose a closed form formula of e-values in the discrete setting.

Second, power results can be desirable to quantify rigorously the amount of power improvement and help assess whether there is further room for enhancement, particularly in the online and discrete settings. In the canonical setting, theoretical analyses of power have been conducted by Donoho and Jin (2004) and Abraham et al. (2021). They investigate the signal strength, deriving boundaries between regimes where achieving good power while maintaining low risk is possible and where no detection is possible without causing high risk. These kinds of studies could be relevant in the online setting but might be much more demanding as the results cannot depend on the unknown number of tests m. Additionally, since the online setting models a context with an ever-growing number of hypotheses tested, an exhaustive model might need

to encompass a dynamic signal trend over time. Nonetheless, some power results have been presented in the online literature. For instance, Tian and Ramdas (2021) conducted a power comparison for online FWER controlling procedures, while Javanmard and Montanari (2018) proposed optimal spending sequence $\{\gamma_t\}_{t\geq 1}$ for the LORD procedure in the context of Gaussian mean testing. Extending these investigations to the online discrete setting could be valuable for determining the optimal spending and investing sequences $\{\gamma_t\}_{t\geq 1}$, $\{\gamma'_t\}_{t\geq 1}$ or quantifying the power gains of rewarded procedures. However, obtaining power results in the discrete context is challenging, and as of now, there are no such available results. This is because conducting a power study in a general discrete context (not necessarily limited to the context of the FET) requires modeling the discreteness of the p-values, and there are currently no established guidelines on how to model this mechanism. One potential option might involve using a Poisson process to generate jumps for a function that acts as a proxy mimicking the discrete F_i in different regimes.

Bibliography

Abraham, K., Castillo, I., and Roquain, E. (2021). Sharp multiple testing boundary for sparse sequences. *arXiv preprint arXiv:2109.13601*.

Aharoni, E. and Rosset, S. (2014). Generalized alpha-investing: definitions, optimality results and application to public databases. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(4):771–794.

Ahmed, I., Dalmasso, C., Haramburu, F., Thiessard, F., Broët, P., and Tubert-Bitter, P. (2010). False discovery rate estimation for frequentist pharmacovigilance signal detection methods. *Biometrics*, 66(1):301–309.

Bacon, R., Accardo, M., Adjali, L., Anwand, H., Bauer, S., Biswas, I., Blaizot, J., Boudon, D., Brau-Nogue, S., Brinchmann, J., et al. (2010). The muse second-generation vlt instrument. In *Ground-based and Airborne Instrumentation for Astronomy III*, volume 7735, pages 131–139. SPIE.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Bates, S., Candès, E., Lei, L., Romano, Y., and Sesia, M. (2023). Testing for outliers with conformal p-values. *The Annals of Statistics*, 51(1):149–178.

Benjamini, Y. (2010). Simultaneous and selective inference: Current successes and future challenges. *Biometrical Journal*, 52(6):708–721.

Benjamini, Y. and Braun, H. (2002). John w. tukey's contributions to multiple comparisons. *Annals of Statistics*, pages 1576–1594.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B*, 57(1):289–300.

Benjamini, Y. and Hochberg, Y. (1997). Multiple hypotheses testing with weights. *Scandinavian Journal of Statistics*, 24(3):407–418.

Benjamini, Y., Krieger, A. M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, 93(3):491–507.

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

Berman, R., Pekelis, L., Scott, A., and Van den Bulte, C. (2018). P-hacking and false discovery in a/b testing. *B Testing (December 11, 2018)*.

Berry, G. and Armitage, P. (1995). Mid-p confidence intervals: a brief review. *Journal of the Royal Statistical Society Series D: The Statistician*, 44(4):417–423.

Biswas, A. and Chattopadhyay, G. (2020). Modified estimator for the proportion of true null hypotheses under discrete setup with proven fdr control by the adaptive benjamini-hochberg procedure. *arXiv preprint arXiv:2009.03803*.

Blanchard, G., Delattre, S., and Roquain, E. (2014). Testing over a continuum of null hypotheses with false discovery rate control. *Bernoulli*, 20(1):304 – 333.

Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281–1303.

Blanchard, G. and Roquain, E. (2008). Two simple sufficient conditions for FDR control. *Electron. J. Stat.*, 2:963–992.

Blanchard, G. and Roquain, É. (2009). Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(12).

Bogdan, M., Chakrabarti, A., Frommlet, F., and Ghosh, J. K. (2011). Asymptotic bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–1579.

Candès, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-x' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Chavant, F., Favrelière, S., Lafay-Chebassier, C., Plazanet, C., and Pérault-Pochat, M.-C. (2011). Memory disorders associated with consumption of drugs: updating through a case/noncase study in the french pharmacovigilance database. *British journal of clinical pharmacology*, 72(6):898–904.

Chen, S. and Arias-Castro, E. (2021). On the power of some sequential multiple testing procedures. *Ann. Inst. Stat. Math.*, 73(2):311–336.

Chen, S. and Kasiviswanathan, S. (2020). Contextual online false discovery rate control. In *International Conference on Artificial Intelligence and Statistics*, pages 952–961. PMLR.

Chen, X. and Doerge, R. (2015). fdrdiscretenull: False discovery rate procedure under discrete null distributions. *R package version*, 1.

Chen, X., Doerge, R. W., and Heyse, J. F. (2018). Multiple testing with discrete data: proportion of true null hypotheses and two adaptive FDR procedures. *Biom. J.*, 60(4):761–779.

Chen, X., Doerge, R. W., and Sarkar, S. K. (2015). A weighted FDR procedure under discrete and heterogeneous null distributions. *arXiv e-prints*, page arXiv:1502.00973.

Chi, Z., Ramdas, A., and Wang, R. (2022). Multiple testing under negative dependence. *arXiv preprint arXiv:2212.09706*.

Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C. (2021). *Handbook of multiple comparisons*. CRC Press.

Dickhaus, T., Straßburger, K., Schunk, D., Morcillo-Suarez, C., Illig, T., and Navarro, A. (2012). How to analyze many contingency tables simultaneously in genetic association studies. *Statistical applications in genetics and molecular biology*, 11(4).

Ditzhaus, M. and Janssen, A. (2019). Variability and stability of the false discovery proportion. *Electronic Journal of Statistics*, 13(1):882–910.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2009). *Multiple testing problems in pharmaceutical statistics*. CRC press.

Döhler, S. (2016). A discrete modification of the Benjamini—Yekutieli procedure. *Econometrics and Statistics*.

Döhler, S., Durand, G., and Roquain, E. (2018). New fdr bounds for discrete and heterogeneous tests. *Electronic Journal of Statistics*, 12(1):1867–1900.

Döhler, S. and Roquain, E. (2020). Controlling the false discovery exceedance for heterogeneous tests. *Electronic Journal of Statistics*, 14(2):4244 – 4272.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics*, 32(3):962 – 994.

Dümbgen, L. and Wellner, J. A. (2023). A new approach to tests and confidence bands for distribution functions. *The Annals of Statistics*, 51(1):260–289.

Dumusque, X., Pepe, F., Lovis, C., Ségransan, D., Sahlmann, J., Benz, W., Bouchy, F., Mayor, M., Queloz, D., Santos, N., et al. (2012). An earth-mass planet orbiting α centauri b. *Nature*, 491(7423):207–211.

Durand, G. (2019). Adaptive p-value weighting with power optimality. *Electron. J. Statist.*, 13(2):3336–3385.

Durand, G., Blanchard, G., Neuvial, P., and Roquain, E. (2020). Post hoc false positive control for structured hypotheses. *Scandinavian journal of Statistics*, 47(4):1114–1148.

Durand, G., Junge, F., Döhler, S., and Roquain, E. (2019). DiscreteFDR: An R package for controlling the false discovery rate for discrete test statistics. *arXiv e-prints*, page arXiv:1904.02054.

Embrechts, P. and Hofert, M. (2013). A note on generalized inverses. *Mathematical Methods of Operations Research*, 77(3):423–432.

Foster, D. P. and Stine, R. A. (2008). Alpha-investing: a procedure for sequential control of expected false discoveries. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(2):429–444.

Freedman, D. A. (1975). On tail probabilities for martingales. *The Annals of Probability*, pages 100–118.

Gang, B., Sun, W., and Wang, W. (2020). Structure-adaptive sequential testing for online false discovery rate control.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian data analysis*. Chapman and Hall/CRC Texts in Statistical Science Series. CRC, Boca Raton, Florida, third edition.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *The annals of statistics*, 32(3):1035–1061.

Genovese, C. R., Roeder, K., and Wasserman, L. (2006). False discovery control with p-value weighting. *Biometrika*, 93(3):509–524.

Genovese, C. R. and Wasserman, L. (2006). Exceedance control of the false discovery proportion. *Journal of the American Statistical Association*, 101(476):1408–1417.

Gilbert, P. B. (2005). A modified false discovery rate multiple-comparisons procedure for discrete data, applied to human immunodeficiency virus genetics. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):143–158.

Giraud, C. (2021). *Introduction to high-dimensional statistics*. CRC Press.

Goeman, J. J., Hemerik, J., and Solari, A. (2021). Only closed testing procedures are admissible for controlling false discovery proportions. *The Annals of Statistics*, 49(2):1218 – 1238.

Goeman, J. J., Meijer, R. J., Krebs, T. J., and Solari, A. (2019). Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856.

Goeman, J. J. and Solari, A. (2011). Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597.

Gorgolewski, K. J., Varoquaux, G., Rivera, G., Schwarz, Y., Ghosh, S. S., Maumet, C., Sochat, V. V., Nichols, T. E., Poldrack, R. A., Poline, J.-B., et al. (2015). Neurovault. org: a web-based repository for collecting and sharing unthresholded statistical maps of the human brain. *Frontiers in neuroinformatics*, 9:8.

Guo, W. and Romano, J. (2007). A generalized sidak-holm procedure and control of generalized error rates under independence. *Statistical applications in genetics and molecular biology*, 6(1).

Habiger, J. D. (2015). Multiple test functions and adjusted p-values for test statistics with discrete distributions. *Journal of Statistical Planning and Inference*, 167:1–13.

Habiger, J. D. and Pena, E. A. (2011). Randomised p-values and nonparametric procedures in multiple testing. *Journal of nonparametric statistics*, 23(3):583–604.

Heesen, P. and Janssen, A. (2016). Dynamic adaptive multiple tests with finite sample fdr control. *Journal of Statistical Planning and Inference*, 168:38–51.

Heller, R. and Gur, H. (2011). False discovery rate controlling procedures for discrete tests. *ArXiv e-prints*.

Heller, R. and Gur, H. (2011). False discovery rate controlling procedures for discrete tests. *arXiv preprint arXiv:1112.4627*, pages 00092–17.

Heller, R., Gur, H., and Yaacoby, S. (2012). discretemtp: Multiple testing procedures for discrete test statistics. *R package version 0.1-2*.

Hemerik, J., Solari, A., and Goeman, J. J. (2019). Permutation-based simultaneous confidence bounds for the false discovery proportion. *Biometrika*, 106(3):635–649.

Heyse, J. F. (2011). A false discovery rate procedure for categorical data. In *Recent Advances in Biostatistics: False Discovery Rates, Survival Analysis, and Related Topics*, pages 43–58.

Hirji, K. F. (2005). *Exact analysis of discrete data*. CRC Press.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Statist.*, 6(2):65–70.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. (2021). Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055 – 1080.

Hu, J. X., Zhao, H., and Zhou, H. H. (2010). False discovery rate control with groups. *J. Amer. Statist. Assoc.*, 105(491):1215–1227.

Ignatiadis, N., Klaus, B., Zaugg, J. B., and Huber, W. (2016). Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580.

Ignatiadis, N., Wang, R., and Ramdas, A. (2022). E-values as unnormalized weights in multiple testing. *arXiv preprint arXiv:2204.12447*.

James, N. D., Sydes, M. R., Clarke, N. W., Mason, M. D., Dearnaley, D. P., Anderson, J., Popert, R. J., Sanders, K., Morgan, R. C., Stansfeld, J., et al. (2008). Stampede: Systemic therapy for advancing or metastatic prostate cancer—a multi-arm multi-stage randomised controlled trial. *Clinical Oncology*, 20(8):577–581.

Jamieson, K., Malloy, M., Nowak, R., and Bubeck, S. (2014). lil'UCB: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, pages 423–439. PMLR.

Javanmard, A. and Montanari, A. (2018). Online rules for control of false discovery rate and false discovery exceedance. *The Annals of statistics*, 46(2):526–554.

Johari, R., Pekelis, L., and Walsh, D. J. (2019). Always valid inference: Bringing sequential analysis to a/b testing.

Johnson, K. D., Stine, R. A., and Foster, D. P. (2020). Fitting high-dimensional interaction models with error control.

Johnson, N. L., Kotz, S., and Johnson, N. L. (1970). Continuous univariate distributions.

Karp, N. A., Mason, J., Beaudet, A. L., Benjamini, Y., Bower, L., Braun, R. E., Brown, S. D., Chesler, E. J., Dickinson, M. E., Flenniken, A. M., et al. (2017). Prevalence of sexual dimorphism in mammalian phenotypic traits. *Nature communications*, 8(1):15475.

Katsevich, E. and Ramdas, A. (2020). Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, 48(6):3465–3487.

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, page 1168?1176, New York, NY, USA. Association for Computing Machinery.

Kohavi, R., Tang, D., Xu, Y., Hemkens, L. G., and Ioannidis, J. P. (2020). Online randomized controlled experiments at scale: lessons and extensions to medicine. *Trials*, 21(1):1–9.

Lark, R. M. (2017). Controlling the marginal false discovery rate in inferences from a soil dataset with α-investment. *European Journal of Soil Science*, 68(2):221–234.

Lehmann, E. and Romano, J. P. (2022). *Testing Statistical Hypotheses*. Springer, Cham.

Lehmann, E. L. and Romano, J. P. (2005). Generalizations of the familywise error rate. *The Annals of Statistics*, 33(3):1138 – 1154.

Lei, L. and Fithian, W. (2016). Power of ordered hypothesis testing. PMLR.

Li, A. and Barber, R. F. (2017). Accumulation tests for fdr control in ordered hypothesis testing. *Journal of the American Statistical Association*, 112(518):837–849.

Li, J., Maathuis, M. H., and Goeman, J. J. (2022). Simultaneous false discovery proportion bounds via knockoffs and closed testing. *arXiv preprint arXiv:2212.12822*.

MacDonald, P. W., Liang, K., and Janssen, A. (2019). Dynamic adaptive procedures that control the false discovery rate.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Machine learning meets false discovery rate. *arXiv preprint arXiv:2208.06685*.

Mary, D., Bacon, R., Conseil, S., Piqueras, L., and Schutz, A. (2020). Origin: Blind detection of faint emission line galaxies in muse datacubes. *Astronomy & Astrophysics*, 635:194.

Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283.

Meinshausen, N. (2006). False discovery control for multiple tests of association under general dependence. *Scandinavian Journal of Statistics*, 33(2):227–237.

Meinshausen, N. and Bühlmann, P. (2005). Lower bounds for the number of false null hypotheses for multiple testing of associations under general dependence structures. *Biometrika*, 92(4):893–907.

Meinshausen, N. and Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *The Annals of Statistics*, 34(1):373 – 393.

Muñoz-Fuentes, V., Cacheiro, P., Meehan, T. F., Aguilar-Pimentel, J. A., Brown, S. D., Flenniken, A. M., Flicek, P., Galli, A., Mashhadi, H. H., Hrabě de Angelis, M., et al. (2018). The international mouse phenotyping consortium (impc): a functional catalogue of the mammalian genome that informs conservation. *Conservation genetics*, 19:995–1005.

Neuvial, P. (2008). Asymptotic properties of false discovery rate controlling procedures under independence. *Electron. J. Statist.*, 2:1065–1110.

Neuvial, P. (2013). Asymptotic results on adaptive false discovery rate controlling procedures based on kernel estimators. *Journal of Machine Learning Research*, 14:1423–1459.

Neuvial, P. and Roquain, E. (2012). On false discovery rate thresholding for classification under sparsity. *The Annals of Statistics*, 40(5):2572–2600.

Nowinski, W. L. (2021). Evolution of human brain atlases in terms of content, applications, functionality, and availability. *Neuroinformatics*, 19(1):1–22.

Perrot-Dockès, M., Blanchard, G., Neuvial, P., and Roquain, E. (2021). Post hoc false discovery proportion inference under a hidden markov model. *arXiv preprint arXiv:2105.00288*.

Pounds, S. and Cheng, C. (2006). Robust estimation of the false discovery rate. *Bioinformatics*, 22(16):1979–1987.

Ramdas, A. (2019). Foundations of large-scale sequential experimentation. T16 tutorial at the 25th ACM SIGKDD conference on knowledge discovery and data mining.

Ramdas, A., Yang, F., Wainwright, M. J., and Jordan, M. I. (2017). Online control of the false discovery rate with decaying memory. *Advances in neural information processing systems*, 30.

Ramdas, A., Zrnic, T., Wainwright, M. J., and Jordan, M. I. (2018). SAFFRON: an adaptive algorithm for online control of the false discovery rate. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 4283–4291. PMLR.

Ramdas, A. K., Barber, R. F., Wainwright, M. J., and Jordan, M. I. (2019). A unified treatment of multiple testing with prior knowledge using the p-filter. *The Annals of Statistics*, 47(5):2790–2821.

Robbins, H. (1954). A one-sided confidence interval for an unknown distribution function. *Annals of Mathematical Statistics*, 25(2):409–409.

Robertson, D. S., Wason, J., and Ramdas, A. (2022). Online multiple hypothesis testing for reproducible research. *arXiv preprint arXiv:2208.11418*.

Robertson, D. S., Wildenhain, J., Javanmard, A., and Karp, N. A. (2019). onlineFDR: an R package to control the false discovery rate for growing data repositories. *Bioinformatics*, 35(20):4196–4199.

Roeder, K. and Wasserman, L. (2009). Genome-wide significance levels and weighted hypothesis testing. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 24(4):398.

Roquain, E. and Van de Wiel, M. (2008). Optimal weighting for false discovery rate control. *Electronic Journal of Statistics*, 3.

Rousson, V. (2013). *Statistique appliquée aux sciences de la vie*, chapter 11. Springer.

Rubin, D., Dudoit, S., and van der Laan, M. (2006). A method to increase the power of multiple testing procedures through sample splitting. *Stat. Appl. Genet. Mol. Biol.*, 5:Art. 19, 20 pp. (electronic).

Rubin-Delanchy, P., Heard, N. A., and Lawson, D. J. (2019). Meta-analysis of mid-p-values: some new results based on the convex order. *Journal of the American Statistical Association*, 114(527):1105–1112.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, pages 135–168.

Shaked, M. and Shantikumar, J. G. (2007). *Stochastic orders*. Springer.

Shorack, G. R. and Wellner, J. A. (2009). *Empirical processes with applications to statistics*. SIAM.

Šidák, Z. (1967). Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association*, 62(318):626–633.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Storey, J. D., Taylor, J. E., and Siegmund, D. (2004). Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205.

Sun, W. and Tony Cai, T. (2009). Large-scale multiple testing under dependence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):393–424.

Tan, X., Liu, G., Zeng, D., Wang, W., Diao, G., Heyse, J., and Ibrahim, J. (2019). Controlling false discovery proportion in identification of drug-related adverse events from multiple system organ classes. *Statistics in Medicine*, 38.

Tarone, R. E. (1990). A modified bonferroni method for discrete data. *Biometrics*, 46(2):515–522.

Tian, J. and Ramdas, A. (2019). ADDIS: an adaptive discarding algorithm for online FDR control with conservative nulls. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 9383–9391.

Tian, J. and Ramdas, A. (2021). Online control of the familywise error rate. *Statistical Methods in Medical Research*, 30(4):976–993. PMID: 33413033.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Uffelmann, E., Huang, Q. Q., Munung, N. S., De Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):59.

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B., and Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):1006565.

Vesely, A., Finos, L., and Goeman, J. J. (2021). Permutation-based true discovery guarantee by sum tests. *arXiv preprint arXiv:2102.11759*.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

Wasserman, L. and Roeder, K. (2006). Weighted hypothesis testing. Technical report, Dept. of statistics, Carnegie Mellon University.

Weinstein, A. and Ramdas, A. (2020). Online control of the false coverage rate and false sign rate. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10193–10202. PMLR.

Westfall, P. and Wolfinger, R. (1997). Multiple tests with discrete distributions. *The American Statistician*, 51(1):3–8.

Wu, X. and Ye, Y. (2006). Exploring gene causal interactions using an enhanced constraint-based method. *Pattern Recognition*, 39(12):2439–2449.

Xu, Z. and Ramdas, A. (2021). Dynamic algorithms for online multiple testing.

Xu, Z., Wang, R., and Ramdas, A. (2022). Post-selection inference for e-value based confidence intervals. *arXiv preprint arXiv:2203.12572*.

Yang, F., Ramdas, A., Jamieson, K., and Wainwright, M. J. (2017a). A framework for multi-a(rmed)/b(andit) testing with online fdr control.

Yang, H., Li, S., Cao, H., Zhang, C., and Cui, Y. (2017b). Predicting disease trait with genomic data: a composite kernel approach. *Briefings in Bioinformatics*, 18(4):591–601.

Zeisel, A., Zuk, O., and Domany, E. (2011). Fdr control with adaptive procedures and fdr monotonicity. *The Annals of applied statistics*, pages 943–968.

Zhang, W., Kamath, G., and Cummings, R. (2020). Paprika: Private online false discovery rate control.

Zhao, H. and Zhang, J. (2014). Weighted p-value procedures for controlling FDR of grouped hypotheses. *J. Statist. Plann. Inference*, 151/152:90–106.

Zrnic, T., Jiang, D., Ramdas, A., and Jordan, M. (2020). The power of batching in multiple hypothesis testing. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 3806–3815. PMLR.

Zrnic, T., Ramdas, A., and Jordan, M. I. (2021). Asynchronous online testing of multiple hypotheses. *J. Mach. Learn. Res.*, 22:33:1–33:39.

Appendix A

Supplementary material for Chapter 2

Outline of the current chapter

A.1 Proofs

A.1.1 Proofs for online FWER control

We start by proving Theorem 2.3.3 and then deduce Theorems 2.3.1 and 2.3.2.

Proof A.1.1 (Proof of Theorem 2.3.3) *First, let us show that for any critical values* $(\alpha_t, t \geq 1)$*, a sufficient condition for FWER control under* (2.2) *is given by*

$$\alpha_T + \sum_{t=1}^{T-1} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t) \leq (1-\lambda)\alpha \qquad \textit{(a.s.)} \tag{A.1}$$

if either (2.3) *or if* $(\alpha_t, t \geq 1)$ *are deterministic for all* $T \geq 1$*. This comes from Markov's inequality combined with Lemma A.1.1:*

$$\begin{aligned}
\mathrm{FWER}(T, \mathcal{A}, P) &\leq \mathbf{E}_{X\sim P}\Big(\sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0(P), p_t \leq \alpha_t\} \Big) \\
&\leq (1-\lambda)^{-1}\, \mathbf{E}\left(\sum_{t=1}^{T} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t) \right) \\
&\leq (1-\lambda)^{-1}\, \mathbf{E}\left(\alpha_T + \sum_{t=1}^{T-1} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t) \right),
\end{aligned}$$

which gives the announced sufficient condition. Now, we obtain statement (i) of the theorem by verifying the above criterion (A.1) *for* α_t^0 *using the (crude) bound* $F_t(x) \leq x$ *and assumption* (2.19)*. Next, we obtain statement (ii) of the theorem by verifying the above criterion* (A.1) *for* α_t*. This is done by reducing this to a statement on* α_t^0 *via Lemma A.1.2. More precisely, with* $a_T = \sum_{t=1}^{T} \gamma'_t$*, we have*

$$\begin{aligned}
\alpha_T + \sum_{t=1}^{T-1} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t) &\leq \alpha_T + \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\} \left[(1 - a_{T-t})\alpha_t + a_{T-t} F_t(\alpha_t) \right] \\
&= \alpha_T^0 + \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\} \alpha_t^0 \leq \alpha,
\end{aligned}$$

where the equality above is true provided that the following recursion holds for all $T \geq 1$*,*

$$\alpha_T = \alpha_T^0 + \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\} \alpha_t^0 - \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\} \left[(1 - a_{T-t})\alpha_t + a_{T-t} F_t(\alpha_t) \right].$$

This is true by Lemma A.1.2 because of the expression (2.20) *of* α_t*. This concludes the proof.*

Proof A.1.2 (Proof of Theorems 2.3.1 and 2.3.2) *Theorems 2.3.1 and 2.3.2 are corollaries of Theorem 2.3.3, by considering* $\mathcal{A}^0 = \mathcal{A}^{OB}$ *(*$\lambda = 0$*) and* $\mathcal{A}^0 = \mathcal{A}^{AOB}$*, respectively. Indeed, checking* (2.19) *is straightforward for* $\mathcal{A}^{OB}$ *from the spending sequence definition or comes from Lemma A.1.3 for* $\mathcal{A}^{AOB}$*.*

A.1.2 Proofs for online mFDR control

The global proof strategy is similar to the one used for FWER: we start by proving Theorem 2.4.3 and then deduce Theorem 2.4.1 and Theorem 2.4.2.

Proof A.1.3 (Proof of Theorem 2.4.3) *First, we establish that mFDR control is provided*

under (2.2) *and* (2.3) *for any procedure* $\mathcal{A} = (\alpha_t, t \geq 1)$ *if*

$$\alpha_T + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} F_t(\alpha_t) \leq (1-\lambda)\alpha\,(1 \vee R(T)), \quad \textit{(a.s.)}. \tag{A.2}$$

Indeed, by Lemma A.1.1, we have

$$\begin{aligned}\mathbf{E}_{X\sim P}\Big(\sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0, p_t \leq \alpha_t\}\Big) &\leq (1-\lambda)^{-1}\,\mathbf{E}\left(\sum_{t=1}^{T} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t)\right) \\ &\leq \alpha\,\mathbf{E}(1 \vee R(T)),\end{aligned}$$

by using (A.2)*, which is exactly the desired mFDR control. Now, statement (i) holds because* (A.2) *holds for* $(\alpha_t^0, t \geq 1)$ *from* (2.30) *and* (2.2)*. Finally, we establish statement (ii). By* (2.30) *and* (2.2)*, condition* (A.2) *holds for* $(\alpha_t, t \geq 1)$ *if for all* $T \geq 1$*,*

$$\alpha_T + \sum_{t=1}^{T-1} \mathbf{1}\{p_t(X) \geq \lambda\}\left[(1-a_{T-t})\alpha_t + a_{T-t}F_t(\alpha_t)\right] = \alpha_T^0 + \sum_{p_t \geq \lambda, 1 \leq t \leq T-1} \alpha_t^0,$$

where $a_T = \sum_{t=1}^{T} \gamma_t'$*. Now the last display holds true by Lemma A.1.2 because of* (2.20)*, which concludes the proof.*

Proof A.1.4 (Proof of Theorems 2.4.1 and 2.4.2) *Theorem 2.4.1 and Theorem 2.4.2 can be derived from Theorem 2.4.3 for* $\mathcal{A}^0 = \mathcal{A}^{LORD}$ *(using* $\lambda = 0$*) and* $\mathcal{A}^0 = \mathcal{A}^{ALORD}$*, respectively, by checking* (2.30) *in both cases. First, for* $\mathcal{A}^{LORD}$*, we have*

$$\begin{aligned}\sum_{t=1}^{T} \alpha_t^{LORD} &= \sum_{t=1}^{T}\left(W_0\gamma_t + (\alpha - W_0)\gamma_{t-\tau_1} + \alpha\sum_{j\geq 2}\gamma_{t-\tau_j}\right) \\ &= W_0\sum_{t=1}^{T}\gamma_t + (\alpha - W_0)\sum_{t=1}^{T}\gamma_{t-\tau_1} + \alpha\sum_{j\geq 2}\mathbf{1}\{T - \tau_j \geq 1\}\sum_{t=1}^{T}\gamma_{t-\tau_j} \\ &\leq \alpha(1 + 0 \vee (R(T-1) - 1)) \leq \alpha(1 \vee R(T)),\end{aligned} \tag{A.3}$$

because $\tau_j \leq T-1$ *is equivalent to* $R(T-1) \geq j$ *by definition. Second, for* $\mathcal{A}^{ALORD}$*, we proceed similarly with the help of Lemma A.1.3: by definition* (2.28)*, we have*

$$\begin{aligned}&(1-\lambda)^{-1}\left(\alpha_T^{ALORD} + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \alpha_t^{ALORD}\right) \\ &= W_0\left(\gamma_{\mathcal{T}_0(T)} + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \gamma_{\mathcal{T}_0(t)}\right) + (\alpha - W_0)\left(\gamma_{\mathcal{T}_1(T)} + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \gamma_{\mathcal{T}_1(t)}\right) \\ &\quad + \alpha\sum_{j\geq 2}\mathbf{1}\{T \geq \tau_j + 1\}\left(\gamma_{\mathcal{T}_j(T)} + \sum_{\substack{1 \leq t \leq T-1, \\ p_t \geq \lambda}} \gamma_{\mathcal{T}_j(t)}\right).\end{aligned}$$

Finally, by using (A.6) *and* (A.7)*, the latter is equal to*

$$W_0 \sum_{t=1}^{\mathcal{T}_0(T)} \gamma_t + (\alpha - W_0) \sum_{t=1}^{\mathcal{T}_1(T)} \gamma_t + \alpha \sum_{j\geq 2} \mathbf{1}\{T \geq \tau_j + 1\} \sum_{t=1}^{\mathcal{T}_j(T)} \gamma_t$$
$$\leq W_0 + \alpha - W_0 + \alpha \sum_{j\geq 2} \mathbf{1}\{T \geq \tau_j + 1\} = \alpha(1 + 0 \vee (R(T-1) - 1)) \leq \alpha\,(1 \vee R(T)),$$

because $T \geq \tau_j + 1$ *if and only if* $R(T-1) \geq j$.

A.1.3 Auxiliary lemmas

The following lemma provides a tool for controlling both online FWER and mFDR.

Lemma A.1.1 *For any procedure* $\mathcal{A} = (\alpha_t, t \geq 1)$*, we have for all* $\lambda \in [0, 1)$*,*

$$\mathbf{E}_{X\sim P}\Big(\sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0, p_t \leq \alpha_t\}\Big) \leq (1-\lambda)^{-1}\, \mathbf{E}\left(\sum_{t=1}^{T} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t)\right), \tag{A.4}$$

provided that (2.2) *holds and if either* (2.3) *holds or if the critical values* $(\alpha_t, t \geq 1)$ *are deterministic.*

Proof A.1.5 *Recall* α_t *is either deterministic or* $\mathcal{F}_{t-1}$*-measurable (in which case it is independent of* $p_t(X)$ *under* (2.3)*). Therefore, under the conditions of the lemma, we have in any case: for all* $t \in \mathcal{H}_0$*, both*

$$\mathbf{E}\left(\frac{\mathbf{1}\{p_t(X) > \lambda\}}{1-\lambda}\middle| \alpha_t\right) \geq 1, \quad \mathbf{P}\left(p_t(X) \leq \alpha_t \mid \alpha_t\right) \leq F_t(\alpha_t).$$

This entails

$$\begin{aligned}
\mathbf{E}_{X\sim P}\Big(\sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0, p_t \leq \alpha_t\}\Big) &= \sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0\}\, \mathbf{E}\left(\mathbf{P}(p_t(X) \leq \alpha_t \mid \alpha_t)\right) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0\}\, \mathbf{E}\left(F_t(\alpha_t)\right) \\
&\leq \sum_{t=1}^{T} \mathbf{1}\{t \in \mathcal{H}_0\}\, \mathbf{E}\left(F_t(\alpha_t)\mathbf{E}\left(\frac{\mathbf{1}\{p_t(X) \geq \lambda\}}{1-\lambda} \mid \alpha_t\right)\right) \\
&\leq (1-\lambda)^{-1}\, \mathbf{E}\left(\sum_{t=1}^{T} \mathbf{1}\{p_t(X) \geq \lambda\} F_t(\alpha_t)\right).
\end{aligned}$$

The following representation lemma is the key tool for building the new rewarded critical values.

Lemma A.1.2 *Let* $(\alpha_t^0, t \geq 1)$ *be any nonnegative sequence. Let* $(\tilde{\alpha}_t, t \geq 1)$ *be the sequence defined by the recursive relation*

$$\tilde{\alpha}_T = \alpha_T^0 + \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\}\alpha_t^0 - \sum_{t=1}^{T-1} \mathbf{1}\{p_t \geq \lambda\}\left[(1 - a_{T-t})\tilde{\alpha}_t + a_{T-t} F_t(\tilde{\alpha}_t)\right], \quad T \geq 1, \tag{A.5}$$

where $a_T = \sum_{t=1}^{T} \gamma'_t$, $T \geq 1$ *for any real values* γ'_t, p_t, λ *and functions* F_t. *Let* $(\bar{\alpha}_t, t \geq 1)$ *be the sequence defined by the recursive relation*

$$\bar{\alpha}_T = \alpha_T^0 + \sum_{\substack{1 \leq t \leq T-1 \\ p_t \geq \lambda}} \gamma'_{T-t}(\bar{\alpha}_t - F_t(\bar{\alpha}_t)) + \mathbf{1}\{p_{T-1} < \lambda\}(\bar{\alpha}_{T-1} - \alpha_{T-1}^0), \quad T \geq 1.$$

Then we have $\tilde{\alpha}_t = \bar{\alpha}_t$ *for all* $t \geq 1$. *Moreover,* $\bar{\alpha}_t \geq \bar{\alpha}_t^0$ *for all* $t \geq 1$ *under* (2.2). *In particular, these critical values are nonnegative.*

Proof A.1.6 *Clearly,* $\tilde{\alpha}_1 = \alpha_1^0 = \bar{\alpha}_1$ *so the result is satisfied for* $T = 1$. *For* $T \geq 2$, *by using* (A.5) *for* $\tilde{\alpha}_T$ *and* $\tilde{\alpha}_{T-1}$, *we have*

$$\begin{aligned}\tilde{\alpha}_T - \tilde{\alpha}_{T-1} = \alpha_T^0 - \alpha_{T-1}^0 &+ \mathbf{1}\{p_{T-1} \geq \lambda\}\alpha_{T-1}^0 \\ &- \mathbf{1}\{p_{T-1} \geq \lambda\}\left[(1-a_1)\tilde{\alpha}_{T-1} + a_1 F_{T-1}(\tilde{\alpha}_{T-1})\right] \\ &+ \sum_{t=1}^{T-2} \mathbf{1}\{p_t \geq \lambda\}\left[(a_{T-t} - a_{T-t-1})\tilde{\alpha}_t - (a_{T-t} - a_{T-t-1})F_t(\tilde{\alpha}_t)\right].\end{aligned}$$

Hence, by using $\tilde{\alpha}_{T-1} = \tilde{\alpha}_{T-1}\mathbf{1}\{p_{T-1} < \lambda\} + \tilde{\alpha}_{T-1}\mathbf{1}\{p_{T-1} \geq \lambda\}$, *we obtain*

$$\begin{aligned}\tilde{\alpha}_T = \alpha_T^0 &- \mathbf{1}\{p_{T-1} < \lambda\}\alpha_{T-1}^0 + \tilde{\alpha}_{T-1}\mathbf{1}\{p_{T-1} < \lambda\} \\ &+ \mathbf{1}\{p_{T-1} \geq \lambda\}\left[\gamma'_1\tilde{\alpha}_{T-1} - \gamma'_1 F_{T-1}(\tilde{\alpha}_{T-1})\right] \\ &+ \sum_{t=1}^{T-2} \mathbf{1}\{p_t \geq \lambda\}\left[\gamma'_{T-t}\tilde{\alpha}_t - \gamma'_{T-t}F_t(\tilde{\alpha}_t)\right],\end{aligned}$$

because $\gamma'_1 = a_1$, *and we recognize the expression given in the lemma.*

Let us finally prove that $\bar{\alpha}_T \geq \bar{\alpha}_T^0$ *for all* $T \geq 1$. *This is true for* $\bar{\alpha}_1$ *because* $\bar{\alpha}_1 = \alpha_1^0$. *Now, if* $\bar{\alpha}_1 \geq \alpha_1^0, \ldots, \bar{\alpha}_{T-1} \geq \alpha_{T-1}^0$ *then we also have*

$$\bar{\alpha}_T = \alpha_T^0 + \sum_{\substack{1 \leq t \leq T-1 \\ p_t \geq \lambda}} \gamma'_{T-t}(\bar{\alpha}_t - F_t(\bar{\alpha}_t)) + \mathbf{1}\{p_{T-1} < \lambda\}(\bar{\alpha}_{T-1} - \alpha_{T-1}^0) \geq \alpha_T^0,$$

because $\bar{\alpha}_t \geq F_t(\bar{\alpha}_t)$ *by* (2.2). *This finishes the proof.*

We now establish a result for the functionals $\mathcal{T}(\cdot)$ and $\mathcal{T}_j(\cdot)$, $j \geq 1$, which are used by the adaptive procedures $\mathcal{A}^{\text{AOB}}$ and $\mathcal{A}^{\text{ALORD}}$, respectively.

Lemma A.1.3 *Consider the functional* $\mathcal{T}(\cdot)$ *defined by* (2.16) *for some realization of the p-values and some* $\lambda \in [0, 1)$. *Then for any sequence* $(\gamma_t)_{t \geq 1}$ *and for any* $T \geq 1$, *we have*

$$\sum_{t=1}^{T} \mathbf{1}\{p_t \geq \lambda\}\gamma_{\mathcal{T}(t)} = \sum_{t=1}^{\mathcal{T}(T+1)-1} \gamma_t. \tag{A.6}$$

In addition, for any $j \geq 1$, *consider the* τ_j *defined by* (B.14) *and the functional* $\mathcal{T}_j(\cdot)$ *defined by* (2.26). *Then for all* $T \geq \tau_j + 1$,

$$\sum_{\substack{1 \leq t \leq T \\ p_t \geq \lambda}} \gamma_{\mathcal{T}_j(t)} = \sum_{t=1}^{\mathcal{T}_j(T+1)-1} \gamma_t. \tag{A.7}$$

Proof A.1.7 *Let us first prove* (A.6). *Since* $\mathcal{T}(t+1) = \mathcal{T}(t) + 1$ *when* $p_t \geq \lambda$ *from definition* (2.16)*, we can write*

$$\sum_{t=1}^{T} \mathbf{1}\{p_t \geq \lambda\}\gamma_{\mathcal{T}(t)} = \sum_{t=1}^{T} \mathbf{1}\{p_t \geq \lambda\}\gamma_{\mathcal{T}(t+1)-1} = \sum_{t=2}^{T+1} \mathbf{1}\{p_{t-1} \geq \lambda\}\gamma_{\mathcal{T}(t)-1}.$$

Additionally, it is clear that $\mathcal{T}(\cdot)$ *is a bijection mapping* $\{1,\, 2 \leq t \leq T+1 \,:\, p_{t-1} \geq \lambda\}$ *into* $\{1, 2, \ldots, \mathcal{T}(T+1)\}$*. Hence, the latter sum can be rewritten as* $\sum_{t=2}^{\mathcal{T}(T+1)} \gamma_{t-1} = \sum_{t=1}^{\mathcal{T}(T+1)-1} \gamma_t$ *which provides* (A.6).

Second, for proving (A.7)*, the crucial point is that according to the definition of* $\mathcal{T}_j(T)$ (2.26)*, the functional* $\mathcal{T}_j : \{\tau_j + 1, \ldots\} \to \{1, \ldots\}$ *is a bijection from* $\{\tau_j + 1\} \cup \{t \in \{\tau_j + 2, \ldots, T+1\} : p_{t-1} \geq \lambda\}$ *to* $\{1, \ldots, \mathcal{T}_j(T+1)\}$*, for any* $j \geq 1$ *and* $T \geq \tau_j + 1$*. In particular, this entails*

$$\begin{aligned}\sum_{p_t \geq \lambda, 1 \leq t \leq T} \gamma_{\mathcal{T}_j(t)} &= \sum_{p_t \geq \lambda, \tau_j + 1 \leq t \leq T} \gamma_{\mathcal{T}_j(t)} = \sum_{p_t \geq \lambda, \tau_j + 1 \leq t \leq T} \gamma_{\mathcal{T}_j(t+1)-1} \\ &= \sum_{p_{t-1} \geq \lambda, \tau_j + 2 \leq t \leq T+1} \gamma_{\mathcal{T}_j(t)-1} = \sum_{t=2}^{\mathcal{T}_j(T+1)} \gamma_{t-1} = \sum_{t=1}^{\mathcal{T}_j(T+1)-1} \gamma_t.\end{aligned}$$

This proves (A.7).

A.2 Delayed spending approach

In this section we present another way of incorporating super-uniformity into OMT which we refer to as *delayed spending* (in the sequel abbreviated as DS). We are grateful to Aaditya Ramdas for this suggestion.

The new procedure is introduced in Section A.2.1, while we highlight some mathematical and practical differences with our approach in Sections A.2.2 and A.2.3. In order to make the new procedure more efficient we also present a hybrid version in Section A.2.4. For simplicity, we restrict ourselves to FWER controlling procedures for discrete data throughout this section.

A.2.1 Definition

Let us start with the critical value $\alpha_1 = \alpha\gamma_1$. While the OB procedure would choose $\alpha_2 = \alpha\gamma_2$, the idea is that if the super-uniformity is strong enough to ensure $F_1(\alpha\gamma_1) + F_2(\alpha\gamma_1) \leq \alpha\gamma_1$, we can still use $\alpha_2 = \alpha\gamma_1$ in the second round. This process can be continued until $F_1(\alpha\gamma_1) + \cdots + F_{b_1+1}(\alpha\gamma_1) > \alpha\gamma_1$, in which case we switch to $\alpha_{b_1} = \alpha\gamma_2$, and so on. This way, we can incorporate the super-uniformity directly by 'delaying' the γ sequence.

More formally, consider the setting of Section 4.2, where a null bounding family $\mathcal{F} = \{F_t, t \geq 1\}$ satisfying (2.2) is at hand. The above strategy reads:

$$\alpha_t^{\mathrm{DS}} = \alpha\gamma_{\mathcal{C}(t)}, \text{ where } \mathcal{C}(t) = \min\{j \geq 1 \,:\, b_j \geq t\}, \quad t \geq 1; \tag{A.8}$$

$$b_j = \max\Big\{T \geq b_{j-1} + 1 \,:\, \sum_{t=b_{j-1}+1}^{T} F_t(\alpha\gamma_j) \leq \alpha\gamma_j\Big\}, \quad j \geq 1, \tag{A.9}$$

(with the convention $b_0 = 0$ and $b_j = +\infty$ if the set in (A.9) is empty), so that $j = \mathcal{C}(t)$ for $b_{j-1} + 1 \leq t \leq b_j$. Thus, the DS method processes each sub-budget $\alpha\gamma_j$ one at a time, until the

Table A.1: Number of discoveries for SURE online Bonferroni (2.15) (bandwidth $h = 10$) and the DS approach (A.8). Here $\mathcal{C}(30\,000) = 5083$ as defined in (A.8). These numbers are obtained by running the procedures on the first 30 000 genes for male (second row) and female (third row) mice in the IMPC data.

Procedures	OB	ρOB	Delayed
# discoveries (male)	229	377	293
# discoveries (female)	267	481	355

stopping rule in (A.9) is met and the transition to the next sub-budget $\alpha\gamma_{j+1}$ is made. Since $\mathcal{C}(t) \leq t$ we can interpret $\alpha_t^{\mathrm{DS}} = \alpha\gamma_{\mathcal{C}(t)}$ as a 'slowed-down' variant of the original OB procedure.

The procedure (A.8) controls the online FWER under (2.2) because by (2.13), a sufficient condition is given by $\sum_{t=1}^{T} F_t(\alpha_t) \leq \alpha$, $T \geq 1$, and we indeed have

$$\sum_{t\geq 1} F_t(\alpha_t^{\mathrm{DS}}) \leq \sum_{j\geq 1} \sum_{t=b_{j-1}+1}^{b_j} F_t(\alpha\gamma_j) \leq \sum_{j\geq 1} \alpha\gamma_j \leq \alpha,$$

by definition of the b_j's. Note that the b_j's are based on local averages (in time) of the $F_t(x)$'s at certain points x. This shares similarity to the approach of Westfall and Wolfinger (1997) for offline FWER control.

A.2.2 Comparison to SUR for real data

Both the DS and the SUR approaches use super-uniform rewarding. In a nutshell, the DS approach slows down the clock whereas the SUR approach augments the critical values of existing OMT procedures in an additive way. While a more detailed comparison can be found in the following Section A.2.3, we may say that no method dominates the other one uniformly. The examples given in Section A.2.3 (delayed start, long/infinite delay, ineffective delay) suggest that the DS method could be more efficient at the very start of the stream but may suffer from conservativeness afterwards.

To assess the behaviour of the procedures in a practical setting, we reanalyse the IMPC data from Section 2.5.3 using the DS procedure defined by (A.8) and (A.9) and compare it with the OB and ρOB from Section 2.3. The results for FWER control at level $\alpha = 0.2$ are displayed in Table A.1 and Figure A.1. As Figure A.1 (right panel) shows, the rejection process $\{R(T), T \geq 1\}$, is almost identical at the very start. However, for larger T, the delayed approach makes less discoveries than the ρOB procedure and this, uniformly in time for this data set. This conservative behaviour is probably caused by under-utilization of wealth as described in Section A.2.3. More specifically, the non-utilized component of $\alpha = 0.2$ accumulates up to time $T = 1500$ approximately to 0.077, so that approximately 38.5% of $\alpha = 0.2$ are effectively neglected. Accordingly, the wealth plot displayed in Figure A.1 shows that the delayed approach manages to spend more wealth than the OB procedure, but still deviates strongly from the nominal wealth curve. Figure A.2 illustrates the same phenomenon for the critical values. (This replaces the old section D.2)

A.2.3 Formal properties

From the definition of the DS approach we obtain the following comparison to OB and ρOB:

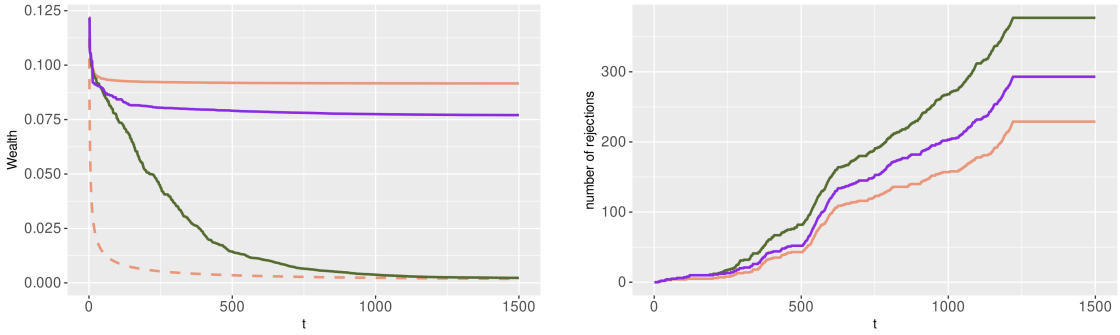

Figure A.1: Comparison with DS. Left: nominal wealth for OB (dashed orange curve), effective wealth for OB (solid orange curve), effective wealth for ρOB (solid green curve) and effective wealth for DS (solid purple curve), plot similar to Figure 2.1. Right: rejection numbers, cumulated over time, for the same procedures (same color code). Both plots are computed from the male IMPC data.

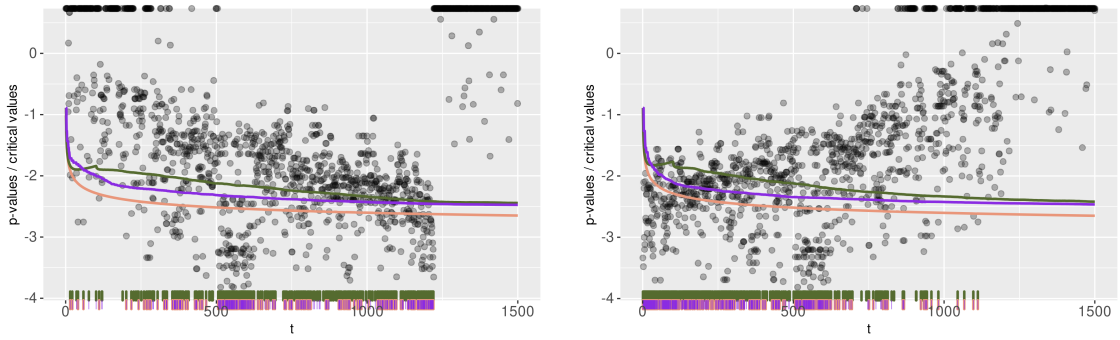

Figure A.2: Critical values of OB (orange), ρOB (green) and DS (purple) for the IMPC data (left panel is for male, right panel is for female).

- the DS approach improves OB uniformly when γ_t is nonincreasing: indeed $\mathcal{C}(t) \leq t$, so that $\alpha_t^{\mathrm{DS}} = \alpha\gamma_{\mathcal{C}(t)} \geq \alpha\gamma_t = \alpha_t^{\mathrm{OB}}$.
- the DS approach does not depend on any other tuning parameter such as the bandwidth. By contrast, choosing this parameter badly in the ρOB procedure may adversely affect its performance.
- the DS approach is another way of using the super uniformity reward. For instance, if there is no super uniformity reward, that is, $F_t(\alpha\gamma_t) = \alpha\gamma_t$ for all t, then $b_t = t$ and the DS procedure reduces to OB.

In addition, we have the following observations:

- Delayed start: If $F_t(x) = 0$ for all $x < 1$ and $t \leq T_0$ and $F_t(x) = x$ for $t \geq T_0 + 1$, the DS procedure is much more intuitive: it yields $b_1 = T_0 + 1$ by (A.9) and $\alpha_t^{\mathrm{DS}} = \alpha\gamma_{t-T_0}$ for $t \geq T_0 + 1$ which is the most natural way to proceed (just start the testing process at time $T_0 + 1$). By contrast, ρOB (with rectangular kernel of bandwidth r) collects some reward in $\alpha_t^{\rho\mathrm{OB}}$, $1 \leq t \leq T_0$, spends the reward in the following r time points, but continues with $\alpha_t^{\rho\mathrm{OB}} = \alpha\gamma_t$ for $t \geq T_0 + r + 1$. Hence, delaying spends the super-uniformity more intuitively

than ρOB in that situation. More generally, in practice, we may therefore expect DS to be more efficient in the beginning of the stream.

- Long/infinite delay: Conversely, if there exists $T_0 \geq 1$ such that for all $t \geq b_{T_0} + 1$, $F_t(\alpha\gamma_{\mathcal{C}(T_0)+1}) = 0$, then we have $b_{T_0+1} = +\infty$ from (A.9), which in turn implies $\mathcal{C}(t) \leq T_0 + 1$. But for $t \geq b_{T_0} + 1$, we have $\mathcal{C}(t) \geq T_0 + 1$ by (A.8). Hence, for $t \geq b_{T_0} + 1$, $\mathcal{C}(t) = T_0 + 1$ and the 'spending clock' freezes. On the one hand, we have $\alpha_t^{\mathrm{DS}} = \alpha\gamma_{T_0+1}$ so the delaying works perfectly to effectively improve the OB critical values. On the other hand, this effectively stops the spending of any further budget and thus a large part of the wealth is left unspent. This is in contrast to the SUR approach which uses a reward of an additive nature and thus always has a chance to spend the budget.

- Under-utilization of wealth. The DS method processes each sub-budget $\alpha\gamma_j$ one at a time, until the transition to the next sub-budget $\alpha\gamma_{j+1}$ is made. In most cases, however, the inequality (A.9) defining the transition time b_j will be a strict inequality, meaning that when we move on to the next sub-budget we will have used $\sum_{t=b_{j-1}+1}^{b_j} F_t(\alpha\gamma_j) < \alpha\gamma_j$. Thus, this method does not exhaust the available sub-budgets. Moreover, since it neglects these 'alpha-gaps', they accumulate over time. This under-utilized wealth leads to unnecessary conservatism. Removing such gaps was precisely the primary motivation for introducing our SUR method, see Section 2.2.3.

  The most disadvantageous scenario occurs when $b_t = t$ for all $t \leq T$, so that the DS procedure reduces to the original OB procedure up to time T. As an example consider $\epsilon \in (0, \alpha\gamma_T)$ for some large $T \geq 1$ and assume that the support of each p_t is given by $S_t = \{\epsilon, A_t, \alpha\gamma_{t-1}\} \cup \{1\}$ (convention $\alpha\gamma_0 = 1$), where A_t is a finite subset of $(\alpha\gamma_t, \alpha\gamma_{t-1})$. Then we have $F_1(\alpha\gamma_1) + F_2(\alpha\gamma_1) = \alpha\gamma_1 + \epsilon$ hence $b_1 = 1$, and more generally $F_t(\alpha\gamma_t) + F_{t+1}(\alpha\gamma_t) = \alpha\gamma_t + \epsilon$ for all $t \leq T$, which implies $b_t = t$ for all $t \leq T$. However, we know that OB does not allow to spend all the budget in such a discrete situation, see Figure 2.1.

  A potential remedy for the conservatism of the DS method could be to combine it with our SUR method. We describe such a hybrid approach in more detail in Section A.2.4.

In summary, it may be said that the delaying method is particularly appealing in terms of simplicity and elegance, while the primary aim of the SUR approach is on efficiency.

A.2.4 Hybrid approach

In this section, we describe a hybrid approach, combining the ideas underlying DS and SUR, in order to improve the utilization of wealth of DS.

The method starts as follows: first let $\alpha_1^{\mathrm{Hyb}} = \alpha\gamma_1, \ldots, \alpha_{b_1}^{\mathrm{Hyb}} = \alpha\gamma_1$ as long as $F_1(\alpha\gamma_1) + \cdots + F_{b_1}(\alpha\gamma_1) \leq \alpha\gamma_1$. Then consider the reward $\rho_1 = \alpha\gamma_1 - (F_1(\alpha\gamma_1) + \cdots + F_{b_1}(\alpha\gamma_1))$ and let $\alpha_{b_1+1}^{\mathrm{Hyb}} = \alpha\gamma_2 + \rho_1, \ldots, \alpha_{b_2}^{\mathrm{Hyb}} = \alpha\gamma_2 + \rho_1$ as long as $F_{b_1+1}(\alpha\gamma_2 + \rho_1) + \cdots + F_{b_2}(\alpha\gamma_2 + \rho_1) \leq \alpha\gamma_2 + \rho_1$. More generally, let $b_0 = 0$, $\rho_0 = 0$, and for all $j \geq 1$,

$$\begin{aligned}
&\alpha_{b_{j-1}+1}^{\mathrm{Hyb}} = \alpha\gamma_j + \rho_{j-1}, \ldots, \alpha_{b_j}^{\mathrm{Hyb}} = \alpha\gamma_j + \rho_{j-1} \\
&\qquad b_j = \max\left\{ T \geq 1 \,:\, \sum_{t=b_{j-1}+1}^{T} F_t(\alpha\gamma_j + \rho_{j-1}) \leq \alpha\gamma_j + \rho_{j-1} \right\} \\
&\qquad \rho_j = \alpha\gamma_j + \rho_{j-1} - \left(\sum_{t=b_{j-1}+1}^{b_j} F_t(\alpha\gamma_j + \rho_{j-1}) \right).
\end{aligned}$$

Then the online FWER control holds because for all $j_0 \geq 1$, we have

$$\sum_{t\geq 1} F_t(\alpha_t^{\mathrm{Hyb}}) = \sum_{j=1}^{j_0} \left(\sum_{t=b_{j-1}+1}^{b_j} F_t(\alpha\gamma_j + \rho_{j-1}) \right) + F_{b_{j_0}+1}(\alpha\gamma_{j_0+1} + \rho_{j_0})$$
$$\leq \sum_{j=1}^{j_0} (\alpha\gamma_j + \rho_{j-1} - \rho_j) + \alpha\gamma_{j_0+1} + \rho_{j_0} = \sum_{j=1}^{j_0+1} \alpha\gamma_j \leq \alpha,$$

because $\sum_{j=1}^{j_0}(\rho_{j-1} - \rho_j) = -\rho_{j_0}$ (telescopic sum). When $\rho_t = 0$ for all $t \geq 1$, the hybrid approach reduces to the DS approach. When $b_j = j$, the hybrid approach reduces to the greedy SUR procedure.

We can also combine the DS with smoothed SUR rewarding, which gives us the following, slightly more involved, procedure. For some SUR spending sequence $\gamma' = (\gamma'_t)_{t\geq 1}$ (nonnegative and such that $\sum_{t\geq 1} \gamma'_t \leq 1$), let $b_0 = 0$, $\rho_0 = 0$ and for all $j \geq 1$,

$$\alpha_{b_{j-1}+1}^{\mathrm{Hyb}} = \alpha\gamma_j + \sum_{i=1}^{j-1} \gamma'_{j-i}\rho_i, \quad \ldots, \quad \alpha_{b_j}^{\mathrm{Hyb}} = \alpha\gamma_j + \sum_{i=1}^{j-1} \gamma'_{j-i}\rho_i$$
$$b_j = \max\left\{ T \geq 1 \,:\, \sum_{t=b_{j-1}+1}^{T} F_t\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i\right) \leq \alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i \right\}$$
$$\rho_j = \alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i - \left(\sum_{t=b_{j-1}+1}^{b_j} F_t\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i\right)\right).$$

The online FWER control holds because for all $j_0 \geq 1$, we have

$$\sum_{t\geq 1} F_t(\alpha_t^{\mathrm{Hyb}}) \leq \sum_{j=1}^{j_0} \left(\sum_{t=b_{j-1}+1}^{b_j} F_t\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i\right)\right) + \alpha\gamma_{j_0+1} + \sum_{i=1}^{j_0}\gamma'_{j_0+1-i}\rho_i.$$

Now letting $a_T = \sum_{t=1}^{T}\gamma'_t$, we obtain

$$\sum_{j=1}^{j_0} \left(\sum_{t=b_{j-1}+1}^{b_j} F_t\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i\right)\right)$$
$$\leq \sum_{j=1}^{j_0} a_{j_0-j+1}\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i - \rho_j\right) + \sum_{j=1}^{j_0}(1 - a_{j_0-j+1})\left(\alpha\gamma_j + \sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i\right)$$
$$= \sum_{j=1}^{j_0}\alpha\gamma_j + \sum_{j=1}^{j_0}\sum_{i=1}^{j-1}\gamma'_{j-i}\rho_i - \sum_{j=1}^{j_0} a_{j_0-j+1}\rho_j$$
$$= \sum_{j=1}^{j_0}\alpha\gamma_j + \sum_{i=1}^{j_0-1} a_{j_0-i}\rho_i - \sum_{j=1}^{j_0} a_{j_0-j+1}\rho_j,$$

and the latter is equal to $\sum_{j=1}^{j_0}\alpha\gamma_j - \sum_{j=1}^{j_0-1}\gamma'_{j_0-j+1}\rho_j - a_1\rho_{j_0} = \sum_{j=1}^{j_0}\alpha\gamma_j - \sum_{j=1}^{j_0}\gamma'_{j_0-j+1}\rho_j$. Combining this with the above bound for the FWER concludes the proof.

To compare the performance of the hybrid approach with the SUR and DS approaches, we use the simulation setting from Section 2.5.2 in the case where the signal is positioned at the beginning of the stream for each simulation run, which is the most favorable position of the signal for any procedure (see Section 2.5.2 for more details). We consider both procedures based on the uniform kernel (bandwidth $h = 100$) and those based on the greedy spending sequence (denoted by 'greedy').

Figure A.3 shows that taking super-uniformity into account is always beneficial, regardless of the specific approach used. The base DS method performs similarly to the greedy ρOB and the greedy hybrid. In contrast, the hybrid approach based on a uniform kernel improves DS, with performance close to ρOB. Hence, we conclude that closing the alpha-gaps by smoothing with an adequate kernel can make the hybrid approach as powerful as the smoothed ρOB method. However, given the added complexity of the hybrid approach, we prefer to stick with the smoothed SUR.

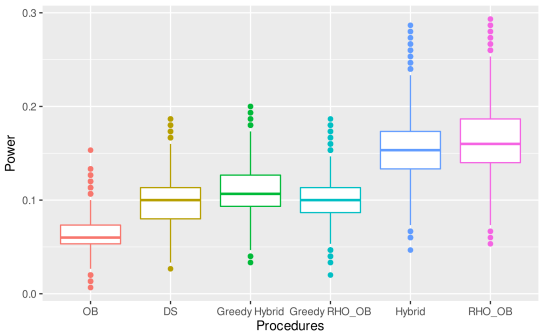

Figure A.3: Power of several online FWER controlling approaches for simulated data (see text): online Bonferroni (OB), Delayed spending (DS), greedy hybrid, greedy ρOB, hybrid, ρOB.

A.3 Complements on generalized α-investing rules

A.3.1 SUR-GAI++ rules

GAI++ rules have been introduced in Ramdas et al. (2017) to control the (m)FDR. Here, we can extend them to our super-uniform setting as follows. Let us consider the following recursive constraints: for $t \geq 1$,

$$\begin{aligned}
R_t &= \mathbf{1}\{p_t \leq \alpha_t\} \\
W(t) &= W(t-1) - \phi_t + R_t\psi_t \quad \text{'wealth available at time } t+1\text{'} \\
\phi_t &\in [0, W(t-1)] \quad \text{'spent at time } t\text{'} \\
\psi_t &\leq b_t + \min\left(\phi_t, \phi_t/F_t(\alpha_t) - 1\right) \quad \text{'reward at time } t\text{'} \\
\psi_t &\geq 0 \\
b_t &= \alpha - W_0\mathbf{1}\{t \leq \tau_1\},
\end{aligned}$$

where $W(0) = W_0 \in [0, \alpha]$. Any choice of W_0 and α_t, ϕ_t, ψ_t that are $\mathcal{F}_{t-1}$ measurable and satisfying the above constraints defines a SUR-GAI++ procedure. Here, the only difference with the original GAI++ rule is the presence of $F_t(\alpha_t)$ instead of α_t in the definition of Ψ_t.

Proposition A.3.1 *Consider the setting of Section 4.2 where a null bounding family* $\mathcal{F} =$

$\{F_t, t \geq 1\}$ satisfying (2.4) is at hand. Then any SUR-GAI++ procedure controls the mFDR at level α.

The proof is totally analogous to the one of Theorem 1 in Ramdas et al. (2017) (adapted to the mFDR, so without using any monotonicity).

A.3.2 GAI++ weighting

Consider (continuous) p-values satisfying (2.1)-(2.3) and weights $w_t \geq 0$ that are $\mathcal{F}_{t-1}$ measurable for all t. In Section 5 of Ramdas et al. (2017), the following (implicit) GAI++ weighting scheme has been proposed:

$$\begin{aligned} R_t &= \mathbf{1}\{p_t \leq w_t \alpha_t\} \\ W(t) &= W(t-1) - \phi_t + R_t \psi_t \\ \phi_t &\in [0, W(t-1)] \\ \psi_t &\leq b_t + \min\left(\phi_t, \phi_t/(w_t\alpha_t) - 1\right) \\ \psi_t &\geq 0 \\ b_t &= \alpha - W_0 \mathbf{1}\{t \leq \tau_1\}. \end{aligned}$$

Note that the latter constraints are similar to the constraints given in Section A.3.1 for $F_t(x) = (w_t x) \wedge 1$ (up to the '$\wedge 1$' which makes the constraints here slightly more stringent) so that this weighting case is a particular SUR-GAI++ procedure.

For given raw weights $r_t \geq 0$ ($\mathcal{F}_{t-1}$ measurable), an explicit procedure which is used in Ramdas et al. (2017)[1], is obtained by choosing α_t, w_t, ϕ_t, ψ_t as follows:

$$\begin{aligned} w_t &= r_t \wedge \frac{1}{1-b_t} \\ \phi_t &= \alpha_t = W_0 \gamma_t + \sum_{j \geq 1} \gamma_{t-\tau_j} \psi_{\tau_j} \\ \psi_t &= b_t + \min\left(\phi_t, 1/w_t - 1\right). \end{aligned}$$

This choice is valid because $\alpha_t \leq W(t-1)$ for all t. Indeed,

$$W(t-1) = W_0 + \sum_{i=1}^{t-1} (-\alpha_i + R_i \psi_i),$$

so $\alpha_t \leq W(t-1)$ if and only if $\sum_{i=1}^{t} \alpha_i \leq W_0 + \sum_{i=1}^{t-1} R_i \psi_i$, which is true.

A.3.3 Our ρ-LORD is a SUR-GAI++ rule

We claim here that the procedure ρ-LORD corresponds to a SUR-GAI++ rule with the choice $\phi_t = F_t(\alpha_t)$, $\psi_t = b_t$, and

$$\alpha_t = W_0 \gamma_t + (\alpha - W_0)\gamma_{t-\tau_1} + \alpha \sum_{j \geq 2} \gamma_{t-\tau_j} + \sum_{i=1}^{t-1} \gamma'_{t-i} \rho_i \quad t \geq 1. \tag{A.10}$$

[1] This procedure is available at `https://github.com/fanny-yang/OnlineFDRCode`

To establish this, we check that all constraints given in Section A.3.1 are satisfied. The only non-trivial one is $\phi_t = F_t(\alpha_t) \leq W(t-1)$. Let us now prove it. Recall that $W(t) = W(t-1) - \phi_t + R_t b_t$ and $W(0) = W_0$. Hence $\alpha_1 = W_0\gamma_1 \leq W_0$. Moreover, for $t \geq 2$,

$$W(t-1) = W_0 + (\alpha - W_0)\mathbf{1}\{t-1 \geq \tau_1\} + \alpha \sum_{j \geq 2} \mathbf{1}\{t-1 \geq \tau_j\} - \sum_{i=1}^{t-1} F_i(\alpha_i).$$

So we have $\bar{\alpha}_t \leq W(t-1)$ for the critical value

$$\bar{\alpha}_t = \left(\sum_{i=1}^{t} \gamma_i\right) W_0 + \sum_{i=1}^{t-1} \left((\alpha - W_0)\gamma_{i-\tau_1+1}\mathbf{1}\{i \geq \tau_1\} + \alpha \sum_{j \geq 2} \gamma_{i-\tau_j+1}\mathbf{1}\{i \geq \tau_j\} \right) - \sum_{i=1}^{t-1} \left[a_{t-i} F_i(\bar{\alpha}_i) + (1 - a_{t-i})\bar{\alpha}_i \right],$$

by letting $a_t = \sum_{i=1}^{t} \gamma'_i$. But now, we have that $\bar{\alpha}_t = \alpha_t$ for all t, for α_t defined by (A.10). Indeed, this can be seen from Lemma A.1.2, applied with $\lambda = 0$ and α_T^0 being the LORD critical values.

A.4 Additional numerical experiments

A.4.1 Sample size

Figure A.4 illustrates results when the sample size N, i.e., the subjects number per group, takes values in the set $\{25, 50, \ldots, 150\}$. As expected, the power plots show that the detection problem becomes easier when N increases. In fact, for large N the power of all procedures converge to 1. We see that our rewarded procedures do well on the whole range of N values and improve substantially on existing OMT procedures for small and moderate values of N, including our default value $N = 25$.

A.4.2 Signal strength

Here, we vary the strength of the signal p_3 in the set $\{0.1, 0.2, \ldots, 1\}$. We see that the SUR procedures dominate their base counterparts, as expected. In addition, depending on the signal strength, the gain in power can be considerable. Also note that, perhaps surprisingly, all curves exhibit a decrease in power for p_3 near 1. Since this happens even for the original OB procedure, this is not due to the super-uniformity reward, but could perhaps be caused by the behavior of the power function of multiple Fisher exact tests taken at different levels.

A.4.3 Local alternatives

As Figure A.4 demonstrates, for a fixed value of the signal strength p_3, the detection problem becomes easier as N increases, so that all procedures attain a power of 1. In this section we are interested in obtaining a more refined analysis of the various power curves when N is large. To this end, we introduce local alternatives, i.e. we now model p_3 as a function of the sample size N. To be more specific, we take $N \in \{5, 10, \ldots, 30\} \times 1000$ and set $p_3 = p_1 + \frac{1}{\sqrt{N}}$ for mFDR procedures and, $p_3 = p_1 + \frac{1.5}{\sqrt{N}}$ for FWER procedures, we fix $p_1 = p_2 = 0.1$, and generate simulated

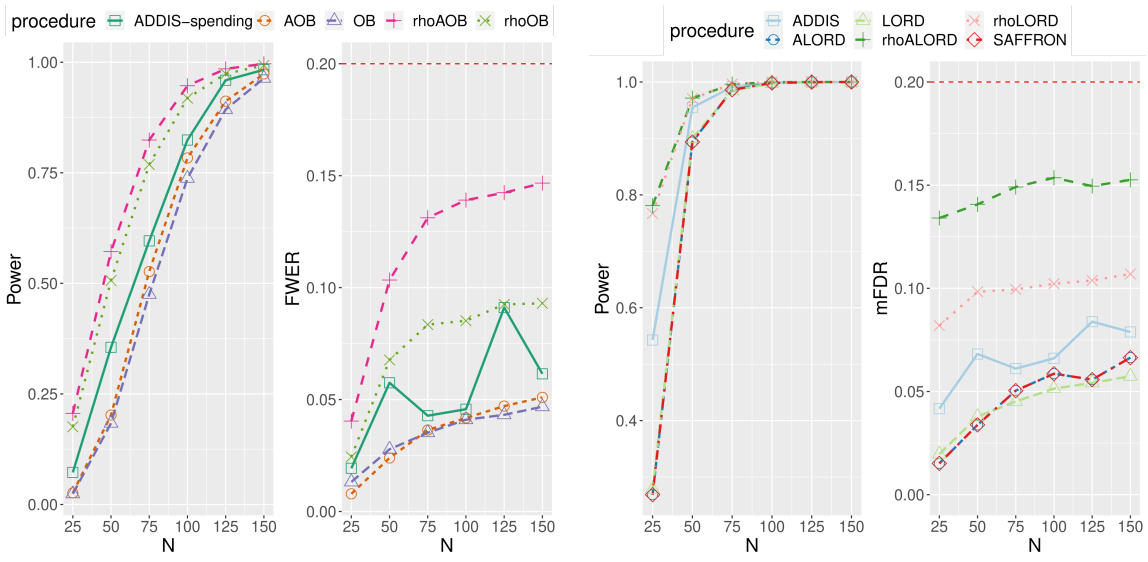

Figure A.4: Power and type I error rates of the considered procedures versus $N \in \{25, 50, \ldots, 150\}$, the number of subjects in the groups.

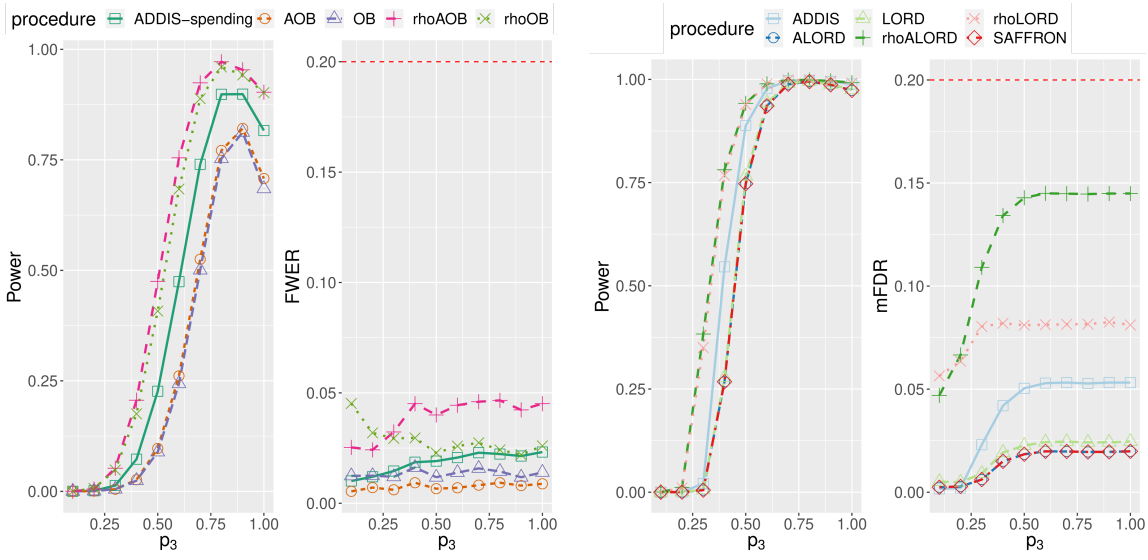

Figure A.5: Power and type I error rates of the considered procedures versus the strength of the signal $p_3 \in \{0.1, 0.2, \ldots, 0.9, 1\}$.

data as in Section 5.2. Figure A.6 displays power and error rates for this data. Taking N as a (crude) proxy for discreteness, we observe that even with a low discreteness (say $N \leq 30000$) the SUR methods still provide some degree of improvement. Finally, for FWER procedures, ADDIS-spending provides the best power performance over the whole range of the experiment.

This might be explained by the setting causing very conservative nulls p-values (*i.e.* very close to 1), thus allowing the discarding scheme to redistribute and spend a large part of the wealth on testing alternative hypotheses. Using the SUR method along with the discarding scheme (Tian and Ramdas, 2019, 2021) might provide an interesting avenue for further improvement, but this would define yet another class of procedures, which is outside of the scope of this paper.

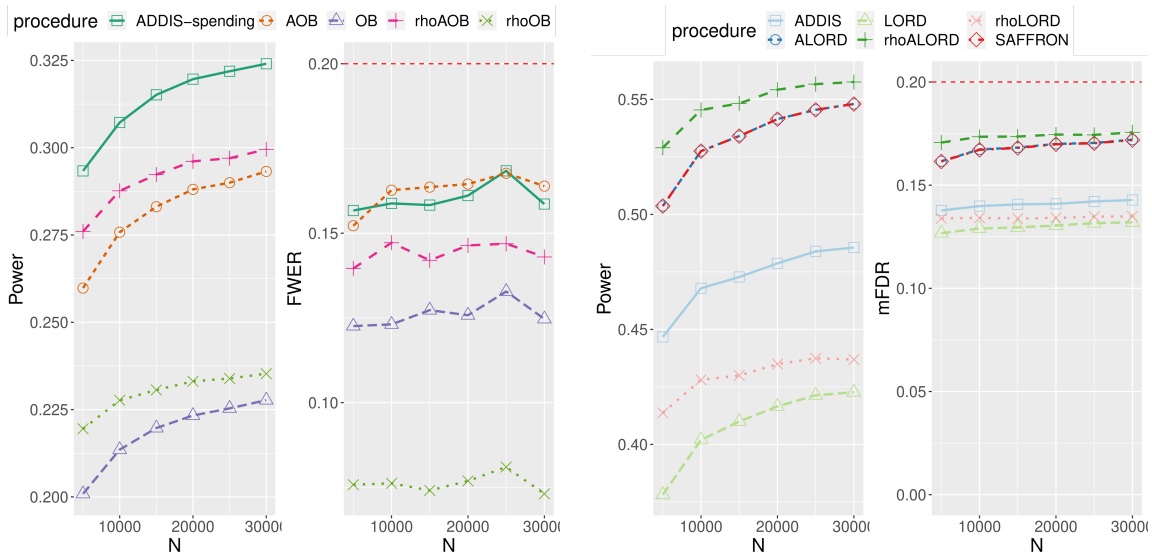

Figure A.6: Power and type I error rates of the considered procedures versus $N \in \{5, 10, \ldots, 30\} \times 1000$, with local alternatives.

A.4.4 Adaptivity parameter

We study the choice of λ for the procedures using adaptivity. It seems that $\lambda = 0.5$ is a reasonable choice for the adaptive procedures.

A.4.5 Rectangular kernel bandwidth

Finally, we study the choice of the bandwidth parameter for the rectangular kernel used for the rewarded procedures. As we can see, using a smaller bandwidth provides the best performance for the mFDR controlling rewarded procedures, whereas FWER controlling procedures require a larger bandwidth. The choices $h = 100$ for FWER controlling procedures, and $h = 10$ for mFDR controlling procedures seem reasonable although not necessarily optimal.

A.5 Additional figures for the analysis of IMPC data

A.5.1 Localization of small p-values

Figures A.9 and A.10 show that small p-values mostly occur at the beginning of the data set, both for male and female mice.

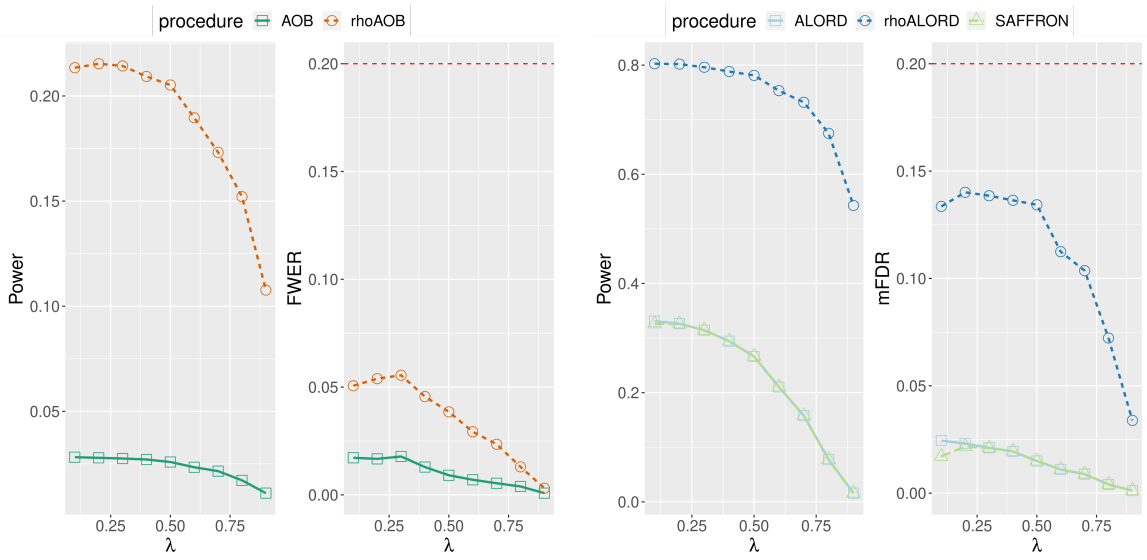

Figure A.7: Power and type I error rates, for the considered procedures, versus the adaptivity parameter λ.

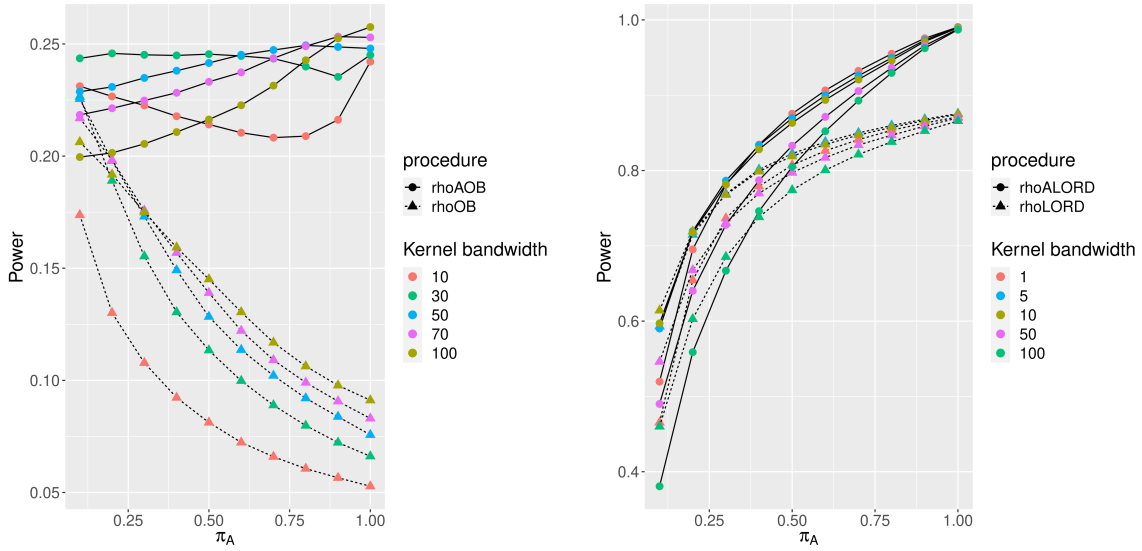

Figure A.8: Power for FWER (left) and mFDR (right) rewarded procedures versus the proportion of signal π_A, for different kernel bandwidths.

A.5.2 Figures for female mice in the IMPC data

Figures A.11 and A.12 display the critical values of the studied online procedures when applied to the IMPC data in the case of female mice.

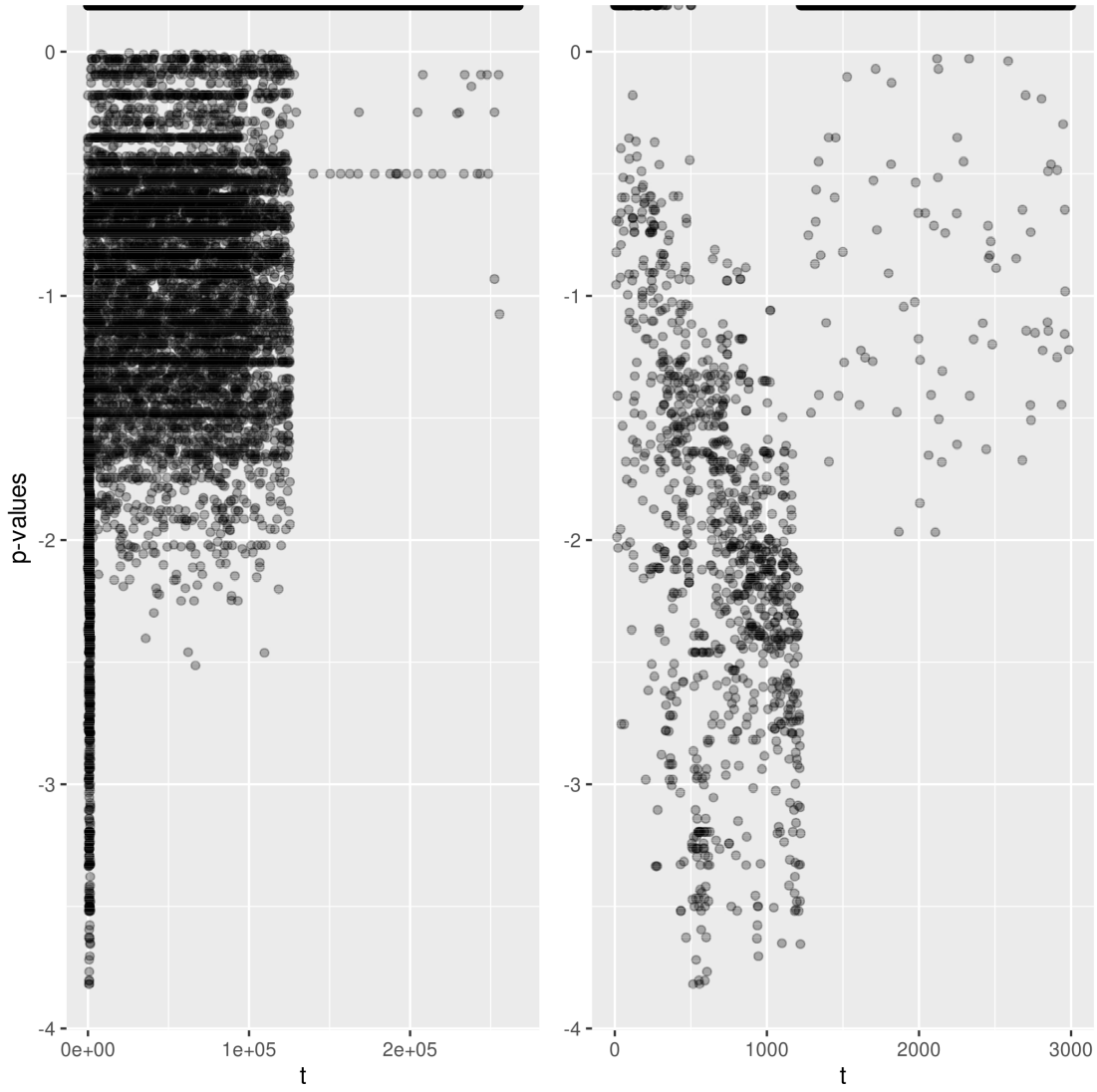

Figure A.9: p-values for male mice in the IMPC data of Section 2.5.3. The left panel presents all p-values, the right panel the first 3000 p-values. The p-values have been transformed as in Figure 2.3.

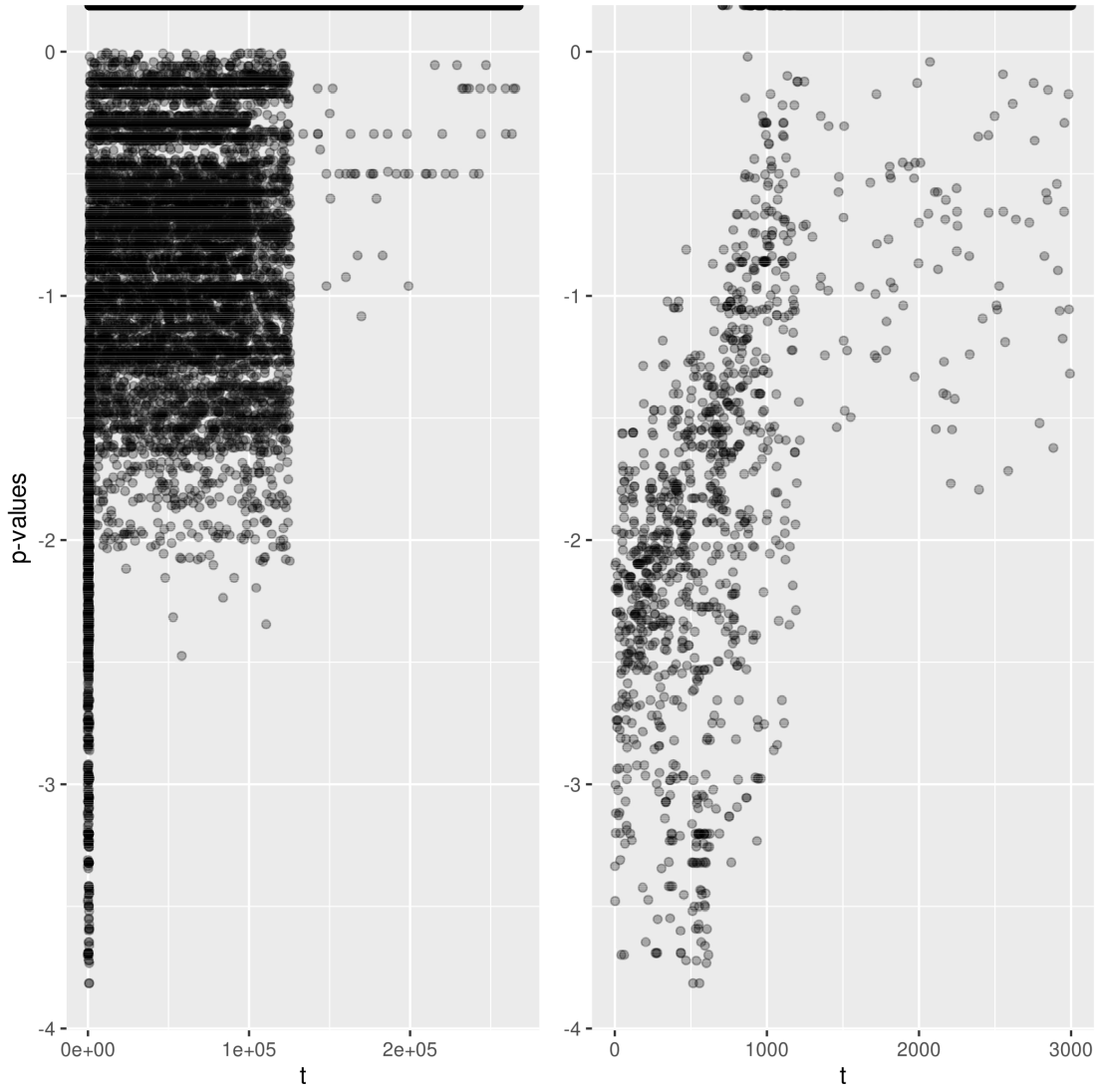

Figure A.10: p-values for female mice in the IMPC data of Section 2.5.3. The left panel presents all p-values, the right panel the first 3000 p-values. The p-values have been transformed as in Figure 2.3.

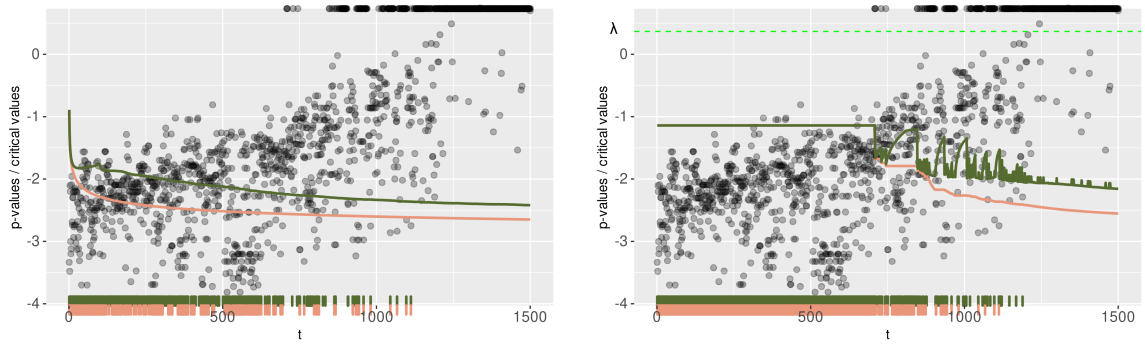

Figure A.11: Same as Figure 2.7 but for female mice of IMPC data (see Section 2.5.3).

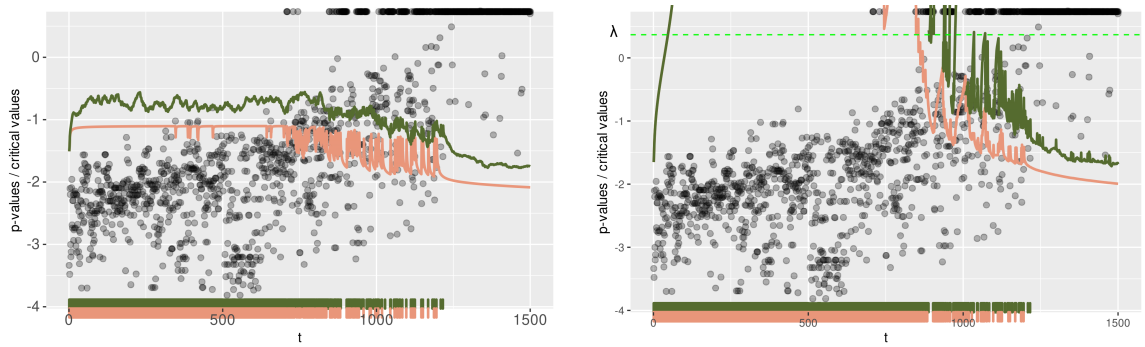

Figure A.12: Same as Figure 2.8 for female mice of IMPC data (see Section 2.5.3).

Appendix B

Supplementary material of Chapter 3

Outline of the current chapter

B.1 Power results

B.1.1 Top-k setting

Definition B.1.1 *The sparse one-sided Gaussian location model of parameter m, b, c, β, denoted as $\mathbf{P}^{(m)}_{b,c,\beta}$, is defined as follows: $p_i = \overline{\Phi}(X_i)$, $1 \leq i \leq m$, the X_i's are independent, with $X_i \sim \mathcal{N}(0,1)$ for $i \in \mathcal{H}_0$ and $X_i \sim \mathcal{N}(\mu_m, 1)$ otherwise, for $\mu_m = \sqrt{2\beta \log m} + b$, $b > 0$, and $m_1 = |\mathcal{H}_1| = cm^{1-\beta}$, $c \in (0,1)$, $\beta \in [0,1)$.*

Note that $\beta = 0$ is the dense case for which the alternative mean $\mu_m = b > 0$ is a fixed quantity, whereas $\beta > 0$ in the sparse case, for which $\mu_m = \sqrt{2\beta \log m} + b$ tends to infinity. In both case, the magnitude of alternative mean is defined to be on the 'verge of detectability' where the BH procedure has some non-zero power, see Bogdan et al. (2011); Neuvial and Roquain (2012); Abraham et al. (2021) for instance.

Theorem B.1.1 *Let $\alpha \in (0,1)$. In the above one-sided Gaussian location model $\mathbf{P}^{(m)}_{b,c,\beta}$, the number of rejections $\hat{k}_\alpha$ of the BH procedure is such that, as m grows to infinity,*

$$\mathbf{P}^{(m)}_{b,c,\beta}(t^*_m \leq \alpha \hat{k}_\alpha/m \leq t^\sharp_m) \geq 1 - 2e^{-dm_1}, \text{ for } m_1/m \lesssim t^*_m \leq t^\sharp_m \lesssim m_1/m, \tag{B.1}$$

for some constant $d > 0$ (depending on α, β, b), where $t^*_m \in (0,1)$ is the unique solution of $G_m(t) = 2t/\alpha$, $t^\sharp_m \in (0,1)$ is the unique solution of $G_m(t) = 0.5t/\alpha$, and where $G_m(t) = (m_0/m)t + (m_1/m)\overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu_m)$, with $\overline{\Phi}(z) = \mathbf{P}(Z \geq z)$, $z \in \mathbb{R}$.

Proof B.1.1 *First let $F_m(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - \mu_m)$, $\Psi_m(t) = F_m(t)/t$ and observe that Ψ_m is continuous decreasing on $(0,1]$ with $\lim_0 \Psi_m = +\infty$. This implies that $t^*_m, t^\sharp_m \in (0,1)$ as described in the statement both exist, with*

$$t^*_m = \Psi_m^{-1}(\alpha/2), \;\; t^\sharp_m = \Psi_m^{-1}(\tau_m(2\alpha)), \;\; \tau_m(\alpha) = \frac{m}{m_1}\left(\frac{1}{\alpha} - \frac{m_0}{m}\right).$$

We first establish

$$t^*_m \gtrsim m_1/m \tag{B.2}$$

$$t^\sharp_m \lesssim m_1/m. \tag{B.3}$$

If $\beta = 0$, then $m_0/m = 1 - c$, $m_1/m = c$, $\mu_m = b$, $\tau_m > 0$, $F_m(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - b)$, $\Psi_m(t) = \overline{\Phi}(\overline{\Phi}^{-1}(t) - b)/t$, $\tau_m(\alpha)$ all do not depend on m. Hence, t^*_m and $t^\sharp_m$ are both constant, which establishes (B.2) *and* (B.3)*. Let us now turn to the sparse case, for which $\beta \in (0,1)$. The inequality* (B.3) *follows from the upper bound*

$$0.5t^\sharp_m/\alpha = G_m(t^\sharp_m) \leq t^\sharp_m + m_1/m.$$

For (B.2)*, the analysis is slightly more involved. We first prove that for m large enough*

$$\overline{\Phi}^{-1}(t^*_m) \leq \mu_m - b. \tag{B.4}$$

This will establish (B.2)*, since it implies $F_m(t^*_m) \geq F_m(\overline{\Phi}(\mu_m - b)) = \overline{\Phi}(-b) > 0$ and also $t^*_m = (\tau_m(\alpha/2))^{-1}F_m(t^*_m) \gtrsim m_1/m$. On the one hand,*

$$\Psi_m(\overline{\Phi}(\mu_m - b)) = \frac{\overline{\Phi}(-b)}{\overline{\Phi}(\mu_m - b)} \geq \overline{\Phi}(-b)\frac{\mu_m - b}{\phi(\mu_m - b)} = \Phi(-b)m^\beta\sqrt{2\beta\log m}$$

because $\mu_m - b = \sqrt{2\beta\log m}$ and $\phi(\mu_m - b) = m^{-\beta}$, and by using $\Phi(x) \leq \phi(x)/x$ for all $x > 0$. On the other hand,

$$\Psi_m(t^*_m) = \tau_m(\alpha/2) \leq \frac{2}{\alpha}m^\beta.$$

Hence, for m large enough, we have $\Psi_m(\overline{\Phi}(\mu_m - b)) \geq \Psi_m(t^*_m) = \Psi_m(\overline{\Phi}(\overline{\Phi}^{-1}(t^*_m)))$, which in turn implies (B.4)*.*

We now turn to prove the result (B.1) *and follow for a classical concentration argument. Let*

$$\hat{G}_m(t) = m^{-1}\sum_{i=1}^m \mathbf{1}\{p_i \leq t\}, \quad t \in [0,1],$$

so that $G_m(t) = \mathbf{E}\hat{G}_m(t)$ *for all* $t \in [0,1]$. *Hence, for all* $t \in (0,1)$,

$$\begin{aligned}\mathbf{P}(\alpha\hat{k}_\alpha/m < t) &\leq \mathbf{P}\left(\hat{G}_m(t) \leq t/\alpha\right) \\ &= \mathbf{P}\left(\hat{G}_m(t) - G_m(t) \leq t/\alpha - G_m(t)\right),\end{aligned}$$

because $\alpha\hat{k}_\alpha/m = \max\{t \in (0,1) : \hat{G}_m(t) \geq t/\alpha\}$ *by definition of* $\hat{k}_\alpha$. *Applying this with* $t = t^*_m$, *this gives*

$$\begin{aligned}\mathbf{P}(\alpha\hat{k}_\alpha/m < t^*_m) &= \mathbf{P}\left(\hat{G}_m(t^*_m) - G_m(t^*_m) \leq -G_m(t^*_m)\right) \\ &\leq \exp\left(-cmG_m(t^*_m)\right) \leq \exp\left(-Cm_1F_m(t^*_m)\right),\end{aligned}$$

for some constant $C > 0$, *by applying Bernstein's inequality. Since* $F_m(t^*_m) \geq \overline{\Phi}(-b) > 0$, *this gives* $\mathbf{P}(\alpha\hat{k}_\alpha/m < t^*_m) \leq e^{-dm_1}$ *for* m *large enough and some constant* $d > 0$.

Next, for all $t \in [t^\sharp_m, 1)$, *still applying Bernstein's inequality,*

$$\begin{aligned}&\mathbf{P}(\alpha\hat{k}_\alpha/m > t) \\ &\leq \sum_{k=1}^m \mathbf{1}\{\alpha k/m > t\}\mathbf{P}\left(\hat{G}_m(\alpha k/m) - G_m(\alpha k/m) \geq k/m - G_m(\alpha k/m)\right) \\ &\leq \sum_{k=1}^m \mathbf{1}\{\alpha k/m > t\}\exp\left(-m\frac{(k/m - G_m(\alpha k/m))^2}{G_m(\alpha k/m) + (1/3)(k/m - G_m(\alpha k/m))}\right) \leq m\exp\left(-Cmt^\sharp_m\right),\end{aligned}$$

because for all $\alpha k/m \geq t^\sharp_m$, $k/m - G_m(\alpha k/m) \geq G_m(\alpha k/m) \geq G_m(t^\sharp_m) = 0.5t^\sharp_m/\alpha$ *(given the monotonicity of* $t \mapsto G_m(t)/t$*). Applying this for* $t = t^\sharp_m \in (0,1)$, *we obtain*

$$\mathbf{P}(\alpha\hat{k}_\alpha/m > t^\sharp_m) \leq e^{-dm_1},$$

because $t^\sharp_m \geq t^*_m \gtrsim m_1/m$. *This proves the result.*

B.1.2 Pre-ordered setting

We introduce below a model generalizing the one of Lei and Fithian (2016) to the possibly sparse case. Here, without loss of generality we assume that the ordering π is identity, that is, $\pi(i) = i$ for all $i \in \{1, \ldots, m\}$. Below, with some abuse, the notation π will be re-used to stick with the notation of Lei and Fithian (2016).

Definition B.1.2 *The sparse VCT model of parameters* m, π, β, F_0, F_1, *denoted as* $\mathbf{P}^{(m)}_{\pi,\beta,F_0,F_1}$, *is the p-value mixture model where* $(p_k, H_k) \in [0,1] \times \{0,1\}$, $1 \leq k \leq m$, *are independent and generated as follows:*

- *the* H_k, $1 \leq k \leq m$, *are independent and* $\mathbf{P}(H_k = 1) = \pi_m(k/m)$, $1 \leq k \leq m$, *with* $\pi_m(x) = \pi(m^\beta x)$, $x \geq 0$, *where* $\pi : [0,\infty) \to [0,1)$ *is some measurable function (instantaneous signal probability function) with* $\pi(0) > 0$ *and* $\pi(x) = \pi(1)$ *for* $x \geq 1$ *and* $\beta \in [0,1)$ *is a sparsity parameter.*
- *conditionally on* $H_1, \ldots, H_k$, *the p-values* p_k, $1 \leq k \leq m$, *are independent, with a marginal distribution super-uniform under the null:* $p_k \mid H_k = 0 \sim F_0$, $1 \leq k \leq m$, *where* F_0 *is a c.d.f. with* $F_0(t) \leq t$ *for all* $t \in [0,1]$*; and* $p_k \mid H_k = 1 \sim F_1$, $1 \leq k \leq m$, *where* F_1 *is some alternative c.d.f.*

We denote $\Pi(t) := t^{-1}\int_0^t \pi(s)ds$, *with* $\Pi(0) = \pi(0)$ *and*

$$\Pi_m(t) := t^{-1}\int_0^t \pi_m(s)ds = t^{-1}\int_0^t \pi(m^\beta s)ds = m^{-\beta}t^{-1}\int_0^{m^\beta t} \pi(s)ds = \Pi(m^\beta t)$$

the expected fraction of signal before time mt. *We also let* $\pi_1 := \Pi_m(1) = \int_0^1 \pi_m(s)ds = m^{-\beta}\Pi(1)$ *the overall expected fraction of signal. We consider the asymptotic where* m *tends to infinity and* F_0, F_1 *are fixed.*

When $\beta = 0$, π_m, Π_m are fixed and we recover the dense VTC model introduced in Lei and Fithian (2016) (also noting that we are slightly more general because F_0 is possibly non-uniform and F_1 not concave). Interestingly, the above formulation can also handle the sparse case for which $\beta \in (0,1)$ and the probability to generate a signal is shrunk to 0 by a factor m^β. For instance, if $\pi(1) = 0$, the model only generates null p-values $p_{k+1}, \dots, p_m$ for $k \geq m^{1-\beta}$.

We now analyze the asymptotic behavior of the number of rejections of the LF procedure. By following the same heuristic than in Lei and Fithian (2016) (which follows by a concentration argument), we have from (3.32) that for $k = \lfloor mt \rfloor$,

$$\begin{aligned}
\widehat{\mathrm{FDP}}_k &= \frac{s}{1-\lambda}\frac{1+\sum_{i=1}^k \mathbf{1}\{p_i > \lambda\}}{1 \vee \sum_{i=1}^k \mathbf{1}\{p_i \leq s\}} \\
&\approx \frac{s}{1-\lambda}\frac{\left(\sum_{i=1}^k (1-\pi_m(i/m))\right)(1-F_0(\lambda)) + \left(\sum_{i=1}^k \pi_m(i/m)\right)(1-F_1(\lambda))}{\left(\sum_{i=1}^k (1-\pi_m(i/m))\right)F_0(s) + \left(\sum_{i=1}^k \pi_m(i/m)\right)F_1(s)} \\
&\approx \frac{1+\Pi_m(t)\left(\frac{1-F_1(\lambda)}{1-\lambda}-1\right)}{1+\Pi_m(t)\left(\frac{F_1(s)}{s}-1\right)} = \mathrm{FDP}^\infty(m^\beta t),
\end{aligned}$$

by assuming $F_0(s) = s$, $F_0(\lambda) = \lambda$, $F_1(s) > s$, $F_1(\lambda) > \lambda$ and by letting

$$\mathrm{FDP}^\infty(t) = \frac{1+\Pi(t)\left(\frac{1-F_1(\lambda)}{1-\lambda}-1\right)}{1+\Pi(t)\left(\frac{F_1(s)}{s}-1\right)}, \quad t \geq 0. \tag{B.5}$$

By (3.32), the quantity $\hat{k}_\alpha/m^{1-\beta}$ should be asymptotically close to

$$t^*_\alpha = \max\{t \in [0,+\infty) \,:\, \mathrm{FDP}^\infty(t) \leq \alpha\}, \tag{B.6}$$

with the convention $t^*_\alpha = +\infty$ if the set is not upper bounded. We should however ensure that the latter set is not empty. For this, we let

$$\underline{\alpha} = \frac{1+\pi(0)\left(\frac{1-F_1(\lambda)}{1-\lambda}-1\right)}{1+\pi(0)\left(\frac{F_1(s)}{s}-1\right)}. \tag{B.7}$$

Hence, $\hat{r}_\alpha = \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_i \leq s\}$, the number of rejections of LF procedure, should be close to $\left(\sum_{i=1}^{\hat{k}_\alpha}(1-\pi_m(i/m))\right)F_0(s) + \left(\sum_{i=1}^{\hat{k}_\alpha}\pi_m(i/m)\right)F_1(s) \gtrsim \hat{k}_\alpha s \approx m^{1-\beta}t^*_\alpha s$. This heuristic is formalized in the next result.

Theorem B.1.2 *Consider a sparse VCT model* $\mathbf{P}^{(m)}_{\pi,\beta,F_0,F_1}$ *with parameters* β, π, F_0, F_1 *(see Definition B.1.2) and the LF procedure with parameter* s, λ *(see* (3.32)*), with the assumptions:*

(i) $\Pi : t \in [0,\infty) \to \mathbb{R}_+$ *is continuous decreasing and L-Lipschitz;*

(ii) $F_0(s) = s$, $F_0(\lambda) = \lambda$, $F_1(s) > s$, $F_1(\lambda) > \lambda$;

(iii) $\alpha > \underline{\alpha}$ *where* $\underline{\alpha}$ *is defined by* (B.7).

Let $\alpha' = (\underline{\alpha} + \alpha)/2 \in (\underline{\alpha}, \alpha)$, $t^*_{\alpha'} \in (0, +\infty]$ *given by* (B.6), $t^*_m = t^*_{\alpha'} \wedge m^\beta$ *and let* $a \geq 1$ *be an integer* $a \leq m^{1-\beta} t^*_m$ *such that* $r = \frac{4}{a^{1/4}} \left(\frac{1}{s} + \frac{1}{1-\lambda} \right)$ *is small enough to provide* $r \leq (\alpha - \underline{\alpha})/4$. *Then the number of rejections* $\hat{r}_\alpha = \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_i \leq s\}$ *of the LF procedure* (3.32) *is such that*

$$\mathbf{P}^{(m)}_{\pi,\beta,F_0,F_1}(\hat{r}_\alpha < r^*_m) \leq 2(2 + a^{1/2}) e^{-2a^{1/2}}, \quad r^*_m = \lfloor m^{1-\beta} t^*_m \rfloor s/2. \tag{B.8}$$

In particular, choosing $a = 1 + \lfloor (\log m)^2 \rfloor$, *we have as* m *grows to infinity,* $m^{1-\beta}/\hat{r}_\alpha = O_P(1)$.

Condition (ii) is more general that in Lei and Fithian (2016) and allows to handle binary p-values, like in the 'knockoffs' situation (for which F_0 and F_1 are not continuous). The condition (iii) was overlooked in Lei and Fithian (2016), but it is needed to ensure the existence of t^*_α. It reads equivalently

$$\pi(0) > \frac{1 - \alpha}{1 - \frac{1 - F_1(\lambda)}{1 - \lambda} + \alpha \left(\frac{F_1(s)}{s} - 1 \right)}, \tag{B.9}$$

which provides that the probability to generate a null is sufficiently large at the beginning of the p-value sequence, with a minimum amplitude function of $F_1(s)$ and $F_1(\lambda)$. Note that in the 'knockoffs' case where $s = \lambda = 1/2$, we have $\underline{\alpha} = \frac{1 - \pi(0)M}{1 + \pi(0)M}$ where $M = 2F_1(1/2) - 1 > 0$ can be interpreted as a 'margin'. Hence, the critical level $\underline{\alpha}$ is decreasing in $\pi(0)M$. Hence, the setting is more favorable either when $\pi(0)$ increases, or when the margin M increases.

Proof B.1.2 *First note that* $\mathrm{FDP}^\infty(t)$ *is an decreasing function of* $\Pi(t)$ *because* $\frac{1 - F_1(\lambda)}{1 - \lambda} < 1 < \frac{F_1(s)}{s}$, *see* (B.5). *Since* $\Pi(t)$ *is decreasing from* $\pi(0)$ *to* $\pi(1) = \Pi(+\infty)$, *we have that* $\mathrm{FDP}^\infty : [0, +\infty) \to [\underline{\alpha}, \overline{\alpha}]$ *is continuous increasing, where* $\overline{\alpha} = \left(1 + \pi(1) \left(\frac{1 - F_1(\lambda)}{1 - \lambda} - 1 \right)\right) / \left(1 + \pi(1) \left(\frac{F_1(s)}{s} - 1 \right)\right)$. *Hence, if* $\alpha' < \overline{\alpha}$, *we have* $0 < t^*_{\alpha'} < +\infty$, $t^*_m = t^*_{\alpha'}$ *for* m *large enough, and thus* $\mathrm{FDP}^\infty(t^*_m) = \alpha'$. *If* $\alpha' \geq \overline{\alpha}$, $t^*_{\alpha'} = +\infty$, $t^*_m = m^\beta$ *and* $\mathrm{FDP}^\infty(t^*_m) \leq \alpha'$. *Both cases are considered in what follows. Consider the events*

$$\Omega_1 = \left\{ \sup_{a \leq k \leq m} \left| k^{-1} \sum_{i=1}^k \mathbf{1}\{p_i > \lambda\} - k^{-1} \sum_{i=1}^k \mathbf{P}(p_i > \lambda) \right| \leq 1/a^{1/4} \right\};$$
$$\Omega_2 = \left\{ \sup_{a \leq k \leq m} \left| k^{-1} \sum_{i=1}^k \mathbf{1}\{p_i \leq s\} - k^{-1} \sum_{i=1}^k \mathbf{P}(p_i \leq s) \right| \leq 1/a^{1/4} \right\}.$$

By Lemma B.1.2, the event $\Omega_1 \cap \Omega_2$ *occurs with probability larger than* $1 - 2(2 + a^{1/2}) e^{-2a^{1/2}}$. *Let*

$$e_1 = 1 + \Pi_m(m^{-\beta} t^*_m) \left(\frac{1 - F_1(\lambda)}{1 - \lambda} - 1 \right) = 1 + \Pi(t^*_m) \left(\frac{1 - F_1(\lambda)}{1 - \lambda} - 1 \right);$$
$$e_2 = 1 + \Pi_m(m^{-\beta} t^*_m) \left(\frac{F_1(s)}{s} - 1 \right) = 1 + \Pi(t^*_m) \left(\frac{F_1(s)}{s} - 1 \right),$$

be the numerator and denominator of $\mathrm{FDP}^\infty(t^*_{\alpha'})$*, so that* $e_1/e_2 = \mathrm{FDP}^\infty(t^*_m) \le \alpha'$*. Let* $k_0 = \lfloor m^{1-\beta} t^*_m \rfloor \le m$*. Provided that* $k_0 \ge a$*, we have*

$$\begin{aligned}\left| k_0^{-1} \sum_{i=1}^{k_0} \mathbf{P}(p_i > \lambda) - (1-\lambda) e_1 \right| &\le \left| k_0^{-1} \sum_{i=1}^{k_0} \pi_m(i/m) - \Pi_m(m^{-\beta} t^*_m) \right| |(1 - F_1(\lambda)) - (1-\lambda)| \\ &\le \left| k_0^{-1} \sum_{i=1}^{k_0} \pi_m(i/m) - \Pi_m(k_0/m) \right| + \left| \Pi_m(k_0/m) - \Pi_m(m^{-\beta} t^*_m) \right| \\ &\le 1/a + L/m^{1-\beta},\end{aligned}$$

by applying Lemma B.1.1 and using that $\Pi(\cdot)$ *is L-Lipschitz. Similarly,*

$$\left| k_0^{-1} \sum_{i=1}^{k_0} \mathbf{P}(p_i \le s) - s e_2 \right| \le 1/a + L/m^{1-\beta}.$$

We deduce that on $\Omega_1 \cap \Omega_2$ *and when* $k_0 \ge a$*, we have*

$$\widehat{\mathrm{FDP}}_{k_0} \le \frac{e_1 + \frac{1}{a(1-\lambda)} + \frac{L}{m^{1-\beta}(1-\lambda)} + \frac{1}{k_0(1-\lambda)} + \frac{1}{a^{1/4}(1-\lambda)}}{\frac{1}{as} \vee \left(e_2 - \frac{1}{as} - \frac{L}{m^{1-\beta} s} - \frac{1}{k_0 s} - \frac{1}{a^{1/4} s} \right)} \le \frac{e_1 + r}{e_2 - r} \le \frac{e_1}{e_2} + 4r,$$

provided that $e_2 \ge 2r$*, because* $e_1 \le 1$*,* $e_2 \ge 1$*, and by considering* r *as in the statement. Since* $e_1/e_2 \le \alpha' \le \alpha - 4r$ *and* $e_2 \ge 1 \ge 2r$ *by assumption, we have* $\widehat{\mathrm{FDP}}_{k_0} \le \alpha$ *and thus* $\hat{k}_\alpha \ge k_0$ *on* $\Omega_1 \cap \Omega_2$*. The result is proved by noting that* $\hat{r}_\alpha = \sum_{i=1}^{\hat{k}_\alpha} \mathbf{1}\{p_i \le s\} \ge \sum_{i=1}^{k_0} \mathbf{1}\{p_i \le s\} \ge (e_2 - r) k_0 s \ge k_0 s/2$ *on this event.*

Lemma B.1.1 *In the setting of Theorem B.1.2, we have for all* $a \ge 1$*,* $m \ge a$*,*

$$\sup_{a \le k \le m} \left| k^{-1} \sum_{i=1}^{k} \pi_m(i/m) - \Pi_m(k/m) \right| \le 1/a. \tag{B.10}$$

Proof B.1.3 *First note that because* π_m *is nonnegative continuous decreasing, we have for all* $k \ge 1$*,*

$$(1/k) \sum_{i=1}^{k} \pi_m(i/m) \le \Pi_m(k/m) = (m/k) \int_0^{k/m} \pi_m(s) ds \le (1/k) \sum_{i=0}^{k-1} \pi_m(i/m).$$

Since $\pi_m(0) \le 1$*, the result is clear.*

This following lemma is similar to Lemma 1 in Lei and Fithian (2016).

Lemma B.1.2 *Let* $X_i \sim \mathcal{B}(p_i)$*,* $1 \le i \le m$*, be independent Bernoulli variables for* $p_i \in [0,1]$*,* $1 \le i \le m$*. Then we have for all* $a \ge 1$ *and* $m \ge a$*,*

$$\mathbf{P}\left(\sup_{a \le k \le m} \left| k^{-1} \sum_{i=1}^{k} X_i - p_i \right| \ge 1/a^{1/4} \right) \le (2 + a^{1/2}) e^{-2a^{1/2}}. \tag{B.11}$$

Proof B.1.4 *By Hoeffding's inequality, we have for all $x > 0$,*

$$\mathbf{P}\left(\sup_{1\le k\le a}\left|k^{-1}\sum_{i=1}^{k}(H_i-\pi_m(i/m))\right|\ge x\right)\le 2\sum_{k\ge a}e^{-2kx^2}=\frac{2}{1-e^{-2x^2}}e^{-2ax^2}\le(2+1/x^2)e^{-2ax^2}.$$

We deduce the result by considering $x = 1/a^{1/4}$.

B.1.3 Online setting

Definition B.1.3 *The online one-sided Gaussian mixture model of parameters π_1, F_1, denoted by $\mathbf{P}_{\pi_1,F_1}$, is given by the p-value stream $(p_k, H_k) \in [0,1]\times\{0,1\}$, $k \ge 1$, which is i.i.d. with*

- $\mathbf{P}(H_k = 1) = \pi_1$ *for some fixed $\pi_1 \in (0,1)$;*
- *p-values are uniform under the null: $p_k \mid H_k = 0 \sim U(0,1)$;*
- *p-values have the same alternative distribution: $p_k \mid H_k = 1 \sim F_1$, where F_1 is the c.d.f. corresponding to the one-sided Gaussian problem, that is, $F_1(x) = \bar{\Phi}(\bar{\Phi}^{-1}(x)-\mu)$, $x \in [0,1]$, for some $\mu > 0$.*

Here, we make no sparsity assumption: π_1 is assumed to be constant across time. This will ensure that the online procedure maintains a chance to make discoveries even when the time grows to infinity.

Theorem B.1.3 *Consider the one-sided Gaussian online mixture model and the LORD procedure with $W_0 \in (0,\alpha)$ and a spending sequence $\gamma_k = \frac{1}{k(\log(k))^\gamma}$, $\gamma > 1$. Then its rejection number $R(k)$ at time k satisfies: for all $a \in (0,1)$, $k \ge 1$,*

$$\mathbf{P}(R(k) < k^{1-a}) \le ck^{-a}, \tag{B.12}$$

where c is some constant only depending on α ,W_0, γ, μ and π_1. In particular, $k^{1-a}/R(k) = O_P(1)$ when k tends to infinity.

Proof B.1.5 *We get inspiration from the power analysis of Javanmard and Montanari (2018). Let $c = \min(\alpha - W_0, W_0)$. By definition (3.42), the LORD procedure makes (point-wise) more rejections than the procedure given by the critical values*

$$\alpha_T = c\max\{\gamma_{T-\tau_j}, j \ge 0\}, \tag{B.13}$$

where, for any $j \ge 1$, τ_j is the first time that the procedure makes j rejections, that is,

$$\tau_j = \min\{t \ge 0 \,:\, R(t) \ge j\} \quad (\tau_j = +\infty \text{ if the set is empty}), \tag{B.14}$$

(note that $\tau_0 = 0$) for $R(T) = \sum_{t=1}^{T}\mathbf{1}\{p_t \le \alpha_t\}$. Let $\Delta_j = \tau_j - \tau_{j-1}$ the time between the j-th rejection and the $(j-1)$-th rejection. It is clear that $(R(t))_{t\ge1}$ is a renewal process with holding times $(\Delta_j)_{j\ge1}$ and jump times $(\tau_j)_{j\ge1}$. In particular, the Δ_j's are i.i.d. As a result, we have for all $r, k \ge 1$,

$$\mathbf{P}(R(k) < r) \le \mathbf{P}(\tau_r \ge k) = \mathbf{P}(\Delta_1 + \cdots + \Delta_r \ge k) \le r\mathbf{E}\Delta_1/k,$$

where

$$\mathbf{E}\Delta_1 = \sum_{m\geq 1} \mathbf{P}(\Delta_1 \geq m) = \sum_{m\geq 1} \prod_{\ell=1}^{m} (1 - G(c\gamma_\ell)) \leq \sum_{m\geq 1} e^{-mG(c\gamma_m)}.$$

In addition, since G is concave,

$$\frac{G(x)}{x} \geq g'(x) = \pi_0 + \pi_1 c\, e^{\mu \bar{\Phi}^{-1}(x)} \geq e^{c'\sqrt{2\log(1/x)}} \geq (\log(1/x))^{\gamma+2},$$

for x small enough and $c, c' > 0$ some constants. This gives for large $m \geq M$, $e^{-mG(c\gamma_m)} \leq e^{-cm\gamma_m(\log(1/(c\gamma_m)))^{2+\gamma}} \leq e^{-2\log m}$, for some $M > 0$, by the choice made for γ_m. As a result,

$$\mathbf{E}\Delta_1 \leq C + \sum_{m\geq M} e^{-mG(c\gamma_m)} \leq C + \sum_{m\geq M} e^{-cm\gamma_m(\log(1/(c\gamma_m)))^{\gamma+2}} \leq C + \sum_{m\geq 1} e^{-2\log m} = C + \pi^2/6,$$

for some constant $C > 0$. This gives

$$\mathbf{P}(R(k) < r) \leq r(C + \pi^2/6)/k.$$

and taking $r = k^{1-a}$ gives (B.12).

B.2 Proofs

B.2.1 Proof of Proposition 3.2.1

For $j \geq 1$, let $\delta_j = \delta j^{-2}$, $\tau_j = 2^{-j}$ and

$$A_j = \left\{ \forall t \in [\tau_j, 1],\ n^{-1} \sum_{i=1}^{n} \mathbf{1}\{p_i \leq t\} \leq t\, \lambda_j \right\};$$
$$\lambda_j = h^{-1}\left(\frac{\log(1/\delta_j)}{n\tau_j/(1-\tau_j)} \right),$$

so that by Wellner's inequality, we have $\mathbf{P}(A_j) \geq 1 - \delta_j$ and with a union bound $\mathbf{P}(\cap_{j\geq 1} A_j) \geq 1 - \delta\pi^2/6$. Now let $t \in (0,1)$ and $j_0 = \min\{j \geq 1 \,:\, t \geq \tau_j\} = \min\{j \geq 1 \,:\, j \geq \log_2(1/t)\}$, so that $j_0 = \lceil \log_2(1/t) \rceil \geq 1$. This yields

$$\log(1/\delta_{j_0}) = \log(1/\delta) + 2\log\left(\lceil \log_2(1/t) \rceil\right).$$

On the event $\cap_{j\geq 1} A_j$, we have, since $t \in [\tau_{j_0}, 1]$ by definition,

$$n^{-1} \sum_{i=1}^{n} \mathbf{1}\{p_i \leq t\} \leq t\, \lambda_{j_0} = t h^{-1}\left(\frac{\log(1/\delta_{j_0})}{n\tau_{j_0}/(1-\tau_{j_0})} \right) = t h^{-1}\left(\frac{\log(1/\delta) + 2\log\left(\lceil \log_2(1/t) \rceil\right)}{n g(t)} \right),$$

because $\tau_{j_0} = 2^{-\lceil \log_2(1/t) \rceil}$. The result then comes from replacing δ by $\delta 6/\pi^2$.

B.2.2 Proof of Proposition 3.2.3

Let us prove it for the adaptive uniform Wellner envelope (the other ones being either simpler or provable by using a similar argument). The idea is to prove that on an event where the

(non-adaptive) Wellner envelope (3.14) is valid, we also have $m_0 \leq \hat{m}_0^{\text{Well}}$. The result is implied just by monotonicity (Lemma B.4.1).

For this, we come back to apply (3.13) with $(U_1, \ldots, U_n) = (p_i, i \in \mathcal{H}_0)$, $n = m_0$. Hence, on an event with probability at least $1 - \delta$, we have for all $t \in (0, 1)$,

$$m_0^{-1} \sum_{i \in \mathcal{H}_0} \mathbf{1}\{p_i \leq t\} \leq t\, h^{-1}\left(\frac{C_t}{t m_0}\right) \leq t\, \left(1 + \sqrt{C_t/(2tm_0)}\right)^2,$$

where we apply an upper bound coming from Lemma B.4.1. This gives

$$V_t/m_0 \geq 1 - t\, \left(1 + \sqrt{C_t/(2tm_0)}\right)^2 = 1 - t - \sqrt{2tC_t/m_0} - C_t/(2m_0).$$

As a result, $V_t \geq m_0(1-t) - \sqrt{2tC_t m_0} - C_t/2$ and thus $(1-t)m_0 - \sqrt{2tC_t} m_0^{1/2} - C_t/2 - V_t \leq 0$, which gives

$$m_0 \leq \left(\frac{\sqrt{2tC_t} + \sqrt{2tC_t + 4(1-t)(C_t/2 + V_t)}}{2(1-t)}\right)^2 = \left(\sqrt{\frac{tC_t}{2(1-t)^2}} + \sqrt{\frac{C_t}{2(1-t)^2} + \frac{V_t}{1-t}}\right)^2.$$

Since this is uniform in t, we can take the minimum over t, which gives the m_0 confidence bound $\hat{m}_0^{\text{Well}}$.

B.3 Tools of independent interest

B.3.1 A general envelope for a sequence of tests

An important basis for our work is the following theorem, which has the flavor of Lemma 1 of Katsevich and Ramdas (2020), but based on a different martingale inequality, derived from a Freedman type bound (see Section B.3.2).

Theorem B.3.1 *Consider a potentially infinite set of null hypotheses $H_1, H_2, \ldots$ for the distribution P of an observation X, with associated p-values $p_1, p_2, \ldots$ (based on X). Consider an ordering $\pi(1), \pi(2), \ldots$ (potentially depending on X) and a set of critical values $\alpha_1, \alpha_2, \ldots$ (potentially depending on X). Let $\lambda \in [0, 1)$ be a parameter and assume that there exists a filtration*

$$\mathcal{F}_k = \sigma\left((\pi(i))_{1 \leq i \leq k}, (\mathbf{1}\{p_{\pi(i)} \leq \alpha_i\})_{1 \leq i \leq k}, (\mathbf{1}\{p_{\pi(i)} > \lambda\})_{1 \leq i \leq k}\right), \ k \geq 1,$$

such that for all $k \geq 2$,

$$\mathbf{P}(p_{\pi(k)} \leq t \,|\, \mathcal{F}_{k-1}, H_{\pi(k)} = 0) \leq t \text{ for all } t \in [0, 1]. \tag{B.15}$$

Then, for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, it holds

$$\forall k \geq 1, \ \sum_{i=1}^{k} (1 - H_{\pi(i)}) \mathbf{1}\{p_{\pi(i)} \leq \alpha_i\} \leq \overline{V}_k,$$

for

$$\overline{V}_k = \sum_{i=1}^{k}(1-H_{\pi(i)})\mathbf{1}\{p_{\pi(i)} > \lambda\}\frac{\alpha_i}{1-\lambda} + \Delta\left(\sum_{i=1}^{k}(1-H_{\pi(i)})\nu_i\right), \tag{B.16}$$

where $\Delta(u) = 2\sqrt{\varepsilon_u}\sqrt{u \vee 1} + \frac{1}{2}\varepsilon_u$, $\varepsilon_u = \log((1+\kappa)/\delta) + 2\log(1+\log_2(u \vee 1))$, $u > 0$, $\kappa = \pi^2/6$. *and* $\nu_i = \alpha_i(1 + \min(\alpha_i, \lambda)/(1-\lambda))$, *for* $i \geq 1$.

Proof B.3.1 *By Lemma B.3.1, we can apply Corollary B.3.2 (it self coming from Freedman's inequality) with*

$$\xi_i = (1-H_{\pi(i)})\left(\mathbf{1}\{p_{\pi(i)} \leq \alpha_i\} - F_i(\alpha_i)\frac{\mathbf{1}\{p_{\pi(i)} > \lambda\}}{1-F_i(\lambda)}\right),$$

where $F_i(\alpha_i)$ *and* $F_i(\lambda)$ *are defined by* (B.18). *First note that* $\xi_i \leq 1 =: B$ *almost surely. Let us now prove*

$$\mathbf{E}(\xi_i^2 \mid \mathcal{F}_{i-1}) \leq (1-H_{\pi(i)})\nu_i. \tag{B.17}$$

Indeed, assuming first $\alpha_i \leq \lambda$, *we have by* (B.15),

$$\begin{aligned}\mathbf{E}(\xi_i^2 \mid \mathcal{F}_{i-1}) &= (1-H_{\pi(i)})\left(\mathbf{E}(\mathbf{1}\{p_{\pi(i)} \leq \alpha_i\} \mid \mathcal{F}_{i-1}) + (F_i(\alpha_i))^2\frac{\mathbf{E}(\mathbf{1}\{p_{\pi(i)} > \lambda\} \mid \mathcal{F}_{i-1})}{(1-F_i(\lambda))^2}\right)\\ &\leq (1-H_{\pi(i)})(\alpha_i + \alpha_i^2/(1-\lambda)) = (1-H_{\pi(i)})\nu_i.\end{aligned}$$

which gives (B.17). *Now, if* $\alpha_i > \lambda$, *still by* (B.15),

$$\begin{aligned}\mathbf{E}(\xi_i^2 \mid \mathcal{F}_{i-1}) &= (1-H_{\pi(i)})\left(\mathbf{E}(\mathbf{1}\{p_{\pi(i)} \leq \alpha_i\} \mid \mathcal{F}_{i-1}) + (F_i(\alpha_i))^2\frac{\mathbf{E}(\mathbf{1}\{p_{\pi(i)} > \lambda\} \mid \mathcal{F}_{i-1})}{(1-F_i(\lambda))^2}\right.\\ &\qquad \left. -2\frac{F_i(\alpha_i)}{1-F_i(\lambda)}\mathbf{E}(\mathbf{1}\{\lambda < p_{\pi(i)} \leq \alpha_i\} \mid \mathcal{F}_{i-1})\right)\\ &= (1-H_{\pi(i)})\big[F_i(\alpha_i) + (F_i(\alpha_i))^2/(1-F_i(\lambda)) - 2F_i(\alpha_i)(F_i(\alpha_i) - F_i(\lambda))/(1-F_i(\lambda))\big]\\ &= (1-H_{\pi(i)})F_i(\alpha_i)\big[1 + (2F_i(\lambda) - F_i(\alpha_i))/(1-F_i(\lambda))\big]\\ &\leq (1-H_{\pi(i)})F_i(\alpha_i)\big[1 + F_i(\lambda)/(1-F_i(\lambda))\big] \leq (1-H_{\pi(i)})\nu_i,\end{aligned}$$

which implies (B.17) *also in that case. Finally,* (B.17) *is established, which yields*

$$\forall k \geq 1,\ \ S_k \leq 2\sqrt{\varepsilon_k(\delta)}\sqrt{\sum_{i=1}^{k}(1-H_{\pi(i)})\nu_i + 4\varepsilon_k(\delta)}$$

and thus (B.16).

Lemma B.3.1 *In the setting of Theorem B.3.1, let*

$$F_k(\alpha_k) = \mathbf{P}(p_{\pi(k)} \leq \alpha_k \mid \mathcal{F}_{k-1}, H_{\pi(k)} = 0),\ \ F_k(\lambda) = \mathbf{P}(p_{\pi(k)} \leq \lambda \mid \mathcal{F}_{k-1}, H_{\pi(k)} = 0) \tag{B.18}$$

the process $(S_k)_{k\geq 1}$ *defined by*

$$S_k = \sum_{i=1}^{k}(1-H_{\pi(i)})\left(\mathbf{1}\{p_{\pi(i)} \leq \alpha_i\} - F_i(\alpha_i)\frac{\mathbf{1}\{p_{\pi(i)} > \lambda\}}{1-F_i(\lambda)}\right), \quad k \geq 1,$$

is a martingale with respect to the filtration $(\mathcal{F}_k)_{k\geq 1}$.

Proof B.3.2 *First,* S_k *is clearly* $\mathcal{F}_k$ *measurable. Second, we have for all* $k \geq 2$,

$$\begin{aligned}\mathbf{E}(S_k \mid \mathcal{F}_{k-1}) &= \mathbf{E}\left(S_{k-1} + (1 - H_{\pi(k)})\left(\mathbf{1}\{p_{\pi(k)} \leq \alpha_k\} - F_k(\alpha_k)\frac{\mathbf{1}\{p_{\pi(k)} > \lambda\}}{1 - F_k(\lambda)}\right) \mid \mathcal{F}_{k-1}\right) \\ &= S_{k-1} + (1 - H_{\pi(k)})(F_k(\alpha_k) - F_k(\alpha_k)) = S_{k-1}.\end{aligned}$$

B.3.2 Uniform-Empirical version of Freedman's inequality

We establish a time-uniform, empirical Bernstein-style confidence bound for bounded martingales. Various related inequalities have appeared in the literature, in particular in the online learning community. The idea is based on 'stitching' together time-uniform bounds that are accurate on different segments of (intrinsic) time. The use of the stitching principle has been further pushed and developed into many refinements by Howard et al. (2021), who also propose a uniform empirical Bernstein bound as a byproduct. The version given here, based on a direct stitching of Freedman's inequality, has the advantage of being self-contained with an elementary proof (though the numerical constants may be marginally worse than Howard et al.'s).

We first recall Freedman's inequality in its original version (Freedman, 1975). Let $(\xi_i, \mathcal{F}_i)_{i\geq 1}$ be a supermartingale difference sequence, i.e. $\mathbb{E}[\xi_i|\mathcal{F}_{i-1}] \leq 0$ for all i. Define $S_n := \sum_{i=1}^n \xi_i$ (then $(S_n, \mathcal{F}_n)$ is a supermartingale), and $V_n := \sum_{i=1}^n \mathrm{Var}\xi_i|\mathcal{F}_{i-1}$.

Theorem B.3.2 (Freedman's inequality; Freedman, 1975, Theorem 4.1) *Assume* $\xi_i \leq 1$ *for all* $i \geq 1$. *Then for all* $t, v > 0$*:*

$$\mathbb{P}\left[S_n \geq t \text{ and } V_n \leq v \text{ for some } n \geq 1\right] \leq \exp\left(-\varphi(v,t)\right), \tag{B.19}$$

where

$$\varphi(v,t) := (v+t)\log\left(1 + \frac{t}{v}\right) - t. \tag{B.20}$$

We establish the following corollary (deferring the proof for now):

Corollary B.3.1 *Assume* $\xi_i \leq 1$ *for all* $i \geq 1$. *Then for all* $\delta \in (0,1)$ *and* $v > 0$*:*

$$\mathbb{P}\left[S_n \geq \sqrt{2v\log\delta^{-1}} + \frac{\log\delta^{-1}}{2} \text{ and } V_n \leq v \text{ for some } n \geq 1\right] \leq \delta. \tag{B.21}$$

Following the stitching principle applied to the above we obtain the following.

Corollary B.3.2 *Assume* $\xi_i \leq B$ *for all* $i \geq 1$, *where* B *is a constant. Put* $\widetilde{V}_k := (V_k \vee B^2)$ *and* $\kappa = \pi^2/6$. *Then for all* $\delta \in (0, 1/(1+\kappa))$, *with probability at least* $1 - (1+\kappa)\delta$ *it holds*

$$\forall k \geq 1 : S_k \leq 2\sqrt{\widetilde{V}_k \varepsilon(\delta,k)} + \frac{1}{2}B\varepsilon(\delta,k),$$

where $\varepsilon(\delta,k) := \log\delta^{-1} + 2\log(1 + \log_2(\widetilde{V}_k/B^2))$.

Proof B.3.3 *Denote* $v_j^2 := 2^j B^2$, $\delta_j := (j \vee 1)^{-2}\delta$, $j \geq 0$, *and define the nondecreasing sequence of stopping times* $\tau_{-1} = 1$ *and* $\tau_j := \min\left\{k \geq 1 : V_k > v_j^2\right\}$ *for* $j \geq 0$. *Define the events for*

$j \geq 0$:

$$A_j := \left\{ \exists k \geq 1 : S_k \geq \sqrt{2v_j^2 \log \delta_j^{-1}} + \frac{1}{2} B \log \delta_j^{-1} \text{ and } V_k \leq v_j^2 \right\},$$
$$A'_j := \left\{ \exists k \text{ with } \tau_{j-1} \leq k < \tau_j : S_k \geq 2\sqrt{\widetilde{V}_k \varepsilon(\delta, k)} + \frac{1}{2} B \varepsilon(\delta, k) \right\}.$$

From the definition of v_j^2, δ_j, we have $j = \log_2(v_j^2/B^2)$ for $j \geq 1$. For $j \geq 1$, $\tau_{j-1} \leq k < \tau_j$ implies $\widetilde{V}_k = V_k$, $v_{j-1}^2 = v_j^2/2 < \widetilde{V}_k \leq v_j^2$, and further

$$\log \delta_j^{-1} = \log \delta^{-1} + 2 \log \log_2(v_j^2/B^2) \leq \varepsilon(\delta, k).$$

Therefore it holds $A'_j \subseteq A_j$. Furthermore, for $j = 0$, we have $v_0^2 = B^2, \delta_0 = \delta$. Further, if $k < \tau_0$ it implies $V_k < B^2$ and therefore $\widetilde{V}_k = B^2$, thus $\varepsilon(\delta, k) = \log \delta^{-1}$. Hence

$$A'_0 \subseteq \left\{ \exists k \text{ with } k < \tau_0 : S_k \geq 2\sqrt{B^2 \log \delta_0^{-1}} + \frac{1}{2} B \log \delta_0^{-1} \right\}$$
$$\subseteq \left\{ \exists k \geq 1 : S_k \geq \sqrt{2v_0^2 \log \delta_0^{-1}} + \frac{1}{2} B \log \delta_0^{-1} \text{ and } V_k \leq v_0^2 \right\} = A_0.$$

Therefore, since by (B.21) *it holds $\mathbb{P}[A_j] \leq \delta_j$ for all $j \geq 0$:*

$$\mathbb{P}\left[\exists k \leq n : S_k \geq 2\sqrt{V_k \varepsilon(\delta, k)} + B\varepsilon(\delta, k) \right] = \mathbb{P}\left[\bigcup_{j \geq 0} A'_j \right] \leq \mathbb{P}\left[\bigcup_{j \geq 0} A_j \right] \leq \delta \sum_{j \geq 0} (j \vee 1)^{-2} \leq 3\delta.$$

Proof B.3.4 (Proof of Corollary B.3.1) *It can be easily checked that $\varphi(v, t)$ is increasing in t (for $v, t > 0$). Thus $S_n \geq t \Leftrightarrow \varphi(p, (S_n)_+) \geq \varphi(p, t)$. Since $\varphi(v, 0) = 0$, and $\lim_{t \to \infty} \varphi(v, t) = \infty$, it follows that for any $\delta \in (0, 1]$, there exists a unique real $t(v, \delta)$ such that $\varphi(v, t(v, \delta)) = -\log \delta$. It follows that* (B.19) *is equivalent to:*

$$\forall v > 0, \forall \delta \in (0, 1] : \qquad \mathbb{P}[A_{v,\delta}] \leq \delta, \tag{B.22}$$

where

$$A_{v,\delta} := \{ \varphi(v, (S_n)_+) \geq -\log \delta \text{ and } T_n \leq v \text{ for some } n \geq 1 \}.$$

Observe that $\varphi(v, t) = vh\left(\frac{v+t}{v}\right)$, where h is the function defined by (3.11). *Since $h(\lambda) \geq 2(\sqrt{\lambda} - 1)^2$ from Lemma B.4.1, we deduce $\varphi(v, t) \geq 2(\sqrt{v + t} - \sqrt{v})^2$ thus, whenever $\varphi(v, (S_n)_+) \leq -\log \delta$, we have:*

$$\sqrt{v + (S_n)_+} \leq \sqrt{v} + \sqrt{\frac{\log \delta^{-1}}{2}};$$

taking squares on both sides entails

$$S_n \leq \sqrt{2v \log \delta^{-1}} + \frac{\log \delta^{-1}}{2},$$

proving (B.21).

B.4 Auxiliary results

Lemma B.4.1 *The function h defined by (3.11) is increasing strictly convex from $(1,\infty)$ to $(0,\infty)$, while h^{-1} is increasing strictly concave from $(0,\infty)$ to $(1,\infty)$. The functions h and h^{-1} satisfy the following upper/lower bounds:*

$$2(\sqrt{\lambda}-1)^2 \leq h(\lambda) \leq (\lambda-1)^2/2, \quad \lambda > 1$$
$$1+\sqrt{2y} \leq h^{-1}(y) \leq (1+\sqrt{y/2})^2, \quad y > 0$$

In particular, $h^{-1}(y) - 1 \leq \sqrt{2y} + \mathcal{O}(y)$ as $y \to 0$. In addition, for any $c > 0$, $x \in (1,+\infty) \mapsto xh^{-1}(c/x)$ is increasing.

Proof B.4.1 *Clearly, $h' = \log$, which is positive and increasing on $(1,\infty)$. This gives the desired property for h and h^{-1}. Next, the bounds can be easily obtained by studying the functions $\lambda \mapsto (\lambda-1)^2/2 - h(\lambda)$ and $\lambda \mapsto h(\lambda) - 2(\sqrt{\lambda}-1)^2$. For the last statement, since h^{-1} is strictly concave and $h^{-1}(0) = 1$, we have that $y \in (0,\infty) \mapsto (h^{-1}(y)-1)/y$ is decreasing. Since $y \in (0,\infty) \mapsto 1/y$ is also decreasing, this gives that $y \in (0,\infty) \mapsto h^{-1}(y)/y$ is decreasing. This gives the last statement.*

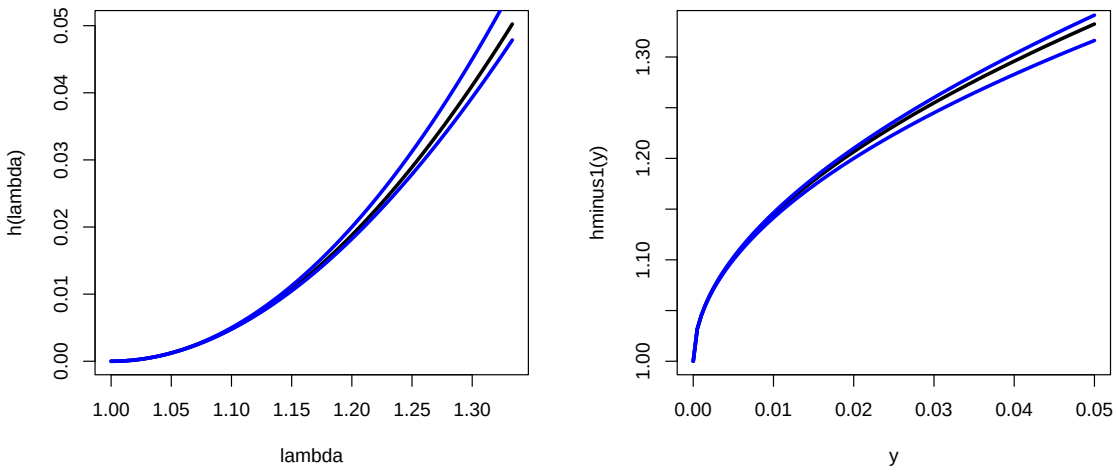

Figure B.1: Displaying h (left) and h^{-1} (right). Bounds of Lemma B.4.1 are displayed in blue.

Lemma B.4.2 (Wellner's inequality, Inequality 2, page 415, with the improvement of Exercise 3 page 418 *Let $U_1,\dots,U_n$ be $n \geq 1$ i.i.d. uniform random variables. For all $\lambda \geq 1$, $a \in [0,1)$, we have*

$$\mathbf{P}\left(\exists t \in [a,1], n^{-1}\sum_{i=1}^{n} \mathbf{1}\{U_i \leq t\}/t \geq \lambda\right) \leq e^{-nah(\lambda)/(1-a)},$$

for $h(\cdot)$ defined by (3.11).

Lemma B.4.3 *The KR constants in* (3.34) *and* (3.44) *satisfy, as* $a \to \infty$*,*

$$\frac{\log(1/\delta_a)}{a\log(1+\frac{1-\delta_a^{B/a}}{B})} = 1 + O\left(\frac{\log(a)}{a}\right);$$
$$\frac{\log(1/\delta_a)}{a\log(1+\log(1/\delta_a)/a)} = 1 + O\left(\frac{\log(a)}{a}\right),$$

where $\delta_a = c\delta/a$*,* $c = \pi^2/6$ *and the* $O(\cdot)$ *depends only on the constants* $\delta > 0$ *and* $B > 0$*.*

B.5 Additional experiments

We reproduce here the figures of the numerical experiments in the top-k and preordered settings, by adding the interpolated bounds. On each graph, the median of the generated interpolated bound is marked by a star symbol, which is given in addition to the former boxplot (of the non-interpolated bound). By doing so, we can evaluate the gain brought by the interpolation operation in each case. Note that the interpolated bound is not computed for $m \geq 10^5$ for computational cost reasons.

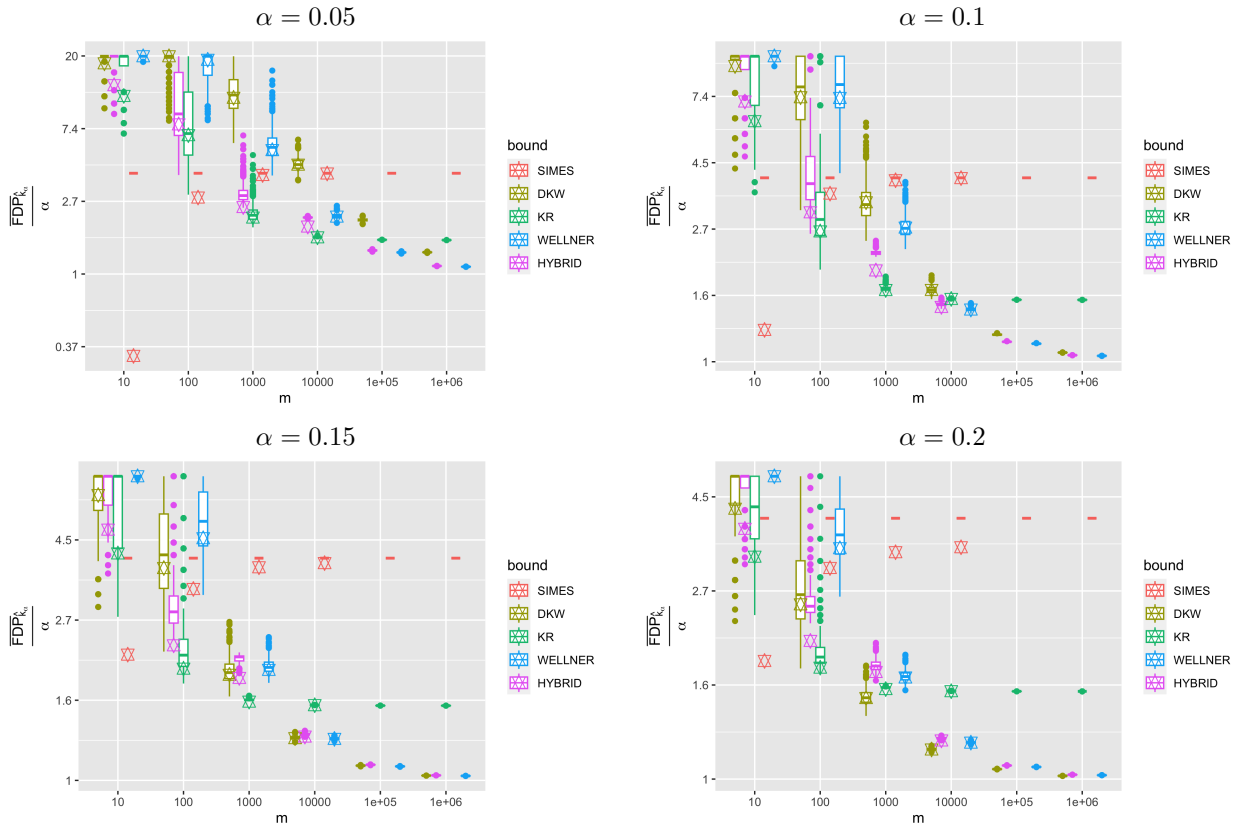

Figure B.2: Figure 3.1 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Top-k dense case ($\pi_0 = 0.5$, $\mu = 1.5$).

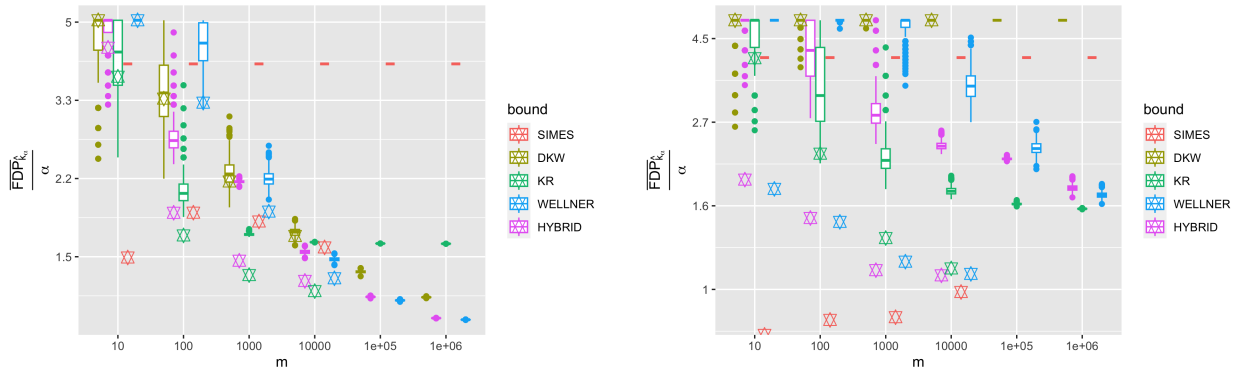

Figure B.3: Figure 3.2 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Top-k sparse case $\pi_0 = 1 - 0.5m^{-0.25}$, $\mu = \sqrt{2\log m}$ (left) $\pi_0 = 1 - 0.5m^{-0.55}$, μ =10 (right), $\alpha = 0.2$.

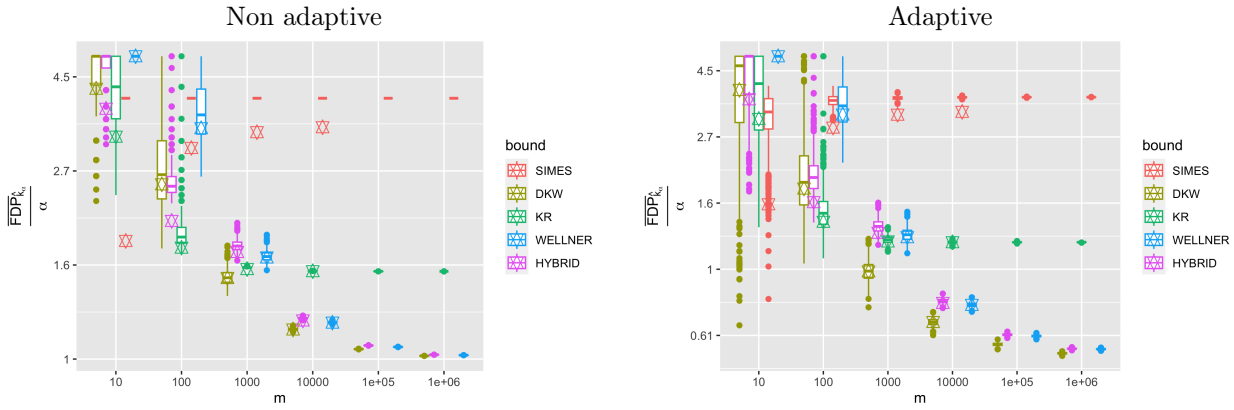

Figure B.4: Figure 3.3 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Top-k dense case with nonadaptive bounds (left) and adaptive bounds (right) ($\pi_0 = 0.5$, $\alpha = 0.2$).

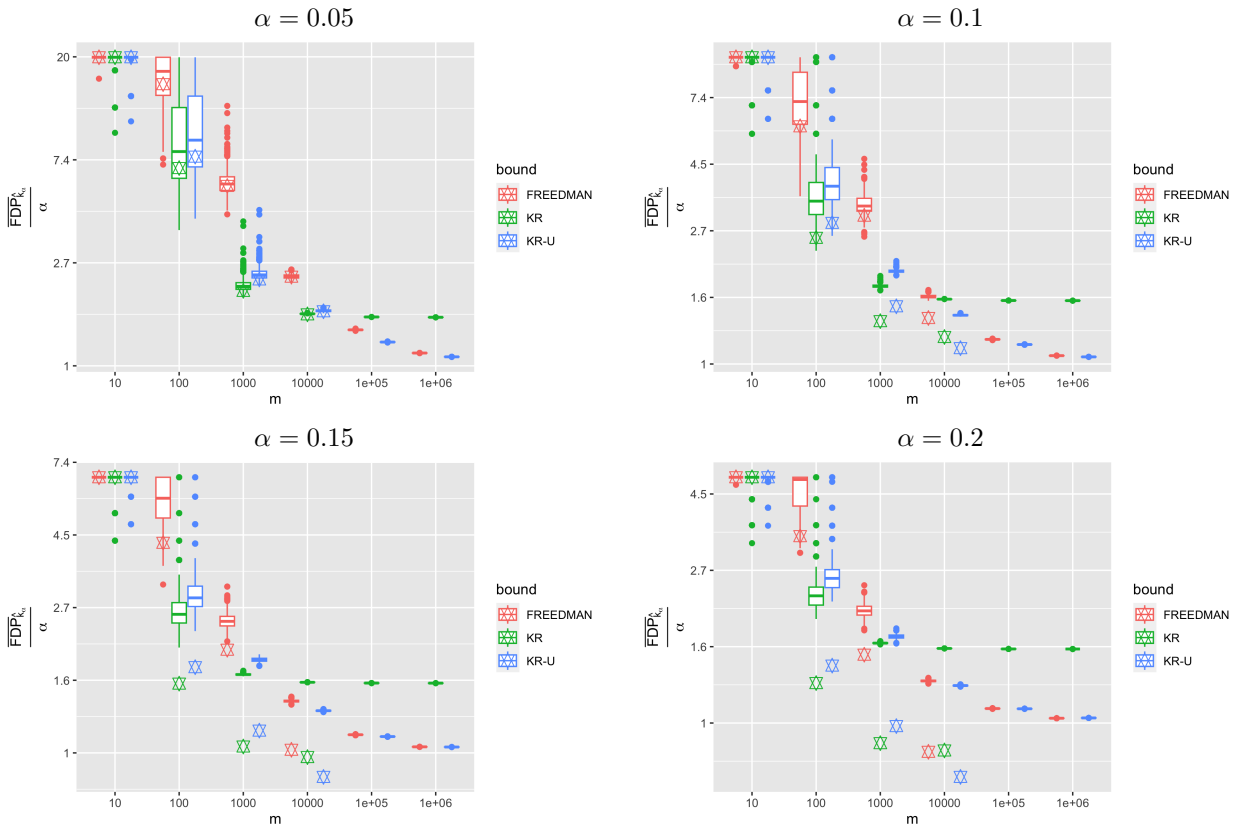

Figure B.5: Figure 3.5 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Preordered dense ($\beta = 0$) LF setting with LF procedure ($s = 0.1\alpha$, $\lambda = 0.5$).

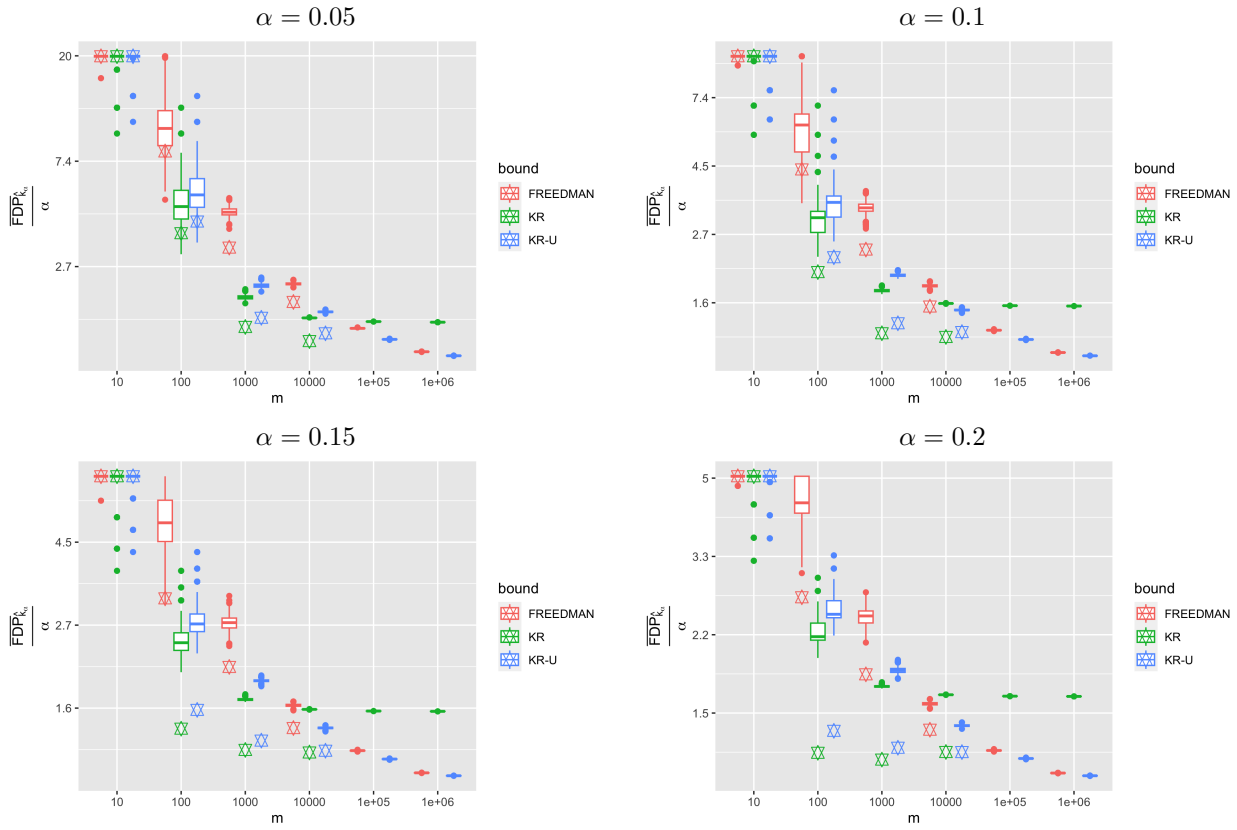

Figure B.6: Figure 3.6 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Preordered sparse ($\beta = 0.25$) LF setting with LF procedure ($s = 0.1\alpha$, $\lambda = 0.5$).

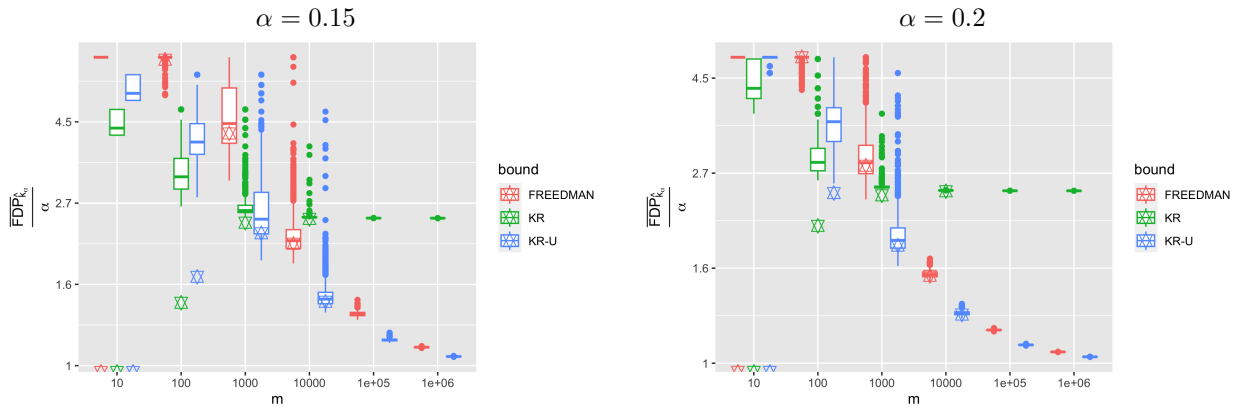

Figure B.7: Figure 3.7 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Pre-ordered dense ($\beta = 0$) knockoff setting with BC procedure (i.e., LF procedure with $s = \lambda = 0.5$).

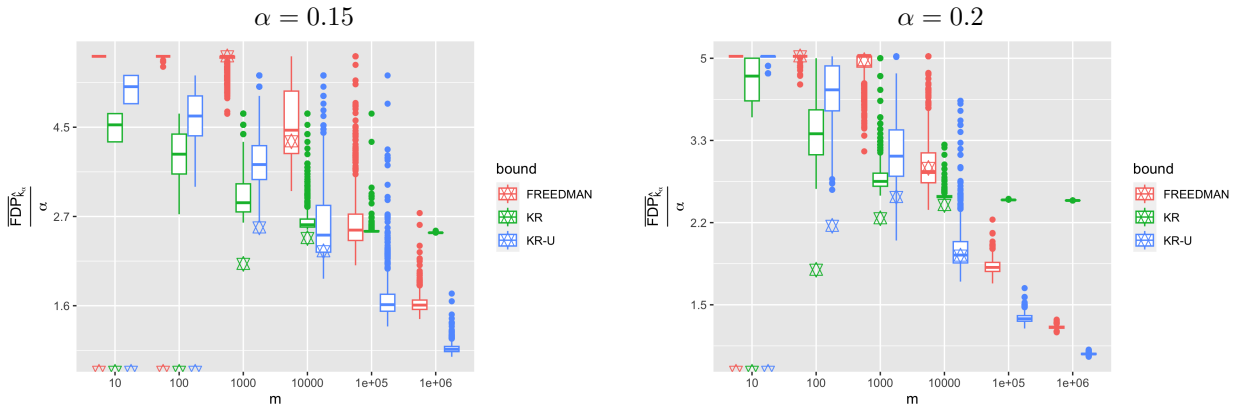

Figure B.8: Figure 3.8 where we have superposed in each case the (median of the) interpolated bounds (star symbols). Pre-ordered sparse ($\beta = 0.25$) knockoff setting with BC procedure (i.e., LF procedure with $s = \lambda = 0.5$).

Appendix C

Supplementary material of Chapter 4

Outline of the current chapter

C.1 Auxiliary definitions and results

In this appendix we recall some definitions and results of stochastic ordering following the presentation in [SS], to which we also refer the reader for further details. We also recall a well-known bound on the inverse moment of the Binomial distribution.

Definition C.1.1 (Stochastic order) *Let X and Y be two random variables such that*

$$\mathbf{P}(X > x) \leq \mathbf{P}(Y > x) \quad \textit{for all } x \in (-\infty, \infty),$$

Then X is said to be smaller than Y in the usual stochastic order denoted by $X \leqslant_{st} Y$.

An equivalent characterization of the stochastic order is that $X \leqslant_{\mathrm{st}} Y \Leftrightarrow \mathbf{E}[g(X)] \leq \mathbf{E}[g(Y)]$, for all non-decreasing functions $g : \mathbb{R} \to \mathbb{R}$ for which the expectations exist (see (1.A.7) in [SS])).

Definition C.1.2 (Convex order) *Let X and Y be two random variables such that*

$$\mathbf{E}(\phi(X)) \leq \mathbf{E}(\phi(Y)) \quad \textit{for all convex functions } \phi : \mathbb{R} \to \mathbb{R},$$

provided the expectations exist. Then X is said to be smaller than Y in the convex order denoted as $X \leqslant_{cx} Y$.

The next results follows from the definition of convex ordering, see Chapter 3 of [SS].

Lemma C.1.1 (Theorem 3.A.24 in [SS]) *Let X be a random variable with mean $\mathbf{E}X$. Denote the left (right) endpoint of the support of X by l_X $[u_X]$. Let Z be a random variable such that $\mathbf{P}\{Z = l_X\} = (u_X - \mathbf{E}X)/(u_X - l_X)$ and $\mathbf{P}\{Z = u_X\} = (\mathbf{E}X - l_X)/(u_X - l_X)$. Then*

$$\mathbf{E}X \leqslant_{cx} X \leqslant_{cx} Z$$

where $\mathbf{E}X$ denotes a random variable that takes on the value $\mathbf{E}X$ with probability 1 (the left handside just restates Jensen's inequality).

Lemma C.1.2 (Theorem 3.A.44 in [SS]) *Let X and Y be two random variables with equal means, density functions f and g, distribution functions F and G, and survival functions $\bar{F}$ and $\bar{G}$, respectively. Denote by $S^-(a)$ the number of sign changes for function a. Then $X \leq_{\mathrm{cx}} Y$ if any of the following conditions hold:*

$$S^-(g - f) = 2 \text{ and the sign sequence is } +, -, +;$$
$$S^-(\bar{F} - \bar{G}) = 1 \text{ and the sign sequence is } +, -;$$
$$S^-(G - F) = 1 \text{ and the sign sequence is } +, -.$$

Proposition C.1.1 (Theorem 3.A.12 d) in [SS]) *Let $X_1, X_2, \ldots, X_m$ be a set of independent random variables and let $Y_1, Y_2, \ldots, Y_m$ be another set of independent random variables. If $X_i \leqslant_{cx} Y_i$ for $i = 1, 2, \ldots, m$, then*

$$\sum_{j=1}^{m} X_j \leqslant_{cx} \sum_{j=1}^{m} Y_j.$$

That is, the convex order is closed under convolutions.

Lemma C.1.3 (Example 3.A.48 in [SS]) *Let X and Y be Bernoulli random variables with parameters p and q, respectively, with $0 < p \leq q \leq 1$. Then*

$$\frac{X}{p} \geq_{cx} \frac{Y}{q}.$$

Lemma C.1.4 (Inverse moment for the Binomial distribution) *Let $B_1, \ldots, B_k \sim \mathbf{Bin}(1, q)$. Then $\mathbf{E}[1/(1 + \sum_{i=1}^{k} B_i)] \leq 1/((k+1)q)$.*

Proof C.1.1 *See e.g. Benjamini et al. (2006).*

C.2 Complements to Section 4.4.1

In the context of Gaussian one-sided testing described in Section 4.4.1, let $\widehat{m}_0(p_1, \ldots, p_m) = \frac{1}{\nu}(1 + \sum_{i=1}^{m} g(p_i)) \in \mathcal{F}_0$. Define $X_0 \sim g(p_i)$ for $i \in \mathcal{H}_0$ and $X_1 \sim g(p_i)$ where $i \in \mathcal{H}_1$. Then we

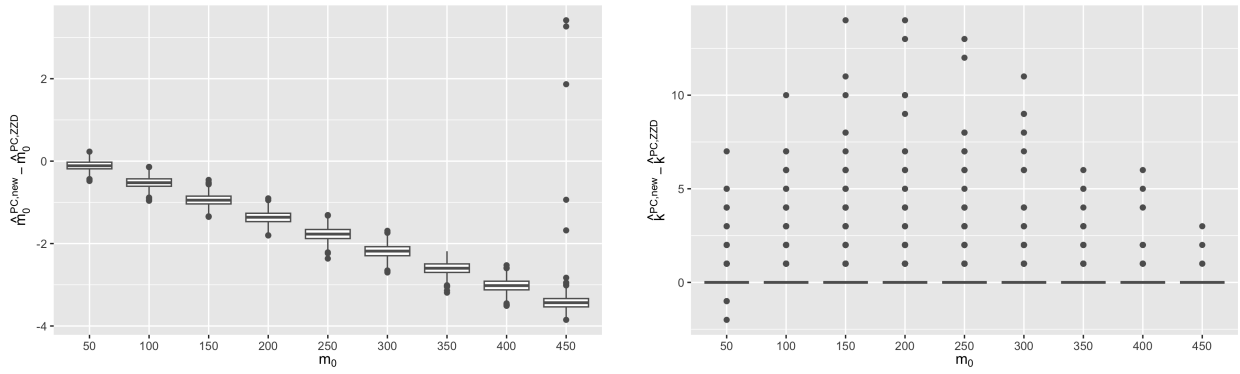

Figure C.1: Box plots for the difference between $\widehat{m}_0^{\mathrm{PC,new}}$ and $\widehat{m}_0^{\mathrm{PC,ZZD}}$ for point estimation (left) and rejection numbers for the plug-in BH procedures (right) against a range of true $m_0 = 50, 100, \ldots, 450$.

have

$$\mathrm{Bias}(\widehat{m}_0) = \frac{1}{\nu}\mathbf{E}\left(1 + \sum_{i\in\mathcal{H}_1} g(p_i)\right) = (1 + (m - m_0)\cdot \mathbf{E}X_1)/\nu,$$

$$\mathrm{Var}(\widehat{m}_0) = \frac{1}{\nu^2}\mathrm{Var}\left(\sum_{i\in\mathcal{H}_0} g(p_i) + \sum_{i\in\mathcal{H}_1} g(p_i)\right) = (m_0\cdot\mathrm{Var}(X_0) + (m - m_0)\cdot\mathrm{Var}(X_1))/\nu^2, \qquad \text{with}$$

$$\mathrm{Var}(X_0) = \int_0^1 g(u)^2 du - \left[\int_0^1 g(u)du\right]^2,$$

$$\mathrm{Var}(X_1) = \int_0^1 g(u)^2 f_1(u)du - \left[\int_0^1 g(u)f_1(u)du\right]^2.$$

where $f_1(t) = \exp\left(-\mu\cdot\Phi^{-1}(t) - \mu^2/2\right)$ denotes the density of the p-values under the alternative.

C.3 Complements to Section 4.4.2

Here we present some numerical results, comparing the performance of $\widehat{m}_0^{\mathrm{PC,new}}$ (see (4.8)) and $\widehat{m}_0^{\mathrm{PC,ZZD}}$ (see (4.15)) for $m = 500$, where the correction factors $C(500) = 1.011709$ and $s(500) = 98$ are taken from Table S1 in Zeisel et al. (2011).

We first analyze the two estimators on simulated data in a one-sided Gaussian testing setting where we observe realizations of independent rv's $X_1, \ldots, X_{m_0} \sim N(0,1)$ and $X_{m_0+1}, \ldots, X_{500} \sim N(1.5, 1)$ for 1000 Monte-Carlo simulation runs and a varying range of $m_0 = 50, 100, \ldots, 450$. We obtain 500 p-values by testing the null hypotheses $H_{0,i} : \mu = 0$ vs. the alternatives $H_{1,i} : \mu > 0$ simultaneously for all $i \in \{1, \ldots, 500\}$ and calculate $\widehat{m}_0^{\mathrm{PC,new}}$ and $\widehat{m}_0^{\mathrm{PC,ZZD}}$ as well as the number of rejections obtained from the plug-in BH procedure in (4.2) with $\alpha = 0.05$.

Figure C.1 shows that over a wide range of true m_0 values, $\widehat{m}_0^{\mathrm{PC,new}}$ and $\widehat{m}_0^{\mathrm{PC,ZZD}}$ yield comparable results both w.r.t. the point estimates and for the number of rejections. In fact, $\widehat{m}_0^{\mathrm{PC,new}}$

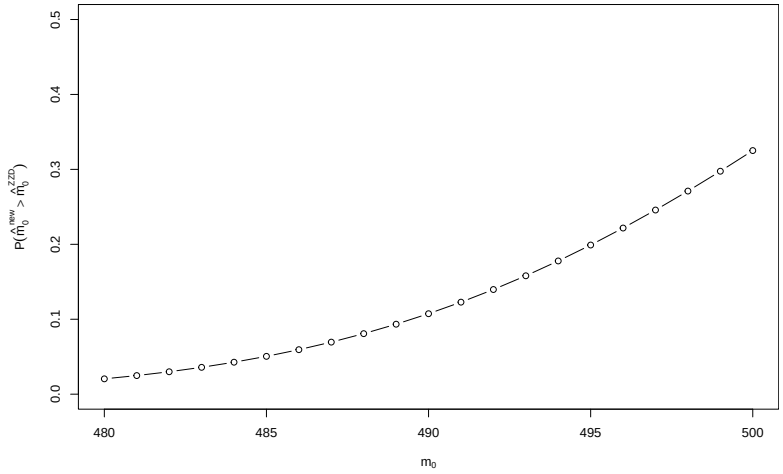

Figure C.2: Approximate probabilities $\mathbf{P}(\widehat{m}_0^{\mathrm{PC,new}} > \widehat{m}_0^{\mathrm{PC,ZZD}})$ for various values of true m_0, with fixed $m = 500$.

appears to be slightly more efficient than $\widehat{m}_0^{\mathrm{PC,ZZD}}$.

Another comparison can be obtained when we assume that the signal under the alternative is strong and that most hypotheses are nulls. In this case we have $2\sum_{i=1}^{m} p_i \approx 2\sum_{i\in\mathcal{H}_0} p_i =: S$ so that we can use the Central Limit Theorem to quantify the probability that $\widehat{m}_0^{\mathrm{PC,new}}$ is more conservative than $\widehat{m}_0^{\mathrm{PC,ZZD}}$

$$\mathbf{P}(\widehat{m}_0^{\mathrm{PC,new}} > \widehat{m}_0^{\mathrm{PC,ZZD}}) = \mathbf{P}(S > m \cdot C(m) - 2) \approx \overline{\Phi}\left(\sqrt{\frac{3}{m_0}} \cdot (m \cdot C(m) - (m_0 + 2))\right).$$

Figure C.2 shows that this probability, for various values of the true m_0, is quite small and even under the complete null ($m_0 = 500$) it is bounded by 1/3.

C.4 Additional Figures for simulated data of Section 4.4

We provide additional results on simulated data in the Gaussian one-sided testing setting described in Section 4.4.1, with $m = 10000$ and $\mu = 1.5$. Figure C.3 displays estimation results for π_0 over 1000 Monte-Carlo replications. They are in line with the analytical comparisons of the MSE provided in Figure 4.1. Alongside, we also provide results on power, defined as the ratio of the number of true discoveries to the number of alternatives, for the corresponding plug-in BH (abbreviated in ABH for adaptive BH) procedures using each of the estimators, the raw BH and oracle plug-in BH (using the true m_0). The procedures are run for a fixed level $\alpha = 0.05$. The power enhancement among the different plug-in estimators' is not striking except perhaps for very small values of π_0 were where we recover the same performance ranking as in Figure 4.1. For larger values of π_0, the differences in power is not perceptible anymore, every procedure behaves poorly as there is less and less signal.

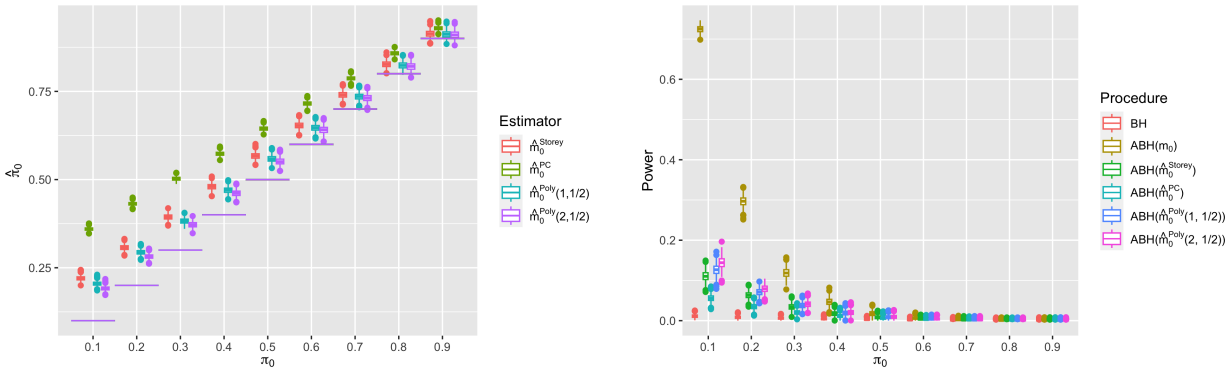

Figure C.3: Estimation results (left panel) for $\widehat{m}_0^{\text{Storey}}$, $\widehat{m}_0^{\text{PC,new}}$, $\widehat{m}_0^{\text{Poly}}(1, 1/2)$, and $\widehat{m}_0^{\text{Poly}}(2, 1/2)$, and power results (right panel) for the associated plug-in BH procedures on simulated data.

C.5 Upper and lower bounds for the inverse moment of the uniform sum distribution

The Pounds and Cheng estimator is closely related to the sum of independent uniform random variables. This distribution plays a role in various contexts and is also known as the *Irwin-Hall* distribution (for more details, see Johnson et al. (1970)). As an auxiliary result, we give lower and upper bounds for the inverse moment of this distribution.

Lemma C.5.1 (Inverse moments for Erlang distributions) *Let $E_1, \ldots, E_k \sim \mathcal{E}(1)$ be independent exponentially distributed random variables. Then $\mathbf{E}[1/\sum_{i=1}^k E_i] \leq 1/(k-1)$.*

Proof C.5.1 *Since $X = \sum_{i=1}^k E_i$ is Gamma-distributed with shape $\alpha = k$ and inverse scale parameter $\beta = 1$ then $1/X$ is Inverse-gamma distributed with mean $\beta/(\alpha - 1)$, see Gelman et al. (2013).*

Proposition C.5.1 *(Inverse moment for sums of uniforms) For $k \geq 2$ let $U_1, U_2, \ldots, U_k \sim \mathcal{U}[0,1]$ iid. Then we have*

$$\frac{2}{k} \leq \mathbf{E}\left(\frac{1}{\sum_{i=1}^k U_i}\right) \leq \frac{2}{k-1} \tag{C.1}$$

Proof C.5.2 *Let $E_1, E_2, \ldots, E_k \sim \mathcal{E}(1)$ iid. From Theorems 3.A.24 and 3.A.46 in [SS] we have for $i = 1, \ldots, , k$*

$$1 \leqslant_{cx} 2U_i \leqslant_{cx} E_i$$

and since the convex ordering is preserved under convolutions (see [SS], Theorem 3.A.12.) we obtain

$$k \leqslant_{cx} \sum_{i=1}^k 2U_i \leqslant_{cx} \sum_{i=1}^k E_i.$$

Together with the convexity of the mapping $x \mapsto 1/x$ on $(0,1)$ this yields

$$\frac{1}{k} \leq \mathbf{E}\left(\frac{1}{\sum_{i=1}^{k} 2U_i}\right) \leq \mathbf{E}\left(\frac{1}{\sum_{i=1}^{k} E_i}\right) \leq \frac{1}{k-1},$$

where the last inequality follows from Lemma C.5.1.

Controlling False Discovery Proportion in Structured Data Sets

Abstract

The present work proposes new methodologies for controlling the False Discovery Proportion (FDP) while accommodating different types of data structures arising from the underlying scientific context. Since the seminal work of Benjamini and Hochberg (1995) (BH) introducing the FDP, multiple testing procedures have found widespread applications across diverse domains. The BH procedure has facilitated the identification of significant variables within large data sets, providing insights to scientific questions in fields such as biology, medicine, or marketing research, by ensuring guarantees on the proportion of false discoveries. However, the BH procedure has several limitations, among which e.g. the fact that it is most effective for uniform p-values under the null; it is developed within a batch framework requiring simultaneous availability of all p-values; the false discoveries control guarantee is only in expectation. These limitations can lead to a range of unfavorable outcomes – spanning from reduced interpretability, loss of statistical power, to potential inflation of the Type I error rate – particularly in contexts where we perceive the data as possessing inherent "structure." This work aims to push back those limits by providing new procedures and methodologies that adapt to settings where p-values can be discrete, online, preordered, or weighted. This ultimately gives the practitioner more effective tools for identifying significant variables in structured data sets as we illustrate in various numerical experiments.

Keywords: multiple testing, discrete p-values, online p-values, weighted p-values, preordered p-values, (m)FDR control, FDP confidence bounds, plug-in FDR control

---

Résumé

Ce travail propose de nouvelles méthodologies pour contrôler la proportion de fausses découvertes (FDP) tout en prenant en compte différentes types de structures de données résultant du contexte scientifique sous-jacent. Depuis le travail fondamental de Benjamini and Hochberg (1995) (BH) introduisant le FDP, les procédures de tests multiples ont trouvé une application dans de nombreux domaines. La procédure de BH a facilité l'identification de variables significatives dans de grands ensembles de données, permettant de répondre à des questions scientifiques dans des domaines tels que la biologie, la médecine ou le marketing, tout en fournissant des garanties sur la proportion de fausses découvertes. Toutefois, la procédure de BH présente plusieurs limites : elle est plus efficace pour des p-valeurs uniformes sous l'hypothèse nulle ; elle est développée dans un cadre *offline* nécessitant la connaissance simultanée de toutes les p-valeurs ; la garantie de contrôle des fausses découvertes est en espérance. Ces limitations peuvent entraîner une perte de puissance, une réduction de l'interprétabilité, voire même une inflation de l'erreur de Type I dans différents contextes où les données sont considérées comme "structurées". Ce travail vise à combler ces lacunes en fournissant de nouvelles procédures et méthodologies qui s'adaptent à des contextes structurels où les p-valeurs peuvent être discrètes, en ligne, pré-ordonnées ou pondérées. Cela donne, in fine, au praticien des outils plus efficaces pour identifier les variables significatives dans un ensemble de données structurées, comme nous l'illustrons dans diverses expériences numériques.

Mots clés : tests multiples, p-valeurs discrètes, p-valeurs en ligne, p-valeurs pondérées, p-valeurs ordonnées, contrôle du (m)FDR, bornes de confiance pour le FDP, contrôle du plug-in FDR

Laboratoire de Probabilités, Statistique et Modélisation
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France