Contributions to reliable machine learning via false discovery rate control

Ariane Marandon-Carlhian

Sorbonne Université

LPSM

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

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

Thesis defended by **Ariane Marandon-Carlhian**

Defended on **29th September, 2023**

In order to become Doctor from Sorbonne Université

Academic Field **Applied mathematics**

Speciality **Statistics**

Contributions to reliable machine learning via false discovery rate control

Thesis supervised by	Etienne Roquain	Supervisor
	Tabea Rebafka	Co-Monitor
	Nataliya Sokolovska	Co-Monitor

Committee members

Referees	Julien Chiquet	Senior Researcher at INRAE	
	Ruth Heller	Professor at Tel Aviv University	
Examiners	Gérard Biau	Professor at Sorbonne Université	
	Gilles Blanchard	Professor at Université Paris Saclay	Committee President
	Pierre Latouche	Professor at Université Clermont Auvergne	
Supervisors	Etienne Roquain	Associate Professor at Sorbonne Université	
	Tabea Rebafka	Associate Professor at Sorbonne Université	
	Nataliya Sokolovska	Professor at Sorbonne Université	

This work was supported by the Paris-Ile-de-France Region via the DIM Math Innov program.

Remerciements

Alors que cette thèse touche à sa fin, je mesure à quel point celle-ci représente pour moi un chapitre de ma vie très important. Ainsi, j'éprouve beaucoup de gratitude envers tous ceux qui m'ont accompagnée d'une manière ou d'une autre. Les remerciements qui vont suivre ne sont qu'un petit aperçu de cette reconnaissance.

Un grand merci tout d'abord, à mes trois encadrants, Etienne, Tabea, et Nataliya, pour les nombreuses heures que vous avez investies, pour votre confiance, et pour votre accompagnement de manière générale. Je suis très reconnaissante de la patience dont vous avez fait preuve tout au long de cette thèse. Un merci particulier à mes encadrants principaux Tabea et Etienne. Merci Tabea pour tous tes encouragements. Merci Etienne pour avoir tout fait pour que cette thèse réussisse.

Je tiens ensuite à remercier les autres acteurs impliqués dans la réalisation de cette thèse, en commençant par ceux qui m'ont conduit à celle-ci: merci Ismaël et Maxime pour m'avoir encouragée à faire une thèse. Merci aussi Gérard pour tes conseils à ce sujet, ainsi que pour avoir accepté d'être membre de mon jury. Je salue le groupe test multiples: merci à Pierre Neuvial, pour l'invitation au workshop à Toulouse, Guillermo et Marie, pour les moments passés ensemble, Gilles pour ta participation à mon comité de mi-thèse et à mon jury, ainsi que pour avoir écouté plusieurs présentations de chacun de mes travaux. Merci à mes collaborateurs David et Lihua. Thank you Lihua for inviting me both as a speaker and as a discussant for your seminar. Merci à Nicolas Verzelen et Alexandra Carpentier pour m'avoir invitée aux workshops à Montpellier et à Munich.

Merci à mes rapporteurs Julien Chiquet et Ruth Heller, pour le temps qu'ils ont consacré à la lecture de mon manuscrit, pour leur engagement dans leurs retours et commentaires à son sujet. Thank you Ruth for coming in person to my defense. Merci également, Julien, pour ton investissement lors de mon comité de mi-thèse. Merci à Pierre Latouche, qui vient compléter mon jury.

Merci au LPSM. Si j'ai autant apprécié ces trois années de thèse, c'est grâce à vous. A ce titre, merci aux permanents du couloir 15-25, Anna, Antoine, Arnaud, Claire, Maxime, Stéphane, pour votre bienveillance et pour les moments partagés dans la salle café et ses environs, ainsi qu'à ceux un peu plus éloignés, Maud, Olivier, Anna B.H., Charlotte, Ismaël, Fanny. Merci aux doctorant.e.s. Merci Iqraa pour toujours savoir rire de nos malheurs et pour les révélations sur la vie. Merci Antonio d'être toi sans réserves. Merci Camila pour ton cœur en or. Merci Miguel de nous faire rêver avec ta confiance en toi. Surtout, merci à tous les quatres pour votre soutien (incluant toutes les séances psy) et pour votre loyauté. Avant de débuter cette thèse, trois ans me paraissait un temps assez long. En fin de compte, je n'ai pas attendu avec impatience cette soutenance, car je suis triste de vous quitter. Merci Yazid pour les fous rires (tu nous as donné matière à moi et Iqraa). Merci Lucas pour ta gentillesse (valable en toute circonstances sauf dans un certain jeu en ligne). Merci Ludovic pour m'avoir appris comment poser des questions en séminaire (que se passe-t-il en grande dimension?). Merci Alexis, Gabriel, Mathis, Pablo, Pierre, Francesco, Grâce, Ulysse, Robin, ainsi qu'aux

anciens Nicklas et Adeline, pour les moments passés ensemble, au labo et hors de celui-ci.

Merci à mes amis de longue date: Rojo, Wissal, Claire, Coralie, Mathilde, Sid, Olivier, Etienne. Merci d'être là dans les moments importants, qu'ils soient difficiles ou joyeux, depuis plus de dix ans. Je suis fière de nous.

Pour terminer, merci à mes parents.

Contents

Summary **vii**

1 Introduction **1**
1.1 Motivation . . . 1
1.2 From risk minimization to risk control . . . 3
1.2.1 Background . . . 3
1.2.2 Aim . . . 4
1.3 Brief overview of the considered learning tasks . . . 5
1.4 A primer on multiple testing . . . 7
1.4.1 Preliminaries . . . 10
1.4.2 Existing strategies for FDR control . . . 12
1.5 Contributions . . . 16

2 Adaptive novelty detection with FDR guarantee **19**
2.1 Introduction . . . 20
2.1.1 Novelty detection . . . 20
2.1.2 Existing strategies . . . 21
2.1.3 Contributions . . . 23
2.2 Preliminaries . . . 24
2.2.1 Notation . . . 24
2.2.2 Criteria . . . 24
2.2.3 BH algorithm and its π_0-adaptive variants . . . 25
2.2.4 Our method . . . 25
2.3 FDR control . . . 27
2.3.1 Exchangeability . . . 28
2.3.2 The p-values are PRDS . . . 28
2.3.3 A new FDR expression . . . 29
2.3.4 New FDR bounds for π_0-adaptive procedures . . . 29
2.4 Constructing score functions . . . 31
2.4.1 Assumptions and notation . . . 31
2.4.2 Optimal score function . . . 32
2.4.3 Density estimation . . . 32
2.4.4 PU classification . . . 33
2.4.5 AdaDetect with cross-validation . . . 34
2.5 Power results . . . 36
2.5.1 A constrained ERM score function . . . 37
2.5.2 General score functions . . . 39
2.6 Experiments . . . 41
2.6.1 Simulated data . . . 41

2.6.2 Semi-synthetic data . . . 42
2.7 An astronomy application . . . 43
2.8 Conclusion and discussion . . . 45
2.8.1 Limitations of AdaDetect . . . 45
2.8.2 Other future works . . . 46

3 False membership rate control in mixture models 47
3.1 Introduction . . . 48
3.1.1 Background . . . 48
3.1.2 Aim and approach . . . 48
3.1.3 Presentation of the results . . . 49
3.1.4 Relation to previous work . . . 50
3.1.5 Organization of the paper . . . 51
3.2 Setting . . . 51
3.2.1 Model . . . 51
3.2.2 Procedure and criteria . . . 52
3.2.3 Notation . . . 53
3.3 Methods . . . 53
3.3.1 Oracle procedures . . . 53
3.3.2 Empirical procedures . . . 55
3.4 Theoretical guarantees for the plug-in procedure . . . 56
3.4.1 Additional notation and assumptions . . . 56
3.4.2 Results . . . 58
3.5 Experiments . . . 59
3.5.1 Synthetic data set . . . 59
3.5.2 Real data set . . . 60
3.6 Conclusion and discussion . . . 62

4 Link prediction with FDR control 65
4.1 Introduction . . . 65
4.1.1 Problem and aim . . . 65
4.1.2 Approach . . . 67
4.1.3 Relation to previous work . . . 68
4.2 Methodology . . . 69
4.2.1 Preliminaries . . . 69
4.2.2 Procedure . . . 70
4.2.3 Training the scoring function g . . . 71
4.3 Numerical experiments . . . 74
4.3.1 Simulated data . . . 75
4.3.2 Real data . . . 76
4.4 Discussion . . . 77

5 Discussion 79

A Supplementary material of Chapter 2 93
A.1 Proofs and results for Section 3 . . . 93
A.1.1 Proof of Lemma 2 . . . 93
A.1.2 Key properties . . . 93
A.1.3 Proof of Theorem 3 . . . 94
A.1.4 Proof of Theorem 4 . . . 95

A.1.5 Proof of Theorem 6 . . . 95
A.1.6 Proof of Corollary 7 . . . 96
A.1.7 Proof of Theorem 18 . . . 98
A.1.8 Proof of Theorem 19 . . . 99
A.2 Proofs for Section 4 . . . 100
A.2.1 Proof of Theorem 8 . . . 100
A.2.2 Proof of Lemma 9 . . . 100
A.3 Proofs for Section 5 . . . 101
A.3.1 Proof of Theorem 10 . . . 101
A.3.2 Proof of Theorem 11 . . . 103
A.3.3 Proof of Corollary 12 . . . 106
A.4 Useful lemmas . . . 106
A.5 Auxiliary results for Section 5 . . . 108
A.5.1 Bounding $\zeta_r(\cdot)$ . . . 108
A.5.2 Case of density estimation . . . 113
A.5.3 From the two-sample setting to classical two-group setting . . . 115
A.6 Details of simulation studies in Section 6 . . . 116
A.6.1 Methods . . . 116
A.6.2 Additional experiments with varying k and ℓ . . . 117
A.6.3 Additional experiments with varying n, m, and m_1 . . . 117
A.7 Additional experimental results for the astronomy application . . . 119

B Supplementary material of Chapter 3 121
B.1 Proof of Theorems 16 and 17 . . . 121
B.1.1 A general result . . . 121
B.1.2 An optimal procedure . . . 122
B.1.3 Preliminary steps for proving Theorem 31 . . . 123
B.1.4 Proof of Theorem 31 . . . 124
B.2 Proofs of lemmas . . . 127
B.3 Auxiliary results . . . 129
B.4 Auxiliary results for the Gaussian case . . . 135
B.4.1 Convergence rate for parameter estimation . . . 135
B.4.2 Gaussian computations . . . 136

C Supplementary material of Chapter 4 139
C.1 Additional experimental details . . . 139
C.2 Exchangeability condition for FDR control . . . 139

Summary

The reliability of machine learning (ML) methods is critical in contexts that involve high-stakes decisions. However, while ML methods have achieved impressive results in a wide range of applications (Jordan and Mitchell, 2015), none of them are able to provide a small error guarantee in all settings.

Since models cannot be perfect, they should at least "know that they do not know". While there have been many efforts in the literature to address the issue of uncertainty quantification, whether in the field of probability calibration (Guo et al., 2017), or prediction sets (Angelopoulos and Bates, 2021), these solutions are not satisfactory when an actual decision is required. By contrast, a key to keeping the error rate (or risk) below a certain threshold is to make use of a type of abstention option, which amounts to abstain from making a decision when there is too much uncertainty. The goal of this thesis is to propose new methods for risk control, i.e. for keeping the risk below a certain user-specified threshold α, in several learning tasks. Risk control is a long-standing paradigm in machine learning, in binary classification and related tasks (Rigollet and Tong, 2011; Blanchard et al., 2010a; Bartlett and Wegkamp, 2008; Barber and Candès, 2015; Bates et al., 2023; Angelopoulos et al., 2021), and we aim to either extend existing methods or propose new ones for a certain panel of tasks. Specifically, our idea is to enhance the best existing ML methods by developing an additional layer on top of them that provides an interpretable guarantee on the error rate.

This is achieved by formalizing risk control in a certain task as a type of false discovery rate (FDR) (Benjamini and Hochberg, 1995) control problem. The FDR is a notion that originates from multiple testing (Benjamini and Hochberg, 1995). Single hypothesis testing is the process of choosing between two hypotheses, the null and the alternative, and multiple testing is the field of statistics that is concerned with taking decisions for multiple tests simultaneously. Thus, multiple testing procedures return a set of rejected null hypotheses, and the aim of multiple testing is to build procedures that keep a certain error rate below a level α. The FDR is a popular error rate criterion, defined as the proportion of errors (false discoveries) among the rejected set.

Since risk control involves a type of abstention option, it can be seen in general as a type of FDR control task, where a discovery corresponds to making a decision, and a false discovery corresponds to a decision that is incorrect. A strong advantage of the FDR is that it has a clear interpretation: if $\alpha = 5\%$ and a decision (whose specific form depends on the context) is made for 100 observations, then on average there are less than 5 decisions that are incorrect. Moreover, tools from the multiple testing literature on FDR control can be applied to the problem of controlling an FDR-like criterion.

Our methods can be seen as wrappers that take as input an off-the-shelf ML technique, designed for a certain learning task, and return a set of decisions such that the FDR is controlled at a user-specified level α. Specifically, we focus on the following three learning tasks: novelty detection, clustering and link prediction.

Adaptive novelty detection with FDR guarantee We start with the problem of FDR control for novelty detection in Chapter 2, where the aim is to detect novelties in a test sample based on a sample of "normal" behaviors. This problem amounts to a classical multiple testing setup, so that multiple testing tools for FDR control can be applied directly. Recent seminal works Bates et al. (2023); Yang et al. (2021) from the literature on conformal inference (Angelopoulos and Bates, 2021) and knockoffs (Barber and Candès, 2015), have shown that the test statistic could be learned from the data while maintaining the FDR control guarantee, hereby circumventing crucial limitations of historically-used strategies for FDR control. Our contribution consists in an extension of these previous works by proposing a new way to learn the test statistic that is more powerful than previously, while retaining the control of the FDR. The new method leverages classification algorithms to efficiently detect novelties. In particular, any off-the-shelf classification method, such as random forests or neural networks, can be used. FDR control is established in two main ways. First, we prove that the resulting "conformal" p-values satisfy the PRDS property. The PRDS property (Benjamini and Yekutieli, 2001) is a specific dependence assumption from the multiple testing literature, which, among other things, entails FDR control when using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). Secondly, we also provide new FDR bounds. Finally, we provide a power analysis, where we show that the proposed procedure has a power close to the one of the optimal likelihood ratio test with a suitable level. Numerical experiments on both simulated and real data demonstrate the substantial gain of power with respect to previous work.

Clustering with error rate control The problem of error rate control in a clustering task is studied in Chapter 3. A crucial point is to define the notion of error in clustering, since there is no unique definition of what a cluster is. To overcome this difficulty, we make the choice of considering a mixture model, that defines a natural ground truth clustering. In order to keep the risk smaller than a given level α, even in an unfavorable setting, we consider a procedure with an abstention option, whose output is a set of indices corresponding to the observations to which a cluster label is assigned. The problem is thus formalized in terms of controlling the false membership rate (FMR) criterion, defined as the average proportion of errors in the set of classified observations, up to label-switching. We propose a plug-in procedure inspired from the empirical Bayes approaches in multiple testing (Sun and Cai, 2007). Essentially, this procedure assigns cluster labels to observations for which the maximum class probability is above a data-driven threshold. Our main contribution consists in the theoretical analysis of this procedure: we quantify the FMR deviation with respect to the target level α with explicit remainder terms. In particular, our results imply that if the model estimate is accurate enough, the FMR is close to the target level, with a power close to optimal. Moreover, we develop robust bootstrap procedures for improved empirical performance.

Link prediction with FDR control Finally, we consider the problem of FDR control in a link prediction task in Chapter 4. FDR control in that specific setting does not correspond to a standard multiple testing problem, but we argue that high-level ideas from the literature on conformal inference and knockoffs (Barber and Candès, 2015; Weinstein et al., 2017; Bates et al., 2023; Yang et al., 2021; Mary and Roquain, 2022) and of Chapter 2 can still be applied. We propose a transposition of these ideas to the link prediction setup. The proposed method acts as a wrapper that takes as input an off-the-shelf link prediction technique, whose output consists of connection probabilities, and returns an FDR-controlling method. However, the graph structure induces intricate dependence in the data, which makes the setup markedly different from previous work, that must be taken into account in the procedure. The FDR

control is empirically demonstrated on both simulated and real data.

Chapter 1

Introduction

Contents

1.1 Motivation 1
1.2 From risk minimization to risk control 3
1.2.1 Background 3
1.2.2 Aim 4
1.3 Brief overview of the considered learning tasks 5
1.4 A primer on multiple testing 7
1.4.1 Preliminaries 10
1.4.2 Existing strategies for FDR control 12
1.5 Contributions 16

1.1 Motivation

Machine learning (ML) methods aim at leveraging data to learn various tasks, typically by optimizing a suitable data-based objective, whereas false discovery rate (FDR) control is a staple of multiple testing (Benjamini and Hochberg, 1995) that is concerned with building decision procedures that provide a finite-sample guarantee on the proportion of errors. In this thesis, we make these two fields meet in order to perform reliable machine learning. Reliability is critical in applications that involve high-stakes decisions, such as autonomous driving or medical diagnosis. We identify three issues of modern ML methods in terms of reliability:

Issue n°1: ML methods make mistakes ML methods have achieved impressive results in a wide range of applications (Jordan and Mitchell, 2015), but none of them are able to provide a small error guarantee in all settings: if the task at hand is intrinsically difficult, then there exists a minimal error, even for an "oracle" relying on infinite samples. In many sensitive contexts, however, the potential of error is not acceptable and should be dealt with.

Issue n°2: ML methods (usually) do not provide reliable confidence estimation Since models cannot be perfect, they should at least "know that they do not know". Modern machine learning methods, however, do not provide this information in general, which makes them unreliable. For instance, neural networks (NN) are known to be overly confident (Guo et al., 2017): Figure 1.1 illustrates, in a classification context, that for modern NN architectures the average confidence (i.e. the probability associated with the prediction on the test examples) is substantially higher than the accuracy (the percentage of correct predictions).

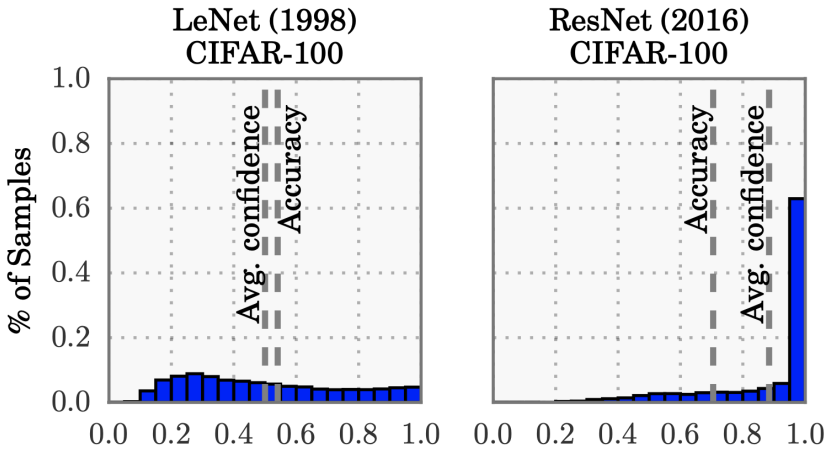


Figure 1.1: Source: (Guo et al., 2017). Histogram of the (estimated) probability associated with the predicted label in a classification task on CIFAR-100, for two NN architectures: a 5-layer LeNet (left) and a 110-layer ResNet (right). The 110-layer ResNet (He et al., 2016) represents a modern-day architecture compared to the 5-layer LeNet (LeCun et al., 1998) that dates back to early works on CNN.

In the case of NN, this concern is made worse by the fact that they are uninterpretable black boxes.

Issue n°3: Confidence estimation is often not sufficient There are a variety of methods dedicated to making the probabilities associated to the predictions of a ML model closer to the 'true' probabilities, see e.g. Guo et al. (2017); Zaidi et al. (2021); Lakshminarayanan et al. (2017). Moreover, conformal prediction (Angelopoulos and Bates, 2021) is a technique that has gathered an important interest in the statistics and machine learning community in the recent years, that provides valid prediction sets, that is, a set of predictions that contain the truth with a pre-specified level of confidence, and for arbitrary machine learning models. However, in any application that involves decision making, confidence estimation or prediction sets are not satisfactory since ultimately, it does not produce a decision. By contrast, a key to keeping the error rate low is to make use of a type of abstention (or selection) option, which amounts to abstain from making a decision when there is too much uncertainty. For instance, in autonomous driving, if the object detector is not confident at some point in time, then the system should rely more on the other sensors for braking. Alternatively, in automated medical diagnosis, if the prediction can not be made reliably, the system should defer to a human expert. However, choosing the abstention rule to be used to get the desired guarantee is difficult and cannot be deduced from confidence bounds in general, such as for instance, the valid prediction sets provided by conformal prediction (Jin and Candès, 2022).

The afore-mentioned issues motivate a shift in paradigm, from risk minimization to risk control, which refers to the goal of keeping the risk below a user-specified threshold. This can be achieved by using a type of abstention option that does not count as an error in the risk. However, to avoid that a method abstains for all observations, the cost of abstaining can be taken into account by seeking a procedure that abstains as little as possible under the constraint on the risk.

1.2 From risk minimization to risk control

1.2.1 Background

Risk control is a long-standing paradigm in binary classification. Early works can be divided according to two main lines of research: Neyman-Pearson (NP) classification and classification with a reject option. In NP classification (Cannon et al., 2002; Scott and Nowak, 2005; Rigollet and Tong, 2011; Blanchard et al., 2010a), risk control is achieved by focusing on one type of error at the expense of the other. In this framework, only a wrong classification into one of two classes, say class "1" (versus class "0"), counts as a mistake. Instead, the decision of classifying into class "0" corresponds to a "safer" option, where it is understood that there may be a potentially large proportion of misclassifications under this label. From this viewpoint, classifying into class "0" can be seen as a kind of abstention decision. Formally, the aim is to ensure that the probability of making a wrongful classification into class "1" is below a fixed margin of error, while minimizing the probability of making a wrong classification into class "0".

By contrast, in classification with a reject option (Herbei and Wegkamp, 2006; Bartlett and Wegkamp, 2008; Grandvalet et al., 2008), no focus is put on a particular type of error. Low overall error is achieved with the use of a reject (or abstention) option, where a classifier can abstain from making a class decision. In this framework, the reject option is accounted for by a particular fixed cost in the risk, which is less than the cost of making a mistake. A main drawback of this type of approach is the use of a cost instead of a confidence level. In practice, it is not always clear how to fix the abstention cost with respect to the cost of making a mistake.

These early works do not provide finite-sample guarantees. Recent breakthrough works have filled the gap, in binary classification and related tasks. To start with, in the field of variable selection, Barber and Candès (2015) introduced a "knockoff" method, with proven finite-sample guarantees on the FDR control. Developed for the Gaussian linear regression model, the core idea is to fabricate "knockoff" features that mimic the original features, in such a way that a knockoff variable satisfies a type of exchangeability with the true noise variables. Since the ground truth is known for the knockoffs, they can be used as benchmarks to evaluate the error (that is, the proportion of false positives) of a certain selection. This seminal work was followed by many extensions and refinements, see, e.g., Barber and Candès (2019); Bates et al. (2021); Weinstein et al. (2017); Barber et al. (2020); Nguyen et al. (2020); Sarkar and Tang (2022).

In the context of novelty (or outlier) detection, Haroush et al. (2022) used "conformal" p-values to perform model-free outlier detection, proving finite-sample control of the probability of making a wrongful novelty class decision for a single test point (which corresponds to a NP paradigm guarantee). Conformal p-values (Vovk et al., 2005) is a concept originating from the conformal prediction framework (Angelopoulos and Bates, 2021; Vovk et al., 2005; Balasubramanian et al., 2014), that measure statistical significance based on a sample of observations that have a "normal" behavior. Conformal p-values and knockoffs start from a similar idea: both techniques rely on having a reference set that is exchangeable with the observations corresponding to an absence of "signal", these being noise variables in the context of variable selection, or observations with normal behavior in the context of novelty detection. Subsequently, in a seminal work, Bates et al. (2023) showed that conformal p-values could be used to obtain a finite-sample guarantee for multiple points, which is formalized in terms of FDR control. Important works that followed include Yang et al. (2021) that consider a test statistic that is more data-adaptive, Liang et al. (2022) that consider a setting where a few examples of outliers are available, and Jin and Candès (2022) that consider a more general

"selection" task, see also Mary and Roquain (2022).

Finally, going back to the classification task, Geifman and El-Yaniv (2017) and Angelopoulos et al. (2021) perform classification with a reject option with finite-sample guarantees on the error, with the later work considering a more general prediction-related task that also includes, for instance, instance segmentation. In both of these works, the guarantee concerns a single test point. Lastly, in the context of fair ML (Barocas et al., 2019), Rava et al. (2021) achieve finite-sample guarantees for an FDR-like criterion, which can in addition take into account a certain protected attribute, and ensure that decision errors are not concentrated in protected groups.

1.2.2 Aim

Risk control involves a type of abstention option, that does not count as an error in the risk. As such, risk control in general can be seen as a type of FDR control task, where a discovery corresponds to making a decision, and a false discovery corresponds to a decision that is incorrect. A strong advantage of FDR control at level α is that it has a clear interpretation: if $\alpha = 5\%$ and a decision (whose specific form depends on the context) is made for 100 observations, then on average there are less than 5 decisions that are incorrect. In this thesis, we consider the problem of controlling an FDR-like criterion in several learning tasks:

- **Novelty detection (ND):** In novelty detection, the aim is to classify observations into 'novelties' and 'nominals'. There are two types of error that can be made and depending on the context, one type is more serious than the other. Here, the context considered is to avoid making a wrong classification into the novelty class (considered as a false positive). This type of concern is typical for applications that seek to screen a very large pool for promising candidates, which are investigated afterwards in a more costly stage. The problem is illustrated in Figure 1.2 on a classical image dataset of handwritten digits, where the aim is to detect digit '9's in the test sample based on a sample of digit '4's. The procedure, that declares as novelties the images with red boxes, can make false discoveries (digit '4') and true discoveries (digit '9'). The FDR is defined as the average proportion of errors among the detections. This problem has been previously investigated in Bates et al. (2023); Yang et al. (2021), and we aim to improve upon these previous works.

- **Clustering:** The aim of clustering is to partition the data into groups that are meaningful in some sense. There are different ways to define a notion of error in that case. In our work, we adopt a model-based viewpoint and use some kind of classification error. It relies on the assumption that there exists a unique ground truth partition, and considers that an error is made when an observation is classified into the wrong group. Moreover, we use an abstention option, which corresponds to not assigning a cluster label. In this set-up, an FDR criterion can be defined as the average proportion of errors among the observations that are chosen to be classified into a cluster.

- **Link prediction (LP):** In link prediction, the aim is to identify missing links in a partially observed graph. In an incomplete graph, the absence of an edge corresponds to either a lack of information or unreported information. In general, adding a false link in the graph is more serious than failing to identify a missing link. In this context, the FDR is defined as the average proportion of errors among the pairs of vertices that are declared to have a true edge, and controlling the FDR is appropriate to complete the graph with a meaningful risk control.

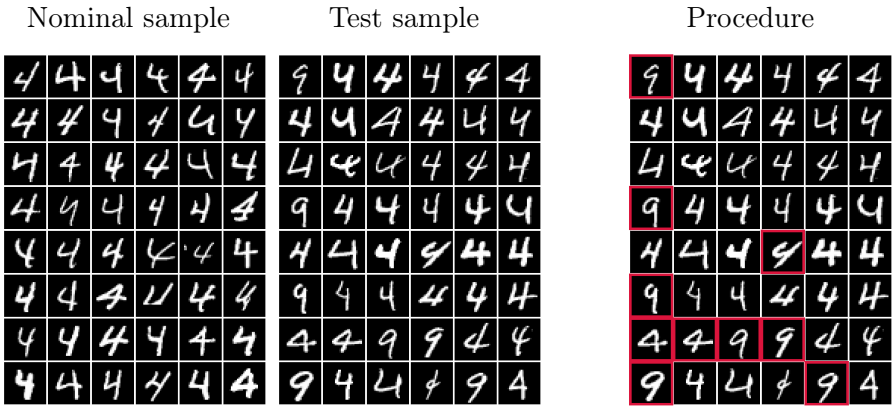


Figure 1.2: Illustration of the novelty detection task on the MNIST data set.

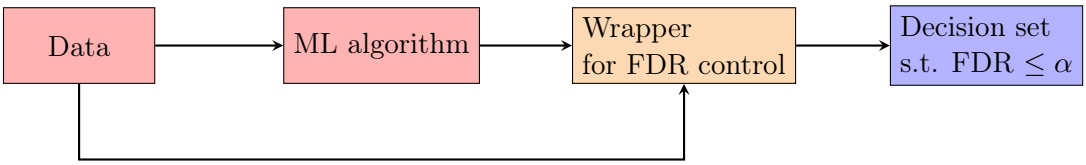


Figure 1.3: High-level illustration of the aim and approach pursued in this thesis.

In a nutshell, the goal of this thesis is to address the tasks above by using tools from multiple testing. To be clear, the new proposed methods do not replace the existing state-of-the-art algorithms in outlier detection, clustering, or link prediction; our methods are meant to be applied on top of those. In other words, our methods can be seen as wrappers that take as input an off-the-shelf technique designed for a given learning task, and return an FDR-controlling technique (see Figure 1.3). Thus, our approach has the full benefits of the best existing ML methods and we add an additional layer to achieve a control of the error rate.

In Section 1.3, we present the learning tasks considered in this thesis, namely novelty detection, clustering, and link prediction. Then Section 1.4 gives a primer on multiple testing. Finally, an overview of our contributions is provided in Section 1.5.

1.3 Brief overview of the considered learning tasks

Machine learning (ML) is a vast research field including various methods and tasks, see Figure 1.4 for a pictorial view. ML can be divided according to three paradigms: supervised learning, unsupervised learning and reinforcement learning. In supervised leaning, the data at hand consists of a set of observations for which the ground truth is known, and can be used to learn a predictor for new observations. In case no observed ground truth is provided, the term of unsupervised learning is used. Moreover, between the two, semi-supervised learning refers to tasks that contain both supervised and unsupervised elements. To complete the picture, reinforcement learning is an area of machine learning that differs from the previous paradigms, where the goal is to enable an agent to learn how to take actions in an interactive environment.

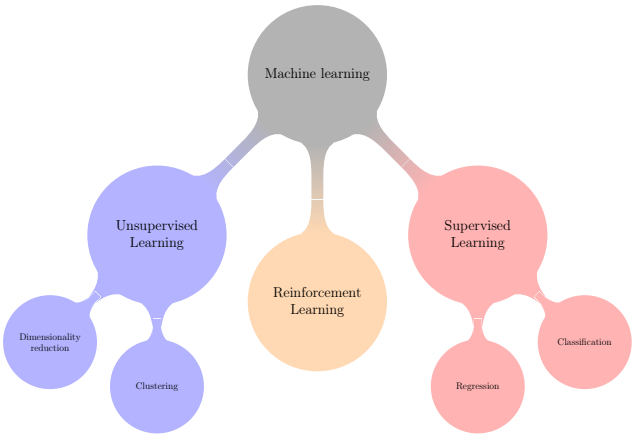


Figure 1.4: A brief look at machine learning tasks.

Here, we will focus on three learning tasks that have a varying degree of supervision: clustering (unsupervised), novelty detection (semi-supervised), and link prediction (supervised). We next proceed to give a brief description of each task.

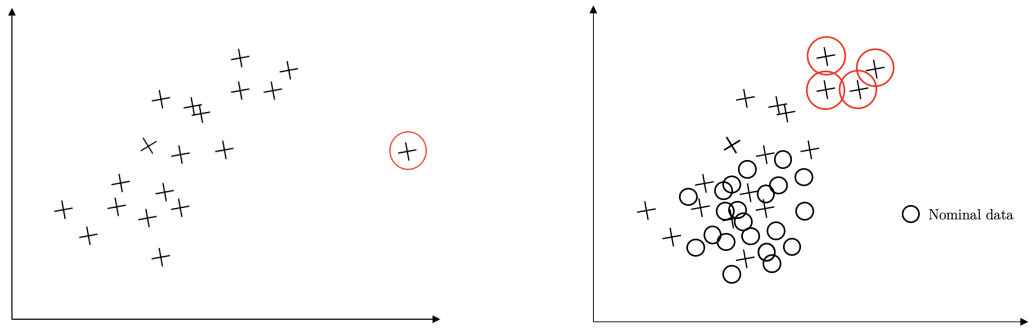
Novelty detection This task (also called anomaly detection, or outlier detection) consists in the identification of observations in a sample that are anomalous in some sense. Examples of definitions include:

- Observations that are far from the majority of the dataset (Liu et al., 2008). In that case, the task is unsupervised because we have no ground truth examples. A typical application is dataset cleaning where outliers must be removed: identifying these observations can help to clean the dataset of erroneous entries. Moreover, since these observations deviate a lot from the rest of the data, they can have a substantial impact when fitting a machine learning model, which is not desirable in some contexts.

- Observations that do not conform to a pre-defined notion of normal behavior (Schölkopf et al., 2001). Here, the task is semi-supervised because ground truth knowledge is partially available: it is assumed that normal behavior corresponds to some distribution P_0, where P_0 is known or a sample from P_0 is available. Applications include fraud detection (Patcha and Park, 2007), medical diagnosis (Tarassenko et al., 1995), galaxy detection (Mary and Roquain, 2022), and out-of-distribution detection (Lee et al., 2018).

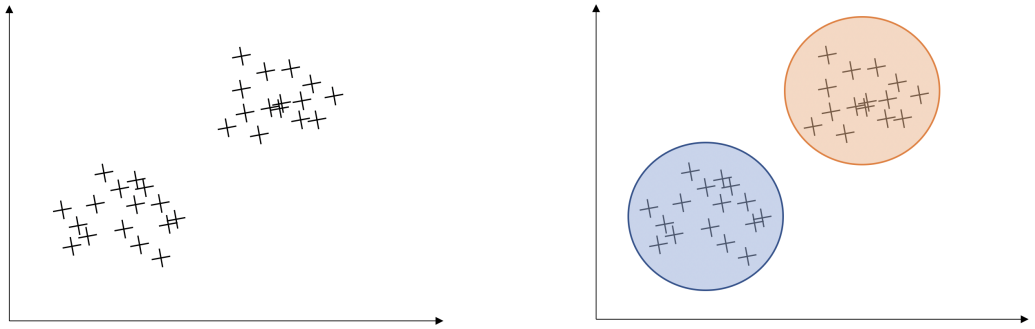
The two settings are illustrated in Figure 1.5. In this thesis, we consider the semi-supervised setting.

Clustering This task consists in partitioning a data sample into groups (called clusters) that are meaningful in some sense. Applications include customer segmentation, e-recommendation, data analysis, or gaining insights into biological processes (Grün, 2019). In general, the idea is that observations of the same cluster are 'similar', and observations in different clusters are 'dissimilar'. The task is illustrated in Figure 1.6. In addition, Figure 1.7 illustrates a key difficulty of this task, which is that outside of this general principle, there is no unique definition for the notion of a cluster. In fact, there could be several interesting partitionings (or structure) present in the data. A popular solution is to use a probabilistic model (see Grün (2019) and references therein), called a mixture model. In a mixture model, the data



(a) Unsupervised (b) Semi-supervised

Figure 1.5: The novelty detection task.



(a) Raw data (b) Clustering

Figure 1.6: The clustering task.

is sampled as follows: first, one samples a group label k for each individual i, then the observation for i is sampled from some distribution P_k. In that case, provided that the model is identifiable, the clustering task is well defined as it amounts to recovering the labels.

Link prediction This task consists in identifying links in a graph that is only partially observed. Graphs (or networks) denote data objects that consists of links (edges) between entities (nodes). Real-world examples are ubiquitous and include social networks, computer networks, food webs, molecules, etc. Examples are given in Figure 1.8 and Figure 1.9. In a partially observed graph, the presence of an edge between a pair of nodes, or lack thereof, is only reported for a part of them, and it is of interest to identify the missing edges. The problem is illustrated in Figure 1.10. Applications include friend or product recommendation (Li and Chen, 2013), and the reconstruction of biological networks, such as protein-protein interaction networks (Kovács et al., 2019), metabolic networks (Bleakley et al., 2007), or food webs (Terry and Lewis, 2020).

1.4 A primer on multiple testing

In this section, we provide an introduction on multiple testing, as well as the main multiple testing tools used in this thesis.

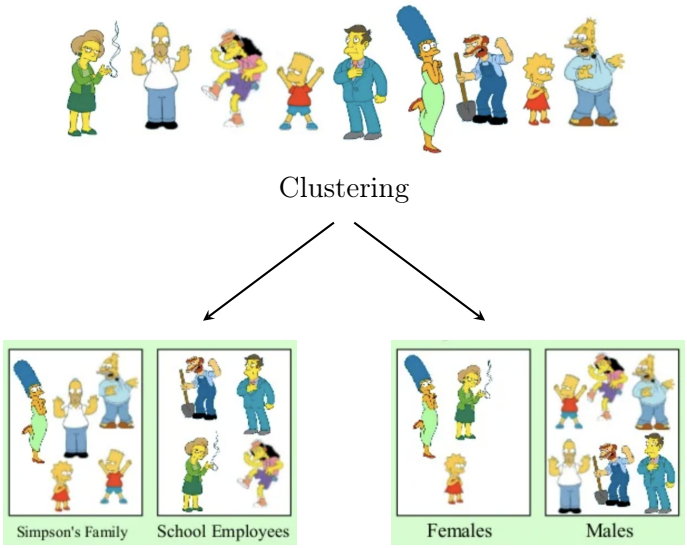


Figure 1.7: The subjectivity of clustering.

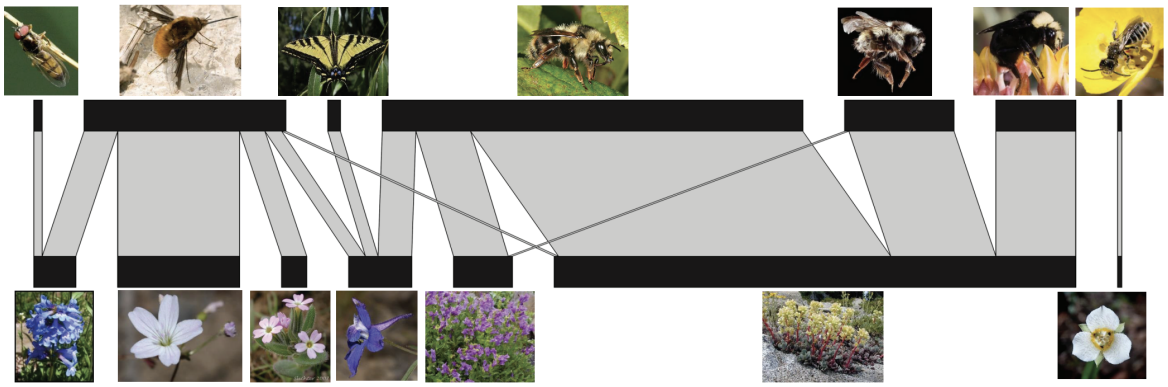


Figure 1.8: A pollination network (Seo and Hutchinson, 2018): nodes represent pollinator or plant species, edges denote interactions (the width indicates the frequency).

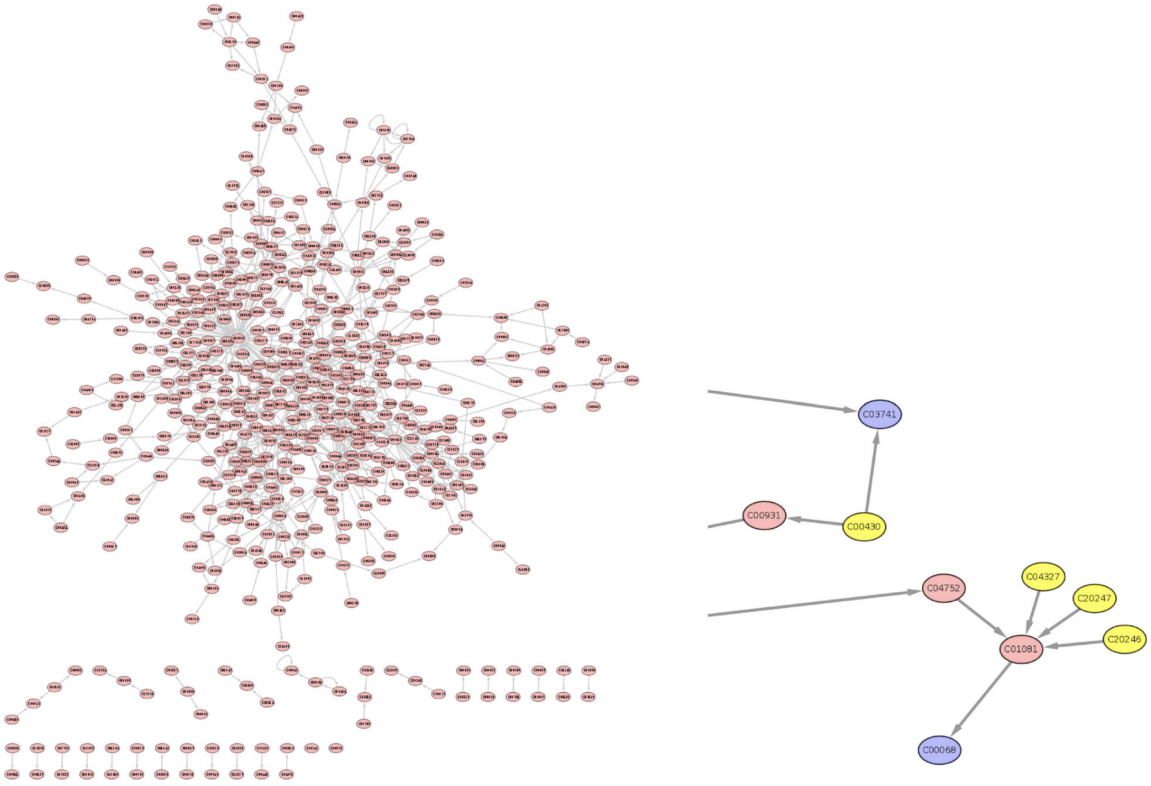


Figure 1.9: A metabolic network (Weber Zendrera et al., 2019): nodes represent molecules, edges denote chemical reactions. The right panel displays a zoom of the network.

(a) Complete graph (b) Observed graph

Figure 1.10: Illustration of the link prediction task. In the left panel, we have the true complete graph which is not observed. The right panel describes our observation: the true edges $(1, 2)$, $(1, 4)$, $(2, 3)$ are observed, along with the non-existent edge $(1, 3)$, but the information concerning the pairs $(2, 4)$ and $(3, 4)$ is missing. We aim to decide for $(2, 4)$ and $(3, 4)$ whether there is a true edge or not.

1.4.1 Preliminaries

Hypothesis testing

We start by recalling some basics about hypothesis testing. Let Z be a random variable (r.v.) denoting the observation at hand, valued in some measurable space $\mathcal{Z}$, and P its distribution, belonging to a family of distributions $\mathcal{P}$ (model). Hypothesis testing is the process of choosing between two hypotheses, H_0 (the null hypothesis) and H_1 (the alternative hypothesis), which are formalized as: H_0: "$P \in \mathcal{P}_0$" and H_1: "$P \in \mathcal{P}_1$", where $\mathcal{P}_0, \mathcal{P}_1$ are disjoint non empty subsets of $\mathcal{P}$. There are two types of errors that can occur:

- Rejecting the null hypothesis when it is true. The probability of making this mistake is called the type I error.
- Accepting the null hypothesis when the alternative is true. The probability of making this mistake is called the type II error.

The goal of hypothesis testing is to control the type I error at a given level α, while making the type II error as small as possible. A central notion in hypothesis testing is that of the p-value.

Definition 1 (p-value)**.** *A r.v. p is a valid p-value (or p-value for short) if it is super-uniform under H_0, i.e. $\forall \alpha \in [0,1]$, $\forall P \in \mathcal{P}_0$, $\mathbb{P}_{Z\sim P}(p \leq \alpha) \leq \alpha$.*

By definition, if p is a p-value, then the test that rejects H_0 when $p \leq \alpha$ controls the type I error at level α. Next, we give two examples of how to construct a p-value.

Let $T : z \in \mathcal{Z} \to \mathbb{R}$ be a test statistic for testing H_0, such that we decide to reject when $T(Z)$ is large. As a first example, the r.v. $p^*(Z)$ with

$$p^* \colon z \in \mathcal{Z} \mapsto \sup_{P \in \mathcal{P}_0} \mathbb{P}_{Z\sim P}(T(Z) \geq T(z))$$

is a valid p-value. In general, when the term of p-value is used in the literature, it refers to the p-value p^*. However p^* is available only under strong model assumptions which is not the case that we will consider. Hence p^* will mostly serves as an oracle. In the sequel, we will refer to p^* as a *theoretical* p-value.

For the second example, let $Z_1, \ldots, Z_n$ be a sample of observations such that $(Z, Z_1, \ldots, Z_n)$ is exchangeable under H_0. Then, a p-value (called here *empirical* p-value) may be obtained using only the computation of the rank of $T(Z)$ among $T(Z_1), \ldots, T(Z_n)$. This is formalized in the following result.

Lemma 1 (Romano and Wolf (2005); Arlot et al. (2010))**.** *If the sequence $(Z, Z_1, \ldots, Z_n)$ is exchangeable under H_0, then the r.v. given by*

$$\hat{p} = \frac{1}{n+1}\left(1 + \sum_{i=1}^{n} \mathbb{1}\{T(Z_i) \geq T(Z)\}\right)$$

is a valid p-value.

Empirical p-values have been historically used, for instance, in two contexts. First, in the context of two-sample testing, a sequence $Z_1, \ldots, Z_n$ that satisfy the exchangeability assumption of Lemma 1 can be constructed by permuting observations in the two samples and considering the resulting pair of data sets Z_j (Romano and Wolf, 2005). In that context the empirical p-value is referred to as a permutation-based p-value. Secondly, in the context of

semi-supervised anomaly detection, $\mathcal{P}_0$ is a singleton: $\mathcal{P}_0 = \{P_0\}$, and a sample of observations $Z_1, \dots, Z_n$ that are marginally distributed according to P_0 is available (representing examples of "normal" behavior). In that case, if exchangeability holds, an empirical p-value can be computed directly, and its use for anomaly detection dates back to Vovk et al. (2005). In that context the empirical p-value is referred to as a conformal p-value.

The multiplicity issue

Let $H_{0,1}, \dots, H_{0,m}$ be a set of m null hypotheses, $\mathcal{H}_0$ be the set of indices i such that $H_{0,i}$ is true (true nulls) with $m_0 = |\mathcal{H}_0|, \pi_0 = m_0/m$, and $\mathcal{H}_1 = \mathcal{H}_0^c$ the set of false nulls with $m_1 = |\mathcal{H}_1|, \pi_1 = m_1/m$. For instance, we have $m = 10000$ of genes and for each gene, we want to test if they are associated with some phenotype, such as a disease. For this, we have at hand observed expression levels on m genes for two groups of subjects, a treatment group (or "affected" group) and a control group (or "unaffected" group). If we consider each test, i.e. each gene, separately, compute a p-value for it and reject when the p-value is below α, then we end up with the following result: among the rejected set of hypotheses, there are on average $m_0\alpha$ that are false, which gives 450 false rejections for $m_0 = 9000$ inactive genes and $\alpha = 5\%$. Hence, using as a selection rule a non-corrected p-value thresholding at level $\alpha = 5\%$ will lead to a non-reliable list of genes.

The issue is that the type I error guarantee is valid only in a marginal sense, whereas when doing multiple tests we should think in terms of the joint distribution of the p-values. Thus the notion of type I error must be extended to a suitable error criterion taking into account the multiplicity aspect.

Selective inference The multiplicity issue is a part of a larger group of issues that come from mis-using hypothesis testing and which are at the root of the "reproducibility" crisis (Ioannidis, 2005; Zeevi et al., 2020; Shenhav et al., 2015; Heller et al., 2014), referring to the failure of reproducing published scientific results. In particular, a very common mis-use is data snooping, that refers to performing testing after looking at the data (i.e. deciding which hypotheses to test based on the data) or manipulating the data (such as removing variables from the data). Selective inference is the field of statistics that is dedicated to solving these issues. It encompasses two main paradigms. The first is multiple testing and it consists of producing a set of selected items with a guarantee on the error. Alternatively, if the user has its own selection rule due to, e.g., common practice in the field, or budget constraints, selective inference can be used to identify confidence bounds on the error rate for the selected set. This second paradigm is known as post-hoc selective inference (Genovese and Wasserman, 2006; Goeman and Solari, 2011; Lee et al., 2016; Benjamini et al., 2019; Blanchard et al., 2021).

Multiple testing

We define a multiple testing procedure as a (measurable) function $R = R(Z)$ that returns a subset of $\{1, \dots, m\}$ corresponding to the indices i of the rejected null hypotheses $H_{0,i}$. The main error rate criteria in the multiple testing literature are:

- Family-Wise Error Rate (FWER): the probability of making at least one false rejection, i.e.
$$\mathrm{FWER}(R) = \mathbb{P}_{Z\sim P}(|R| \geq 1).$$
- False Discovery Rate (FDR): the expected proportion of false rejections among the

rejections, i.e.

$$\mathrm{FDR}(R) = \mathbb{E}_{Z\sim P}\left[\frac{\sum_{i\in\mathcal{H}_0}\mathbb{1}_{i\in R}}{1\vee|R|}\right].$$

- Marginal False Discovery Rate (mFDR): the expected number of false rejections over the expected number of rejections, i.e.

$$\mathrm{mFDR}(R) = \frac{\mathbb{E}_{Z\sim P}\left[\sum_{i\in\mathcal{H}_0}\mathbb{1}_{i\in R}\right]}{\mathbb{E}_{Z\sim P}\left[|R|\right]},$$

with the convention $0/0 = 0$.

How to choose among these criteria ? For very sensitive contexts it may be required to control the probability of making any error, in which case the goal of controlling the FWER is appropriate. However this criterion is very stringent: if m is large, in practical cases the number of hypotheses that are rejected with this approach will be very few or none, which can be problematic. The FDR criterion introduced by Benjamini and Hochberg (1995) fills this gap by allowing for false rejections, in an amount proportional to the number of rejections. This provides a clear interpretation: if we make, for instance, 100 rejections, then an FDR below $\alpha = 5\%$ guarantees that on average, there are at most 5 errors in the rejected set. The FDR approach is well suited for "exploratory" research, that is, cases where we are screening a very large pool for a set of promising candidates (such as genes associated to a phenotype of interest), with the idea that in a later stage, these candidates are to be investigated carefully for confirmation. By contrast, the FWER is more appropriate for "confirmatory" research, when a definite result is desired. Finally, the mFDR (Genovese and Wasserman, 2002) is a substitute of the FDR that involves a ratio of expectations rather than the expectation of a ratio, which is easier to work with in some cases.

Remark 1 (mFDR versus FDR). The mFDR can be understood as a "single-point guarantee", in contrast to the FDR that is a guarantee over multiple points. Indeed, assuming independence of the rejections in the mFDR definition (admittedly this is an over-simplification, practical procedures are data-driven), the mFDR reduces to the probability of making a mistake for a single hypothesis point given that it is rejected.

We also need to extend the notion of type II error, that is, define a notion of power. In general, we work with the True Discovery Rate (TDR), see e.g. (Dickhaus, 2014), defined as the expected proportion of rejections among the false nulls, i.e.

$$\mathrm{TDR}(R) = \mathbb{E}_{Z\sim P}\left[\frac{\sum_{i\in\mathcal{H}_1}\mathbb{1}_{i\in R}}{1\vee|m_1|}\right].$$

Another criterion is the expected number of rejections, $\mathbb{E}_{Z\sim P}(|R|)$, used for convenience when the TDR is out of reach. The aim in multiple testing is to build a procedure R that controls the FDR (or FWER, or mFDR) at the level α, while having a power as large as possible.

1.4.2 Existing strategies for FDR control

We present a panel of existing strategies for FDR control. We start with two main types of historically-used approaches: the Benjamini-Hochberg (BH) (Benjamini and Hochberg, 1995) procedure and its variants, and the empirical Bayes approaches (Efron et al., 2001; Sun and Cai, 2007; Sun and Cai, 2009; Cai et al., 2019; Heller and Rosset, 2021; Jin and Cai, 2007; Cai and Jin, 2010; Heller and Yekutieli, 2014; Roquain and Verzelen, 2022; Abraham et al., 2022; Rebafka et al., 2022). Then, we move on to recent approaches, namely knockoffs and related methods (Barber and Candès, 2015; Weinstein et al., 2017; Mary and Roquain, 2022; Bates et al., 2023; Yang et al., 2021).

The Benjamini-Hochberg procedure The BH procedure is given in Algorithm 1.

Algorithm 1 BH procedure

Input: p-values $(p_j)_{1\leq j\leq m}$
1. Sort the p-values: $p_{(1)} \leq \cdots \leq p_{(m)}$
2. Find the largest k such that $p_{(k)} \leq \alpha k/m$
3. Reject the null hypotheses $H_{(1)}, \ldots, H_{(k)}$ (reject nothing if $k = 0$)

Remark 2. The intuition behind the BH procedure is as follows. When we reject all hypotheses with p-value $\leq t$, then it is expected that $m_0 t$ false rejections are made on average. Thus, the false discovery proportion can be estimated by $m_0 t/|R(t)|$ where $R(t)$ denotes the rejection set. Now, the quantity $m_0 t/|R(t)$ cannot be computed, because m_0 is unknown, but $mt/|R(t)|$ is a computable upper-bound. The BH procedure chooses the cut-off t in a data-driven way such that $mt/|R(t)| \leq \alpha$, while rejecting as much as possible under this constraint.

BH is guaranteed to control the FDR at the level $\alpha\pi_0$ under independence of the p-values (Benjamini and Hochberg, 1995). The independence assumption can be relaxed to a specific dependence assumption called the Positive Regression Dependent on a Subset (PRDS) property (Benjamini and Yekutieli, 2001).

Definition 2 (PRDS)**.** *A random vector $\mathbf{X} = (X_i, 1 \leq i \leq m)$ is PRDS on a subset $M \subset \{1, \ldots, m\}$ if, for any $i \in M$ and increasing measurable set D, the function $u \mapsto \mathbb{P}(\mathbf{X} \in D \mid X_i = u)$ is nondecreasing.*

A set D is said to be increasing if for any $x \in D$ and y, we have $y \in D$ provided that $y_i \geq x_i$ for all i. In words, this assumption says if we condition on the value of one X_i, and increase this value, $\mathbf{X}$ is more probable to have large values coordinate-wise. For instance, Benjamini and Yekutieli (2001) give the following example of a PRDS distribution: for $\mathbf{X} \sim \mathcal{N}(\mu, \Sigma)$ and $I_0 = \{1 \leq i \leq m : \ \mu_j = 0\}$, if $\Sigma_{ij} \geq 0$ for each $i \in I_0$ and $j \neq i$, then $\mathbf{X}$ is PRDS on I_0.

Remark 3. In the FDR guarantee of the BH procedure, the control is at the level $\alpha\pi_0$ and not α. If π_0 is not close to one, this results in a loss of power, in the sense that the margin of error is not fully used to make as much rejections as possible. Variants of the BH procedure have been proposed to estimate π_0 (Storey, 2002; Benjamini et al., 2006; Blanchard and Roquain, 2009). In particular, a well-known result from the literature is the following: for the Storey-BH procedure, which consists in applying BH at the level $\alpha/\hat{\pi}_0$ with $\hat{\pi}_0 = (n/2)^{-1}(1+n-|R(1/2)|)$, the FDR control is guaranteed under independence of the p-values (Storey et al., 2004).

Empirical Bayes approaches The BH procedure requires p-values, which may not be available in practice when the distribution of the null tests statistics is unknown or misspecified, as argued in a series of papers by Efron, see Efron et al. (2001) and Efron (2004, 2007, 2008, 2009) (see also Figure 1 in Roquain and Verzelen (2022) and references therein). To avoid using p-values, the use of empirical Bayes procedures (Sun and Cai, 2007; Sun and Cai, 2009; Cai et al., 2019; Heller and Rosset, 2021) is an approach popularized by Efron et al. (2001) and widely used afterwards, that is based on so-called local FDR quantities (Efron et al., 2001). More formally, Efron et al. (2001) introduced a two-group mixture model, where the observation consists of m univariate measurements $Z_1, \ldots, Z_m$, assumed to be generated independently as

$$\begin{aligned} H_i &\sim \mathcal{B}(\pi_0), && i = 1, \ldots, m \\ Z_i | H_i &\sim f_0 \, \mathbb{1}_{H_i=0} + f_1 \, \mathbb{1}_{H_i=1}, && i = 1, \ldots, m. \end{aligned}$$

The local FDR is defined as the density ratio $\text{Lfdr} : z \mapsto \frac{\pi_0 f_0(z)}{\pi_0 f_0(z)+\pi_1 f_1(z)}$, where f_0 and f_1 are the unknown densities of Z_i under the null and the alternative, respectively. The local FDR value $\text{Lfdr}(Z_i)$ is equal to the probability that H_i is 0 given the observation Z_i, and can be used to act as a test statistic (that is more refined than Z_i). We focus on the procedure proposed in the seminal work Sun and Cai (2007), which paved the way for many studies that followed (Heller and Rosset, 2021). The approach of Sun and Cai (2007) consists in identifying the optimal mFDR-controlling procedure, which is obtained by thresholding the local FDR values $\text{Lfdr}(Z_i)$. Since the densities f_0 and f_1 and the null proportion π_0 are unknown in practice, this procedure is called an "oracle" procedure. Relying on estimates $\hat{f}_0, \hat{f}_1, \hat{\pi}_0$, Sun and Cai (2007) proposes a type of plug-in procedure based on the oracle, given in Algorithm 2.

Algorithm 2 Plug-in local FDR procedure (Sun and Cai, 2007)

Input: Estimates $\hat{f}_0, \hat{f}_1, \hat{\pi}_0$
1. Sort the local FDR values: $\widehat{\text{Lfdr}}_{(1)} \leq \cdots \leq \widehat{\text{Lfdr}}_{(m)}$
2. Find the largest k such that $\sum_{i=1}^{k} \widehat{\text{Lfdr}}_{(i)} \leq \alpha$
3. Reject the null hypotheses $H_{(1)}, \ldots, H_{(k)}$ (reject nothing if $k = 0$)

The procedure is shown to control the mFDR asymptotically in Sun and Cai (2007), and in a HMM set-up in Sun and Cai (2009), provided that the model parameters can be estimated consistently. Moreover, under the same assumptions, it is power-optimal in an asymptotic sense. Recently, in a general two-group mixture model where the measurements can be dependent, Heller and Rosset (2021) identified the optimal procedure in terms of FDR control, showing that it consists of a data-driven thresholding local FDR values, and gave an algorithm based on linear programming to compute it.

Historically, the usual strategies for FDR control relied on the BH procedure with theoretical p-values and the empirical Bayes approaches. However, both types of strategies display serious practical limitations. On the one hand, theoretical p-values require to know the distribution of the test statistic under the null hypothesis. However, the null distribution can be unknown or mis-specified in practice. Moreover, it constrains the test statistic to be prespecified, while it should be learned from data for the sake of power. On the other hand, empirical Bayes avoids the use of p-values and learns the test statistic from the data, but the control is only accurate as long as the quality of the estimates is good, and can be severely violated otherwise. Recently, there has been a series of seminal works Barber and Candès (2015); Weinstein et al. (2017); Bates et al. (2023); Yang et al. (2021); Mary and Roquain (2022) that partially circumvent these limitations. We next proceed to give an overview of these works, known in the literature as either knockoffs methods or conformal inference-based methods, the later being strongly related to the former.

Knockoffs and related methods, part I: variable selection In a breakthrough work, Barber and Candès (2015) introduced a "knockoff" method that comes with a finite-sample FDR guarantee, while making use of a data-driven test statistic, in the specific context of variable selection in the Gaussian linear regression model. Specifically, we observe $\mathbf{X}, \mathbf{Y}$ in the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, with $\beta \in \mathbb{R}^m$, $\epsilon \sim \mathcal{N}(0, \sigma^2 I_N)$, in which $\mathbf{X}$ is a non-random matrix valued in $\mathbb{R}^{N\times m}$, σ is unknown, and $m \leq N$. The aim is to test $H_{0,j} \colon \beta_j = 0$ against the alternative $H_{1,j} \colon \beta_j \neq 0$, simultaneously for all $1 \leq j \leq m$. The general idea of the method is to introduce "fake" variables $\tilde{\mathbf{X}} \in \mathbb{R}^{N\times m}$ designed to be exchangeable with the true null variables, so that they can be used as benchmarks to estimate the proportion of

false positives in a certain selection. To be more precise, the knockoff variables are used to construct test statistics that obey a "sign-flipping" property, in which the signs of null test statistics are i.i.d. random. In particular, such test statistics can be constructed from the LASSO (Tibshirani, 1996) solution fitted on the augmented matrix $[\mathbf{X}\tilde{\mathbf{X}}]$. Subsequently, the method was extended in Barber and Candès (2019) to a setting where $\mathbf{X}$ is random and $m \geq N$ ("model-X knockoffs"). Further, Weinstein et al. (2017) considered the particular case where the entries of $\mathbf{X}$ are i.i.d. according to a known distribution G, in which case knockoffs variables can be generated i.i.d. according to G in an arbitrary number n. In that setup, Weinstein et al. (2017) proposed another FDR-controlling procedure, the counting knockoffs (CK), with the aim of increasing power. Here, the null test statistics do not have to verify the sign-flipping property, but instead, they should satisfy exchangeability with the statistics computed for knockoffs; the LASSO coefficient is given as a possible choice.

Knockoffs and related methods, part 2: out-of-distribution testing A series of subsequent works (Mary and Roquain, 2022; Bates et al., 2023; Yang et al., 2021) considered out-of-distribution testing and provided model-free procedures that come with finite-sample FDR control, and they are closely related with the knockoff methodology. In these works, we have at hand a first sample of observations $Z_1, \ldots, Z_n$ sharing a common marginal distribution P_0, and a second sample of observations $Z_{n+1}, \ldots, Z_{n+m}$ (test sample), such that the aim is to test $H_{0,j} \colon Z_{n+j} \sim P_0$ versus $H_{1,j} \colon Z_{n+j} \nsim P_0$. In the particular case where the measurements Z_j are real-valued (which is equivalent to assume that the test statistics are directly given), Mary and Roquain (2022) studied the "empirical BH" procedure that consists of using the empirical p-values $\hat{p}_j = (1 + \sum_{i=1}^{n} \mathbb{1}\{Z_i \leq Z_{n+j}\})/(n+1)$ into BH and proved FDR control under exchangeability of the null measurements $(Z_1, \ldots, Z_n, Z_j, j \in \mathcal{H}_0)$. In particular, Mary and Roquain (2022) established that empirical BH is equivalent to CK, or in other words, that CK is a particular case of empirical BH with specific test statistics designed for variable selection (and for which the exchangeability assumption holds). Independently, Bates et al. (2023) studied the use of conformal p-values into BH. Conformal p-values rely on reducing each multivariate observation Z_j to a univariate score S_j and computing the rank of the test score among the scores of the nominal data:

$$p_j = \frac{1}{n+1}\left(1 + \sum_{i=1}^{n} \mathbb{1}_{S_i \geq S_{n+j}}\right), \quad 1 \leq j \leq m.$$

In this view, conformal p-values are empirical p-values, and thus, BH with conformal p-values is equivalent to empirical BH with the test statistics S_j. However, with respect to Mary and Roquain (2022), Bates et al. (2023) shows that the score S_j can be *learned* from the nominal data by proceeding as follows: the nominal sample is split into two subsets, $\{Z_1, \ldots, Z_k\}$ and $\{Z_{k+1}, \ldots, Z_n\}$. The first part is used to learn the score (e.g., using one-class classifiers Schölkopf et al., 2001), whereas the second is used to compute the ranks (hence n is replaced by $n-k$ in the above equation). Strikingly, the resulting p-values ("split-conformal" p-values) are shown to satisfy the PRDS assumption (see Definition 2) under i.i.d. data, which entails that FDR control is guaranteed with the BH procedure.

Finally, Yang et al. (2021) proved finite-sample FDR control for empirical BH with a test statistic learned on the test data. Specifically, in their framework, the null distribution is known, and the test statistic is a function of the mixed sample $\{Z_1, \ldots, Z_{n+m}\}$. In particular, they propose to use the estimated local FDR (Efron et al., 2001; Sun and Cai, 2007) as the score function.

1.5 Contributions

Adaptive novelty detection with FDR guarantee In Chapter 2, we consider the aim of FDR control in a novelty detection task, where we have at hand a sample of nominal data $Z_1, \dots, Z_n$ sharing a common marginal distribution P_0, and a test sample of unlabeled data $Z_{n+1}, \dots, Z_{n+m}$ (see Figure 1.2). This task amounts to a standard multiple testing problem, with the null hypotheses $H_{0,i}$: $Z_{n+i} \sim P_0$, $i = 1, \dots, m$. As seen in Section 1.4.2, recent seminal works by Bates et al. (2023); Yang et al. (2021) have shown that the test statistic could be learned from the data while maintaining the FDR control guarantee, hereby circumventing crucial limitations of historically-used strategies for FDR control. The procedure studied in both of these works is the empirical BH procedure (Mary and Roquain, 2022), equivalent to the CK procedure (Weinstein et al., 2017), with certain model-free test statistics. However, on the one hand, in Bates et al. (2023), the test statistic is learned from nominal data only, and does not adapt to the alternatives. On the other hand, Yang et al. (2021) utilize only the test sample: the null distribution is assumed to be known in their framework. Moreover, the test statistic in Yang et al. (2021) is the local FDR, which involves unknown densities, that are difficult to fit in high dimensions.

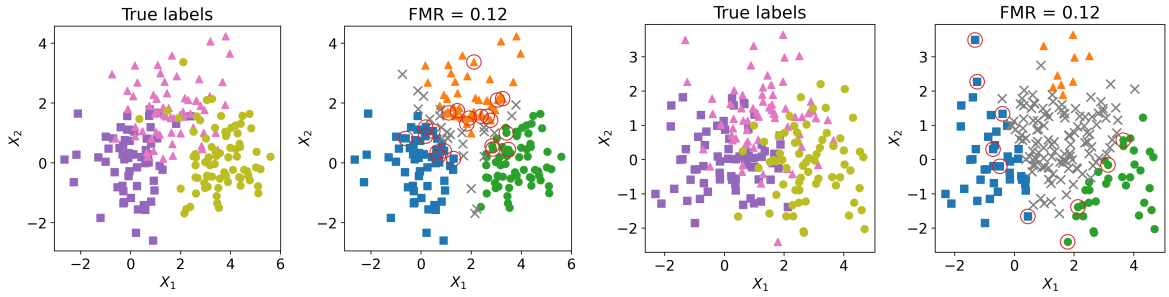
Our contribution consists in an extension of these previous works by proposing a new way to learn the score that utilizes both the nominal sample and the test sample, while retaining the control of the FDR. The idea is to learn a classification of $\{Z_1, \dots, Z_k\}$ against the mixed sample $\{Z_{k+1}, \dots, Z_n, Z_{n+1}, \dots, Z_{n+m}\}$ and to use the probability of being in the class corresponding to the mixed sample as score. When the classification algorithm performs well, the scores tend to be larger for anomalies than for nulls, because anomalies are only present in the mixed sample. Moreover, any classification algorithm can be used, including the state-of-the art, which makes the method both flexible and powerful.

Mixing $Z_{k+1}, \dots, Z_n$ with $Z_{n+1}, \dots, Z_{n+m}$ allows to keep the null scores $(S_k, \dots, S_n, S_j, j \in \mathcal{H}_0)$ exchangeable, which ultimately is the key to FDR control, as shown in Weinstein et al. (2017) in the context of variable selection for the Gaussian linear regression model. We prove that the PRDS property, established under independency in Bates et al. (2023), still holds (which entails FDR control), and we propose new FDR bounds. We also prove FDR control for a version of the procedure that estimates the null proportion π_0. Finally, we provide a power analysis, where we show that the proposed procedure has a power close to the one of the optimal likelihood ratio test with a suitable level. Numerical experiments on both simulated and real data demonstrate the substantial gain of power with respect to previous work. This procedure has already generated an important interest in community, see Bashari et al. (2023); Liang et al. (2023); Gao and Zhao (2023).

Clustering with error rate control In Chapter 3, we consider the problem of error rate control in a clustering task. A main challenge is how to define the notion of error of a clustering, since there is no unique definition of what a cluster is. To overcome this difficulty, we make the choice of considering a mixture model, that defines a natural ground truth clustering. More precisely, we assume that the observed data $\mathbf{X} = (X_1, \dots, X_n)$ is such that every X_i is independently generated according to :

$$Z \sim \Sigma_{q=1}^{Q} \pi_q \delta_q$$
$$X|Z = q \sim F_{\phi_q}, \quad 1 \leq q \leq Q,$$

with $\{F_u, u \in \mathcal{U}\}$ a collection of probability distributions on $\mathbb{R}^d$ and $(\pi_1, \dots, \pi_Q)$, $(\phi_1, \dots, \phi_q)$ some parameters. In this setup, the clustering task amounts to recover the unobserved group labels Z_i. Thus, the clustering risk is naturally defined as the expectation of the difference



(a) Separated clusters (b) Ambiguous clusters

Figure 1.11: Data from Gaussian mixtures with three components ($n = 200$), in a fairly separated case (panel (a)) and an ambiguous case (panel (b)). In each panel, the left part displays the true clustering, while the right part illustrates the new procedure (plug-in procedure at level $\alpha = 10\%$), that does not cluster all items. The points not classified are depicted by grey crosses. Red circles indicate erroneous labels.

between the estimation partition and the true one, that is, $\sum_{i=1}^{n} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$, with a suitable way of taking into account label-switching. In order to keep the risk smaller than a given level α even in an unfavorable setting, we propose a procedure with an abstention option, whose output is a set of indices $S = S(\mathbf{X}) \subset \{1, \dots, n\}$ corresponding to the observations to which a cluster label is assigned. The problem is thus formalized in terms of controlling the false membership rate (FMR) criterion, given by

$$\mathrm{FMR} = \mathbb{E}\left(\min_{\sigma \in [Q]} \mathbb{E}\left(\frac{\sum_{i=1}^{n} \mathbb{1}\{Z_i \neq \sigma(\hat{Z}_i)\}\, \mathbb{1}_{i \in S}}{\max(|S|, 1)} \,\middle|\, \mathbf{X} \right) \right).$$

In contrast to Chapter 2, the setting here is completely unsupervised, making the task of risk control more challenging. We propose a plug-in procedure that is inspired from the empirical Bayes approaches (Sun and Cai, 2007) in multiple testing (see Section 1.4). It works by first computing an estimate $\hat{\theta}$ of the model parameter $\theta = (\pi, \phi)$, which is used to evaluate the class probabilities $\mathbb{P}_{\hat{\theta}}(Z_i = q|X_i)$. Then the procedure assigns the cluster label $\mathrm{argmax}_q\, \mathbb{P}_{\hat{\theta}}(Z_i = q|X_i)$ corresponding to the maximum class probability, provided that this probability is above a data-dependent threshold (otherwise, the procedure abstains). The method is illustrated in Figure 1.11. Our main contribution consists in the theoretical analysis of this procedure: we quantify the FMR deviation of the plug-in procedure, with respect to the target level α, in terms of the estimation error and the model complexity. In particular, our results show that if the model estimate is accurate enough, the FMR is close to the target level. However, the plug-in procedure inherits from the main limitation of the empirical Bayes approaches in multiple testing: it relies on the quality of model estimation, which may not be taken for granted in practice. To remedy this issue, we develop bootstrap procedures and assess their performance in numerical experiments.

Link prediction with FDR control In Chapter 4, we consider the aim of FDR control in a link prediction task (see Figure 1.10). This setting is the most supervised one compared to those of Chapters 2 and 3, in the sense that we observe both a part of the true edges and a part of the non-existing ones. FDR control in this specific setting does not correspond to a standard multiple testing problem, but we argue that high-level ideas from the literature on conformal p-values (Barber and Candès, 2015; Weinstein et al., 2017; Bates et al., 2023; Yang et al., 2021; Mary and Roquain, 2022), namely, the comparison of a test score to scores of a reference set,

can still be applied. We propose a transposition of the procedure of Weinstein et al. (2017); Bates et al. (2023); Yang et al. (2021); Mary and Roquain (2022) and of Chapter 2 to the link prediction setup. More precisely, we use an estimate of the connection probability for an unobserved pair of nodes as a score indicating the relevance of an edge between them. Such an estimate can be obtained using an off-the-shelf link prediction method. We then compare the connection probability for a non-observed pair of nodes to connection probabilities of pairs that are known to be non-existent edges for FDR control. However, the graph structure makes the scores dependent on each other in an intricate way, that must be taken into account in the procedure. The FDR control is empirically demonstrated on both simulated and real data. Compared to previous work on knockoffs and conformal p-values, this problem presents an interesting setup where exchangeability of the scores does not hold.

Outline of the manuscript Each chapter is independent and self-contained. Chapter 2 is a joint work with Lihua Lei (Stanford University), David Mary (Université Nice Côte d'Azur), and Etienne Roquain (Sorbonne Université), that is currently in revision. Chapter 3 is a joint work with Tabea Rebafka (Sorbonne Université), Etienne Roquain (Sorbonne Université), and Nataliya Sokolovska (Sorbonne Université), that has been submitted for publication. Chapter 4 is a personal work and has been submitted for publication.

Chapter 2

Adaptive novelty detection with FDR guarantee

This paper studies the semi-supervised novelty detection problem where a set of "typical" measurements is available to the researcher. Motivated by recent advances in multiple testing and conformal inference, we propose AdaDetect, a flexible method that is able to wrap around any probabilistic classification algorithm and control the false discovery rate (FDR) on detected novelties in finite samples without any distributional assumption other than exchangeability. In contrast to classical FDR-controlling procedures that are often committed to a pre-specified p-value function, AdaDetect learns the transformation in a data-adaptive manner to focus the power on the directions that distinguish between inliers and outliers. Inspired by the multiple testing literature, we further propose variants of AdaDetect that are adaptive to the proportion of nulls while maintaining the finite-sample FDR control. The methods are illustrated on synthetic datasets and real-world datasets, including an application in astrophysics.

Contents

2.1 Introduction 20
2.1.1 Novelty detection 20
2.1.2 Existing strategies 21
2.1.3 Contributions 23
2.2 Preliminaries 24
2.2.1 Notation 24
2.2.2 Criteria 24
2.2.3 BH algorithm and its π_0-adaptive variants 25
2.2.4 Our method 25
2.3 FDR control 27
2.3.1 Exchangeability 28
2.3.2 The p-values are PRDS 28
2.3.3 A new FDR expression 29
2.3.4 New FDR bounds for π_0-adaptive procedures 29
2.4 Constructing score functions 31
2.4.1 Assumptions and notation 31
2.4.2 Optimal score function 32
2.4.3 Density estimation 32
2.4.4 PU classification 33

2.4.5 AdaDetect with cross-validation 34
2.5 Power results 36
2.5.1 A constrained ERM score function 37
2.5.2 General score functions 39
2.6 Experiments 41
2.6.1 Simulated data 41
2.6.2 Semi-synthetic data 42
2.7 An astronomy application 43
2.8 Conclusion and discussion 45
2.8.1 Limitations of AdaDetect 45
2.8.2 Other future works 46

2.1 Introduction

2.1.1 Novelty detection

In this paper, we consider a novelty detection problem (see, e.g., Blanchard et al. (2010b) and references therein) where we observe:

- a null training sample (NTS hereafter) $Y = (Y_1, \ldots, Y_n)$ of "typical" measurements where Y_is share a common marginal distribution P_0 which we refer to as the null distribution;

- and a test sample $X = (X_1, \ldots, X_m)$ of "unlabeled" measurements for which the marginal distribution of X_i is denoted by P_i, which might be different from P_0.

These measurements are assumed to take values in a general space $\mathcal{Z}$ endowed with a prescribed σ-field. For example, the space can be the set of real matrices ($\mathcal{Z} = \mathbb{R}^{d \times d'}$) or real vectors ($\mathcal{Z} = \mathbb{R}^d$), whose dimension is potentially large.

Putting two samples together, we observe

$$Z = (Z_1, \ldots, Z_{n+m}) = (Y_1, \ldots, Y_n, X_1, \ldots, X_m).$$

The aim is to detect novelties, namely X_is with $P_i \neq P_0$. This task is illustrated in Figure 2.1 on a classical image dataset, where we want to detect hand-written digit '9's in the test sample based on an NTS of digits '4'. The procedure, that declares as novelties the images with red boxes, can make false discoveries (digit '4') and true discoveries (digit '9').

To avoid false positives that might be costly in practice, we seek to control the false discovery rate (FDR), defined as the average proportion of errors among the discoveries, while attempting to maximize the true discovery rate (TDR), defined as the average portion of detected novelties. FDR has been a very popular criterion in multiple testing and exploratory analysis since its introduction by Benjamini and Hochberg (1995); see Benjamini (2010) for a detailed discussion and Barber and Candès (2015); Bogdan et al. (2015); Javanmard and Javadi (2019); Barber et al. (2020); Ma et al. (2021a) for recent developments, among others.

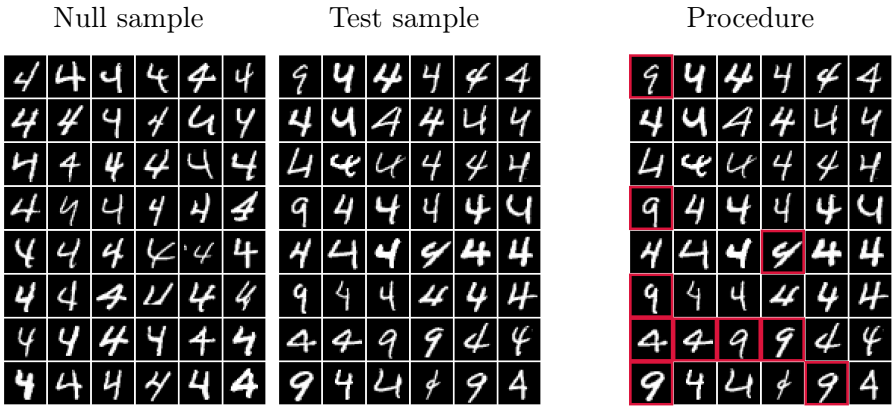


Figure 2.1: Illustration of the novelty detection task on the MNIST dataset (LeCun and Cortes, 2010); see Section 2.6.2 for more details on the setting.

2.1.2 Existing strategies

For a standard multiple testing problem where the null distribution P_0 is known, the celebrated Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) controls the FDR in finite samples uniformly over all alternative distributions, when the test statistics are independent or satisfy the positive regression dependency on each one from a subset (PRDS) property; see Benjamini and Hochberg (1995); Benjamini and Yekutieli (2001). Variants of the BH procedure have been proposed to relax the conservatism when the fraction of true nulls is not close to 1, such as the Storey-BH (Storey et al., 2004) or Quantile-BH procedure (Benjamini et al., 2006; Sarkar, 2008; Blanchard and Roquain, 2009), and to robustify the FDR control under more general dependence structures (see Fithian and Lei, 2020 and references therein).

Despite this generality, BH-like methods have two major limitations in novelty detection problems with multivariate measurements:

(i) it is based on p-values or, more generally, univariate scores with a *known distribution under the null*, which is typically out of reach for such problems;

(ii) the score function that transforms the multivariate measurements into univariate test statistics (e.g., the p-value transformation) is *pre-specified*, while it should be *learned from data* for the sake of power.

We now discuss several existing solutions that partially circumvent these limitations. Table 2.1 provides a summary of the properties of each method, along with the corresponding applicable settings.

A popular solution in the multiple testing literature is the empirical Bayes approach, which operates on the local FDR instead of the p-values. Assuming a two-group mixture model (Efron et al., 2001), the local FDR is defined as the probability of being null conditional on the observed measurement values. The latter can be estimated by estimating the null and alternative densities together with the proportion of nulls; see Efron (2004, 2007, 2008, 2009). Combining local FDRs appropriately controls FDR asymptotically, under the assumptions that allow the model to be consistently estimated, and achieves optimal power, as shown in a series of paper by Sun and Cai (2007); Cai and Sun (2009); Sun and Cai (2009); Cai et al. (2019). We refer to this procedure as the SC procedure hereafter. Despite the appealing

optimality guarantees, the model assumptions tend to be fragile when the dimension d of the test statistics is moderately high. In such cases, accurate model estimation is hard to come by and the FDR of the SC procedure can thus be inflated; see our numerical experiments in Section 2.6 for an illustration.

Another line of research stems from conformal inference. While this technique is designed for prediction inference (see Angelopoulos and Bates (2021) for a recent review), it can also be employed in the novelty detection problem. In particular, it can generate *conformal p-values* that are super-uniform under the null without any model assumption beyond that the data are exchangeable (e.g., Vovk et al., 2005; Balasubramanian et al., 2014; Bates et al., 2023). This approach starts by transforming Z_j into a univariate score S_j, called the *non-conformity score*, that measures the conformity to the data and then computes an *empirical p-value*, also known as the conformal p-value, to evaluate the statistical evidence of being a novelty:

$$p_j = (n+1)^{-1}\Big(1 + \sum_{i=1}^{n} \mathbb{1}_{S_i \geq S_{n+j}}\Big). \tag{2.1}$$

Each p-value is marginally super-uniform under the null due to exchangeability and hence yields a valid test. Nonetheless, since the conformal p-values all use the same null sample, the above operation induces dependence between the p-values, making it unclear whether common multiple testing procedures are guaranteed to control FDR. Bates et al. (2023) carefully study the dependence structure and show that the split (or inductive) conformal p-values are PRDS. As a consequence, BH procedure applied on these conformal p-values controls the FDR. However, the approach limits the construction of the scores to be based solely on null examples and hence cannot learn the patterns of novelties in the mixed samples, unless extra labelled novelties are available (Liang et al., 2022), which are not always possible. Even when labelled novelties are present, they may behave differently than the ones in the mixed sample that we aim to detect. For this reason, Bates et al. (2023) apply the one class classification techniques (e.g. Schölkopf et al., 2001) that are not adaptive to the novelties. In sum, while the method successfully solves the issue (i), it falls short of adequately addressing issue (ii). On the other hand, while other versions of conformal p-values, like full conformal p-values (Vovk et al., 2005) and cross conformal p-values (Vovk, 2015; Barber et al., 2021), can use test samples and yield marginally valid p-values, they generally fail to satisfy the PRDS property, making it unclear whether the BH procedure would control FDR.

A subsequent work by Yang et al. (2021) proposes the Bag Of Null Statistics (BONuS) procedure for multiple testing problems with high dimensional test statistics, which largely motivates our method. The BONuS procedure learns a score function of the form $S_i = g(Z_i, (Z_1, \ldots, Z_n))$ and the method is valid as long as $g(Z_i, \cdot)$ is permutation invariant thereby allowing the transformation to be adapted to novelties. While the framework is flexible, they focus on the parametric setting where the null distribution is known, like Gaussian or multinomial, and the measurements are independent. In these cases, they propose using the estimated local FDR as the score function for which the alternative distribution and null proportion are learned by an empirical Bayes approach. The BONuS procedure controls the FDR in finite samples regardless of the quality of the estimates, even if the working model is completely wrong. However, for novelty detection problems, the local FDR involves unknown null and alternative densities, which are difficult to fit in high dimensions. Hence, point (ii) mentioned earlier remains partially addressed.

Lastly, we briefly review other related work that study different settings. The “counting knockoffs" procedure introduced by Weinstein et al. (2017) is designed for multiple testing for high-dimensional linear models with random design matrices. Mary and Roquain (2022) show that it is equivalent to applying the BH procedure to the scores $S_1, \ldots, S_{n+m}$ and closely

Method	Finite sample FDR control	Adaptative score	Learning alternative	Unknown null
Benjamini and Hochberg (1995)	yes	no	no	no
Sun and Cai (2007)	no	yes	yes	yes
Weinstein et al. (2017) Mary and Roquain (2022)	yes	no	no	yes
Bates et al. (2023)	yes	yes	no	yes
Yang et al. (2021)	yes	yes	yes	no
AdaDetect (our approach)	yes	yes	yes	yes

Table 2.1: Properties of different methods and the specific settings in which they can be applied for novelty detection.

related to the BONuS procedure. More recently, Rava et al. (2021) develop a method that is equivalent to applying the BH procedure on the conformal p-values to obtain a finite sample control of the false selection rate (FSR) for the task of (supervised) classification.

2.1.3 Contributions

In this work we introduce AdaDetect, an extension[1] of the BONuS procedure for novelty detection problems. In particular, we show how to leverage flexible off-the-shelf classification algorithms in machine learning to address both issues (i) and (ii) without compromising the FDR-controlling guarantees. In a nutshell, AdaDetect operates by initially splitting the null sample in two parts, $(Y_1, \dots, Y_k)$ and $(Y_{k+1}, \dots, Y_n)$, generating a membership label $A_j = -1$ if $Z_j \in \{Y_1, \dots, Y_k\}$ and $A_j = 1$ otherwise, and subsequently calculating a score function using a binary classifier trained on $(Z_i, A_i)_{i=1}^{n+m}$ and applying the BH procedure on the empirical p-values. For the example illustrated in Figure 2.1, Adadetect would split the null samples (digits '4') into two subsets and train a probabilistic classifier using both the null and test samples to distinguish the first subset of the null sample and the mix of the second subset of the null sample and the test sample (digits '4' and '9'). The predicted probability to be in the mixed sample is taken as the score. When the classification algorithm performs well, the scores tend to be larger for novelties than for nulls, because novelties are only present in the mixed sample. A comprehensive description of the procedure can be found in Section 2.2.4.

We summarize our main results below.

- In Section 2.3, we revisit the theoretical guarantees in Weinstein et al. (2017); Mary and Roquain (2022); Bates et al. (2023) and provide new FDR bounds based on an extension of the leave-one-out technique in the multiple testing literature. The bounds show that AdaDetect, as well as its π_0-adaptive variants Storey-AdaDetect and Quantile-AdaDetect, controls the FDR in finite samples with *arbitrary* classification algorithms even if the algorithm performs poorly. This is in sharp contrast to the SC procedure which heavily relies on correct model specification and consistent estimation.

- In addition, we extend the result in Bates et al. (2023) to show that the empirical p-values are PRDS under a more general exchangeability assumption, even if the score function depends on both null and test samples. For instance, our condition covers the Gaussian distributions with equi-correlation (Example 1). This PRDS property suggests

[1] More precisely, we extend the version of BONuS where the score function is fit only in the initial stage; see the discussion in Section 2.8 for more details.

that the resulting p-values can be applied in other contexts beyond the FDR control (e.g., Goeman and Solari, 2011).

- In Section 2.4, we show that *any* score function that is monotone in the ratio between the average density of novelties and the null density yields the optimal power. In particular, the optimal classifier to distinguish between the null and mixed samples is efficient despite that the null training is split and that the mixed sample is contaminated by nulls. The optimal score function can be obtained by minimizing certain loss function such as the cross-entropy loss that is commonly used in neural networks (NN hereafter).

- We provide non-asymptotic power analyses for AdaDetect in Section 2.5. First, we investigate AdaDetect with the score function given by a constrained empirical risk minimizer (ERM) of the 0-1 loss and show it approaches the optimal likelihood ratio test in an appropriate sense. Next, we provide an oracle inequality for general score functions and conditions under which the procedure mimics its oracle version. We apply the results to analyze power for AdaDetect procedures based on NN and on non-parametric kernel density estimation.

- We demonstrate the efficiency, flexibility, and robustness of AdaDetect[2] in Sections 2.6 and 2.7 on synthetic, semi-synthetic, and real datasets, including the MNIST image dataset and an astronomy dataset from the 'Sloan Digital Sky Survey'.

2.2 Preliminaries

2.2.1 Notation

As in Section 2.1.1, we let $Y = (Y_1, \dots, Y_n)$ denote the null training sample (NTS) with a common marginal distribution P_0, $X = (X_1, \dots, X_m)$ the test sample with $X_i \sim P_i$ $(1 \leq i \leq m)$, $Z = (Z_1, \dots, Z_{n+m}) = (Y_1, \dots, Y_n, X_1, \dots, X_m)$ the full sample, $\mathcal{H}_0 = \{1 \leq i \leq m : P_i = P_0\}$ the set of nulls in the test sample with $m_0 = |\mathcal{H}_0|, \pi_0 = m_0/m$, and $\mathcal{H}_1 = \{1, \dots, m\} \setminus \mathcal{H}_0$ the set of novelties with $m_1 = |\mathcal{H}_1|, \pi_1 = m_1/m$. For notational convenience, we write $n + \mathcal{H}_0$ for the set $\{n + i, i \in \mathcal{H}_0\}$. Furthermore, we denote by P the joint distribution of Z, which belongs to a family of distributions $\mathcal{P}$ (model).

Throughout the paper, we consider the semi-supervised setting (Mary and Roquain, 2022) where the null distribution P_0 is unknown and one can access it only through the measurements in the NTS. In practice, the NTS can be obtained from external data, past experiments or black-box samplers.

2.2.2 Criteria

A novelty detection procedure is a measurable function $R(\cdot)$ that takes Z as input and returns a subset of $\{1, \dots, m\}$ corresponding to the indices of detected novelties within $\{X_1, \dots, X_m\}$. Throughout the paper, we will slightly abuse the notation by using R to refer to both the procedure and the rejection set given by the procedure. Ideally, we want $R(Z)$ to capture novelties (i.e., alternative hypotheses in $\mathcal{H}_1$) and avoid inliers (i.e., null hypotheses in $\mathcal{H}_0$). Given a procedure R, the false discovery rate (FDR) is defined as the expectation of the false

[2]The code is publicly available at `https://github.com/arianemarandon/adadetect`https://github.com/arianemarandon/adadetect

discovery proportion (FDP) with respect to the distribution $P \in \mathcal{P}$:

$$\mathrm{FDR}(P,R) = \mathbb{E}_{Z\sim P}[\mathrm{FDP}(P,R)], \quad \mathrm{FDP}(P,R) = \frac{\sum_{i\in\mathcal{H}_0} \mathbb{1}_{i\in R}}{1 \vee |R|}. \tag{2.2}$$

Similarly, the true discovery rate (TDR) is defined as the expectation of the true discovery proportion (TDP):

$$\mathrm{TDR}(P,R) = \mathbb{E}_{Z\sim P}[\mathrm{TDP}(P,R)], \quad \mathrm{TDP}(P,R) = \frac{\sum_{i\in\mathcal{H}_1} \mathbb{1}_{i\in R}}{1 \vee m_1(P)}. \tag{2.3}$$

Note that $m_1(P) = 0$ implies $\mathrm{TDP}(P,R) = 0$. Our goal is to build a procedure R that controls the FDR and maximizes the TDR to the fullest extent.

2.2.3 BH algorithm and its π_0-adaptive variants

Suppose a set of p-values $(p_i, 1 \le i \le m)$ is available, the BH algorithm (Benjamini and Hochberg, 1995) returns

$$R = \{i \in \{1,\dots,m\} \,:\, p_i \le \alpha\hat{k}/m\},$$

where α is the target FDR level and

$$\hat{k} = \max\left\{k \in \{0,\dots,m\} \,:\, \sum_{i=1}^{m} \mathbb{1}_{p_i\le\alpha k/m} \ge k\right\}. \tag{2.4}$$

When the null p-values $(p_i, i \in \mathcal{H}_0)$ are independent, super-uniform, and independent of alternative p-values $(p_i, i \in \mathcal{H}_1)$, the BH procedure is proved to control the FDR at level $\alpha\pi_0$ in finite samples (Benjamini and Hochberg, 1995). The independence assumption can be further relaxed to the PRDS condition (Benjamini and Yekutieli, 2001).

When π_0 is not close to 1, the BH procedure is conservative because $\alpha\pi_0 < \pi_0$. When π_0 is known, it can be applied at level α/π_0 to close the gap. In practice, π_0 is usually unknown though. Nonetheless, there exists estimators $\hat{\pi}_0$ of π_0 such that the BH procedure with level $\alpha/\hat{\pi}_0$ continues to control the FDR under independence. Two celebrated estimators are introduced by Storey et al. (2004) and Benjamini et al. (2006):

$$\widehat{\pi}_0^{Storey} = \frac{1 + \sum_{i=1}^{m} \mathbb{1}_{p_i\ge\lambda}}{m(1-\lambda)}, \quad \lambda > 0; \tag{2.5}$$

$$\text{or} \quad \widehat{\pi}_0^{Quant} = \frac{m - k_0 + 1}{m(1 - p_{(k_0)})}, \quad k_0 \in \{1,\dots,m\}, \tag{2.6}$$

where $p_{(k_0)}$ is the k_0-th smallest[3] p-value. These procedures are often called the π_0-adaptive versions of the BH algorithm.

2.2.4 Our method

In this paper, we propose a method called AdaDetect. It is an adaptive novelty detection procedure that extends the existing strategies described in Weinstein et al. (2017), Yang et al. (2021), Mary and Roquain (2022), and Bates et al. (2023). It starts by splitting the null

[3] In this paper, a convention is to order the p-values from the smallest to the largest, while the test statistics are ordered from the largest to the smallest.

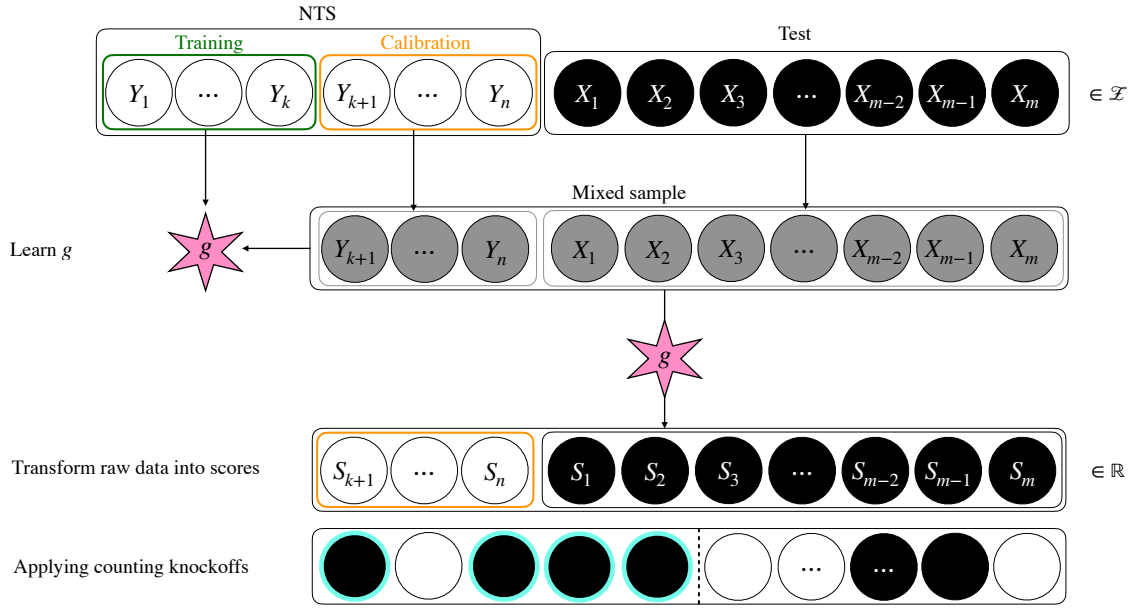


Figure 2.2: A schematic illustration of AdaDetect: •/∘ stands for a test/null observation, respectively. The vertical dashed line corresponds to the largest threshold t for which $\widehat{\mathrm{FDP}}(t) \leq \alpha$ and the • circled in blue correspond to the discoveries of AdaDetect procedure.

sample $(Y_1, \ldots, Y_n)$ in two samples $(Y_1, \ldots, Y_k)$ and $(Y_{k+1}, \ldots, Y_n)$ with $k \geq 0$. To avoid cluttering our notation, we define ℓ as the size of the second null sample, i.e.,

$$\ell = n - k.$$

It proceeds with the following steps.

1. Compute a data-driven score function of form

$$g(z) = g(z, (Z_1, \ldots, Z_k), (Z_{k+1}, \ldots, Z_{n+m})), \quad z \in \mathcal{Z}, \tag{2.7}$$

which satisfies the following invariance property: for any permutation π of $\{k+1, \ldots, n+m\}$ and $z, z_1, \ldots, z_{n+m} \in \mathcal{Z}$, we have

$$g(z, (z_1, \ldots, z_k), (z_{\pi(k+1)}, \ldots, z_{\pi(n+m)})) = g(z, (z_1, \ldots, z_k), (z_{k+1}, \ldots, z_{n+m})). \tag{2.8}$$

2. Transform the raw data into univariate scores

$$S_i = g(Z_i; (Z_1, \ldots, Z_k), (Z_{k+1}, \ldots, Z_{n+m})), \quad i \in \{k+1, \ldots, n+m\}. \tag{2.9}$$

Here, we assume that novelties typically have large scores.

3. For each test point X_j, generate the empirical p-value by comparing S_i with the scores in the NTS:

$$p_j = \frac{1}{\ell + 1} \left(1 + \sum_{i=k+1}^{n} \mathbb{1}_{S_i > S_{n+j}} \right), \quad j \in \{1, \ldots, m\}. \tag{2.10}$$

4. Apply the BH algorithm to $(p_1, \ldots, p_m)$ at the target level α.

We will call this procedure AdaDetect$_\alpha$ in the sequel to emphasize the target level. By simple algebra, the last two steps together are equivalent to the "counting knockoff" algorithm proposed by Weinstein et al. (2017) applied to the scores $S_{k+1}, \ldots, S_{n+m}$. Specifically, the method declares i as a novelty if $S_i \geq \hat{t}$ where

$$\hat{t} = \min\left\{t \in \{S_i : k+1 \leq i \leq n+m\} \ : \ \widehat{\mathrm{FDP}}(t) \leq \alpha\right\};$$
$$\widehat{\mathrm{FDP}}(t) = \frac{m}{\ell+1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}_{S_i \geq t}\right) / \sum_{i=n+1}^{n+m} \mathbb{1}_{S_i \geq t}\,.$$

Therefore, the counting knockoff procedure can be seen as a shortcut that avoids computing the empirical p-values explicitly. The pipeline for AdaDetect is illustrated in Figure 2.2.

AdaDetect offers greater flexibility than existing methods in the types of score functions that can be employed.

- Prespecified p-value transformations are score functions that do not depend on $(Z_1, \ldots, Z_k)$ and $(Z_{k+1}, \ldots, Z_{n+m})$. For example, when $\mathcal{Z} = \mathbb{R}^d$, the χ^2 test chooses the non-adaptive score $g(z) = \sum_{j=1}^{d} z_j^2$, $z \in \mathbb{R}^d$.
- The one-class classification approach considered by Bates et al. (2023) corresponds to score functions that only depend on $(Z_1, \ldots, Z_k)$, but not $(Z_{k+1}, \ldots, Z_{n+m})$.
- The BONuS procedure (Yang et al., 2021) considers empirical Bayes-based score functions that depend on the pooled sample $\{Z_1, \ldots, Z_{n+m}\}$ without distinguishing between the null and mixed samples.
- Our proposed method constructs the score function $g(\cdot, (Z_1, \ldots, Z_k), (Z_{k+1}, \ldots, Z_{n+m}))$ as the estimated probability by any probabilistic classifier that distinguishes between $(Z_1, \ldots, Z_k)$ and $(Z_{k+1}, \ldots, Z_{n+m})$; see Section 2.4 for details.

Lastly, we propose the Storey-AdaDetect and Quantile-AdaDetect as the π_0-adaptive versions of AdaDetect applied at level $\alpha/\widehat{\pi}_0^{Storey}$ and $\alpha/\widehat{\pi}_0^{Quant}$, respectively, in which the p-values have been replaced by the empirical ones.

Remark 4. An appealing property of Adadetect and its adaptive versions is that the rejection is invariant to strictly increasing transformations of score function. This feature proves useful in the power analysis of AdaDetect, see Section 2.4.

Remark 5. By construction, the empirical p-values are multiples of $1/(\ell+1)$. As Mary and Roquain (2022) point out, the number of null samples ℓ needs to be larger than $m/(\alpha(1 \vee M))$ in order to guarantee sufficient resolution of the p-values for the BH procedure, where $M \geq 0$ is some high-probability lower bound on the number of rejections. Typically, if M is of the order of m, a constant ℓ would suffice, while if $M = 0$ (i.e., without any prior knowledge on the number of rejections), ℓ should be larger than m/α. In general practical situations where $n \gtrsim m$, we recommend setting $\ell = m$ and this choice works reasonably well in our numerical experiments. When $m > n$, it might be more appropriate to impose further assumptions on the distribution (e.g., the knowledge of M).

2.3 FDR control

In this section, we prove that AdaDetect and its π_0-adaptive variants control the FDR. In Section 2.3.1, we state the key assumption of exchangeability and show it translates to the scores

as long as g satisfies the condition (2.8). Based on this observation, we prove in Section 2.3.2 that the empirical p-values are PRDS, which is a highly non-trivial extension of the results by Bates et al. (2023). Though the PRDS property implies the FDR control of AdaDetect as a result of Benjamini and Yekutieli (2001), we present in Section 2.3.3 an alternative proof based on a new FDR expression that unify and extend the previous FDR bounds. Lastly, in Section 2.3.4, we prove the FDR control for Storey-AdaDetect and Quantile-AdaDetect based on an FDR bound for general π_0-adaptive versions of AdaDetect.

2.3.1 Exchangeability

We make the following assumption on the raw measurements throughout the paper.

Assumption 1. *$(Y_1, \ldots, Y_n, X_i, i \in \mathcal{H}_0)$ are exchangeable conditional on $(X_i, i \in \mathcal{H}_1)$.* [4]

Clearly, Assumption 1 holds when the measurements are independent, as assumed by Yang et al. (2021) and Bates et al. (2023). In general, Assumption 1 allows for dependencies among the measurements.

Example 1. Consider the observation where $Z_i = \mu_i + \rho^{1/2}\xi + (1-\rho)^{1/2}\varepsilon_i$, $1 \le i \le n+m$, with the variables $\xi, \varepsilon_1, \ldots, \varepsilon_{n+m}$ being i.i.d. $\sim \mathcal{N}(0, I_d)$, ρ being a nonnegative correlation coefficient, and $\mu_i = 0$ for $i \in \{1, \ldots, n\} \cup (n + \mathcal{H}_0)$ (hence $\mathcal{Z} = \mathbb{R}^d$). Then Assumption 1 holds. The case $d = 1$ corresponds to the Gaussian equi-correlated case, which is widely studied in the multiple testing literature (e.g., Korn et al., 2004).

For our results, a necessary assumption is exchangeability of the scores under the null:

Assumption 2. *$(S_{k+1}, \ldots, S_n, S_{n+i}, i \in \mathcal{H}_0)$ is exchangeable conditionally on $(S_{n+i}, i \in \mathcal{H}_1)$.*

It turns out the exchangeability of the raw measurements translates to the scores.

Lemma 2. *Under Assumption 1, the adaptive scores defined by* (2.9) *satisfy Assumption 2 for any score function that satisfies the condition* (2.8).

This result substantially simplifies the FDR analysis presented in the next section. To avoid unnecessary mathematical complications, we make the following mild assumption.

Assumption 3. *$(S_{k+1}, \ldots, S_{n+m})$ have no ties almost surely.*

2.3.2 The p-values are PRDS

Following Benjamini and Yekutieli (2001), we say a family of p-values $(p_i, 1 \le i \le m)$ is PRDS on $\mathcal{H}_0$ if, for any $i \in \mathcal{H}_0$ and nondecreasing[5] measurable set $D \subset [0,1]^m$, the function $u \in [0,1] \mapsto \mathbb{P}((p_j, 1 \le j \le m) \in D \mid p_i = u)$ is nondecreasing.

Theorem 3. *For any family of scores $(S_{k+1}, \ldots, S_{n+m})$ satisfying Assumptions 2 and 3, the empirical p-values defined in* (2.10) *are PRDS on $\mathcal{H}_0$ and the null p-values are super-uniform. In particular, under Assumptions 1 and 3, this result holds for the p-values generated by AdaDetect with a score function satisfying* (2.8).

[4] Note that such an assumption implicitly assumes that such a conditional distribution exists, which is always the case for instance when $\mathcal{Z} = \mathbb{R}^d$ or $\mathcal{Z}$ is discrete.

[5] A set $D \subset [0,1]^m$ is said to be nondecreasing if for any $x \in D$ and $y \in [0,1]^m$, we have $y \in D$ provided that $y_i \ge x_i$ for all i.

We present a proof of Theorem 3 in Section A.1.3. It extends Theorem 2 in Bates et al. (2023) to dependent scores. Notably, the AdaDetect scores are dependent in general even if the measurements Z_i's are independent because the data-adaptive score function depends on the entire dataset.

Theorem 3 has interesting consequences. First, the celebrated result for the BH procedure (Benjamini and Yekutieli, 2001; Romano and Wolf, 2005) implies that AdaDetect strongly controls the FDR at level $\alpha\pi_0$. Second, the PRDS property is also useful for other purposes, such as post hoc inference (Goeman and Solari, 2011), FDR control with structural constraints (Ramdas et al., 2019a; Loper et al., 2019), online FDR control (Zrnic et al., 2021; Fisher, 2021), hierarchical FDR control (Foygel Barber and Ramdas, 2015) and weighted FDR control with prior knowledge (Ramdas et al., 2019b). Hence, our result paves the way for developing similar AdaDetect-style procedures in these contexts.

2.3.3 A new FDR expression

While the PRDS property implies the FDR control for AdaDetect, we pursue an alternative way based on a new expression for the FDR of the BH procedure in our setting, which would also yield a lower bound for FDR that is not implied by the PRDS property.

Theorem 4. *Consider any family of scores $(S_{k+1}, \dots, S_{n+m})$ satisfying Assumptions 2 and 3. Let R_α denote the rejection set of BH procedure applied to p-values defined in* (2.10) *at level α. Then, for any distribution $P \in \mathcal{P}$,*

$$\mathrm{FDR}(P, R_\alpha) = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left(\frac{\lfloor \alpha(\ell+1)K_i/m \rfloor}{(\ell+1)K_i} \right), \tag{2.11}$$

where K_i is a random variable that takes values in $\{1, \dots, m\}$ for any $i \in \mathcal{H}_0$. In particular, under Assumptions 1 and 3, (2.11) *holds with $R_\alpha = AdaDetect_\alpha$, the AdaDetect procedure at level α.*

The proof of Theorem 4 is presented in Section A.1.4. It is similar to the classical leave-one-out technique to prove the FDR control for step-up procedure (e.g. Ferreira and Zwinderman, 2006; Roquain and Villers, 2011; Ramdas et al., 2019b; Giraud, 2022), though it is non-trivial to handle empirical p-values. Since for any $x > 0$ and integer k, we have $\lfloor x \rfloor k \le \lfloor xk \rfloor \le xk$, expression (2.11) immediately implies the following bounds.

Corollary 5. *Under Assumptions 1 and 3, the following holds, for any values of $k, \ell, m \ge 1$ and any parameter $P \in \mathcal{P}$:*

$$m_0 \lfloor \alpha(\ell+1)/m \rfloor / (\ell+1) \le \mathrm{FDR}(P, AdaDetect_\alpha) \le \alpha m_0/m. \tag{2.12}$$

In particular, $\mathrm{FDR}(P, AdaDetect_\alpha) = \alpha\pi_0$ *when $\alpha(\ell+1)/m$ is an integer.*

Corollary 5 recovers Theorem 3.1 in Mary and Roquain (2022) which imposes a slightly more restrictive condition than Assumption 2. Their proofs are based on martingale techniques and the proof for the lower bound is particularly involved. Here, we rely instead on the exact expression (2.11), which is arguably simpler and more comprehensible.

2.3.4 New FDR bounds for π_0-adaptive procedures

For each $i \in \mathcal{H}_0$, let $\mathcal{D}_i$ be the distribution of $(p'_j, 1 \le j \le m)$, where

$$\begin{cases} p'_j = 0,\ j \in \mathcal{H}_1,\ p'_i = 1/(\ell+1); \\ p'_j,\ j \in \mathcal{H}_0 \backslash \{i\} \text{ are i.i.d. conditionally on } U \text{ with a common c.d.f. } F^U; \\ U = (U_1, \dots, U_{\ell+1}) \text{ has i.i.d. } U(0,1) \text{ components,} \end{cases} \tag{2.13}$$

where F^U denotes the discrete c.d.f. $F^U(x) = (1 - U_{(\lfloor x(\ell+1) \rfloor + 1)})\, \mathbb{1}_{1/(\ell+1) \leq x < 1} + \mathbb{1}_{x \geq 1}$, $x \in \mathbb{R}$, and $U_{(1)} > \cdots > U_{(\ell+1)}$ denote the order statistics of the vector U. Note that the distribution $\mathcal{D}_i$ only depends on i, m, ℓ and $\mathcal{H}_0$. The following general result holds.

Theorem 6. *In the setting of Theorem 4, denote $p = (p_i, 1 \leq i \leq m)$ the family of empirical p-values defined in (2.10) and consider any function $G : [0,1]^m \to (0, \infty)$ that is coordinate-wise nondecreasing. Then the procedure, denoted by $R_{\alpha m / G(p)}$, combining the BH algorithm at level $\alpha m / G(p)$ with these empirical p-values is such that, for any parameter $P \in \mathcal{P}$,*

$$\mathrm{FDR}(P, R_{\alpha m / G(p)}) \leq \alpha \sum_{i \in \mathcal{H}_0} \mathbb{E}_{p' \sim \mathcal{D}_i} \left(\frac{1}{G(p')} \right), \tag{2.14}$$

where $\mathcal{D}_i$ is defined by (2.13). In particular, this FDR expression holds for $R_{\alpha m / G(p)} = \mathit{AdaDetect}_{\alpha m / G(p)}$ under Assumptions 1 and 3.

Theorem 6 is proved in Section A.1.5. In a nutshell, the distribution $\mathcal{D}_i$ is a least favorable distribution for the FDR of the adaptive BH procedure applied to empirical p-values defined in (2.10). It can be seen as an adaptation of the classical leave-one-out technique for adaptive BH procedures; see Benjamini et al. (2006) and Theorem 11 of Blanchard and Roquain (2009).

This result generalizes Theorem 6 of Bates et al. (2023) which only works for the Storey-BH procedure. Our proof technique is fundamentally different and works for a broad class of estimators of π_0. Applying Theorem 6 to the estimators defined in (2.5) and (2.6), we obtain the following result.

Corollary 7. *Under Assumptions 1 and 3, the following holds:*

- *Storey-AdaDetect controls the FDR at level α for any $\lambda = K/(\ell+1)$ and $K \in \{2, \ldots, \ell\}$.*
- *Quantile-AdaDetect controls the FDR at level α for any $k_0 \in \{1, \ldots, m\}$.*

The proof of Corollary 7 is presented in Section A.1.6. It bounds the RHS of (2.14) via combinatoric arguments. The result for Quantile-AdaDetect is novel. The result for Storey-AdaDetect was proved in Yang et al. (2021) for BONuS, with a different proof technique, in the special case where the scores are independent. Hence, we extend it to the exchangeable case.

To illustrate the robustness of π_0-adaptive AdaDetect under dependence, consider Example 1 with common alternative means $\mu_i \equiv \mu \in \mathbb{R}^d$ and a fixed score function $S_i = \mu^T Z_i$, $1 \leq i \leq n+m$. One alternative approach to Storey-AdaDetect is to apply the Storey-BH procedure on the marginal p-values $p_i = \bar{\Phi}(S_{n+i}/\|\mu\|)$, $1 \leq i \leq m$. Interestingly, Figure 2.3 shows that the Storey-BH procedure inflates the FDR substantially in the presence of high correlation while Storey-AdaDetect with $k = 0$ controls the FDR for any correlation ρ (as implied by Corollary 7). Hence, while Storey-AdaDetect is only based on an NTS without the knowledge of the true null distribution, it is more robust to dependence than Storey-BH that requires more information. Furthermore, Storey-AdaDetect is more powerful than Storey-BH because the effect of the common variable ξ is cancelled out in the calculation of empirical p-values.

Remark 6. Assumptions 2 and 3 hold true in other contexts. For example, this is the case for LASSO-based scores in the Gaussian linear model where the design matrix has i.i.d. entries with a known distribution (Weinstein et al., 2017). Hence, the FDR bounds we developed also hold in those cases.

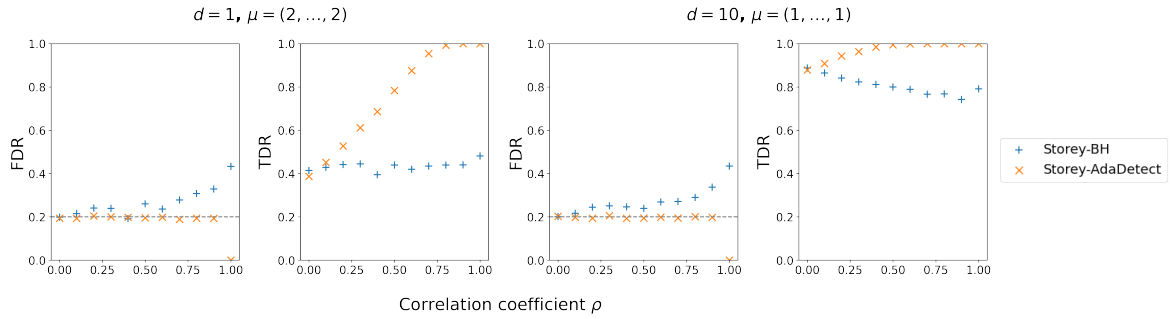


Figure 2.3: FDR and TDR for Storey-BH and Storey-AdaDetect (both with oracle test statistics/scores) in Example 1 with varying correlation $\rho \in [0, 1]$. The dimension $d = 1$ in the two left panels and $d = 10$ in the two right panels. In all settings, $m = 100$, $n = \ell = 1000$, $\alpha = 0.2$, $\pi_0 = 0.9$, and $\lambda = 500/1001$.

2.4 Constructing score functions

While any score function satisfying (2.8) can be used in AdaDetect, we discuss principles and various techniques to construct score functions that yield high power. Section 2.4.1 introduces the assumptions and notation. In Section 2.4.2 we show that the optimal score function is given by any monotone function of the ratio between the average density of novelties and the average density of all points. We proceed by discussing two methods to approach the optimal score based on direct density estimation (Section 2.4.3) and classification (Section 2.4.4). The latter is more scalable and flexible in the sense that it is able to wrap around any probabilistic classification algorithms. In Section 2.4.5, we discuss a cross-validation approach for hyperparameter tuning and model selection without compromising the finite-sample FDR control.

2.4.1 Assumptions and notation

In this section, we make the following two assumptions:

Assumption 4. *$Y_1, \ldots, Y_n, X_1, \ldots, X_m$ are mutually independent.*

Given the setting of Section 2.1.1, we thus have under Assumption 4 that $(Y_1, \ldots, Y_n, X_i, i \in \mathcal{H}_0)$ are i.i.d. $\sim P_0$ and independent of $(X_i, i \in \mathcal{H}_1)$ which are mutually independent.

Assumption 5. *For each $i \in \{0\} \cup \mathcal{H}_1$, P_i has a positive density f_i w.r.t. a measure ν.*

Let

$$f = \pi_0 f_0 + \pi_1 \bar{f}_1, \tag{2.15}$$

$$\bar{f}_1 = m_1^{-1} \sum_{i \in \mathcal{H}_1} f_i. \tag{2.16}$$

Under Assumptions 4 and 5, f_0 is the average density of $(Z_1, \ldots, Z_k)$, $\bar{f}_1$ is the average alternative density, f is the average density of the test sample $(X_1, \ldots, X_m)$. Similarly the average density of $(Z_{k+1}, \ldots, Z_{n+m})$ is f_γ where

$$\gamma = \frac{m_1}{\ell + m}; \tag{2.17}$$

$$f_\gamma = (1 - \gamma) f_0 + \gamma \bar{f}_1 = \frac{\ell}{\ell + m} f_0 + \frac{m}{\ell + m} f, \tag{2.18}$$

Compared to f, the mixture f_γ is contaminated by more nulls, that is, $\pi_0 \leq 1-\gamma = \frac{\ell+m_0}{\ell+m}$. Lastly, we define the density ratio

$$r(x) = \frac{\pi_1 \bar{f}_1(x)}{f(x)}, \quad x \in \mathcal{Z}. \tag{2.19}$$

Note that $r(x) \in (0,1)$ for ν-almost every $x \in \mathcal{Z}$ by Assumption 5.

2.4.2 Optimal score function

Recall that AdaDetect is equivalent to applying the counting knockoff on the scores which relies on an estimator $\widehat{\mathrm{FDP}}$ (Section 2.2.4). For each given t, when ℓ and m is large,

$$\widehat{\mathrm{FDP}}(t) \approx \frac{m\,\mathbb{P}_{S_i \sim P_0}(S_i \geq t)}{\mathbb{E}[|R(t)|]} \approx \frac{\mathbb{E}[|R(t) \cap \{k+1,\ldots,n\}|]}{\mathbb{E}[|R(t)|]},$$

where $R(t)$ is set of rejections at threshold t. The RHS is called the marginal FDR (mFDR), an error metric that is close to FDR when $|R(t)|$ is large and often used for asymptotic power analysis of FDR-controlling procedures (e.g. Sun and Cai, 2007; Lei and Fithian, 2018). The following theorem derives the optimal score function among all procedures that reject hypotheses with S_i above some thresholds subject to mFDR control (see Weinstein, 2021; Rosset et al., 2022 for results for FDR instead of mFDR).

Theorem 8. *Assume Assumptions 4 and 5 hold. The likelihood ratio function $r(\cdot)$ defined in (2.19) is an optimal score function in the sense that the rejection set $R = \{i \in \{1,\ldots,m\} : r(X_i) \geq c(\alpha)\}$, where $c(\alpha) \in (0,1)$ is chosen such that $mFDR(R) = \alpha$ (assuming it exists), has a higher TPR than any rejection set $R' = \{i \in \{1,\ldots,m\} : r'(X_i) \geq c'\}$ where $c' \in \mathbb{R}$ and $r' : \mathcal{Z} \mapsto \mathbb{R}$ is measurable with mFDR at most α.*

The proof can be found in Section A.2.1. Theorem 8 suggests the following oracle procedure.

Definition 3. *The oracle AdaDetect procedure, denoted by AdaDetect*, is defined as the AdaDetect procedure with the score function $r(\cdot)$ defined in (2.19).*

Since AdaDetect is invariant under any strictly monotone transformation of the score function (see Remark 4), AdaDetect* can be realized as any AdaDetect procedure with a score function of the form

$$g^* = \Psi \circ r, \text{ for some increasing continuous } \Psi : (0,1) \to \mathbb{R}, \tag{2.20}$$

where Ψ could depend on unknown parameters. This is a crucial property of AdaDetect that enables flexible classification methods to construct score functions without concerning about the composition of nulls and novelties that may change the oracle score r.

Since r (or g^*) is unknown, the oracle procedure AdaDetect* is not directly accessible in practice. Our goal is to learn a g^* in the form of (2.7) that satisfies the constraint (2.8).

2.4.3 Density estimation

A first example of score function is built from density estimation. From (2.15) and (2.18), the following score

$$g^*(x) = f_\gamma(x)/f_0(x) = 1 - \gamma/\pi_1 + (\pi_0\gamma/\pi_1)(1-r(x))^{-1} \tag{2.21}$$

is indeed of the form (2.20). A straightforward approach is to directly estimate the densities as follows.

- Estimate f_0 by a density estimator $\widehat{f_0}$ based on the sample $(Z_1, \ldots, Z_k)$
- Estimate f_γ by a density estimator $\widehat{f_\gamma}$ based on the mixed sample $(Z_{k+1}, \ldots, Z_{n+m})$ via a mixture estimation approach.
- Estimate $g^*(x)$ by $\widehat{g}(x) = \widehat{f_\gamma}(x)/\widehat{f_0}(x)$ assuming that $\widehat{f_0}(Z_i) > 0$.

Above, the density estimators can be either parametric or non-parametric. Both versions will be considered in the sequel (see Section 2.4.3 and the numerical experiments in Section 2.6). Note that Yang et al. (2021) applies this approach when f_0 is known.

2.4.4 PU classification

While density estimation is straightforward, it is not scalable when the dimension d is large; see the numerical experiments in Section 2.6 for an illustration. In this section, we consider a different strategy that estimate density ratios through probabilistic classification (e.g. Friedman, 2003; Sugiyama et al., 2012; Lei et al., 2021; Wang et al., 2022).

Define $(Z_1, \ldots, Z_k)$ as the "positive sample" and the sample $(Z_{k+1}, \ldots, Z_{n+m})$ as the "unlabeled sample", and let $(A_1, \ldots, A_k) = (-1, \ldots, -1)$ and $(A_{k+1}, \ldots, A_{n+m}) = (1, \ldots, 1)$ the corresponding labels. In this context, the classification task is typically referred to as the PU (positive unlabeled) classification, which is an active research area; see Du Plessis et al. (2014); Calvo et al. (2007); Guo et al. (2020); Ivanov (2020) among others and Bekker and Davis (2020) for a recent review. Here, we are considering a slightly different setting where the unlabeled samples are independent but not identically distributed.

Usually, the classifier is learned by empirical risk minimization (ERM) where the objective function is in the form of

$$\widehat{J}_\lambda(g) = \sum_{i=1}^{n+m} \lambda_{A_i} \ell(A_i, g(Z_i)) = \sum_{i=1}^{k} \ell(-1, g(Z_i)) + \lambda \sum_{i=k+1}^{n+m} \ell(1, g(Z_i)),$$

where $\ell : \{-1, +1\} \times \mathbb{R} \to \mathbb{R}_+$ is a loss function and $\lambda_a = \lambda \mathbb{1}_{a \geq 0} + \mathbb{1}_{a \leq 0}$ with $\lambda > 0$ measuring the relative cost misclassifying a positive sample to misclassifying an unlabeled sample. Here, g is a function that belongs to $\mathcal{G}$, a class of measurable functions from $\mathcal{Z}$ to $\mathbb{R}$ and the classifier corresponds to the sign of g. Typical choices of the loss function include the hinge loss $\ell(a, u) = 0.5(1 - au)_+$ and the cross entropy loss $\ell(a, u) = -\log(1 - u)\, \mathbb{1}_{a=-1} - \log(u)\, \mathbb{1}_{a=+1}$. The population objective function is given by

$$J_\lambda(g) = \mathbb{E}\, \widehat{J}_\lambda(g) = k\, \mathbb{E}_{Z \sim f_0}\, \ell(-1, g(Z)) + \lambda(\ell + m)\, \mathbb{E}_{Z \sim f_\gamma}\, \ell(1, g(Z)), \tag{2.22}$$

where f_γ is defined in (2.18). The following result shows that the minimizer of (2.22) over all measurable functions yields an optimal score in the form of (2.20) when the loss function ℓ is appropriately chosen.

Lemma 9. *Let $g^\sharp$ denote the minimizer of (2.22) over all measurable functions.*

(i) When $\ell(\cdot, \cdot)$ is the hinge loss, assuming that the set $\{x \in \mathcal{Z} : f_\gamma(x) = c f_0(x)\}$ is of ν-measure zero for any $c > 0$, where ν is defined in Assumption 5,

$$g^\sharp(x) = \operatorname{sign}\left(\frac{\lambda(\ell + m)}{k} \frac{f_\gamma(x)}{f_0(x)} - 1\right) = \operatorname{sign}\left(\frac{\lambda \ell}{k} + \frac{\lambda m_0}{k}(1 - r(x))^{-1} - 1\right),$$

and the minimum is unique ν-almost everywhere.

(ii) When $\ell(\cdot,\cdot)$ is the cross entropy,

$$g^\sharp(x) = \frac{\lambda(\ell+m)f_\gamma(x)}{\lambda(\ell+m)f_\gamma(x)+kf_0(x)} = \left(1+\left\{\frac{\lambda\ell}{k}+\frac{\lambda m_0}{k}(1-r(x))^{-1}\right\}^{-1}\right)^{-1},$$

and the minimum is unique ν-almost everywhere.

The proof is presented in Section A.2.2. Clearly, $g^\sharp$ is an optimal score function in the form of (2.20) with the cross-entropy loss but not so with the hinge loss because the sign function is not strictly monotone. For cross-entropy loss, when $\lambda = 1$,

$$g^\sharp(x) = \frac{\frac{\ell+m}{n+m}f_\gamma(x)}{\frac{\ell+m}{n+m}f_\gamma(x)+\frac{k}{n+m}f_0(x)}, \tag{2.23}$$

which can be roughly interpreted as the posterior probability to be in class 1.

In practice, it is computationally infeasible and statistically inefficient to optimize over all measurable functions. Instead, we often choose a function class $\mathcal{G}$ and estimate the score function by

$$\widehat{g} \in \operatorname{argmin}_{g\in\mathcal{G}} \widehat{J}_\lambda(g). \tag{2.24}$$

By construction, $\widehat{J}_\lambda(g)$ is invariant to permutations of $(Z_{k+1},\ldots,Z_{n+m})$, $\widehat{g}$ always satisfies the condition (2.8). When $\mathcal{G}$ has low complexity, we should expect $\widehat{g} \approx g^\sharp_{\mathcal{G}}$ where

$$g^\sharp_{\mathcal{G}} \in \operatorname{argmin}_{g\in\mathcal{G}} J_\lambda(g). \tag{2.25}$$

On the other hand, when $\mathcal{G}$ is sufficiently rich, we can expect $g^\sharp_{\mathcal{G}} \approx g^\sharp$. In summary, when the function class $\mathcal{G}$ and the loss function $\ell(\cdot,\cdot)$ are chosen appropriately, $\hat{g} \approx g^\sharp_{\mathcal{G}} \approx g^\sharp$, which is an optimal score function.

We illustrate the roles of function classes and loss functions in a simple setting where the positive class and the unlabeled class are generated from two gaussian distributions with dimension 1 or 2. The results are presented in Figure 2.4, with each row corresponding to a data-generating process. For all settings, the first panel displays the null and alternative distributions and the second panel displays the distributions of the positive and unlabeled classes. In all settings, we plot $\hat{g}$ and $g^\sharp$ for hinged loss (SVM) and cross-entropy losses with two function classes, an inaccurate one (Logistic Regression) and an accurate one (Neural Networks). In the two-dimensional settings, we display the functions by contour plots. For instance, for the cross entropy loss, we can observe the NN function class outperforms the logistic function class for approximating $g^\sharp$.

In conclusion, both the loss function $\ell(\cdot,\cdot)$ and the function class $\mathcal{G}$ are pivotal. Among the two loss functions we discuss, the cross entropy loss with a sufficiently rich function class (e.g., fully-connected neural networks) is particularly suitable for AdaDetect in that it is computationally feasible and approximately optimal. In contrast to the classification literature, hinged loss is undesirable for our purpose since the estimator does not converge to an optimal score.

2.4.5 AdaDetect with cross-validation

In previous sections, we focus on a single score function. Nevertheless, most density estimation and classification algorithms involve hyperparameters that require data-driven tuning to

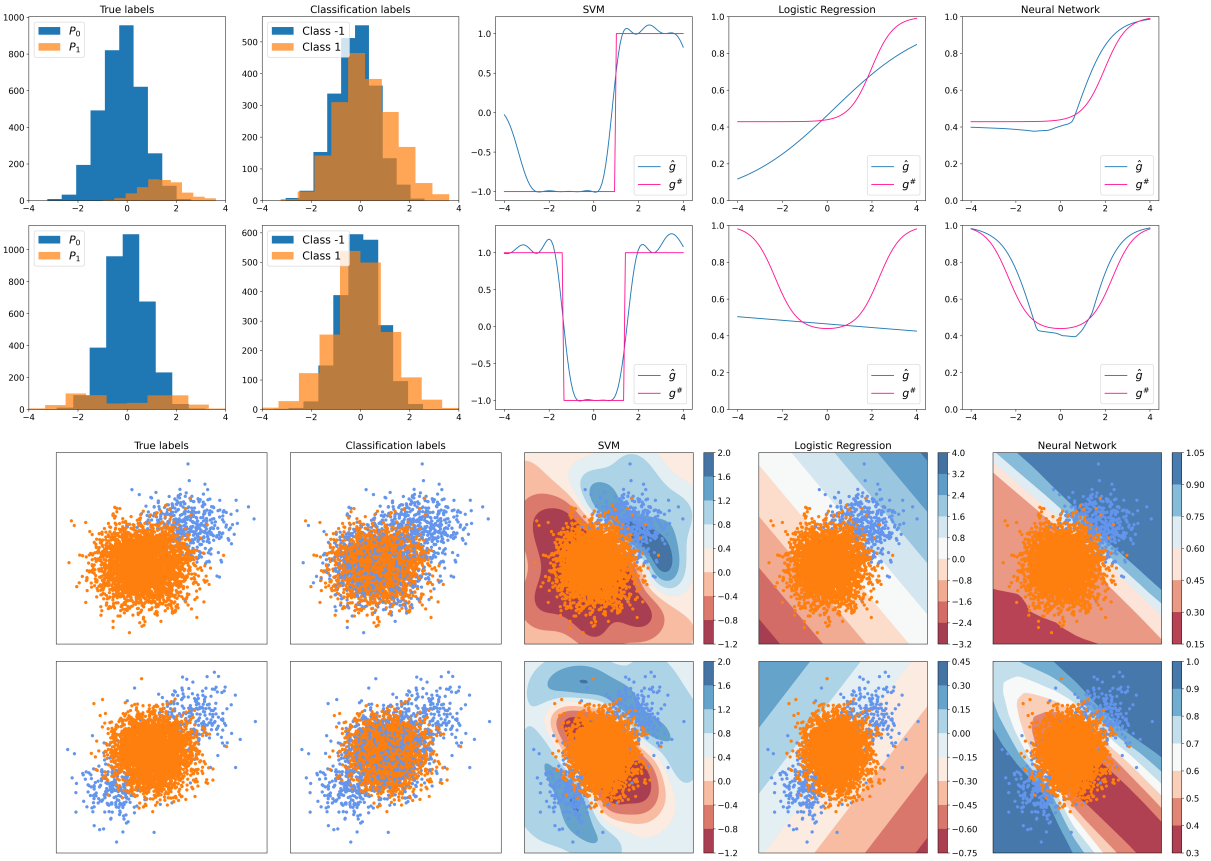


Figure 2.4: Plot of g^* and $g^\sharp$ in different settings (rows) with different loss functions $\ell(\cdot,\cdot)$ and function classes $\mathcal{G}$ (with default parameters in `scikit-learn`). In all settings, $m = 1000$, $m_0 = 500$, $m_1 = 500$, $n = 3000$, and $k = 2000$. The top two rows correspond to $d = 1$ and the bottom two rows correspond to $d = 2$. In all cases, $P_0 = \mathcal{N}(0, I_d)$. For the first and third rows, $P_1 = \mathcal{N}((2, \ldots, 2), I_d)$ (one-sided alternatives); for the second and fourth rows, $P_1 = 0.5\mathcal{N}((2, \ldots, 2), I_d) + 0.5\mathcal{N}((-2, \ldots, -2), I_d)$ (two-sided alternatives).

maximize the power. Examples include the bandwidth for kernel density estimation, the maximum depth for random forests, the width and number of hidden layers for neural networks, and the numerical algorithm to optimize the loss.

Formally, we assume the researcher has a class of candidate score functions $\{g_\upsilon, \upsilon \in \mathcal{U}\}$ indexed by the hyper-parameter υ. The goal is to choose $\hat{\upsilon}$ based on data and use $g_{\hat{\upsilon}}$ as the score function without breaking the FDR guarantee. By Theorem 4, the FDR is controlled so long as $g_{\hat{\upsilon}}$ satisfies the condition (2.8). Motivated by the "double BONuS" procedure proposed in Yang et al. (2021), we propose the following version of AdaDetect with cross-validation, which we abbreviate as the AdaDetect cv procedure.

1. Split $(Y_1, \ldots, Y_k)$ further into two parts $(Y_1, \ldots, Y_s)$ and $(Y_{s+1}, \ldots, Y_k)$ for some $s < k$.
2. Generate a class of score functions g_υ that satisfy a stronger condition than (2.8):
$$g_\upsilon(z, (z_1, \ldots, z_s), (z_{\pi(s+1)}, \ldots, z_{\pi(n+m)})) = g(z, (z_1, \ldots, z_s), (z_{s+1}, \ldots, z_{n+m})).$$
3. For each g_υ, apply AdaDetect with $(Y_{k+1}, \ldots, Y_n, X_1, \ldots, X_m)$ being the test sample and $(Y_1, \ldots, Y_k) = (Y_1, \ldots, Y_s; Y_{s+1}, \ldots, Y_k)$ being the NTS. Denote by r_υ the number of rejections.
4. Choose $\hat{\upsilon} \in \operatorname{argmax}_{\upsilon \in \mathcal{U}} r_\upsilon$
5. Apply AdaDetect with score function $g_{\hat{\upsilon}}$ to the original problem (with $(X_1, \ldots, X_m)$ being the test sample and $(Y_1, \ldots, Y_n) = (Y_1, \ldots, Y_k; Y_{k+1}, \ldots, Y_n)$ being the NTS).

The pipeline to compute $g_{\hat{\upsilon}}$ is illustrated in Figure 2.5. By definition, each g_υ is invariant to permutation of $(Y_{s+1}, \ldots, Y_n, X_1, \ldots, X_m)$ and hence invariant to permutation of the mixed sample $(Y_{k+1}, \ldots, Y_n, X_1, \ldots, X_m)$. Thus, r_υ is also invariant to $(Y_{k+1}, \ldots, Y_n, X_1, \ldots, X_m)$, implying that $\hat{\upsilon}$ is so as well. As a result, $g_{\hat{\upsilon}}$ satisfies the condition (2.8). Therefore, the results in Section 2.3 all carry over to the AdaDetect cv procedure.

In principle, we can use any other objective function that is invariant to the mixed sample than the number of rejections r_υ. Nonetheless, r_υ tends to be a good proxy for the number of rejections in the last step and hence a better objective to optimize than the indirect ones like classification accuracy.

Remark 7. When fitting the hyper-parameter υ, the sample sizes $s, k-s, \ell+m, \ell$ do not maintain the same proportions as the original sizes k, ℓ, m. Our recommendation, following the guidelines in Remark 5, is to choose s such that $k-s$ is of the same order as $\ell+m$ and s is of the same order as m (e.g., $\ell = m$, $s = 3m$, $k = 4m$).

Remark 8. The cross-validation can rule out overfitted models that performs well in training data but does poorly out of sample. By including nonsophisticated baseline models that likely generalize, the power of AdaDetect becomes less sensitive to overfitting of other complicated models or other failure modes that we have yet discovered. For example, the researcher can always add a non-adaptive score that cannot incur overfitting and might be underpowered.

2.5 Power results

In this section, we analyze the power of AdaDetect with appropriately chosen score functions. Throughout this section we assume that the measurements take values in $\mathcal{Z} = \mathbb{R}^d$. We start in Section 2.5.1 with a specific score function given by a constrained empirical risk minimizer (ERM) with the 0-1 loss and show it is as powerful as the classification approach based on the

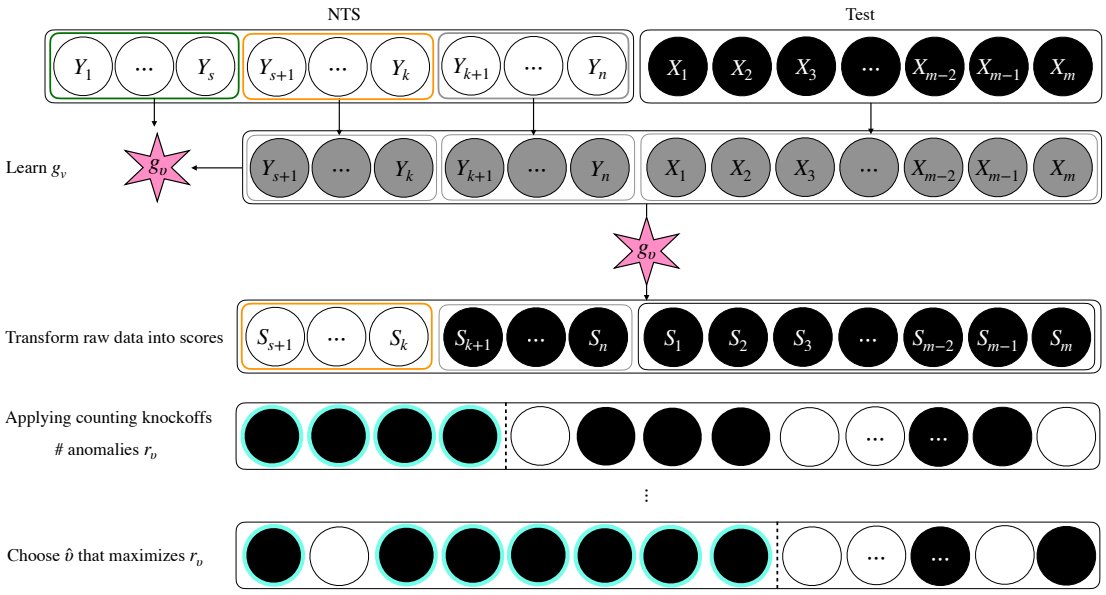


Figure 2.5: The pipeline to compute score function $g_{\hat{v}}$ for AdaDetect cv. Same pictural conventions as in Figure 2.2.

optimal score functions defined in (2.20) when the function class is sufficiently flexible and up to asymptotically vanishing remainder terms. In Section 2.5.2, we turn to a general estimated score function that is close to an oracle (deterministic) score function on all measurements in the mixed sample. When the latter is sufficiently smooth, we show that AdaDetect with the estimated score function is as efficient as AdaDetect with the oracle score function, up to explicit remainder terms that are asymptotically vanishing.

2.5.1 A constrained ERM score function

For the convenience of theoretical analysis, we study a constrained empirical risk minimizer (ERM) score function with 0-1 loss motivated by the Neyman-Pearson (NP) formulation of classification problems given in Blanchard et al. (2010b); see also Cannon et al. (2002) and Scott and Nowak (2005). Define

$$\hat{R}_0(g) = k^{-1}\sum_{i=1}^{k} \mathbb{1}_{g(Z_i)\geq 0}, \quad R_0(g) = \mathbb{E}\,\hat{R}_0(g) = \mathbb{P}_{Z\sim f_0}(g(Z)\geq 0), \tag{2.26}$$

$$\hat{R}_\gamma(g) = (m+\ell)^{-1}\sum_{i=k+1}^{n+m} \mathbb{1}_{g(Z_i)<0}, \quad R_\gamma(g) = \mathbb{E}\,\hat{R}_\gamma(g) = (1-\gamma)(1-R_0(g)) + \gamma R_1(g),$$

$$R_1(g) = \mathbb{P}_{Z\sim \bar{f}_1}(g(Z)<0),$$

where γ, f_0, and $\bar{f}_1$ are defined in (2.17) and (2.18), respectively. We consider a function class $\mathcal{G}$ with a finite Vapnik-Chervonenkis (VC) dimension $V(\mathcal{G})$ (Vapnik, 1998) and the following constrained ERM score function

$$\hat{g} \in \operatorname{argmin}_{g\in\mathcal{G}} \left\{ \hat{R}_\gamma(g) \;:\; \hat{R}_0(g) \leq \beta + \epsilon_0 \right\}, \tag{2.27}$$

for some $\epsilon_0 > 0$, as well as its population version

$$g_{\mathcal{G}}^{\sharp} \in \operatorname{argmin}_{g \in \mathcal{G}} \{R_\gamma(g) \,:\, R_0(g) \leq \beta\}. \tag{2.28}$$

Theorem 10. *Consider the setting of Theorem 8. Assume $\alpha, \beta \in (0,1)$, $k, m_1 \geq 1$, and $g_{\mathcal{G}}^{\sharp}$, defined in (2.28), satisfies $R_0(g_{\mathcal{G}}^{\sharp}) = \beta$. Fix any $\delta \in (0, 1/2)$. Then there exist constants $C, C' > 0$ that only depend on δ such that, if*

$$\epsilon_0 = C\sqrt{\frac{V(\mathcal{G}) + \log(1/\delta)}{k}}, \quad \Delta = C'\gamma^{-1}\sqrt{\frac{V(\mathcal{G}) + \log(1/\delta)}{k \wedge \ell}}, \tag{2.29}$$

where γ is defined in (2.17), the following results hold.

(i) With probability at least $1-\delta$, $R_0(\hat{g}) \leq \beta + \Delta$ and $R_1(\hat{g}) \leq R_1(g_{\mathcal{G}}^{\sharp}) + \Delta$.

(ii) Let $M = \lceil (1 - R_1(g_{\mathcal{G}}^{\sharp}) - \Delta) m_1 \rceil$. Assume that

$$1 - R_1(g_{\mathcal{G}}^{\sharp}) \geq (1 + \alpha^{-1})\Delta, \quad \ell \geq \frac{2m}{\alpha M}, \quad \beta \leq \frac{0.4\alpha M}{m}. \tag{2.30}$$

Then, with probability at least $1-\delta$,

$$\mathit{AdaDetect}_\alpha \supset \{i \in \{1, \ldots, m\} \,:\, \hat{g}(X_i) \geq 0\}, \tag{2.31}$$

$$|\mathit{AdaDetect}_\alpha \cap \mathcal{H}_1|/m_1 \geq 1 - R_1(g_{\mathcal{G}}^{\sharp}) - \Delta, \tag{2.32}$$

where $\mathit{AdaDetect}_\alpha$ denotes the rejection set of AdaDetect with score function $\hat{g}$.

The proof of Theorem 10 is presented in Section A.3.1. The idea is to show that there are many alternatives with a nonnegative score, while there are only a few true nulls with nonnegative scores. This yields small empirical p-values for hypotheses with nonnegative scores, which implies that the procedure AdaDetect_α detects these nonnegative scores, see Lemma 24.

Theorem 10 (i) shows that $\hat{g}$ has a similar classification accuracy to $g_{\mathcal{G}}^{\sharp}$ on both the NTS and mixed sample. It is analogous to Theorem 2 in Blanchard et al. (2010b), though Blanchard et al. (2010b) considers a different setting where the proportion of nulls is random. Theorem 10 (ii) entails that, with high probability, all hypotheses with nonnegative scores will be rejected and the power of AdaDetect with $\hat{g}$ is nearly as large as the power of the classification procedure given by $g_{\mathcal{G}}^{\sharp}$.

Note that the Lagrangian form of the above problem is in the form of the weighted loss defined in (2.22). Thus, by Lemma 9 (i), there exists $\lambda_\beta > 0$ such that

$$g_{\mathcal{G}}^{\sharp}(x) = g^*(x) = \frac{\lambda_\beta(\ell + m)}{k}\frac{f_\gamma(x)}{f_0(x)} - 1,$$

if the constraint is feasible and $\mathcal{G}$ is sufficiently rich to include the above function. Above, $g^*(x)$ satisfies (2.20) and hence yields the optimal power. If we define $b = R_1(g_{\mathcal{G}}^{\sharp}) - R_1(g^*)$ as the bias due to the constraint, Theorem 10 (ii) implies that, with probability $1-\delta$, the power of AdaDetect with $\hat{g}$ is at most $\Delta + b$ below the optimal power. Thus, the function class $\mathcal{G}$ incurs a tradeoff that a richer class yields a smaller b but a larger Δ and vice versa.

Here, we aim at making $\mathcal{G}$ as flexible as possible while ensuring $\Delta = o(1)$. When k, ℓ, and m are of the same order and δ is a constant, $\Delta \asymp \frac{m}{m_1}\sqrt{\frac{V(\mathcal{G})}{m}}$. Hence, $\Delta = o(1)$ if

$$\frac{m_1}{m} \gg \sqrt{\frac{V(\mathcal{G})}{m}}. \tag{2.33}$$

For illustration, consider the class $\mathcal{G}_{N,L,s}$ of ReLU feed forward neural networks with fixed topology, maximum width $N \asymp m^c$ ($c \in (0,1)$), depth $L \asymp \log m$ and sparsity $s \asymp N \log m$. Bartlett et al. (2019) show that

$$V(\mathcal{G}_{N,L,s}) \leq 2sL \log(4eN) \lesssim m^c (\log m)^3.$$

Hence, condition (2.33) reads in this case

$$m_1/m \gg m^{\frac{c-1}{2}} (\log m)^{3/2}.$$

This implies that $\Delta = o(1)$ unless the novelties are too sparse. On the other hand, given the approximation ability of class of neural networks, we should expect $1 - R_1(g^\sharp_{\mathcal{G}_{N,L,s}}) \approx 1 - R_1(g^*)$. Thus, Theorem 10 (ii) implies the resulting score function is nearly optimal.

Remark 9 (Choice of β). Since AdaDetect$_\alpha$ controls FDR at level α, it is necessary to impose an upper bound on β in Theorem 10 (ii). Roughly speaking, our condition on β guarantees that the classifier $g^\sharp_{\mathcal{G}}$ controls the FDR at level α, up to remainder terms.

Remark 10. The condition on ℓ in (2.30) is needed to ensure that the minimum value $1/(1+\ell)$ that p-values can take is sufficiently small so that the BH procedure can reject. A similar condition was introduced in Mary and Roquain (2022), see also Remark 5.

2.5.2 General score functions

Now we move to general score functions. Let g^* be any measurable function $\mathbb{R}^d \to \mathbb{R}$ in the form of (2.20) and

$$\overline{G}_0(s) = \mathbb{P}_{X \sim P_0}(g^*(X) \geq s), \quad s \in \mathbb{R}; \tag{2.34}$$

$$\zeta_r(\eta) = \max_{u \in [\alpha(r \vee 1)/m, \alpha]} \left\{ \frac{\overline{G}_0(\overline{G}_0^{-1}(u) - 2\eta) - u}{u} \right\}, \quad \eta > 0, r \in \{0, \dots, m\}. \tag{2.35}$$

Here, $\zeta_r(\cdot)$ measures the local fluctuation of $\overline{G}_0$. We suppress the dependence on α and m to ease notation. Furthermore, consider any data-driven score function $\hat{g}$ satisfying the condition (2.8) and let

$$\hat{\eta} = \max_{k+1 \leq i \leq n+m} |\hat{g}(Z_i; (Z_1, \dots, Z_k), (Z_{k+1}, \dots, Z_{n+m})) - g^*(Z_i)|, \tag{2.36}$$

which measures the maximal discrepancy of scores in the mixed sample. In the following, AdaDetect$_\alpha$ denotes the procedure with score function $\hat{g}$ and AdaDetect$^*_\alpha$ denotes the procedure with score function g^*.

Theorem 11. *Fix any $r \in \{0, \dots, m\}$ and let $\mathcal{R} = \{|AdaDetect^*_\alpha| \geq r\}$. Assume $m \geq 1$, $\ell, k \geq 0$, $n = k + \ell \geq 1$, and $\overline{G}_0$ (2.34) is continuous and strictly decreasing. Under Assumptions 4 and 5, for any $\delta, \eta \in (0,1)$ such that $(\ell+1)\delta\alpha(r \vee 1)/m \geq 2$,*

$$\mathbb{P}\left(\mathcal{R} \cap \{AdaDetect^*_\alpha \subset AdaDetect_{\alpha'}\}^c\right) \leq \mathbb{P}\left(\hat{\eta} > \eta\right) + 2me^{-(3/28)(\ell+1)\delta^2\alpha(r\vee 1)/m}, \tag{2.37}$$

where $\alpha' = \alpha(1+3\delta)(1+\zeta_r(\eta))$ and $\zeta_r(\cdot)$ and $\hat{\eta}$ are defined in (2.35) and (2.36), respectively. Furthermore, (2.37) is also true with AdaDetect$^*_\alpha$ replaced by BH$^*_\alpha$, the BH algorithm applied to the oracle p-values $p^*_i = \overline{G}_0(g^*(X_i))$, $1 \leq i \leq m$.

The proof is presented in Section A.3.2. The condition $(\ell+1)\delta\alpha(r\vee 1)/m \geq 2$ is analogous to the one studied (Mary and Roquain, 2022) for fixed score functions. When we choose $r = 0$, we have $\mathbb{P}(\mathcal{R}) = 1$ and thus (2.37) implies

$$\mathbb{P}\left(\text{AdaDetect}^*_\alpha \subset \text{AdaDetect}_{\alpha'}\right) \geq 1 - \mathbb{P}\left(\hat{\eta} > \eta\right) - 2me^{-(3/28)(\ell+1)\delta^2\alpha/m},$$

for any δ with $(\ell+1)\delta\alpha/m \geq 2$. If $\ell/m \gg \log m$, we can choose $\delta = o(1)$ such that $(\ell+1)\delta^2\alpha/m \gg \log m$ and η such that $\mathbb{P}(\hat{\eta} \geq \eta) = o(1)$, in which case

$$\mathbb{P}\left(\text{AdaDetect}^*_\alpha \subset \text{AdaDetect}_{\alpha'}\right) = 1 - o(1),$$

where $\alpha' = \alpha(1+\zeta_0(\eta))(1+o(1))$. Thus, when $\zeta_0(\eta)$ is small, we show that AdaDetect with the estimated score function and slight inflation of the target level is strictly more powerful than its oracle version.

In general, when $|\text{AdaDetect}^*_\alpha|$ is larger with high probability, we can choose a larger r to relax the condition on δ, reduces $\zeta_r(\eta)$ (and hence α'), and improve the RHS of (2.37). In particular, we can set r appropriately to obtain the following result on the asymptotic TDR.

Corollary 12. *Consider the setting of Theorem 11. Fix any $\epsilon > 0$. Assume $m_1 \geq 1$ and $(\ell+1)\delta\alpha\lceil m_1\epsilon\rceil/m \geq 2$. Then*

$$\text{TDR}(\mathit{AdaDetect}_{\alpha'}) \geq \text{TDR}(\mathit{AdaDetect}^*_\alpha) - \mathbb{P}\left(\hat{\eta} > \eta\right) - 2me^{-(3/28)(\ell+1)\delta^2\alpha\lceil m_1\epsilon\rceil/m} - \epsilon, \tag{2.38}$$

where $\alpha' = \alpha(1+3\delta)(1+\zeta_{\lceil m_1\epsilon\rceil}(\eta))$. In particular, if there exist sequences $\delta = \delta(k,\ell,m,m_1)$, $\epsilon = \epsilon(k,\ell,m,m_1)$, and $\eta = \eta(k,\ell,m,m_1)$ such that, as ℓ, m, m_1 tend to infinity,

$$\delta, \epsilon \to 0,\ \ell\delta^2\epsilon m_1/m \to \infty,\ \mathbb{P}\left(\hat{\eta} > \eta\right) \to 0 \text{ and } \zeta_{\lceil m_1\epsilon\rceil}(\eta) \to 0, \tag{2.39}$$

then

$$\liminf_{\ell,m,m_1} \left\{\text{TDR}(\mathit{AdaDetect}_{\tilde{\alpha}}) - \text{TDR}(\mathit{AdaDetect}^*_\alpha)\right\} \geq 0, \quad \text{for any fixed } \tilde{\alpha} > \alpha. \tag{2.40}$$

Furthermore, these results hold with $\mathit{AdaDetect}^*_\alpha$ replaced by BH^*_α defined in Theorem 11.

The proof is presented in Section A.3.3. Corollary 12 shows that AdaDetect is nearly as powerful as the oracle version, as well as the BH procedure with the optimal score.

Now we discuss the choice of η. Note that η is a parameter that only shows up in the bound but not in the algorithm. It incurs a tradeoff that a larger η would improve the tail bound by decreasing $\mathbb{P}\left(\hat{\eta} > \eta\right)$ but inflate α' through increasing $\zeta_r(\eta)$. Ideally, we would want η so that $\mathbb{P}\left(\hat{\eta} > \eta\right)$ and $\zeta_r(\eta)$ are both negligible. For illustration, assume

$$\mathbb{P}\left(\hat{\eta} > (n+m)^{-\kappa}\right) = o(1), \tag{2.41}$$

for some $\kappa \in (0, 1/2)$ and $\zeta_r(\eta) \lesssim \eta/\gamma$, where γ is defined in (2.17). In this case, $\mathbb{P}\left(\hat{\eta} > \eta\right)$ and $\zeta_r(\eta)$ are both $o(1)$ if

$$(n+m)^{-\kappa} = o(\gamma) = o\left(\frac{m_1}{m+\ell}\right).$$

Again, this would hold unless the novelties are too sparse.

We show in Lemma 29 that the score function given by density estimation satisfies (2.41) under regularity conditions. Another example is given by Theorem 3.2 in Audibert and Tsybakov (2007) in the case where g^* is the posterior probability under a different setting; see Section A.5.3). For $\zeta_r(\eta)$, we provide bounds in Section A.5.1 for two examples. In the Gaussian example, we show that $\zeta_r(\eta) \lesssim \eta/\gamma$, where γ is defined in (2.17).

Remark 11. Theorem 3 in Yang et al. (2021) provides another asymptotic power analysis showing that the symmetric difference between the rejection set for the data-driven score function and its oracle version has a size $o_P(m)$. Unlike Theorem 11 and Corollary 12, it does not have implications when the oracle procedure can only reject $o_P(m)$ hypotheses, as in the case where $m_1/m = o(1)$.

2.6 Experiments

In this section, we examine the performance of AdaDetect on both simulated data (Section 2.6.1) and real data (Section 2.6.2). We apply AdaDetect with various score functions, including the oracle score defined in (2.20) (`AdaDetect oracle`), the density estimation-based score (`AdaDetect parametric` and `AdaDetect KDE`), the PU classification-based score (`AdaDetect SVM`, `AdaDetect RF`, `AdaDetect NN`, and `AdaDetect NN cv`). We also include the SC procedures proposed by Sun and Cai (2007) (`SC parametric` and `SC KDE`) and the conformal novelty detection procedures proposed by Bates et al. (2023) (`CAD SVM` and `CAD IForest`). Note that both `CAD SVM` and `CAD IForest` are instances of AdaDetect with one-class classification-based scores. See Section A.6.1 for a full description of these methods. For all our experiments, we use the Python package `scikit-learn` for Expectation Maximization (EM) algorithm, kernel density estimation, random forests, and neural networks, with the default hyper-parameters from the packages unless otherwise specified.

Note that these procedures all provably control the FDR under Assumption 4 (see Corollary 5) except for `SC parametric` and `SC KDE`, which only control the FDR asymptotically with a consistent estimator of the density ratio (Sun and Cai, 2007).

2.6.1 Simulated data

In all experiments considered this section, we generate measurements under Assumption 4 with all novelties generated from the same distribution P_1. Unless otherwise specified, we set $n = 3000, m = 1000$, $\pi_0 = m_0/m = 0.9$, and calculate FDR and TDR based on 100 Monte-Carlo simulations. Following Remark 5, we set $k = 2m$ and $\ell = m$ for all AdaDetect methods.

Gaussian setting We start with a setting where $P_0 = \mathcal{N}(0, I_d)$ and $P_1 = \mathcal{N}(\mu, I_d)$, where $\mu \in \mathbb{R}^d$ is a sparse vector with the first 5 coordinates equal to $\sqrt{2\log(d)}$ and the remaining ones equal to 0. The results are presented in Figure 2.6 with the dimension d varying. First, we note that neither `SC parametric` nor `SC KDE` control the FDR, even if the model is correctly specified for the former, and the FDR inflation is substantial in high dimensions. By contrast, as implied by our theory, all other procedures control the FDR at level $\pi_0\alpha$. Next, we compare the TDR of procedures which control the FDR. In low dimensions ($d \leq 100$), `AdaDetect parametric` has the highest power that is close to `AdaDetect oracle`, which is expected since the model is correctly specified and parametric estimation is accurate when the dimension is low. In high dimensions ($d = 500$), `AdaDetect parametric` becomes much more noisy while `AdaDetect RF` maintains a stable and high power.

Non-gaussian setting We now consider a non-gaussian setting, where the first two coordinates of nulls and novelties are independent draws from $\text{Beta}(5, 5)$ and $\text{Beta}(1, 3)$, respectively, and the other coordinates are independent draws from $\text{Beta}(1, 1)$ for both nulls and novelties. Note that `AdaDetect parametric` is now based on a misspecified parametric model; see Section A.6.1 for detail. Figure 2.7 presents the results with the dimension d varying.

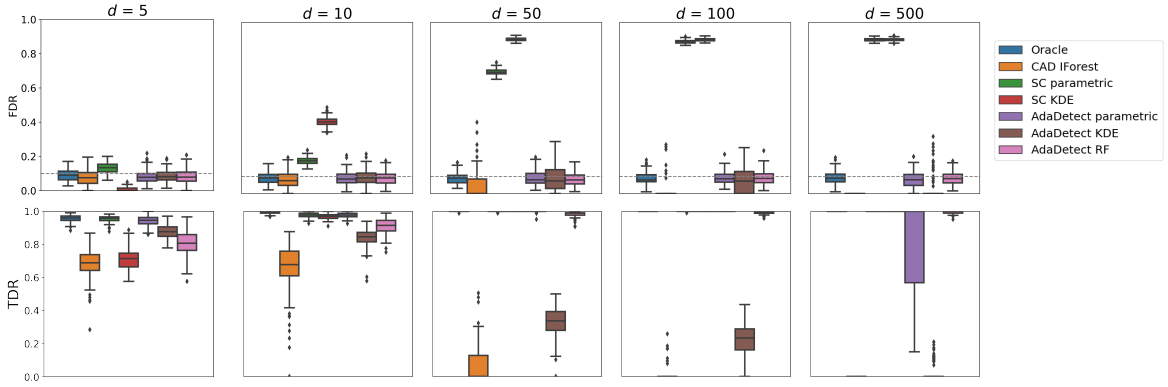


Figure 2.6: Gaussian setting. FDR (top) and TDR (bottom) as a function of d. The dashed line indicates the target level.

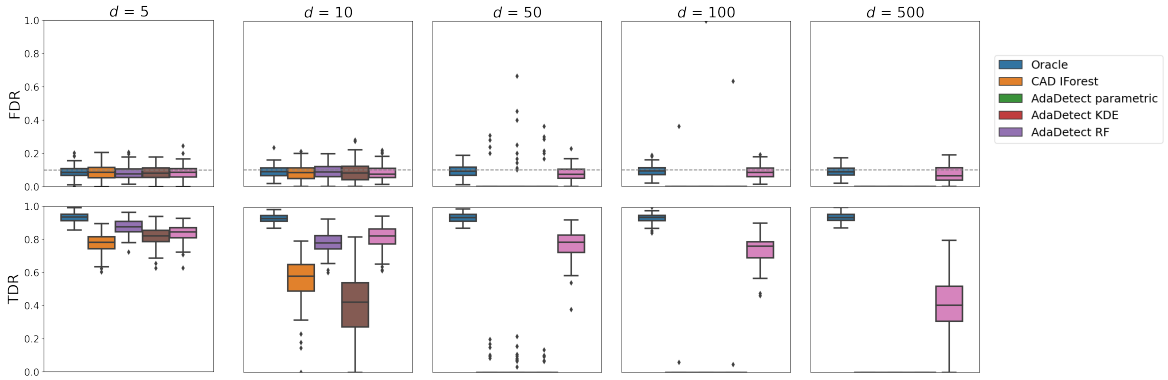


Figure 2.7: Non-gaussian (beta) setting. FDR (top) and TDR (bottom) as a function of d. The dashed line indicates the target level.

`AdaDetect parametric` is now clearly dominated by `AdaDetect RF`, especially in high dimensions. This shows that machine learning-based classification methods are more robust to model misspecification when combined with AdaDetect.

2.6.2 Semi-synthetic data

In this section we study the performance of AdaDetect on real datasets. Each dataset contains measurements that are labeled as either typical or novelty. We summarize the datasets in Table 2.2. The first four datasets are also used in Section 5.3 of Bates et al. (2023); see the descriptions and references therein. The Musk data contains a set of molecules that are identified as either musk (nulls) or non-musk (novelties). The MNIST data (LeCun and Cortes, 2010) contains a set of labeled images of size 28×28 of handwritten digits from '0' to '0'. We restrict the analysis to '4' (nulls) and '9' (novelties). The categorical features are converted via one-hot encoding. We construct test samples and null training samples by subsampling the dataset with $n = 5000$, $m = 1000$, and a fixed null proportion $\pi_0 = m_0/m = 0.9$. For AdaDetect, we choose $k = 4m$, $\ell = m$, $s = 3m$ for cross-validation, and the target level $\alpha = 0.1$. For the MNIST dataset, we consider two more methods based on a convolutional neural network (CNN), with two convolution layers and one fully connected layer. The first method `CAD SVDD CNN` is the conformal novelty detection procedure of Bates et al. (2023) with a special one-class classifier, given by the Support Vector Data Description (SVDD) method introduced in Ruff et al. (2018) used with a family of functions given by the CNN. The second

Table 2.2: Summary of datasets.

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
Dimension d	9	30	40	6	166	28×28
Feature type	Real	Real	Real, categorical	Real	Real	Real
Inliers	45586	284315	47913	10923	5581	5842
Novelties	3511	492	200	260	1017	5949

Table 2.3: FDR (top) and TDR (bottom) of AdaDetect with different score functions on real datasets. The target FDR level is $\alpha = 0.1$. We report the mean value and the standard deviation (in brackets) over 100 runs. The two best-performing methods are highlighted in bold.

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
			FDR			
CAD SVM	0.04 (0.08)	0.00 (0.00)	0.00 (0.00)	0.05(0.10)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.10 (0.07)	0.09 (0.06)	0.08 (0.07)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.01 (0.05)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.07 (0.07)	0.05 (0.08)	0.02 (0.06)	0.08 (0.07)	0.02 (0.08)	0.00 (0.00)
AdaDetect SVM	0.08 (0.04)	0.07 (0.05)	0.07 (0.05)	0.07 (0.06)	0.08 (0.06)	0.02 (0.03)
AdaDetect RF	0.08 (0.04)	0.09 (0.04)	0.08 (0.04)	0.04 (0.10)	0.03 (0.06)	0.03 (0.07)
AdaDetect NN	0.07 (0.05)	0.09 (0.04)	0.06 (0.07)	0.09 (0.06)	0.06 (0.09)	0.06 (0.08)
AdaDetect cv NN	0.08 (0.04)	0.09 (0.05)	0.08 (0.11)	0.08 (0.05)	0.06 (0.08)	0.01 (0.03)
CAD SVDD + CNN	-	-	-	-	-	0.03 (0.14)
AdaDetect CNN	-	-	-	-	-	0.09 (0.05)
			TDR			
CAD SVM	0.10 (0.18)	0.00 (0.00)	0.00 (0.00)	0.03 (0.06)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.45 (0.09)	0.39 (0.22)	0.56 (0.35)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.02 (0.07)	0.00 (0.00)	0.00 (0.00)	0.07 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.44 (0.33)	0.12 (0.20)	0.11 (0.24)	0.22 (0.17)	0.02 (0.06)	0.00 (0.00)
AdaDetect SVM	**0.85 (0.17)**	0.68 (0.28)	0.66 (0.32)	0.43 (0.13)	**0.40 (0.17)**	**0.52 (0.21)**
AdaDetect RF	**0.99 (0.01)**	**0.85 (0.03)**	**0.99 (0.01)**	**0.48 (0.10)**	0.04 (0.09)	0.03 (0.08)
AdaDetect NN	0.76 (0.15)	**0.80 (0.07)**	0.52 (0.41)	**0.47 (0.14)**	0.11 (0.13)	0.01 (0.03)
AdaDetect cv NN	0.84 (0.12)	0.76 (0.13)	**0.74 (0.41)**	0.42 (0.16)	**0.13 (0.12)**	0.01 (0.03)
CAD SVDD + CNN	-	-	-	-	-	0.03 (0.15)
AdaDetect CNN	-	-	-	-	-	**0.93 (0.06)**

method is AdaDetect with the two-class classifier based on the CNN, denoted by `AdaDetect CNN`.

The FDR and TDR for the methods are evaluated by using 100 runs and the results are reported in Table 2.3. As expected, all methods control the FDR. Compared to Bates et al. (2023), AdaDetect with classification-based scores substantially boosts the power because it incorporates the novelties in learning the score function. Overall, the best performing method is `AdaDetect RF`, with `AdaDetect NN` (possibly cross-validated) coming in second. `AdaDetect CNN` is particularly efficient on the classical MNIST dataset, which is unsurprising because CNN-type classifiers are appropriate for such an image dataset (Goodfellow et al., 2016). We however note that the one-class classifier based upon CNN behaves poorly, which shows that two-class classification is the key for the power boost instead of the better representation given by CNN. In addition, further comparisons are made in Appendix A.6.3 for other values of n, m, m_1 in more challenging regimes and the conclusions are qualitatively similar.

To conclude, if a classification method is expected to distinguish between typical and anomalous measurements, combining it with AdaDetect is expected to achieve high power without threatening FDR control.

2.7 An astronomy application

In this section, we apply AdaDetect to detect variable stars using the Sloan Digital Sky Survey (Ivezić et al., 2005), a large labeled dataset with $92,658$ nonvariable (null) and 483

variable (novelties) stars. Each star is encoded as a 4-dimensional vector containing the star's flux in specific bands (colors) of the visible light. This dataset is particularly appealing for demonstrating our method. First, the two classes occupy similar regions in the considered color space, with slight overlap leading to complex decision boundaries. Second, the large number of nonvariable stars allows us to vary the size of the NTS in a large range in the Monte-Carlo simulations. Third, this dataset has been extensively studied by astronomers and has become a standard for benchmarking classification methods (see Chapter 9 of Ivezić et al., 2019). Lastly, we can compute the achieved FDR and TDR for any novelty detection method based on the labeled data.

For each experiment, we sample n nonvariable stars as the NTS along with m_1 variable stars and $m_0 = m - m_1$ additional nonvariable stars as the test sample. We set $m = 100$ and vary n and m_1 across experiments. We apply AdaDetect with Kernel Density Estimation (KDE), Random Forest (RF), and Neural Networks (NN). For comparison, we also include two Empirical BH procedures (Mary and Roquain, 2022), which are special cases of AdaDetect with non-adaptive scores as the squared ℓ_2 norm of the demeaned vectors, where the mean is calculated on all nulls outside of the NTS ("Emp BH full") and on the NTS ("Emp BH current"), respectively. The "Emp BH current" method is closer to the current practice, though it is not granted to control the FDR since the score function does not satisfy (2.8). In addition, we apply the Empirical BH procedure without demeaning the data as well as the SC procedure with estimated local FDR. Neither detects any novelties so we will not report them.

Figure 2.8 presents the results for $m_1 = 50$ and varying n with target FDR level $\alpha = 0.05$. The FDR and TDR are calculated based on 100 Monte-Carlo simulations. To aid visualization, we represents the uncertainty by a shaded area whose width is equal to the standard error of estimated FDR/TDR divided by 10. This can be viewed as an approximation of the standard error with $10,000$ Monte-Carlo simulations. In this setting, $\pi_0 = 0.5$ and thus all methods provably control the FDR at level $\pi_0\alpha = 0.025$ (except "Emp BH current"). This is confirmed in the left panel of Figure 2.8. From the right panel, we observe that AdaDetect with RF achieves the highest power, substantially improving upon AdaDetect with non-adaptive scores (Emp BH). This demonstrates the advantages of utilizing classification-based score functions. In Section A.7, we present results in additional experimental settings that exhibit qualitative similarities to Figure 2.8.

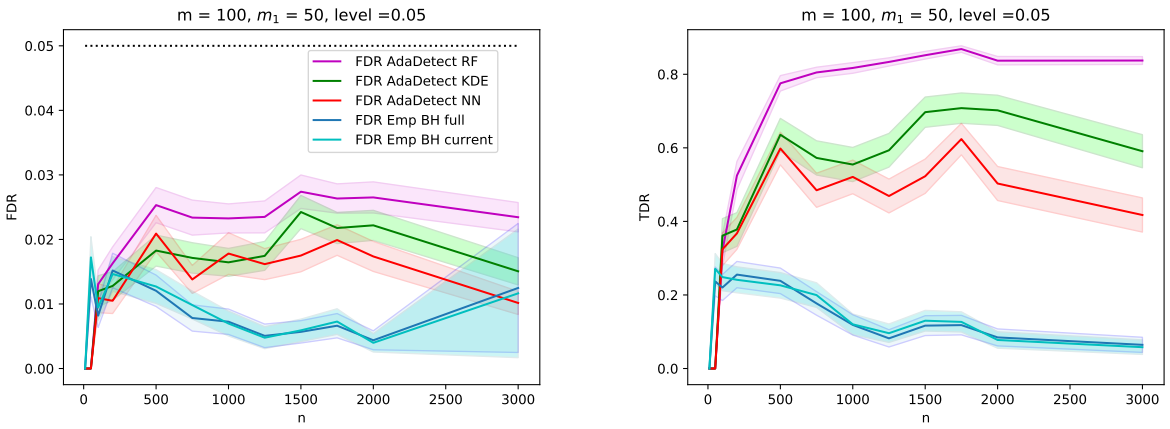


Figure 2.8: Estimated FDR (left) and TDR (right) as a function of n, the size of the NTS, with $m = 100, m_1 = 50$ and $\alpha = 0.05$. All methods shown in the plot provably control the FDR at level $\pi_0\alpha = 0.025$ (except "Emp BH current").

2.8 Conclusion and discussion

In this work, we propose AdaDetect as a generic framework that can wrap around any classification methods and provably control the FDR in finite samples when the null measurements are exchangeable conditional on the novelties. It generalizes and often substantially outperforms previous methods that only work with one-class classification methods which are not adaptive to the novelty distribution. We also develop the π_0-adaptive AdaDetect that further improves the power in the presence of many novelties as well as the cross-validated AdaDetect that allows model selection. The theoretical analysis is based on a novel FDR expression that unifies and generalizes the existing results. In addition, we provide power analysis showing that (1) the optimal score function is given by any monotonic transformation of the ratio between the average density of novelties and the null density and (2) the estimated score function can be asymptotically optimal in terms of power. We demonstrate the versatility of AdaDetect on a variety of tasks.

2.8.1 Limitations of AdaDetect

Here we discuss several limitations of our method and potential solutions.

- Heterogeneous null distributions. A key assumption for the FDR control is that the null distribution P_0 is the same across the NTS and the test sample. This excludes the case where the null can be generated from a bag of distributions $\{P_{0,k}, 1 \leq k \leq K\}$. Under heterogeneity, the empirical p-values can be invalid even marginally since the nulls are no longer exchangeable. One possible way to reconcile this issue is to assume the nulls are generated from a mixture distribution $\sum_{k=1}^{K} \pi_k P_{0,k}$, thereby retaining the exchangeability. We leave this for future research.

- Directional null hypotheses. Throughout the paper we focus on testing whether a new observation has the same distribution as the typical measurements. In some applications, it may be more appropriate to test directional nulls, which are often characterized by the sign of a parameter for parametric models. However, it is unclear how this can be done in nonparametric cases. One possibility is to consider the nulls $P_i \preceq P_0$ where $\preceq$ denotes the stochastic dominance, meaning that there exists a random vector (A, B) such that $A \sim P_0, B \sim P_i$ and $A \geq B$ in an entrywise fashion. By restricting the score function to be entrywise increasing, we may still apply AdaDetect and retain the FDR control.

- Randomness of data splitting. AdaDetect is intrinsically randomized due to the data splitting step. Without carefully documenting random seeds, the researcher can "hack" the results by reporting the best results across different splits. A subsequent work by Bashari et al. (2023) proposes an elegant solution to derandomize AdaDetect by treating the test statistics as e-values and aggregating over all data splits. They show that the e-AdaDetect successfully stabilizes the output of AdaDetect.

- Semi-supervised data. In some applications, labeled novelties are available in the training sample. For example, the researcher may have historical data on fraud transactions recorded in the system and can train a two-class classifier to distinguish between the nulls and labeled novelties. When future novelties are similar to labeled novelties, it should yield an efficient score function. This has been studied by Liang et al. (2022). Combining their approach with ours in a nonstationary setting where future novelties behave differently from the past ones is a promising avenue for future research.

- Sparse novelties: when the novelties are too sparse, two-class classifiers may not be the best at discriminating between nominals and novelties and can be out-performed by simpler one-class classifiers; see Liang et al. (2022). One possible solution is to apply AdaDetect cv by including both one-class and two-class classification methods and let data decide which score function is more efficient. We leave the full examination of this approach for future research.

2.8.2 Other future works

First, we could provide a more detailed power analysis by quantifying the bias term $R_1(g^\sharp_{\mathcal{G}}) - R_1(g^\sharp)$ for a broader class of algorithms. For example, we can consider $\mathcal{G} = \mathcal{G}_{N,L,s}$, the set of realizations of NN with width N, depth L and sparsity s (Bos and Schmidt-Hieber, 2021). Such a quantitative analysis could provide guidelines for choosing hyper-parameters or at least a default range in the cross-validated AdaDetect procedure.

Next, a core assumption of the FDR controlling theory is exchangeability of the null scores conditional on the novelties. This can be satisfied beyond our setting, e.g., the knockoff setting discussed in Remark 6. This suggests a possible path to further improve the knockoffs method.

Lastly, the BONuS algorithm in Yang et al. (2021) can iteratively remove null observations and update the score function correspondingly using a masking technique introduced by Lei and Fithian (2018). While this increases the computation cost, it gradually reduces the attenuation caused by the null sample in the mixed sample and hence improves the accuracy of the estimated score function. It would be interesting to apply their idea in AdaDetect.

Chapter 3

False membership rate control in mixture models

The clustering task consists in partitioning elements of a sample into homogeneous groups. Most datasets contain individuals that are ambiguous and intrinsically difficult to attribute to one or another cluster. However, in practical applications, misclassifying individuals is potentially disastrous and should be avoided. To keep the misclassification rate small, one can decide to classify only *a part* of the sample. In the supervised setting, this approach is well known and referred to as classification with an abstention option. In this paper the approach is revisited in an unsupervised mixture-model framework and the purpose is to develop a method that comes with the guarantee that the false membership rate (FMR) does not exceed a predefined nominal level α. A plug-in procedure is proposed, for which a theoretical analysis is provided, by quantifying the FMR deviation with respect to the target level α with explicit remainder terms. Bootstrap versions of the procedure are shown to improve the performance in numerical experiments.

Contents

3.1 Introduction 48
3.1.1 Background 48
3.1.2 Aim and approach 48
3.1.3 Presentation of the results 49
3.1.4 Relation to previous work 50
3.1.5 Organization of the paper 51
3.2 Setting 51
3.2.1 Model 51
3.2.2 Procedure and criteria 52
3.2.3 Notation 53
3.3 Methods 53
3.3.1 Oracle procedures 53
3.3.2 Empirical procedures 55
3.4 Theoretical guarantees for the plug-in procedure 56
3.4.1 Additional notation and assumptions 56
3.4.2 Results 58
3.5 Experiments 59
3.5.1 Synthetic data set 59
3.5.2 Real data set 60

3.6 Conclusion and discussion 62

3.1 Introduction

3.1.1 Background

Clustering is a standard statistical task that aims at grouping together individuals with similar features. However, it is common that data sets include ambiguous individuals that are inherently difficult to classify, which makes the clustering result potentially unreliable. To illustrate this point, consider a Gaussian mixture model with overlapping mixture components. Then it is difficult, or even impossible, to assign the correct cluster label to data points that fall in the overlap of those clusters, see Figure 3.1. Hence, when the overlap is large (Figure 3.1 panel (b)), the misclassification rate of a standard clustering method is inevitably elevated.

This issue is critical in applications where misclassifications come with a high cost for the user and should be avoided. This is for example the case for medical diagnosis, where an error can have severe consequences on the individual's health. When there is too much uncertainty, a solution is to avoid classification for such individuals, and to adopt a wiser "abstention decision", that leaves the door open for further medical exams.

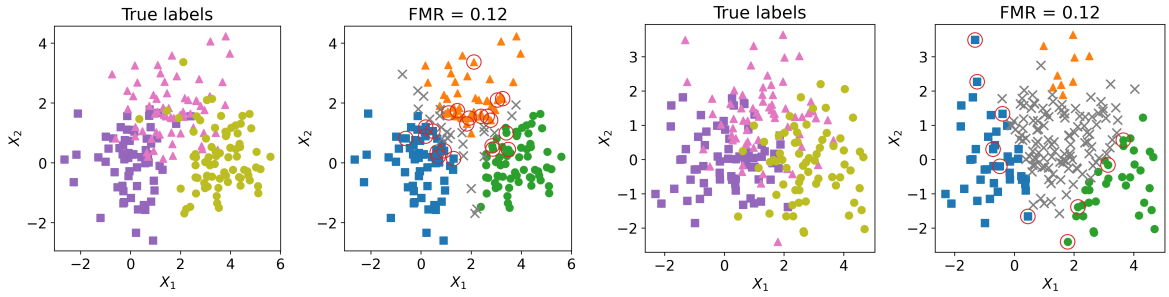
In a supervised setting, classification with a reject (or abstention) option is a long-standing statistical paradigm, that can be traced back to Chow (1970), with more recent works including Herbei and Wegkamp (2006); Bartlett and Wegkamp (2008); Wegkamp and Yuan (2011), among others. In this line of research, rejection is accounted for by adding a term to the risk that penalizes any rejection (i.e., non classification).

Recently, still in the supervised setting, Geifman and El-Yaniv (2017) and Angelopoulos et al. (2021) have considered the problem of having a prescribed control of the classification error among the classified items (those that are not rejected). In these works the proposed method consists of thresholding the estimated class probabilities estimated by a pre-trained classifier, in a data-driven manner. Both of these works provide the guarantee that the resulting selective classifier has its true risk bounded by a prescribed level with high probability.

3.1.2 Aim and approach

The goal of the present work is to propose a labelling guarantee on the classified items in the more challenging unsupervised setting, where no training set is available and data are assumed to be generated from a finite mixture model. This is achieved by the possibility to refuse to cluster ambiguous individuals and by using the false membership rate (FMR), which is defined as the average proportion of misclassifications among the classified objects. Our procedures are devised to keep the FMR below some nominal level α, while classifying a maximum number of items.

It is important to understand the role of the nominal level α in our approach. It is chosen by the user and depends on their acceptance or tolerance for misclassified objects. Since the FMR is the misclassification risk that is allowed on the classified objects, the final interpretation of an FMR control at level α is clear: if, for instance, α is set to 5% and 100 items are finally chosen to be classified by the method, then the number of misclassified items is expected to be at most 5. This high interpretability is similar to the one of the false discovery rate (FDR) in multiple testing, which has known a great success in applications since its introduction by Benjamini and Hochberg (1995). This is a clear advantage of our



(a) Separated clusters (b) Ambiguous clusters

Figure 3.1: Data from Gaussian mixtures with three components ($n = 200$), in a fairly separated case (panel (a)) and an ambiguous case (panel (b)). In each panel, the left part displays the true clustering, while the right part illustrates the new procedure (plug-in procedure at level $\alpha = 10\%$), that does not cluster all items. The points not classified are depicted by grey crosses. Red circles indicate erroneous labels.

approach for practical use compared to the methods with a rejection option that are based on a penalized risk.

In our framework, a procedure is composed of two intertwined decisions:

- a clustering method inferring the labels;
- a selection rule deciding which items to label.

Importantly, the selection rule is only applied *after* a clustering method is fitted on the (entire) sample. In other words, the procedure consists of two subsequent steps: a clustering step, after which cluster labels are kept fixed, and a selection step, that chooses which items to classify – in which case, the label from the previous clustering step is assigned. For the items that are not selected, we discard the cluster label, that is, we effectively abstain to make a classification decision for those items. In particular, we emphasize that the clustering method is not fitted again after selection (which would lead to bias in general).

The quality of the selection heavily relies on the appropriate quantification of the uncertainty of the cluster labels. For this, our approach is model-based, and can be viewed as a method that thresholds the posterior probabilities of the cluster labels with a data-driven choice of the threshold. The performance of the method will depend on the quality of the estimates of these posterior probabilities in the mixture model.

The adaptive character of our method is illustrated in Figure 3.1: when the clusters are well separated (panel (a)), the new procedure only discards few items and provides a clustering close to the correct one. However, when clusters are overlapping (panel (b)), to avoid a high misclassification error, the procedure discards most of the items and only provides few labels, for which the uncertainty is low. In both cases, the proportion of misclassified items among the selected ones is small and in particular close to the target level α (here 10%). Hence, by adapting the amount of labeled or discarded items, our method always delivers a reliable clustering result, inspite of the varying intrinsic difficulty of the clustering task.

3.1.3 Presentation of the results

Let us now describe in more details the main contributions of the paper.

- We introduce three new data-driven procedures that perform simultaneously selection

and clustering: the plug-in procedure (illustrated in Figure 3.1) and two bootstrap procedures (parametric and non-parametric), see Section 3.3.2.

- We provide a theoretical analysis of the plug-in procedure, quantifying the FMR deviation with respect to the target level α with explicit remainder terms, which become small when the sample size grows. In addition, this procedure is shown to satisfy the following optimality property: any other procedure that provides an FMR control necessarily classifies as many or less items than the plug-in procedure, up to a small remainder term (Theorem 17).

- Numerical experiments[1] establish that the bootstrap procedures improve the plug-in procedure, and thus are more reliable for practical use, where the sample size may be moderate, see Section 3.5.1. In particular, the FMR control is shown to be valid in various scenarios, including those where the overall misclassification risk (with no abstention option) is too large.

- Our analysis also shows that a fixed threshold procedure that only labels items with a maximum posterior probability larger than $1 - \alpha$ is generally suboptimal for an FMR control at level α, see Section 3.5.1. To this extent, our procedures can be seen as refined algorithms that classify more individuals while maintaining the FMR control.

- The practical impact of our approach is demonstrated on a real data set, see Section 3.5.2.

3.1.4 Relation to previous work

Other clustering guarantees in unsupervised learning While we provide a specific FMR control guarantee on the clustering, other criteria, not particularly linked to a rejection option, have been previously proposed in an unsupervised setting. Previous works provided essentially two types of guarantees: while early works focused on the probability of exact recovery (Arora and Kannan, 2005; Vempala and Wang, 2004; Abbe, 2018), recent contributions rather considered minimizing the misclassification risk (Lei and Rinaldo, 2015; Lu and Zhou, 2016; Giraud and Verzelen, 2018; Chretien et al., 2019). Other criteria include the probability to make a different decision than the Bayes rule (Azizyan et al., 2013), or the fact that all clusters are mostly homogeneous with high probability (Najafi et al., 2020). All these works provide a guarantee only if the setting is favorable enough. By contrast, providing a rejection option is the key to obtain a guarantee in any setting (in the worst situation, the procedure will not classify any item).

Comparison to Denis and Hebiri (2020) and Mary-Huard et al. (2021) We describe here two recent studies that are related to ours, because they also use a FMR-like criterion. The first one is the work of Denis and Hebiri (2020), which also relies on a thresholding of the (estimated) posterior probabilities. However, the control is different, because it does not provide an FMR control, but rather a type-II error control concerning the probability of classifying an item. Also, the proposed procedure therein requires an additional labeled sample (semi-supervised setting), which is not needed in our context.

The work of Mary-Huard et al. (2021) also proposes a control of the FMR. However, the analysis therein is solely based on the case where the model parameters are known (thus corresponding to the oracle case developed in Section 3.3.1 here). Compared to Mary-Huard

[1] We publicly release the code of these experiments at `https://github.com/arianemarandon/fmrcontrol`. We have also included a Jupyter notebook that demonstrates the use of our procedures.

et al. (2021), the present work provides number of new contributions, which are all given in Section 3.1.3. Let us also emphasize that we handle the label switching problem in the FMR, which seems to be overlooked in Mary-Huard et al. (2021).

Relation to the false discovery rate The FMR is closely related to the false discovery rate (FDR) in multiple testing, defined as the average proportion of errors among the discoveries. In fact, we can roughly view the problem of designing an abstention rule as testing, for each item i, whether the clustering rule correctly classifies item i or not. With this analogy, our selection rule is based on quantities similar to the local FDR values (Efron et al., 2001), a key quantity to build optimal FDR controlling procedures in multiple testing mixture models, see, e.g., Storey (2003); Sun and Cai (2007); Cai et al. (2019); Rebafka et al. (2022). In particular, our final selection procedure shares similarities with the procedure introduced in Sun and Cai (2007), also named cumulative ℓ-value procedure (Abraham et al., 2022). In addition, our theoretical analysis is related to the work of Rebafka et al. (2022), although the nature of the algorithm developed therein is different from here: they use the q-value procedure of Storey (2003), while our method rather relies on the cumulative ℓ-value procedure.

3.1.5 Organization of the paper

The paper is organized as follows: Section 3.2 introduces the model and relevant notation, namely the FMR criterion, with a particular care for the label switching problem. Section 3.3 presents the new methods: the oracle, plug-in and the bootstrap approaches. Our main theoretical results are provided in Section 3.4, after introducing appropriate assumptions. Section 3.5 presents numerical experiments and an application to a real data set, while a conclusion is given in Section 3.6. Proofs of the results and technical details are deferred to appendices.

3.2 Setting

This section presents the notation, model, procedures and criteria that will be used throughout the manuscript.

3.2.1 Model

Let $\mathbf{X} = (X_1, \dots, X_n)$ be an observed random sample of size n. Each X_i is an i.i.d. copy of a d-dimensional real random vector, which is assumed to follow the standard mixture model:

$$\begin{aligned} Z &\sim \mathcal{M}(\pi_1, \dots, \pi_Q), \\ X|Z = q &\sim F_{\phi_q}, \quad 1 \leq q \leq Q, \end{aligned}$$

where $\mathcal{M}(\pi_1, \dots, \pi_Q)$ denotes the multinomial distribution of parameter π (equivalently, $\pi_q = \mathbb{P}(Z = q)$ for each q). The model parameters are given by

- the probability distribution π on $\{1, \dots, Q\}$ that is assumed to satisfy $\pi_q > 0$ for all q. Hence, π_q corresponds to the probability of being in class q;
- the parameter $\phi = (\phi_1, \dots, \phi_Q) \in \mathcal{U}^Q$, where $\{F_u, u \in \mathcal{U}\}$ is a collection of distributions on $\mathbb{R}^d$. Every distribution F_u is assumed to have a density with respect to the Lebesgue measure on $\mathbb{R}^d$, denoted by f_u. Moreover, we assume that the ϕ_q's are all distinct.

The number of classes Q is assumed to be known and fixed throughout the manuscript (see Section 3.6 for a discussion). Thus, the overall parameter is $\theta = (\pi, \phi)$, the parameter set is denoted by Θ, and the distribution of (Z, X) is denoted by P_θ. The distribution family $\{P_\theta, \theta \in \Theta\}$ is the considered statistical model. We also assume that Θ is an open subset of $\mathbb{R}^K$ for some $K \geq 1$ with the corresponding topology.

In this mixture model, the latent vector $\mathbf{Z} = (Z_1, \ldots, Z_n)$ encodes a partition of the n observations into Q classes given by $\{1 \leq i \leq n \ : \ Z_i = q\}$, $1 \leq q \leq Q$. We refer to this model-based, random partition as the true latent clustering in the sequel.

In what follows, the "true" parameter that generates (Z, X) is assumed to be fixed and is denoted by $\theta^* \in \Theta$.

3.2.2 Procedure and criteria

Our approach starts with a given clustering rule, that aims at recovering the true latent clustering for all observed items. In general, a clustering rule is defined as a (measurable) function of the observation $\mathbf{X}$ returning a vector of labels $\widehat{\mathbf{Z}} = (\hat{Z}_i)_{1 \leq i \leq n} \in \{1, \ldots, Q\}^n$ for which the label q is assigned to individual i if and only if $\hat{Z}_i = q$. Note that in the unsupervised setting only the partition of the observations is of interest, not the labels themselves. Switching the labels of $\widehat{\mathbf{Z}}$ does not change the corresponding partition.

The classification error of $\widehat{\mathbf{Z}}$, with respect to specific labels, is given by $\varepsilon(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i=1}^n \mathbb{1}\{Z_i \neq \hat{Z}_i\}$. A label-switching invariant error is the clustering risk of $\widehat{\mathbf{Z}}$ defined by

$$R(\widehat{\mathbf{Z}}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma \in [Q]} \mathbb{E}_{\theta^*}\left(n^{-1}\varepsilon(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \mid \mathbf{X}\right)\right), \tag{3.1}$$

where $[Q]$ denotes the set of all permutations on $\{1, \ldots, Q\}$. The minimum over all permutations σ is the way to handle the aforementioned label-switching problem.

Remark 12. The position of the minimum w.r.t. σ in the risk (3.1) matters: the permutation σ is allowed to depend on X but not on Z. Hence, this risk has to be understood as being computed up to a data-dependent label switching. This definition coincides with the usual definition of the misclassification risk in the situation where the true clustering is deterministic, see Lei and Rinaldo (2015); Lu and Zhou (2016). Hence, it can be seen as a natural extension of the latter to a mixture model where the true clustering is random.

Classically, we aim to find a clustering rule $\widehat{\mathbf{Z}}$ such that the clustering risk is "small". However, as mentioned above, whether this is possible or not depends on the intrinsic difficulty of the clustering problem and thus of the true parameter θ^* (see Figure 3.1). Therefore, the idea is to provide a selection rule, that is, a (measurable) function of the observation $\mathbf{X}$ returning a subset of indices $S \subset \{1, \ldots, n\}$, such that the clustering risk *with restriction to* S is small. Throughout the paper, a *procedure* refers to a couple $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$, where $\widehat{\mathbf{Z}}$ is a clustering rule and S is a selection rule.

Definition 4 (False membership rate)**.** *The false membership rate (FMR) of a procedure* $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ *is given by*

$$\mathrm{FMR}_{\theta^*}(\mathcal{C}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma \in [Q]} \mathbb{E}_{\theta^*}\left(\frac{\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z})}{\max(|S|, 1)} \,\middle|\, \mathbf{X}\right)\right), \tag{3.2}$$

where $\varepsilon_S(\widehat{\mathbf{Z}}, \mathbf{Z}) = \sum_{i \in S} \mathbb{1}\{Z_i \neq \hat{Z}_i\}$ *denotes the misclassification error restricted to subset* S*.*

In this work, the aim is to find a procedure $\mathcal{C}$ such that the false membership rate is controlled at a nominal level α, that is, $\mathrm{FMR}_{\theta^*}(\mathcal{C}) \leq \alpha$. Obviously, choosing S empty implies

$\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) = 0$ a.s. for any permutation σ and thus satisfies this control. Hence, while maintaining the control $\mathrm{FMR}_{\theta^*}(\mathcal{C}) \leq \alpha$, we aim to classify as much individuals as possible, that is, to make $\mathbb{E}_{\theta^*}|S|$ as large as possible.

The definition of the FMR (3.2) involves an expectation of a ratio, which is more difficult to handle than a ratio of expectations. Hence, the following simpler alternative criterion will also be useful in our analysis.

Definition 5 (Marginal false membership rate)**.** *The marginal false membership rate (mFMR) of a procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ is given by*

$$\mathrm{mFMR}_{\theta^*}(\mathcal{C}) = \frac{\mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\varepsilon_S(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right)\right)}{\mathbb{E}_{\theta^*}(|S|)}, \tag{3.3}$$

with the convention $0/0 = 0$.

Note that the mFMR is similar to the criterion introduced in Denis and Hebiri (2020) in the supervised setting.

3.2.3 Notation

We will extensively use the following notation: for all $q \in \{1, \ldots, Q\}$ and $\theta = (\pi, \phi) \in \Theta$, we let

$$\ell_q(X, \theta) = \mathbb{P}_\theta(Z = q|X) = \frac{\pi_q f_{\phi_q}(X)}{\sum_{\ell=1}^{Q} \pi_\ell f_{\phi_\ell}(X)}; \tag{3.4}$$

$$T(X, \theta) = 1 - \max_{q\in\{1,\ldots,Q\}} \ell_q(X, \theta) \in [0, 1 - 1/Q]. \tag{3.5}$$

We can see that $\ell_q(X, \theta)$ is the posterior probability of belonging to class q given the measurement X under the distribution P_θ. The quantity $T(X, \theta)$ is a measure of the risk when classifying X: it is close to 0 when there exists a class q such that $\ell_q(X, \theta)$ is close to 1, that is, when X can be classified with large confidence.

3.3 Methods

In this section, we introduce new methods for controlling the FMR. We start by identifying an *oracle* method, that uses the true value of the parameter θ^*. Substituting the unknown parameter θ^* by an estimator in that oracle provides our first method, called the *plug-in* procedure. We then define a refined version of the plug-in procedure, that accounts for the variability of the estimator and is based on a *bootstrap* approach.

3.3.1 Oracle procedures

MAP clustering Here, we proceed as if an oracle had given us the true value of θ^* and we introduce an oracle procedure $\mathcal{C}^* = (\widehat{\mathbf{Z}}^*, S^*)$ based on this value. As the following lemma shows, the best clustering rule is well-known and given by the Bayes clustering $\widehat{\mathbf{Z}}^* = (\widehat{Z}_1^*, \ldots, \widehat{Z}_n^*)$, which can be written as

$$\widehat{Z}_i^* \in \operatorname*{argmax}_{q\in\{1,\ldots,Q\}} \ell_q(X_i, \theta^*), \;\; i \in \{1, \ldots, n\}, \tag{3.6}$$

where $\ell_q(\cdot)$ is the posterior probability given by (3.4).

Algorithm 3 Oracle procedure

Input: Parameter θ^*, sample $(X_1, \dots, X_n)$, level α.
1. Compute the posterior probabilities $\mathbb{P}_{\theta^*}(Z_i = q|X_i)$, $1 \le i \le n$, $1 \le q \le Q$;
2. Compute the Bayes clustering $\widehat{Z}_i^*$, $1 \le i \le n$, according to (3.6);
3. Compute the probabilities T_i^*, $1 \le i \le n$, according to (3.7);
4. Order these probabilities in increasing order $T_{(1)}^* \le \cdots \le T_{(n)}^*$;
5. Choose k^* the maximum of $k \in \{0, \dots, n\}$ such that $\max(k, 1)^{-1} \sum_{j=1}^{k} T_{(j)}^*(\mathbf{X}) \le \alpha$;
6. Select S_α^*, the index corresponding to the k^* smallest elements among the T_i^*'s.

Output: Oracle procedure $\mathcal{C}_\alpha = (\widehat{\mathbf{Z}}^*, S_\alpha^*)$.

Lemma 13. *We have* $\min_{\widehat{\mathbf{Z}}} R(\widehat{\mathbf{Z}}) = R(\widehat{\mathbf{Z}}^*) = n^{-1} \sum_{i=1}^{n} \mathbb{E}_{\theta^*}(T_i^*)$*, for the Bayes clustering* $\widehat{\mathbf{Z}}^*$ *defined by* (3.6) *and for*

$$T_i^* = T(X_i, \theta^*) = \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i^*|X_i), \;\; i \in \{1, \dots, n\}, \tag{3.7}$$

where $T(\cdot)$ *is given by* (3.5).

In words, Lemma 13 states that the oracle statistics T_i^* correspond to the posterior misclassification probabilities of the Bayes clustering. To decrease the overall misclassification risk, it is natural to avoid classification of points with a high value of the test statistic T_i^*.

Thresholding selection rules In this section, we introduce the selection rule, that decides which items are to be classified. From the above paragraph, it is natural to consider a thresholding-based selection rule of the form $S = \{i \in \{1, \dots, n\} \; : \; T_i^* \le t\}$, for some threshold t to be chosen suitably. The following result gives insights for the choice of such a threshold t.

Lemma 14. *For a procedure* $\mathcal{C} = (\widehat{\mathbf{Z}}^*, S)$ *with Bayes clustering and an arbitrary selection* S*,*

$$\mathrm{FMR}_{\theta^*}(\mathcal{C}) = \mathbb{E}_{\theta^*}\left(\frac{\sum_{i \in S} T_i^*}{\max(|S|, 1)}\right). \tag{3.8}$$

As a consequence, a first way to build an (oracle) selection is to set

$$S = \{i \in \{1, \dots, n\} \; : \; T_i^* \le \alpha\}.$$

Since an average of numbers smaller than α is also smaller than α, the corresponding procedure controls the FMR at level α. This procedure is referred to as the procedure *with fixed threshold* in the sequel. It corresponds to the following naive approach: to get a clustering with a risk of α, we only keep the items that are in their corresponding class with a posterior probability of at least $1 - \alpha$. By contrast, the selection rule considered here is rather

$$S = \{i \in \{1, \dots, n\} \; : \; T_i^* \le t(\alpha)\},$$

for a threshold $t(\alpha) \ge \alpha$ maximizing $|S|$ under the constraint $\sum_{i \in S} T_i^* \le \alpha |S|$. It uniformly improves the procedure with fixed threshold and will in general lead to a (much) broader selection. This gives rise to the *oracle procedure*, that can be easily implemented by ordering the T_i^*'s, see Algorithm 3.

Algorithm 4 Plug-in procedure

Input: Sample $(X_1, \dots, X_n)$, level α.
1. Compute an estimator $\hat{\theta}$ of θ;
2. Run the oracle procedure given in Algorithm 3 with $\hat{\theta}$ in place of θ^*.
Output: Plug-in procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$.

Algorithm 5 Bootstrap procedure

Input: Sample $(X_1, \dots, X_n)$, level α, number B of bootstrap runs.
1. Choose a grid of increasing levels $(\alpha(k))_{1 \le k \le K}$;
2. Compute $\widehat{\mathrm{FMR}}^{B}_{\alpha(k)}$, $1 \le k \le K$, according to (3.10);
3. Choose $\tilde{k}$ according to (3.11).
Output: Bootstrap procedure $\widehat{\mathcal{C}}_\alpha^{\mathrm{boot}} = \widehat{\mathcal{C}}_{\alpha(\tilde{k})}^{\mathrm{PI}}$.

3.3.2 Empirical procedures

Plug-in procedure The oracle procedure cannot be used in practice since θ^* is generally unknown. A natural idea then is to approach θ^* by an estimator $\hat{\theta}$ and to plug this estimate into the oracle procedure. The resulting procedure, denoted $\widehat{\mathcal{C}}^{\mathrm{PI}} = (\widehat{\mathbf{Z}}^{\mathrm{PI}}, \widehat{S}_\alpha^{\mathrm{PI}})$, is called the *plug-in procedure* and is implemented in Algorithm 4.

In Section 3.4, we establish that the plug-in procedure has suitable properties: when n tends to infinity, provided that the chosen estimator $\hat{\theta}$ behaves well and under mild regularity assumptions on the model, the FMR of the plug-in procedure is close to the level α, while it is nearly optimal in terms of average selection number.

Bootstrap procedure Despite the favorable theoretical properties shown in Section 3.4, the plug-in procedure achieves an FMR that can exceed α in some situations, as we will see in our numerical experiments (Section 3.5). This is in particular the case when the estimator $\hat{\theta}$ is too rough. Indeed, the uncertainty of $\hat{\theta}$ near θ^* is ignored by the plug-in procedure.

To take into account this effect, we propose to use a bootstrap approach. It is based on the following result.

Lemma 15. *For a given level $\alpha \in (0,1)$, the FMR of the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{PI}$ is given by*

$$\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{PI}) = \mathbb{E}_{\mathbf{X} \sim P_{\theta^*}} \left(\min_{\sigma \in [Q]} \frac{\sum_{i=1}^n \{1 - \ell_{\sigma(\hat{Z}_i^{PI}(\mathbf{X}))}(X_i, \theta^*)\} \, \mathbb{1}\{i \in \hat{S}_\alpha^{PI}(\mathbf{X})\}}{\max(|\hat{S}_\alpha^{PI}(\mathbf{X})|, 1)} \right). \tag{3.9}$$

The general idea is as follows: since $\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}})$ can exceed α, we choose α' as large as possible such that $\widehat{\mathrm{FMR}}_{\alpha'} \le \alpha$, for which $\widehat{\mathrm{FMR}}_{\alpha'}$ is a bootstrap approximation of $\mathrm{FMR}(\widehat{\mathcal{C}}_{\alpha'}^{\mathrm{PI}})$ based on (3.9).

The bootstrap approximation reads as follows: in the RHS of (3.9), we replace the true parameter θ^* by $\hat{\theta}$ and $\mathbf{X} \sim P_{\theta^*}$ by $\mathbf{X}' \sim \hat{P}$, where $\hat{P}$ is an empirical substitute of P_{θ^*}. This empirical distribution $\hat{P}$ is $P_{\hat{\theta}}$ for the parametric bootstrap and the uniform distribution over the X_i's for the non-parametric bootstrap. This yields the bootstrap approximation of $\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}})$ given by

$$\widehat{\mathrm{FMR}}_\alpha = \mathbb{E}_{\mathbf{X}' \sim \hat{P}} \left(\min_{\sigma \in [Q]} \frac{\sum_{i=1}^n \{1 - \ell_{\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}'))}(X_i', \hat{\theta}(\mathbf{X}))\} \, \mathbb{1}\{i \in \hat{S}_\alpha^{\mathrm{PI}}(\mathbf{X}')\}}{\max(|\hat{S}_\alpha^{\mathrm{PI}}(\mathbf{X}')|, 1)} \,\middle|\, \mathbf{X} \right).$$

Classically, the latter is itself approximated by a Monte-Carlo scheme:

$$\widehat{\mathrm{FMR}}^{B}_{\alpha} = \frac{1}{B}\sum_{b=1}^{B} \min_{\sigma\in[Q]} \frac{\sum_{i=1}^{n}\{1-\ell_{\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}^b))}(X_i^b, \hat{\theta}(\mathbf{X}))\}\,\mathbb{1}\{i\in \hat{S}_{\alpha}^{\mathrm{PI}}(\mathbf{X}^b)\}}{\max(|\hat{S}_{\alpha}^{\mathrm{PI}}(\mathbf{X}^b)|, 1)}, \tag{3.10}$$

with $\mathbf{X}^1, \dots, \mathbf{X}^B$ i.i.d. $\sim \hat{P}$ corresponding to the bootstrap samples of $\mathbf{X}$.

Let $(\alpha(k))_{1\le k\le K} \in (0,1)^K$ be a grid of increasing nominal levels (possibly with restriction to values slightly below the target level α). Then, the bootstrap procedure at level α is defined as $\widehat{\mathcal{C}}_{\alpha}^{\mathrm{boot}} = \widehat{\mathcal{C}}_{\alpha(\tilde{k})}^{\mathrm{PI}}$, where

$$\tilde{k} = \max\left\{k\in\{1,\dots,K\} \,:\, \widehat{\mathrm{FMR}}^{B}_{\alpha(k)} \le \alpha\right\}. \tag{3.11}$$

This procedure is implemented in Algorithm 5.

Remark 13 (Parametric versus non parametric bootstrap). The usual difference between parametric and non parametric bootstrap also holds in our context: the parametric bootstrap is fully based on $P_{\hat{\theta}}$, while the non parametric bootstrap builds an artificial sample (with replacement) from the original sample, which does not come from a P_θ-type distribution. This gives rise to different behaviors in practice: when $\hat{\theta}$ is too optimistic (which will be typically the case here when the estimation error is large), the correction brought by the parametric bootstrap (based on $P_{\hat{\theta}}$) is often weaker than that of the non parametric one. By contrast, when $\hat{\theta}$ is close to the true parameter, the parametric bootstrap approximation is more faithful because it uses the model, see Section 3.5.

3.4 Theoretical guarantees for the plug-in procedure

In this section, we derive theoretical properties for the plug-in procedure: we show that its FMR and mFMR are close to α, while its expected selection number is close to be optimal under some conditions.

3.4.1 Additional notation and assumptions

We make use of an optimality theory for mFMR control, that will be developed in detail in Section B.1.2. This approach extensively relies on the following quantities (recall the definition of $T(X,\theta)$ in (3.5)):

$$\mathrm{mFMR}_t^* = \mathbb{E}_{\theta^*}\left(T(X,\theta^*) \,|\, T(X,\theta^*) < t\right); \tag{3.12}$$

$$t^*(\alpha) = \sup\{t\in[0,1] \,:\, \mathrm{mFMR}_t^* \le \alpha\} \tag{3.13}$$

$$\alpha_c = \inf\{\mathrm{mFMR}_t^* \,:\, t\in(0,1], \mathrm{mFMR}_t^* > 0\}; \tag{3.14}$$

$$\bar{\alpha} = \mathrm{mFMR}_1^*. \tag{3.15}$$

In words, mFMR_t^* is the mFMR of an oracle procedure that selects the T_i^* smaller than some threshold t (Lemma 36). Then, $t^*(\alpha)$ is the optimal threshold such that this procedure has an mFMR controlled at level α. Next, α_c and $\bar{\alpha}$ are the lower and upper bounds for the nominal level α, respectively, for which the optimality theory can be applied.

Now, we introduce our main assumption, which will be ubiquitous in our analysis.

Assumption 6. *For all $\theta\in\Theta$ and $q\in\{1,\dots,Q\}$, under P_{θ^*}, the r.v. $\ell_q(X,\theta)$ given by (3.4) is continuous. In addition, the function $t\mapsto \mathbb{P}_{\theta^*}(T(X,\theta)<t)$ is increasing on $(\alpha_c,\bar{\alpha})$, where $T(X,\theta)$ is given by (3.5).*

Note that Assumption 6 implies the continuity of the r.v. $T(X,\theta)$. Indeed, $\mathbb{P}(T(X,\theta) = t) \leq \sum_{q=1}^{Q} \mathbb{P}(\ell_q(X,\theta) = 1-t)$. Hence, this assumption implies that $t \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) < t)$ is both continuous on $[0,1]$ and increasing on $(\alpha_c, \bar{\alpha})$. This is useful in several regards: first, it prohibits ties in the $T(X_i,\theta)$'s, $1 \leq i \leq m$, so that the selection rule (see Algorithm 3) can be truly formulated as a thresholding rule (see Lemma 37). Second, it entails interesting properties for function $t \mapsto \mathrm{mFMR}_t^*$, see Lemma 36 (this in particular ensures that the supremum in (3.13) is a maximum). Also note that the inequality $0 \leq \alpha_c < \bar{\alpha} < 1 - 1/Q$ holds under Assumption 6.

The next assumption ensures that the density family $\{f_u, u \in \mathcal{U}\}$ is smooth, and will be useful to establish consistency results.

Assumption 7. *For P_{θ^*}-almost all $x \in \mathbb{R}^d$, $u \in \mathcal{U} \mapsto f_u(x)$ is continuous.*

Moreover, we can derive convergence rates under the following additional regularity conditions.

Assumption 8. *There exist positive constants $r = r(\theta^*), C_1 = C_1(\theta^*), C_2 = C_2(\theta^*,\alpha), C_3 = C_3(\theta^*,\alpha)$ such that*

(i) *for $\mathbb{P}_{\theta^*}$-almost all x, $u \in \mathcal{U} \mapsto f_u(x)$ is continuously differentiable, and*

$$\sum_{1 \leq q \leq Q} \mathbb{E}_{\theta^*} \sup_{\substack{\theta \in \Theta \\ \|\theta - \theta^*\| \leq r}} \|\nabla_\theta \ell_q(X,\theta)\| \leq C_1;$$

(ii) *for all $t, t' \in [0,1]$, $|\,\mathbb{P}_{\theta^*}(T(X,\theta^*) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t')| \leq C_2|t - t'|$;*

(iii) *for all $\beta \in [(\alpha_c + \alpha)/2, (\alpha + \bar{\alpha})/2]$, $|t^*(\beta) - t^*(\alpha)| \leq C_3|\beta - \alpha|$.*

Example 2. In Appendix B.4, it is proved that Assumptions 6, 7 and 8 hold true in the homoscedastic two-component multivariate Gaussian mixture model, see Lemma 46.

Next, we consider the following complexity assumption to ensure concentration of the underlying empirical processes. It is given in terms of the VC dimension of specific function classes involving ℓ_q. In the sequel, the VC dimension of a function set $\mathscr{F}$ is defined as the VC dimension of the set family $\{\{x \in \mathbb{R}^d : f(x) \geq u\}, f \in \mathscr{F}, u \in \mathbb{R}\}$, see, e.g., Baraud (2016). We denote

$$\mathscr{V} = \text{ VC dimension of } \{\ell_q(.,\theta), \theta \in \Theta, 1 \leq q \leq Q\}; \tag{3.16}$$

$$\mathscr{V}_- = \text{ VC dimension of } \{\mathbb{1}\{\ell_q(.,\theta) - \ell_{q'}(.,\theta) \geq 0\}, \theta \in \Theta, 1 \leq q, q' \leq Q\}. \tag{3.17}$$

Assumption 9. *The VC dimensions $\mathscr{V}$ and $\mathscr{V}_-$ are finite.*

Example 3. In the two-component case $Q = 2$ where P_θ belongs to an exponential family, we have that $\mathscr{V}, \mathscr{V}_- \lesssim k^2 \log(k)$ (see Lemma 41) with k the dimension of the sufficient statistic vector. For instance, $k = d + d^2$ for the Gaussian family, hence $\mathscr{V}, \mathscr{V}_- \lesssim d^4 \log(d)$ in that case. (For the specific case of the homoscedastic Gaussian family, we have that $\mathscr{V}, \mathscr{V}_- \lesssim d$, see Lemma 47).

Let us now discuss conditions on the estimator $\hat{\theta}$ on which the plug-in procedure is based. We start by introducing the following assumption (used in the concentration part of the proof, see Lemma 39).

Assumption 10. *The estimator $\hat{\theta}$ is assumed to take its values in a countable subset $\mathcal{D}$ of Θ.*

This assumption is a minor restriction, because we can always choose $\mathcal{D} \subset \mathbb{Q}^K$ (recall $\Theta \subset \mathbb{R}^K$). Next, we additionally define a quantity measuring the quality of the estimator: for all $\epsilon > 0$,

$$\eta(\epsilon, \theta^*) = \mathbb{P}_{\theta^*}\left(\min_{\sigma \in [Q]} \|\hat{\theta}^\sigma - \theta^*\|_2 \geq \epsilon\right). \tag{3.18}$$

Example 4. The literature provides several results regarding the estimation of Gaussian mixtures, see e.g. Regev and Vijayaraghavan (2017) for a review. Proposition 44 revisits some of these results, for the estimator derived from EM algorithm (Dempster et al., 1977; Balakrishnan et al., 2017) and the constrained MLE (Ho and Nguyen, 2016).

3.4.2 Results

We now state our main results, starting with the consistency of the plug-in procedure.

Theorem 16 (Asymptotic optimality of the plug-in procedure)**.** *Let Assumptions 6, 7, and 9 be true. Consider an estimator $\hat{\theta}$ satisfying Assumption 10 and which is consistent in the sense that for all $\epsilon > 0$, the probability $\eta(\epsilon, \theta^*)$ given by* (3.18) *tends to* 0 *as n tends to infinity. Then the corresponding plug-in procedure $\widehat{\mathcal{C}}_\alpha^{PI}$ (Algorithm 4) satisfies the following: for any $\alpha \in (\alpha_c, \bar{\alpha})$, we have*

$$\limsup_n \mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{PI}) \leq \alpha, \quad \limsup_n \mathrm{mFMR}(\widehat{\mathcal{C}}_\alpha^{PI}) \leq \alpha,$$

and for any procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ that controls the mFMR at level α, we have

$$\liminf_n \{n^{-1}\,\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{PI}|) - n^{-1}\,\mathbb{E}_{\theta^*}(|S|)\} \geq 0.$$

Next, we derive convergence rates under the additional regularity conditions given by Assumption 8.

Theorem 17 (Optimality of the plug-in procedure with rates)**.** *Consider the setting of Theorem 16, where in addition Assumption 8 holds. Recall $\eta(\epsilon, \theta^*)$ defined by* (3.18) *and $\mathscr{V}, \mathscr{V}_-$ defined by* (3.16), (3.17) *respectively. Let s^* denote the selection rate of the oracle procedure mentioned in Section 3.4.1, with threshold $t^*(\alpha)$ and applied at level $(\alpha + \alpha_c)/2$. With constants $A > 0$ and $B > 0$ only depending on $Q, C_1, C_2, C_3, \mathscr{V}, \mathscr{V}_-$ and s^*, we have for any sequence $\epsilon_n > 0$ tending to zero, for n larger than a constant only depending on α and θ^*,*

$$\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{PI}) \leq \alpha + A\sqrt{\epsilon_n} + B\sqrt{\log n/n} + 5/n^2 + \eta(\epsilon_n, \theta^*) \tag{3.19}$$

$$n^{-1}\,\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{PI}|) - n^{-1}\,\mathbb{E}_{\theta^*}(|S|) \geq -A\sqrt{\epsilon_n} - B\sqrt{\log n/n} - 5/n^2 - \eta(\epsilon_n, \theta^*), \tag{3.20}$$

for any procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ that controls the mFMR at level α.

The proof is based on a more general non-asymptotical result, for which the remainder terms are more explicit, see Theorem 31 and Appendix B.1. It employs techniques that share similarities with the work of Rebafka et al. (2022) developed in a different context. Here, a difficulty is to handle the new statistic $T(X_i, \hat{\theta})$ which is defined as an extremum, see (3.5).

Theorem 17 establishes that, given a model which is regular enough and a consistent estimator, the plug-in procedure controls the FMR and is asymptotically optimal up to remainder terms which are of the order of $\sqrt{\epsilon_n} + \sqrt{\log n/n} + \eta(\epsilon_n, \theta^*)$. Here, ϵ_n dominates the convergence rate of the parameter estimate, and is taken large enough to ensure that $\eta(\epsilon_n, \theta^*)$ vanishes.

For instance, in the multivariate Gaussian mixture model (with further assumptions) and by considering either the EM estimator or the constrained MLE, we have $\eta(\epsilon_n, \theta^*) \leq 1/n$ for $\epsilon_n = C\sqrt{\log(n)/n}$, see Proposition 44. This implies that the remainder terms in (3.19) and (3.20) are at most of order $((\log n)/n)^{1/4}$.

3.5 Experiments

In this section, we evaluate the behavior of the new procedures: plug-in (Algorithm 4), parametric bootstrap and non parametric bootstrap (Algorithm 5). For this, we use both synthetic and real data.

3.5.1 Synthetic data set

The performance of our procedures is studied via simulations in different settings with various difficulties. All of them are Gaussian mixture models, with possible restrictions on the parameter space. For parameter estimation, the classical EM algorithm is applied with 100 iterations and 10 starting points chosen with Kmeans++ (Arthur and Vassilvitskii, 2006). In the bootstrap procedures $B = 1000$ bootstrap samples are generated. The performance of all procedures is assessed via the *sample FMR* and the proportion of classified data points, which is referred to as the *selection frequency*. For every setting and every set of parameters, depicted results display the mean over 100 simulated datasets. As a baseline, we consider the fixed threshold procedure in which one selects data points that have a maximum posterior group membership probability that exceeds $1-\alpha$. The oracle procedure (Algorithm 3) is also considered in our experiments for comparison.

Known proportions and covariances In the first setting, the true mixture proportions and covariance matrices are known and used in the EM algorithm. We consider the case $Q = 2$, $\pi_1 = \pi_2 = 1/2$ and $\Sigma_1 = \Sigma_2 = I_d$ with I_d the $(d \times d)$-identity matrix. For the mean vectors, we set $\mu_1 = 0$ and $\mu_2 = (\epsilon/\sqrt{d}, \ldots, \epsilon/\sqrt{d})$. The quantity ϵ corresponds to the mean separation, that is, $\|\mu_1 - \mu_2\|_2 = \epsilon$ and accounts for the difficulty of the clustering problem.

Figure 3.2 displays the FMR for nominal level $\alpha = 0.1$, sample size $n = 100$, dimension $d = 2$ and varying mean separation $\epsilon \in \{1, \sqrt{2}, 2, 4\}$. Globally, our procedures all have an FMR close to the target level α (excepted for the very well separated case $\epsilon = 4$ for which the FMR is much smaller because a large part of the items can be trivially classified). In addition, the selection rate is always close to the one of the oracle procedure. On the other hand, the baseline procedure is too conservative: its FMR can be well below the nominal level and it selects up to 50% less than the other procedures. This is well expected, because unlike our procedures, the baseline has a fixed threshold and thus does not adapt to the difficulty of the problem.

We also note that the FMR of the plug-in approach is slightly inflated for a weak separation ($\epsilon = 1$). This comes from the parameter estimation, which is difficult in that case. This also illustrates the interest of the bootstrap methods, that allow to recover the correct level in that case, by appropriately correcting the plug-in approach.

Diagonal covariances In this setting, the true parameters are the same as in the previous paragraph, but the true mixture proportions and covariance matrices are unknown. However, to help the estimation, we suppose a diagonal structure for Σ_1 and Σ_2, which is used in the EM algorithm.

Figure 3.3a displays the FMR and the selection frequency as a function of the separation ϵ. The conclusion is qualitatively the same as in the previous case, but with larger FMR values for a weak separation. Overall, it shows that the plug-in procedure is anti-conservative and that the bootstrap corrections are able to recover an FMR and a selection frequency close to the one of the oracle. However, for a weak separation, namely $\epsilon = 1$, the parametric bootstrap correction is not enough and the latter procedure still overshoots the nominal level α. Indeed, in our simulations, it appears that $P_{\hat{\theta}}$ is often a distribution that is more favorable

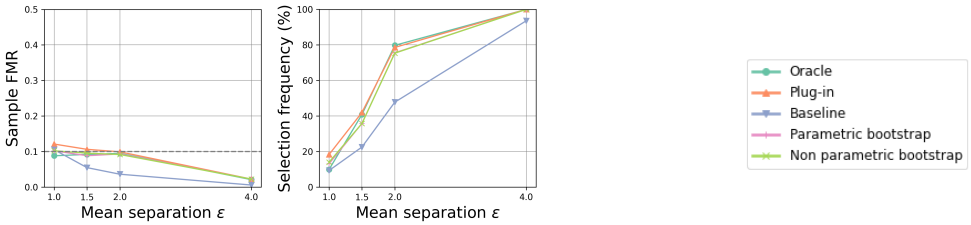


Figure 3.2: FMR (left panel) and selection frequency (right panel) as a function of the mean separation ϵ. Known mixture proportions and covariances setting with $Q = 2$, $n = 100$, $d = 2$, $\alpha = 0.1$.

than P_{θ^*} from a statistical point of view (for instance, with more separated clusters). These conclusions also hold for varying sample size n, see Figure 3.3b.

Figure 3.3c displays the FMR and the selection frequency for varying nominal level α, with $\epsilon = \sqrt{2}$ and $n = 200$. The plug-in is still anti-conservative, while the bootstrap procedures have an FMR that is close to α uniformly on the considered α range. Moreover, we note that for all our procedures (including the plug-in), the gap between the FMR and the nominal level is roughly constant with α: this illustrates the adaptive aspect of our procedures. This is in contrast with the baseline procedure, for which this gap highly depends on α, and which may be either anti-conservative or sub-optimal depending on the α value.

Three-component mixture We next increase the number of classes to $Q = 3$. Figure 3.4 displays the FMR and the selection frequency for varying α, with a mean separation $\|\mu_1 - \mu_2\|_2 = \|\mu_1 - \mu_3\|_2 = \|\mu_2 - \mu_3\|_2 = \epsilon = 3$. The mean separation is chosen so that the selection frequency of the oracle rule is approximately the same as in the previous paragraph. The increase in Q leads to a deterioration of the performances. Specifically, the FMR of the plug-in overshoots the nominal level by a large amount, and when n is too small, the parametric bootstrap procedure can be anti-conservative while the non parametric bootstrap is over-conservative. This deterioration is expected since from the theory established in Section 3.4 (see Theorem 17), the residual terms increase with Q, and since the difficulty of the estimation is also increased. However, for a fairly large sample size ($n = 1000$), both bootstrap procedures are correctly mimicking the oracle.

Larger dimension We now increase the dimension to $d = 10$. In that case, parameter estimation is deteriorated. In particular, the maximum posterior probability for any point tends to be very over-estimated. To remedy this issue, we project the data onto a two-dimensional space using PCA. We then apply the EM algorithm to the projected data. This is similar in spirit to spectral clustering and it has the added benefit of combining the objectives of data reduction with clustering. Results are displayed in Figure 3.5. The conclusions are qualitatively the same as in the previous paragraph.

3.5.2 Real data set

We consider the Wisconsin Breast Cancer Diagnosis (WDBC) dataset from the UCI ML repository. The data consists of features computed from a digitalized image of a fine needle aspirate (FNA) of a breast mass, on a total of 569 patients (each corresponds to one FNA sample) of which 212 are diagnosed as Benign and 357 as Malignant. Ten real-valued measures were computed for each of the cell nucleus present in the images (e.g. radius, perimeter,

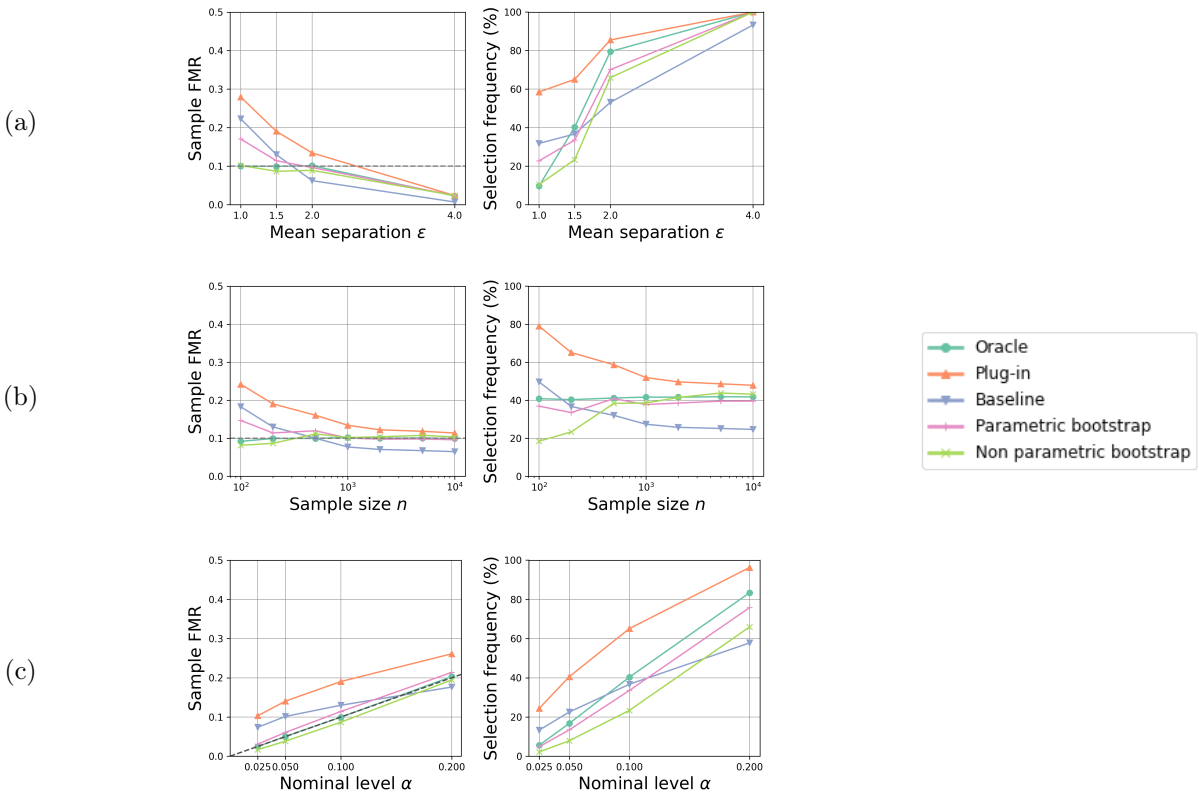
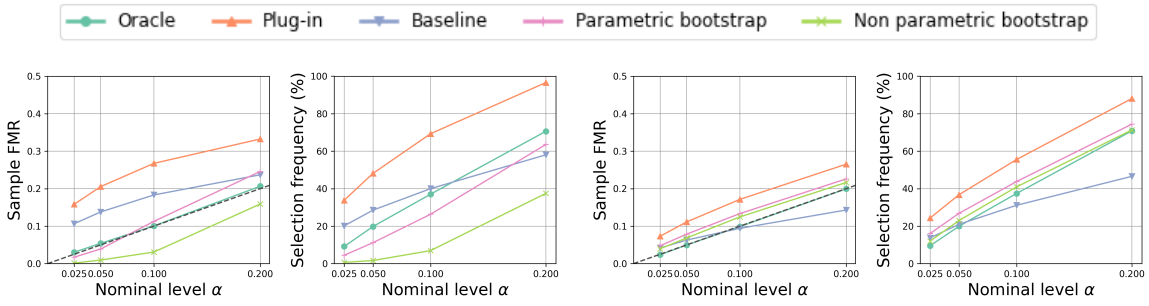
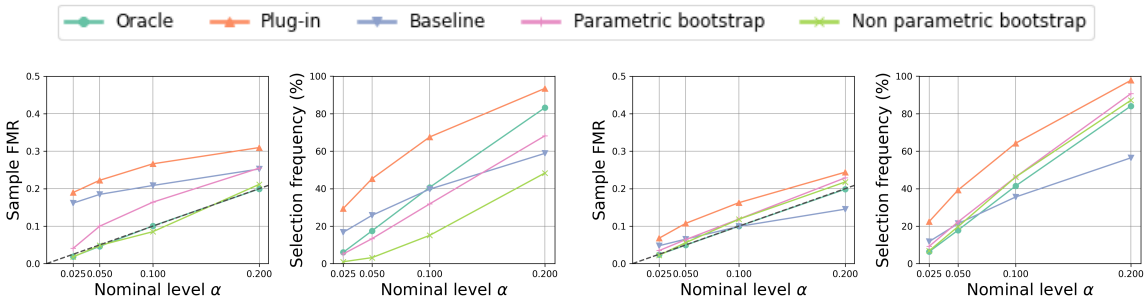


Figure 3.3: FMR (left panel) and selection frequency (right panel) as a function of: (a) the mean separation; (b) the sample size n; (c) the nominal level α. Diagonal covariances setting with $Q = 2$, $d = 2$. Default settings are: $n = 200$, $\alpha = 0.1$, $\epsilon = \sqrt{2}$.



(a) $n = 200$ (b) $n = 1000$

Figure 3.4: FMR (left panel) and selection frequency (right panel) as a function of the nominal level α. Diagonal covariances setting with $Q = 3$, $d = 2$, $\epsilon = 3$.



(a) $n = 200$ (b) $n = 1000$

Figure 3.5: FMR (left panel) and selection frequency (right panel) as a function of the nominal level α. Diagonal covariances setting with $Q = 2$, $d = 10$, $\epsilon = \sqrt{2}$.

texture, etc.). Then, the mean, standard error and mean of the three largest values of these measures were computed for each image, resulting in a total of 30 features. Here, we restrict the analysis to the variables that correspond to the means of these measures.

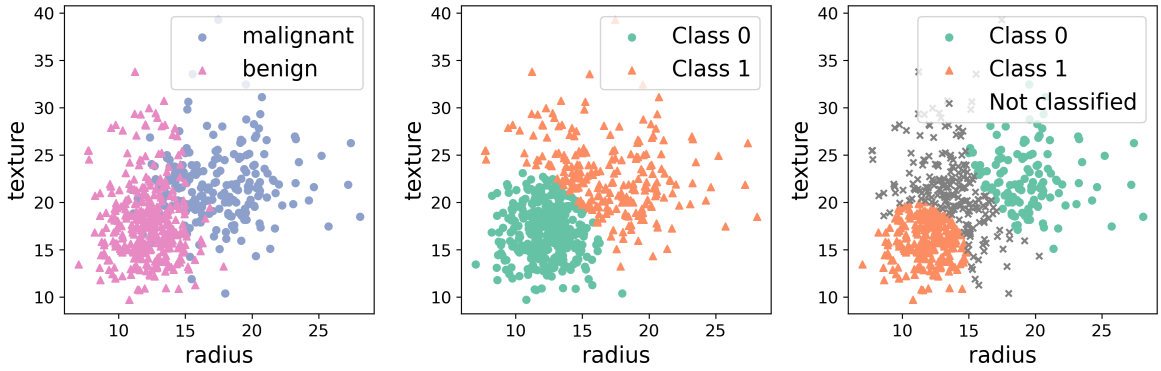
We choose to model the data as a mixture of Student's t-distributions as proposed in Peel and McLachlan (2000). Student mixtures are appropriate for data containing observations with longer than normal tails or atypical observations leading to overlapping clusters. Compared to Gaussian mixtures, Students are less concentrated and thus produce estimates of the posterior probabilities of class memberships that are less extreme, which is favorable for our selection procedures. In our study, the degree of freedom of each component is set to 4, and no constraints are put on the rest of the parameters. The t-mixture is fit via the EM algorithm provided by the Python package `studenttmixture` (Peel and McLachlan, 2000).

To start with, we restrict the analysis to the first two variables of the dataset, the mean radius and the mean texture of the images. For illustration, Figure 3.6 (panel (a)) displays the data. Different colors indicate the ground truth labels (this information is not used in the clustering). One can see that the Student approximation is fairly good for each of the groups, and there is some overlap between them. Figure 3.6 (panel (b)) displays the MAP clustering result for the t-mixture model without any selection. The FMR is computed with respect to the ground truth labels and amounts to 14 %. Figure 3.6 (panel (c)) provides the result of our parametric bootstrap procedure with nominal level $\alpha = 5\%$. The procedure does not classify points that are at the intersection of the clusters, resulting in the classification of 70% of the data, and the FMR equals 3%, which is below the target level.

Finally, Figure 3.7 displays the results when restricting the analysis to the first ten variables of the dataset and applying PCA to reduce the dimension to 2. In that case, the FMR computed with respect to the ground truth labels without selection is 14 %, while using the bootstrap procedures, this reduces to 10 %, with a selection frequency of 80%.

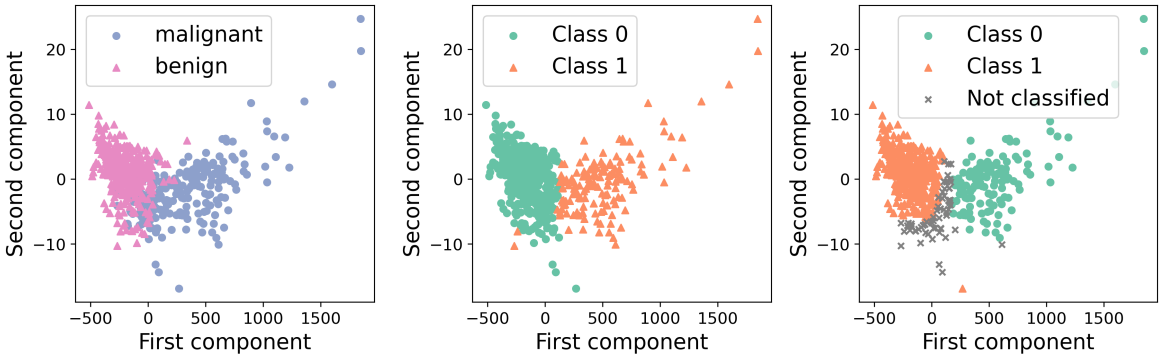
3.6 Conclusion and discussion

We have presented new data-driven methods providing both clustering and selection that ensure an FMR control guarantee in a mixture model. The plug-in approach was shown to be theoretically valid both when the parameter estimation is accurate and the sample size is large enough. When this is not necessarily the case, we proposed two second-order bootstrap corrections that have been shown to increase the FMR control ability on numerical experiments. Finally, applying our unsupervised methods to a supervised data set, our approach has been



(a) Ground truth labels (b) MAP clustering (c) Bootstrap procedure

Figure 3.6: Comparison of the clustering result using t-mixture modelling with ground truth labels on the WDBC dataset, restricted to the variables *radius* and *texture*, with and without selection. With the parametric bootstrap procedure applied at $\alpha = 5\%$, the FMR w.r.t. the ground truth labels is of 3% versus 14% without selection.



(a) Ground truth labels (b) MAP clustering (c) Bootstrap procedure

Figure 3.7: Comparison of the clustering result using PCA and t-mixture modelling with ground truth labels on the WDBC dataset, restricted to the first ten variables, with and without selection. With the parametric bootstrap procedure applied at $\alpha = 5\%$, the FMR w.r.t. the ground truth labels is of 10% versus 14% without selection.

qualitatively validated by considering the attached labels as revealing the true clusters.

We underline that the cluster number Q is assumed to be fixed and known throughout the study. In practice, it can be fitted from the data by using the standard AIC or BIC criteria, using the entire data before application of the selection rule. In addition, if several values of Q make sense from a practical viewpoint, we recommend to provide to the practitioner the collection of the corresponding outputs.

Concerning the pure task of controlling the FMR in the mixture model, our methods provide a correct FMR control in some region of the parameter space, leaving other less favorable parameter configurations with a slight inflation in the FMR level. This phenomenon is well known for FDR control in the two-component mixture multiple testing model (Sun and Cai, 2007; Roquain and Verzelen, 2022), and facing a similar problem in our framework is well expected. On the one hand, in some cases, this problem can certainly be solved by improving on parameter estimation: here the EM algorithm seems to over-estimate the extreme posterior probabilities, which makes the plug-in procedure too anti-conservative. On the other hand, it could be hopeless to expect a robust FMR control uniformly valid over all configurations, while being optimal in the favorable cases. To illustrate that point, we refer to the work Roquain and Verzelen (2022) that shows that such a procedure does not exist in the FDR controlling case, when the null distribution is Gaussian with an unknown scaling parameter (which is a framework sharing similarities with the one considered here). Investigating such a "lower bound" result in the current setting would provide better guidelines for the practitioner and is therefore an interesting direction for future research. In addition, in these unfavorable cases, adding labeled samples and considering a semi-supervised framework can be an appropriate alternative for practical use. This new sample is likely to considerably improve the inference. Studying the FMR control in that setting is another promising avenue.

Chapter 4

Link prediction with FDR control

Most link prediction methods return estimates of the connection probability of missing edges in a graph. Such output can be used to rank the missing edges from most to least likely to be a true edge, but it does not directly provide a classification into true and non-existent. In this work, we consider the problem of identifying a set of true edges with a control of the false discovery rate (FDR). We propose a novel method based on high-level ideas from the literature on conformal inference. The graph structure induces intricate dependence in the data, which we carefully take into account, as this makes the setup different from the usual setup in conformal inference, where exchangeability is assumed. The FDR control is empirically demonstrated for both simulated and real data.

Contents

4.1 Introduction 65
4.1.1 Problem and aim 65
4.1.2 Approach 67
4.1.3 Relation to previous work 68
4.2 Methodology 69
4.2.1 Preliminaries 69
4.2.2 Procedure 70
4.2.3 Training the scoring function g 71
4.3 Numerical experiments 74
4.3.1 Simulated data 75
4.3.2 Real data 76
4.4 Discussion 77

4.1 Introduction

4.1.1 Problem and aim

Graphs (or networks) denote data objects that consist of links (edges) between entities (nodes). Real-world examples are ubiquitous and include social networks, computer networks, food webs, molecules, etc. A fundamental problem in network analysis is link prediction, where the goal is to identify missing links in a partially observed graph. Biological networks such as protein-protein interaction networks (Kovács et al., 2019) or food webs (Terry and Lewis, 2020) are typical examples of incomplete networks: because experimental discovery of interactions is costly, many interactions remain unrecorded. Link prediction can be used to identify

promising pairs of nodes for subsequent experimental evaluations. Other applications include friend or product recommendation (Li and Chen, 2013), or identification of relationships between terrorists (Clauset et al., 2008).

In this work, we consider a link prediction problem, where a graph with a set of vertices $V = \{1, \dots, n\}$ and a set of edges E is only partially observed: namely, we observe a sample of node pairs recorded as interacting (true edges) and a sample of node pairs recorded as non-interacting (false edges). The graph can be directed or undirected and self-loops are allowed. The two observed samples of node pairs make up only a part of the set of all pairs $V \times V$, and the non-observed pairs correspond to missing information, where it is not known whether there is an edge or not. The aim is to identify the true edges among the pairs of nodes for which the interaction status has not been recorded.

There exists a variety of approaches for link prediction and they are mainly divided according to two viewpoints. In Ben-Hur and Noble (2005); Bleakley et al. (2007); Li and Chen (2013); Zhang and Chen (2018), link prediction is treated as a classification problem. That is, examples are constructed by associating the label 1 (or 0) with all true (or false) edges. Then, a classifier is learned by using either a data representation for each edge (Zhang and Chen, 2018), or kernels (Ben-Hur and Noble, 2005; Bleakley et al., 2007; Li and Chen, 2013). Another line of research views link prediction rather as an estimation issue, namely as the problem of estimating the true matrix of the probabilities of connection between node pairs. In this line, Tabouy et al. (2020) propose an estimation procedure for the Stochastic Block Model (SBM) with missing links based on maximum likelihood. In Gaucher et al. (2021), a similar aim is pursued and a technique based on matrix completion tools is proposed, which is also robust to outliers. Finally, Mukherjee and Chakrabarti (2019) give an algorithm for graphon estimation in a missing data set-up.

Concretely, the output of all of these methods are scores for all missing edges, ranking them from most likely to least likely to interact. Such an output is satisfying when the application constrains the number of pairs of vertices to be declared as true edges to be fixed, as e.g. in e-recommendation, where we could have to recommend the top 3-best products most likely to interest the consumer. Alternatively, other practical cases may instead require a classification of the missing edges into true and false edges together with a control of the amount of edges that are wrongly declared as true (false positives). Putting the emphasis on false positives is appropriate in many contexts. For instance, in the reconstruction of biological networks, the edges that are classified as true are then tested experimentally in a costly process, which makes it desirable for the user to avoid false positives in the selection step. This is increasingly true for real-world networks that are in general very sparse. The decision of declaring a missing pair as a false edge can be viewed as a type of abstention option: based on the data, we do not have enough evidence to confidently predict it as a true edge.

How to build a reliable classification procedure? Using an ad hoc rule like declaring as true edges the node pairs with a connection probability above the 50% threshold, may lead to a high number of false positives since a) probabilities may not be estimated correctly and b) even if they were, the probability of making a mistake may still be high if there are many node pairs with moderately elevated connection probability.

In this work, we consider the goal of identifying a subset of the missing pairs of nodes for which we can confidently predict the presence of an edge, with a guarantee on the number of edges that are falsely predicted as true. Our problem can be viewed as finding the appropriate threshold (not 50%) for the connection probabilities such that the number of false positives remains below a prescribed level. The optimal threshold depends on the problem itself. In simple settings a low threshold may be satisfactory, as for instance when most connected triplets are indeed triangles. However, on a graph with much stochasticity, the exact prediction

of links is a very uncertain endeavor.

The problem is formalized in terms of controlling the false discovery rate (FDR), defined as the average proportion of errors among the pairs of vertices declared to be true edges (proportion of false discoveries). More precisely, the goal is to develop a procedure such that the FDR is below a user-specified level α, which is an error margin that represents the acceptance level for the proportion of false edges in the selection. The interpretation for the user is clear: if, for instance, α is set to 5% and the method returns a set of 100 node pairs, then the number of non-existent edges in this set is expected to be at most 5.

4.1.2 Approach

We propose a method that takes as input the partially observed graph and, using an off-the-shelf link prediction method, returns a set of node pairs with an FDR control at level α. The method can be seen as a general wrapper that transforms any link prediction technique into an FDR-controlling procedure. Crucially, even when the quality of the link predictor is not particularly good, our method provides control of the FDR.

Our approach relies on conformal p-values (Bates et al., 2023; Barber and Candès, 2015; Yang et al., 2021), which are a powerful model-free approach in multiple testing that measure statistical significance by comparing the test statistic (or *score*) to a reference set consisting of statistics of observations under the null. This approach comes with guarantees on the FDR control under a suitable exchangeability assumption on the test statistics (Bates et al., 2023; Weinstein et al., 2017; Mary and Roquain, 2022; Marandon et al., 2022), and has been used in novelty detection (Bates et al., 2023; Yang et al., 2021; Marandon et al., 2022; Liang et al., 2022), variable selection (Barber and Candès, 2015; Weinstein et al., 2017), as well as binary classification (Rava et al., 2021).

We propose to use this high-level idea of comparing a score to a set of scores under the null, in order to properly threshold the link prediction probabilities for FDR control. In our link prediction set-up, the connection probability for a pair of nodes (i, j) can be seen as a score indicating the relevance of an edge between i and j. The afore-mentioned score comparison then turns into a comparison of the connection probability for a non-observed pair of nodes to connection probabilities of pairs that are known to be non-existent edges. However, the setup is markedly different from previous literature, making this transposition challenging. In particular, the graph structure makes the scores dependent on each other in an intricate way, which requires to build the scores with care.

Contributions The contributions of this work are summarized as follows:

- We introduce a novel method to obtain FDR control in link prediction (Section 4.2), which extends ideas from the conformal inference literature to graph-structured data. The proposed method is model-free: it does not rely on distributional assumptions, instead, it leverages off-the-shelf link prediction (LP) techniques. It is designed to provide FDR control regardless of the difficulty of the setting and of the quality of the chosen LP technique. Moreover, the ability to use any LP technique of choice, including the state-of-the-art, makes it flexible and powerful.

- Extensive numerical experiments [1] (Section 4.3) assess the excellent performance of the approach and demonstrate its usefulness compared to the state of the art.

[1] We publicly release the code of these experiments at `https://github.com/arianemarandon/linkpredconf`. We have also included a Jupyter notebook that illustrates the use of our procedure.

4.1.3 Relation to previous work

Error rate control in statistical learning Error rate control has notably been considered in novelty detection (Bates et al., 2023; Yang et al., 2021; Marandon et al., 2022; Liang et al., 2022), binary classification (Geifman and El-Yaniv, 2017; Angelopoulos et al., 2021; Rava et al., 2021; Jin and Candès, 2022), clustering (Marandon et al., 2023) and graph inference (Rebafka et al., 2022). The setting closest to ours is that of binary classification, in the sense that here the goal is to classify non-observed pairs of nodes as a 'true' or 'false' edge, given that we observe part of both true edges and non-existent edges. In this line, some approaches (e.g., Zhang and Chen, 2018) view link prediction as a binary classification problem. These approaches use the graph structure to produce edge embeddings, i.e. data objects representing an edge, that are fed to a classifier as learning examples along with labels corresponding to existence or non-existence. The methods introduced in Geifman and El-Yaniv (2017); Angelopoulos et al. (2021); Rava et al. (2021); Jin and Candès (2022) in the context of general binary classification all provide finite-sample guarantees, but the approaches and the type of guarantees vary. To be more precise, the algorithms in Rava et al. (2021); Jin and Candès (2022) control the FDR and are very close to the conformal-based approach of Bates et al. (2023), whereas Geifman and El-Yaniv (2017); Angelopoulos et al. (2021) consider controlling the mis-classification error for a single new point and use certain bounds of the empirical risk with respect to the true risk.

However, these approaches cannot be applied in our situation because here data examples are based on the graph structure and thus depend on each other in a complex way. In particular, we do not have i.i.d. data examples as assumed in the classical binary classification setting. In this regard, our method is related to the work of Marandon et al. (2022) that extends the conformal novelty detection method of Bates et al. (2023) to the case where the learner is not previously trained, but uses the test sample to output class predictions, which makes the class predictions dependent. This is similar to our problem in the sense that here we aim to calibrate connection probabilities that depend on each other through the graph structure.

Conformal inference applied to graph data A few recent works (Huang et al., 2023; Lunde et al., 2023) have considered the application of conformal prediction (Angelopoulos and Bates, 2021) to graph data. Conformal prediction generally refers to the part of conformal inference that is concerned with producing prediction sets that are provably valid in finite samples, rather than error rate control as considered here. Moreover, in these works, the prediction task concerns the nodes: Huang et al. (2023) considers node classification, while Lunde et al. (2023) studies prediction of node covariates (also called network-assisted regression). By contrast, in our work, the prediction task concerns the edges, and therefore, the specific dependency issue that arises differs from Huang et al. (2023); Lunde et al. (2023). Finally, a closely related work is that of Luo et al. (2021), which uses conformal p-values to detect anomalous edges in a graph. However, their method relies on edge-exchangeability, which is a restrictive assumption. Moreover, the guarantee is only for a single edge.

Link with multiple testing The FDR criterion is a staple of multiple testing, where recent works on knockoffs and conformal p-values (Barber and Candès, 2015; Weinstein et al., 2017; Bates et al., 2023; Yang et al., 2021; Marandon et al., 2022) have provided model-free procedures that come with an FDR control guarantee in finite samples. However in this work, while we do use tools of Weinstein et al. (2017); Bates et al. (2023), our setting does not strictly conform to a known multiple testing framework such as the p-value framework (Benjamini and Hochberg, 1995) (the hypotheses being random) or the empirical Bayes framework (Efron

(a) Ground truth A^* (b) Observed graph A

Figure 4.1: Illustration of the learning problem. The left panel shows the true complete graph A^*, which is not observed. The right panel describes our observation: the true edges $(1,2)$, $(1,4)$, $(2,3)$, $(2,5)$, $(3,5)$ are observed, along with the non-existent edges $(1,3)$,$(1,5)$, $(4,5)$ but the information concerning the pairs $(2,4)$ and $(3,4)$ is missing. We aim to decide for $(2,4)$ and $(3,4)$ whether there is a true edge or not.

et al., 2001; Sun and Cai, 2007) (the number of hypotheses being itself random). Hence, previous theory in that area cannot be applied.

4.2 Methodology

4.2.1 Preliminaries

Let $A^* = (A^*_{i,j})_{1\leq i,j\leq n}$ be the adjacency matrix of the true complete graph $\mathcal{G}$, $X \in \mathbb{R}^{n\times d}$ a matrix of node covariates (if available), and $\Omega = (\Omega_{i,j})_{1\leq i,j\leq n}$ the sampling matrix such that $\Omega_{i,j} = 1$ if the interaction status (true/false) of (i,j) is observed, and 0 otherwise. We assume that the entries Ω_{ij} are i.i.d. random variables and that Ω is independent from A^* and X. We denote by A the observed adjacency matrix with $A_{i,j} = \Omega_{i,j}A^*_{i,j}$. Thus, $A_{i,j} = 1$ indicates that there is an observed true edge between i and j, whereas $A_{i,j} = 0$ indicates either the observed lack of an edge or an unreported edge. The sampling matrix Ω is assumed to be observed, so that it is known which zero-entries $A_{i,j} = 0$ correspond to observed false edges and which ones correspond to missing information. The missing information concerns only A^*, and not X. The setting is illustrated in Figure 4.1. We denote by P the joint distribution of $Z^* = (A^*, X, \Omega)$, Z the observation (A, X, Ω) and $\mathcal{Z}$ the observation space.

We are interested in classifying the unobserved node pairs $\{(i,j) : \Omega_{i,j} = 0\}$ into true edges and false edges, or in other words, selecting a set of unobserved node pairs to be declared as true edges, based on the observed graph structure. In order to be consistent with the notation of the literature on conformal p-values (Bates et al., 2023; Liang et al., 2022; Marandon et al., 2022), we use the following notations:

- We denote by $\mathcal{D}_{\text{test}}(Z) = \{(i,j) : \Omega_{i,j} = 0\}$ the set of non-sampled (or missing) node pairs and by $\mathcal{D}(Z) = \{(i,j) : \Omega_{i,j} = 1\}$ the set of sampled pairs, with $\mathcal{D}^0 = \{(i,j) \in \mathcal{D} : A^*_{i,j} = 0\}$ the set of observed non-existent edges and $\mathcal{D}^1 = \{(i,j) \in \mathcal{D} : A^*_{i,j} = 1\}$ the set of observed true edges. We refer to $\mathcal{D}_{\text{test}}(Z)$ as the test set.

- We denote by $\mathcal{H}_0 = \{(i,j) : \Omega_{i,j} = 0, A^*_{i,j} = 0\}$ the (unobserved) set of false edges in the test set and $\mathcal{H}_1 = \{(i,j) : \Omega_{i,j} = 0, A^*_{i,j} = 1\}$ the (unobserved) set of true edges in the test set.

The notations are illustrated in Figure 4.2.

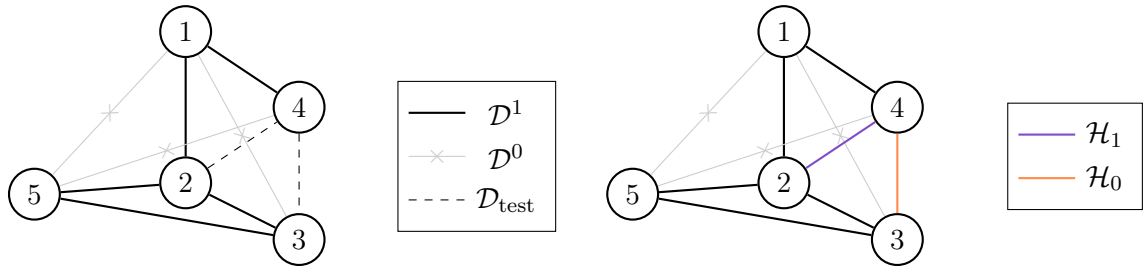


Figure 4.2: Illustration of the notations introduced in Section 4.2.1. The test edges $\mathcal{D}_{\text{test}}$ (left panel) are divided into two subsets (unobserved): true edges $\mathcal{H}_1$, and false edges $\mathcal{H}_0$ (right panel).

In our framework, a selection procedure is a (measurable) function $R = R(Z)$ that returns a subset of $\mathcal{D}_{\text{test}}$ corresponding to the indices (i, j) where an edge is declared. The aim is to design a procedure R close to $\mathcal{H}_1$, or equivalently, with $R \cap \mathcal{H}_0$ (false discoveries) as small as possible. For any such procedure R, the false discovery rate (FDR) of R is defined as the average of the false discovery proportion (FDP) of R under the model parameter $P \in \mathcal{P}$, that is,

$$\text{FDR}(R) = \mathbb{E}_{Z^* \sim P}[\text{FDP}(R)], \quad \text{FDP}(R) = \frac{\sum_{i \in \mathcal{H}_0} \mathbb{1}_{i \in R}}{1 \vee |R|}.$$

Similarly, the true discovery rate (TDR) is defined as the average of the true discovery proportion (TDP), that is,

$$\text{TDR}(R) = \mathbb{E}_{Z^* \sim P}[\text{TDP}(R)], \quad \text{TDP}(R) = \frac{\sum_{i \in \mathcal{H}_1} \mathbb{1}_{i \in R}}{1 \vee |\mathcal{H}_1|}.$$

Our aim is to build a procedure R that controls the FDR while having a TDR (measuring the *power* of the procedure) as large as possible.

4.2.2 Procedure

Let $g : V \times V \times \mathcal{Z} \to \mathbb{R}$ be a *scoring* function, which takes as input a node pair (i, j) and an observation $z \in \mathcal{Z}$ and returns a *score* $S_{i,j} \in \mathbb{R}$, estimating how likely it is that i is connected to j. The score does not have to be in $[0, 1]$: for instance, $S_{i,j}$ can be the number of common neighbors between i and j.

To obtain a set of edges with FDR below α, we borrow from the literature on knockoffs and conformal inference (Weinstein et al., 2017; Bates et al., 2023) to formulate the following idea: some of the observed false edges can be used as a reference set, by comparing the score for a node pair in the test set to scores computed on false edges to determine if it is likely to be a false positive. Effectively, we will declare as edges the pairs that have a test score higher than a cut-off $\hat{t}$ computed from the calibration set and depending on the level α. In detail, the steps are as follows:

1. Use the set of observed node pairs $\mathcal{D}$ to define a reference (or *calibration*) set $\mathcal{D}_{\text{cal}} \subset \mathcal{D}^0$ of false edges and a training set $\mathcal{D}_{\text{train}} \subset \mathcal{D} \setminus \mathcal{D}_{\text{cal}}$ of true and false edges for learning the predictor;

2. Learn a scoring function g on $\mathcal{D}_{\text{train}}$ and compute the scores for the reference set and for the test set;

Algorithm 6 Counting Knockoff (Weinstein et al., 2017)

Input: test scores $(S_w)_{w\in\mathcal{D}_{\text{test}}}$, knockoff scores $(S_w)_{w\in\mathcal{D}_{\text{cal}}}$.
1. Order the scores from lowest to highest, that is $S_{(1)} \geq S_{(2)} \geq \cdots \geq S_{(|\mathcal{D}_{\text{cal}}|+|\mathcal{D}_{\text{test}}|)}$
2. Let s_ℓ be the label $(0/1)$ of $S_{(\ell)}$
3. Set FDP $= 1$, $V = |\mathcal{D}_{\text{cal}}|$, $\ell = |\mathcal{D}_{\text{cal}}| + |\mathcal{D}_{\text{test}}|$, $K = |\mathcal{D}_{\text{test}}|$
4. While FDP $\leq \alpha$ and $K \geq 1$ do $\ell = \ell - 1$
   - if $s_{\ell+1} = 1$, $V = V - 1$
   - else, $K = K - 1$
   - do FDP $= \frac{V+1}{|\mathcal{D}_{\text{cal}}|+1}\frac{|\mathcal{D}_{\text{test}}|}{K}$ (or FDP $= 1$ if $K = 0$)
5. Set $\hat{t} = S_{(K)}$ (or $+\infty$ if $K = 0$).

Output: $\{w \in \mathcal{D}_{\text{test}} : S_w \geq \hat{t}\}$ (return the empty set if $\hat{t} = +\infty$).

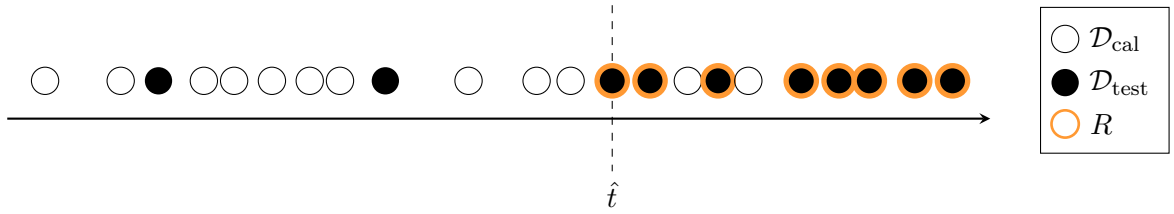


Figure 4.3: Illustration of the CK algorithm (Algorithm 6). The procedure looks for the cut-off $\hat{t}$ by going from the smallest values of the scores (left) to the largest values (right), and stops as soon as the corresponding FDP falls below α. At each step, the FDP is estimated by the proportion of calibration scores among the scores in the left part.

3. Declare as true edges the node pairs in the test set that are returned by the Counting Knockoff (CK) algorithm (Weinstein et al., 2017) given in Algorithm 6.

The CK algorithm (Weinstein et al., 2017) comes from the knockoff literature and is equivalent to the conformal p-value procedure of Bates et al. (2023) (see Mary and Roquain, 2022). It is a stepwise procedure which looks for the appropriate cut-off $\hat{t}$ among a suitable range of values $\mathcal{T}$ by using the calibration scores as benchmarks to evaluate how many false discoveries there are among the selected set of test edges for any cut-off $t \in \mathcal{T}$. The CK algorithm is illustrated in Figure 4.3.

The full procedure is given in Algorithm 7 and Figure 4.5 provides a sketch of the approach. We next describe our proposal for choosing the scoring function g.

Remark 14. This type of procedure is designed to control the FDR at level $\frac{|\mathcal{H}_0|}{|\mathcal{D}_{\text{test}}|}\alpha < \alpha$. To maximize power, we recommend to apply the procedure at level $\alpha/\hat{\pi}_0$ where $\hat{\pi}_0 \in]0,1[$ is an estimate of $\frac{|\mathcal{H}_0|}{|\mathcal{D}_{\text{test}}|}$. In our setting, it is assumed that Ω is independent of A^*, so $\hat{\pi}_0 = \frac{|\mathcal{D}^0|}{|\mathcal{D}|}$ is expected to be a reliable estimate. Alternatively, tools from the multiple testing literature on the estimation of the proportion of null hypotheses may be employed, e.g. by using Storey's estimator (Marandon et al., 2022).

4.2.3 Training the scoring function g

There are two desiderata concerning the properties of the scoring function g:

(i) A key point is that the reference scores mimic the scores $(S_{i,j})_{i,j\in\mathcal{H}_0}$ of the false edges in

Algorithm 7 Conformal link prediction

Input: Adjacency matrix A, node covariate matrix X, sampling matrix Ω, link prediction function g, sample size of the reference set
1. Sample $\mathcal{D}_{\text{cal}}$ uniformly from $\mathcal{D}^0$
2. Learn g on $\mathcal{D}_{\text{train}} \subset \mathcal{D} \setminus \mathcal{D}_{\text{cal}}$
3. For each $(i,j) \in \mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$, compute the score $S_{i,j} = g((i,j), Z)$
4. Apply Algorithm 6 using as input $(S_{i,j})_{(i,j)\in\mathcal{D}_{\text{test}}}$ for the test scores and $(S_{i,j})_{(i,j)\in\mathcal{D}_{\text{cal}}}$ for the reference scores, provinding a set $R(Z) \subset \mathcal{D}_{\text{test}}$.

Output: $R(Z)$

the test set in such a way that, if $(i,j) \in \mathcal{H}_0$ then the rank of $S_{i,j}$ among the reference scores should be uniformly distributed, allowing good estimation of the FDP in the CK algorithm. This property is properly formalized and discussed in Appendix C.2. Here, it entails that the scoring function g must be chosen carefully, because of the dependence structure in the data.

(ii) To fulfill the above property, it is not needed that the quality of the estimates $S_{i,j}$ be particularly good. However, what is important to maximize power is that the ranking provided by the scores is as close as possible to the one given by the true probabilities $\mathbb{P}(A^*_{i,j} = 1 \mid Z)$.

To address the afore-mentioned points, the scoring function g will be taken as the output of a link prediction technique of choice, that will be learned with a suitable subset of the data. Let $\hat{P}\colon (i,j) \in V \times V, z \in \mathcal{Z} \mapsto \hat{P}_{i,j}(z) \in \mathbb{R}$ be a link prediction algorithm , where $\hat{P}_{i,j}$ is the prediction for the node pair (i,j) given the observation z. In a nutshell, given a link prediction algorithm $\hat{P}$, the scoring function $g(\cdot, Z)$ is taken as $\hat{P}(\cdot, Z_{\text{train}})$, with Z_{train} given by

$$Z_{\text{train}} = (A, X, \Omega_{\text{train}}), \quad (\Omega_{\text{train}})_{i,j} = \begin{cases} 0 & \text{if } (i,j) \in \mathcal{D}_{\text{cal}}, \\ \Omega_{i,j} & \text{otherwise.} \end{cases} \tag{4.1}$$

In words, taking $g(\cdot, Z) = \hat{P}(\cdot, Z_{\text{train}}) \neq \hat{P}(\cdot, Z)$ amounts to learn the link predictor by treating the edge examples in the reference set as missing node pairs in the algorithm. We elaborate on this later, after formalizing explicitly the output of LP methods.

Link prediction For simplicity of presentation, let us consider here an undirected graph without node covariates. Let us introduce, for a given $K \in \{1, \dots, n\}$ the r.v.

$$W_{i,j} = (A_{i,\bullet}, A_{j,\bullet}, A^2_{i,\bullet}, A^2_{j,\bullet}, \dots, A^K_{i,\bullet}, A^K_{j,\bullet}), \tag{4.2}$$

where $A^k_{i,\bullet} = (A^k_{i,u})_{1\leq u\leq n}$ for $1 \leq k \leq K$. The r.v. $W_{i,j}$ can be thought of as the "K-hop neighborhood" of (i,j). It represents an embedding for the node pair (i,j), that describes a *pattern* of connection around i and j. If the graph has some structure, it should be observed that the pattern differs when i and j are connected, compared to when they are not. Moreover, there should be some similarity between the patterns observed for true edges, as compared to false edges. Figure 4.4 gives an illustration in the case of a graph with community structure. When there is an edge between i and j (Figure 4.4a), i and j are involved in a same group of nodes that is densely connected (community). Conversely, when there is no edge between them (Figure 4.4b), i and j belong to separate groups of densely connected nodes that share few links between them.

(a) True edge (b) False edge

Figure 4.4: Example of K-hop (K=2) neighborhood of (i, j) (in color), for when (a) i and j are connected and (b) i and j are not connected.

Link prediction methods output a prediction $\hat{P}_{i,j} \in \mathbb{R}$, indicating the relevance of an edge between i and j. In general, the prediction can be written as

$$\hat{P}_{i,j}(Z) = h(W_{i,j}; \{(W_{u,v}, A_{u,v}), (u, v) \in \mathcal{D}_{\text{train}}\}), \tag{4.3}$$

with h some real-valued measurable function and $\mathcal{D}_{\text{train}}$ a certain subset of $\mathcal{D}$. In (4.3), the set $\{(W_{u,v}, A_{u,v}), (u, v) \in \mathcal{D}_{\text{train}}\}$ represents a set of learning examples, where $W_{u,v}$ is an embedding for the pair (u, v) and $A_{u,v}$ is its label. For instance, in the case of the common neighbors (CN) heuristic, the number of common neighbors is given by the scalar product $A^T_{i,\bullet} A_{j,\bullet}$. In that case, the prediction is $\hat{P}_{i,j} = h(W_{i,j}) = A^T_{i,\bullet} A_{j,\bullet}$. Alternatively, when considering supervised approaches such as binary classification (Zhang and Chen, 2018; Bleakley et al., 2007), maximum likelihood (Kipf and Welling, 2016b; Tabouy et al., 2020) or matrix completion (Li et al., 2023; Gaucher et al., 2021), the link prediction function can be written as the minimizer of an empirical risk (ERM):

$$\hat{P}_{i,j}(Z) = \hat{h}(W_{i,j}), \text{ with } \hat{h} \in \left\{ \underset{h \in \mathcal{F}}{\operatorname{argmin}} \sum_{(i,j) \in \mathcal{D}_{\text{train}} \subset \mathcal{D}} \mathcal{L}[P_{i,j}, A_{i,j}] \right\}, \quad P_{i,j} = h(W_{i,j}) \tag{4.4}$$

with $\mathcal{L} : [0, 1] \times \{0, 1\} \to \mathbb{R}$ a loss function and $\mathcal{F}$ a function class. In (4.4), $P_{i,j}$ is an estimate of the probability that there is an edge between i and j, and the error term $\mathcal{L}[P_{i,j}, A_{i,j}]$ quantifies the difference between the prediction $P_{i,j}$ and the true $A_{i,j}$. The ERM formulation for the afore-mentioned supervised approaches can be justified as follows:

- Binary classification approaches (Zhang and Chen, 2018; Bleakley et al., 2007): In that case the ERM formulation (4.4) is obvious. For instance, for SEAL (Zhang and Chen, 2018), $P_{i,j}$ is given by a GNN that takes as input the K-hop subgraph around (i, j), excluding the edge between (i, j) if there is one observed, and augmented with node features that describe the distance of each node in the subgraph to i and to j. The parameters of the GNN are fitted by minimizing the cross-entropy loss over a set $\mathcal{D}_{\text{train}} \subset \mathcal{D}$ of observed true/false edges. In practice, $\mathcal{D}_{\text{train}}$ is subsampled from $\mathcal{D}$ in order to have a $50\% - 50\%$ partitioning between true and false edges.

- Maximum likelihood approaches (Kipf and Welling, 2016b; Tabouy et al., 2020): Maximum likelihood approaches aim to optimize a lower bound on the likelihood (ELBO). This lower bound is an expectation and therefore, using Monte-Carlo approximation, we end up with a function of the form (4.4). For instance, for VGAE (Kipf and Welling, 2016b), $P_{i,j}$ is given by the scalar product $H_i^T H_j$ where H_u is a *node embedding* for node u, the embedding matrix $H \in \mathbb{R}^{n \times n}$ being the output of a GNN. It follows that $H_u = h(A_{u,\bullet}, A^2_{u,\bullet}, \ldots, A^L_{u,\bullet})$ for some function h, with L the number of layers of the GNN.

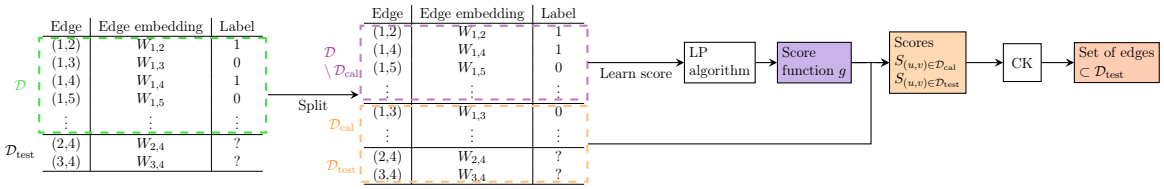


Figure 4.5: Sketch of the procedure proposed in this work.

- Matrix completion (Li et al., 2023; Gaucher et al., 2021): e.g., for Li et al. (2023), one can rewrite the estimated probability matrix $\hat{P}$ as $\hat{P} = \text{argmin}_P\{\sum_{(u,v)\in\mathcal{D}}(A_{u,v}-P_{u,v})^2, P = A_{in}^T \Theta A_{in},\ \text{rank}(\Theta) \leq r\}$ where A_{in} is the sub-matrix of A consisting only of the observed entries. Hence, in that case, $\hat{P}_{i,j}$ is of the form (4.4) with $P_{i,j} = h(A_{i,\bullet}, A_{j,\bullet})$ for some function h.

Construction of the scoring function g Given a link prediction algorithm $\hat{P}$, the scoring function g is learned by training $\hat{P}$ on a subset $\mathcal{D}_{\text{train}}$ of $\mathcal{D}\backslash\mathcal{D}_{\text{cal}}$. Removing $\mathcal{D}_{\text{cal}}$ from the possible set of learning examples allows to enforce that the edge examples (i,j) in the reference set $\mathcal{D}_{\text{cal}}$ are treated as unreported information (i.e., the same as $\mathcal{D}_{\text{test}}$) by the algorithm, which is necessary to fabricate good reference scores and avoid biasing the comparison of the test scores to the reference scores in the CK Algorithm. Otherwise, the scoring g may use the knowledge that the pairs in the reference set are false edges and produce an overfitted score for those. In summary, the scoring function g is thus of the form

$$g((i,j), Z) = h(W_{i,j}; \{(W_{u,v}, A_{u,v}), (u,v) \in \mathcal{D}\backslash\mathcal{D}_{\text{cal}}\}) \tag{4.5}$$

with h some real-valued measurable function.

We outline a key property of our method. Many LP algorithms (e.g. Zhang and Chen, 2018) are not trained on the entire set of observed edges $\mathcal{D}$ but on a subset $\mathcal{D}_{\text{train}} \subset \mathcal{D}$ with a 50-50% distribution of true and false edges. As most real-world networks are sparse, typically all observed edges $\mathcal{D}^1$ are used for training and a randomly chosen subset of false edges in $\mathcal{D}^0$ of the same size as $\mathcal{D}^1$. Then the reference set $\mathcal{D}_{\text{cal}}$ is naturally chosen among the false edges in $\mathcal{D}^0$ that are not used in $\mathcal{D}_{\text{train}}$ for learning the predictor. Consequently, in practice choosing a reference set $\mathcal{D}_{\text{cal}}$ does not diminish the amount of data $\mathcal{D}_{\text{train}}$ on which the predictor is learned.

Remark 15. The sample size of the reference set $\mathcal{D}_{\text{cal}}$ must be large enough to ensure a good power, as pointed out in previous work using conformal p-values in the novelty detection context (Mary and Roquain, 2022; Marandon et al., 2022; Yang et al., 2021). In particular, Mary and Roquain (2022) give a power result under the condition that $|\mathcal{D}_{\text{cal}}| \gtrsim m/(k\alpha)$, where k is the number of "detectable" novelties. Consequently, our recommendation is to choose $|\mathcal{D}_{\text{cal}}|$ of the order of m/α; this choice works reasonably well in our numerical experiments.

4.3 Numerical experiments

In this section, we study the performance of our method both on simulated data (Section 4.3.1) and real data (Section 4.3.2). We consider two choices for the scoring function g, SEAL (Zhang and Chen, 2018) (see Appendix C.1 for details) and CN, yielding the procedures `CN-conf` and `SEAL-conf`. We compare the performance of our method to two "naive" procedures for FDR control (here we assume that $g \in [0,1]$, otherwise, scores are normalized into $[0,1]$ by standardizing the values and applying the sigmoid function):

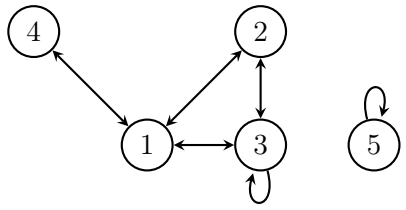


Figure 4.6: Illustration of the SBM model considered in Section 4.3.1. Nodes represent classes, edges indicate connection patterns between classes.

- *Naive thresholding* (NT): We select in $R(Z)$ the edges $(i,j) \in \mathcal{D}_{\text{test}}$ for which $g((i,j),Z) \geq 1-\alpha$. If the probabilities $g((i,j),Z)$ are poorly estimated, this procedure is expected to not control the FDR at level α in general.

- *Cross-validated thresholding* (CVT): We set aside a validation sample $\mathcal{D}_{\text{val}} \subset \mathcal{D}$ and learn g on the remaining data, then compute from a range of values $T \subset [0,1]$ the maximum threshold $\hat{t} \in T$ for which the FDP on $\mathcal{D}_{\text{val}}$ is below α. Then we select for $R(Z)$ the edges $(i,j) \in \mathcal{D}_{\text{test}}$ for which $g((i,j),Z) \geq \hat{t}$. To start with, this procedure is not very convenient because it needs to be fed a range of cut-off values T, whose specification is not obvious. Moreover, it is expected to perform well only if the validation set is large enough. For small n, this is a clear disadvantage compared to our method: ours requires only false edge examples, which are always largely available due to real-world networks being sparse, and the full set of observed true edges is utilized for training. Here, we set $|\,\mathcal{D}_{\text{val}}\,|$ to $20\%|E|$ with a 50%-50% distribution between true and false edges, and $T = \{(1-\alpha)/k, k \in \{1,5,10,20,50,100\}\}$.

In each experiment, the FDR and TDR of the different methods are evaluated by using 100 Monte-Carlo replications.

4.3.1 Simulated data

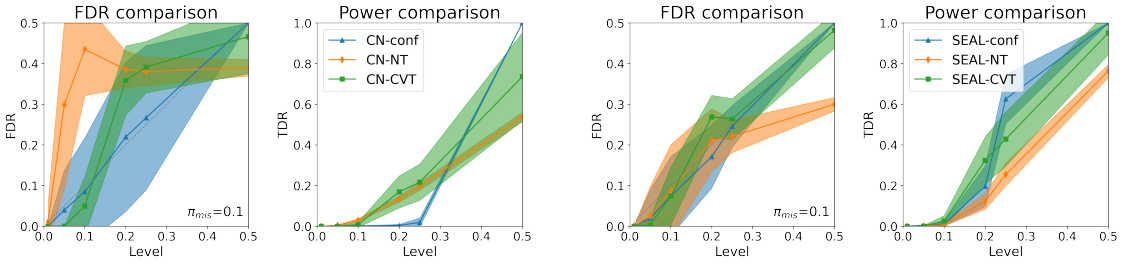
In this section, we evaluate our method on a simulated dataset. We generate a graph A^* of $n = 100$ nodes from a Stochastic Block Model (SBM) with 5 classes, mixture proportions $\pi = (1/5, \dots 1/5)$, and connectivity matrix γ given by

$$\gamma = \begin{pmatrix} \epsilon & p & p & p & \epsilon \\ p & \epsilon & p & \epsilon & \epsilon \\ p & p & p & \epsilon & \epsilon \\ p & \epsilon & \epsilon & \epsilon & \epsilon \\ \epsilon & \epsilon & \epsilon & \epsilon & p \end{pmatrix},$$

with $p = 0.5$ and $\epsilon = 0.05$. The expected number of edges in this setting is ≈ 1150. In this model, the graph displays both community structure and hubs, see Figure 4.6 for an illustration.

We construct training samples $\mathcal{D}(Z)$ and test samples $\mathcal{D}_{\text{test}}(Z)$ by subsampling at random the observed edges and the observed non-existent edges from E and from $(V \times V) \backslash E$ respectively, such that the proportion of missing edges in A is equal to $\pi_{\text{miss}} = 10\%$ and $|\mathcal{H}_0|/|\mathcal{H}_1| = 50\%$. We use $|\,\mathcal{D}_{\text{cal}}\,| = 5000$ for our method. The FDR and TDR are displayed in Figure 4.7a for the choice of CN for the scoring and in Figure 4.7b for SEAL.

When using CN for link prediction, the connection probabilities are not well estimated. This is well-expected since A^* contains nodes that display low probability of connection within



(a) g = CN (b) g = SEAL

Figure 4.7: FDR (left panel) and TDR (right panel) as a function of the nominal level α. Here $\pi_{\text{miss}} = 10\%$, $|\mathcal{H}_0|/|\mathcal{H}_1| = 50\%$, and $|\mathcal{D}_{\text{cal}}| = 5000$. In (a) for the scoring function we use Common Neighbors and in (b) we use SEAL (Zhang and Chen, 2018). The bands indicate the standard deviation.

Table 4.1: Summary of the datasets used. In case covariates are available, the dimension is indicated in parentheses.

	Cora	Yeast	C. ele	T. Albus	Florida Food Web
Number of nodes	2708	2375	297	1056	81
Number of edges	5429	11693	2148	1433	442
Covariates	yes (1433)	no	no	no	no
Directed	no	no	no	yes	yes

their class while being well connected with other classes. In other words, for many pairs of such nodes, it occurs that they share neighbors despite not being connected, so CN is not a good predictor in that case. This leads to NT failing to control the FDR in Figure 4.7a. CVT also fails to control the FDR and it is only our procedure that controls the FDR, with a FDR close to α across all level values. When SEAL is used, we observe again in Figure 4.7b that it occurs that the FDR of CVT exceeds α, while NT can be over-conservative. Our procedure controls the FDR. Moreover, it dominates in terms of power, by having a TDR higher than the others under the FDR constraint. In particular, even when all procedures have an FDR that is close to α, ours displays the most power. Indeed, our method uses an adaptive threshold, unlike NT, and leverages all true edge examples when learning g, unlike CVT.

4.3.2 Real data

In this section, we evaluate our method on real datasets, including benchmarks from Kipf and Welling (2016b); Zhang and Chen (2018); Zhao et al. (2017). A summary of the considered datasets is given in Table 4.1 (see Appendix C.1 for more details).

We construct training samples $\mathcal{D}(Z)$ and test samples $\mathcal{D}_{\text{test}}(Z)$ as in Section 4.3.1 and set $|\mathcal{D}_{\text{cal}}| = 5000$, except for the `Florida Food web` dataset where we use $|\mathcal{D}_{\text{cal}}| = 1000$ because of its smaller size. The FDR and TDR are displayed in Table 4.2 for $\alpha = 0.1$. The results are qualitatively the same as in Section 4.3.1: it occurs that the FDR of the baselines exceed the nominal level while our procedure (using either CN or SEAL for the scoring function) consistently controls the FDR at level α with a substantial power gain.

Table 4.2: Link prediction performance on different data sets: FDR (top) and TDR (bottom). The nominal level is $\alpha = 0.1$. We report the mean value and the standard deviation (in parentheses) over 100 runs. The best performance for each dataset is printed in bold. FDR exceedances are underlined.

	Cora	Yeast	C. ele	T. Albus	Florida Food Web
	FDR				
CN-conf	0.019 (0.008)	0.026 (0.005)	0.088 (0.034)	0.030 (0.099)	0.020 (0.140)
CN-NT	0.001 (0.005)	0.000 (0.001)	0.027 (0.043)	<u>0.705</u> (0.245)	<u>0.500</u> (0.000)
CN-CVT	0.001 (0.004)	0.001 (0.002)	0.042 (0.050)	0.000 (0.000)	0.000 (0.000)
SEAL-conf	0.096 (0.017)	0.098 (0.011)	0.102 (0.025)	0.096 (0.038)	0.097 (0.048)
SEAL-NT	0.050 (0.009)	0.008 (0.003)	0.050 (0.023)	0.037 (0.029)	0.061 (0.071)
SEAL-CVT	0.064 (0.007)	0.017 (0.006)	0.064 (0.030)	0.053 (0.045)	0.058 (0.062)
	TDR				
CN-conf	0.476 (0.022)	0.794 (0.012)	0.339 (0.099)	0.000 (0.000)	0.000 (0.000)
CN-NT	0.062 (0.020)	0.060 (0.004)	0.056 (0.013)	0.590 (0.490)	1.000 (0.000)
CN-CVT	0.026 (0.051)	0.048 (0.073)	0.093 (0.066)	0.000 (0.000)	0.000 (0.000)
SEAL-conf	**0.757** (0.041)	**0.956** (0.010)	**0.603** (0.031)	**0.495** (0.057)	**0.783** (0.088)
SEAL-NT	0.601 (0.033)	0.792 (0.056)	0.314 (0.151)	0.302 (0.051)	0.359 (0.169)
SEAL-CVT	0.620 (0.036)	0.827 (0.030)	0.300 (0.177)	0.249 (0.114)	0.341 (0.223)

4.4 Discussion

We have presented a new method that calibrates the output of a given link prediction technique for FDR control, using recent ideas from the conformal inference literature. The approach is validated using both simulated data and real data and its interest is demonstrated by its superior performance compared to the state of the art.

Let us first mention that if the user wishes to control the probability of making more than a certain number of false positives (Lehmann and Romano, 2005; Janson and Su, 2016), the method can easily be extended in that sense.

In this work, it is assumed that the missing data pattern is independent from the true complete graph and moreover, that the entries of the sampling matrix are i.i.d. It would be interesting to investigate the extension of the method in the general case, where the entries of the sampling are not i.i.d., such as in egocentric sampling (Li et al., 2023).

It is outside of the scope of this work to prove theoretical guarantees, because the setup is markedly different from previous work and the dependence in the data makes the analysis very complex. In particular, we expect that the exchangeability condition on the scores (Marandon et al. (2022)) cannot be verified here outside of trivial cases. Finding a suitable relaxation of this condition for the setup considered here represents a promising avenue for future work.

Chapter 5

Discussion

Conformal FDR control In conformal inference, a key assumption is that the data is exchangeable. A first research direction is moving beyond this setting, in the following two ways. First, by considering the case of structured data. In this line, Chapter 4 proposed a transposition of the conformal p-value procedure to a link prediction setup, where the graph structure induces non-exchangeability. Due to this, the work therein demonstrated FDR control empirically but it lacked theoretical guarantees; it would be desirable to fill the gap. Secondly, distribution shift (which refers to a general phenomenon in learning when a certain data distribution changes over time) is an important practical concern, which has been recently investigated in the context of conformal prediction (Tibshirani et al., 2019; Podkopaev and Ramdas, 2021; Barber et al., 2022). In the context of Chapter 2, distribution shift may occur in the sense that the distribution of the nulls between the training sample and the test sample may be slightly different. In particular, if the null comes from a mixture of distributions, it could be that some of the component distributions are in the test sample and not in the nominal sample, in which case they will be wrongly declared as anomalies. Studying this issue is an interesting question for future investigations.

Next, we have only considered a subpart of selective inference in this thesis, namely how to select a set of items such as to get a pre-specified FDR guarantee (multiple testing). In practical applications, users may manipulate the data, e.g. to follow common practice in the field and/or budget constraints. In that case, the inverse problem arises, of identifying bounds on the FDR in the selected set (Goeman and Solari, 2011; Blanchard et al., 2021). Thus, another research direction concerns the problem of designing FDP bounds for the conformal p-values of Chapter 2, for instance by adapting the work of Katsevich and Ramdas (2020). The main challenge will be to deal with the specific dependence structure that is inherent to the p-value process.

Finally, a future application of the AdaDetect procedure introduced in Chapter 2 could be the task of anomaly detection on graph data (Ma et al., 2021b). This task takes two main forms: graph-level anomaly detection, which is the problem of detecting anomalous graphs in a set of graphs, and node-level anomaly detection, which is concerned with detecting anomalous nodes in a single graph. AdaDetect would allow to leverage state-of-the-art graph or node classification techniques (such as Shervashidze et al. (2011); Kipf and Welling (2016a); Xu et al. (2018), see also Wu et al. (2020) and Errica et al. (2020) for reviews) to perform these tasks with a FDR control guarantee.

Error rate control in unsupervised learning Chapter 3 is only a first step in error rate control in unsupervised learning, which remains largely an unsolved question. To start with, the FMR criterion is constraining since it requires the number of classes to be known.

As a consequence, it cannot compare two clusterings with a different number of clusters. Moreover, from a partitioning point of view, it could be seen as more relevant to measure the error in terms of pairwise clusters assignments, that is, to penalize pairs of elements that are incorrectly labeled as belonging or not to the same cluster. Thus, possible developments include the extension to other criteria relevant for clustering (Grün, 2019), such as the Rand Index (Rand, 1971).

In addition, our approach is a parametric one, where it is assumed that the specification of distribution family for the components distributions is correct; otherwise, there are no guarantees. It would be desirable to take into account model mis-specification in future research.

Lastly, even in a setup where the model specifications are correct, a main limitation of the procedures proposed here is that they rely on the quality of the parameter estimates. This raises the open question of whether finite-sample FMR control can be achieved in this context, and more generally, what are the limits for this problem. In this line, we note the work of Roquain and Verzelen (2022) that shows, in the context of unsupervised novelty detection (which is a framework sharing similarities with the one considered here), that there does not exists an FDR-controlling procedure that is asymptotically power-optimal if the number of false nulls is too large. Alternatively, the extension to a semi-supervised learning setting where a few labeled samples are available, which is likely to considerably improve the inference, is an interesting avenue.

Bibliography

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. Journal of Machine Learning Research, 18(177):1–86.

Abraham, K., Castillo, I., and Roquain, É. (2022). Empirical bayes cumulative ℓ-value multiple testing procedure for sparse sequences. Electronic Journal of Statistics, 16(1):2033–2081.

Angelopoulos, A. N. and Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv preprint arXiv:2107.07511.

Angelopoulos, A. N., Bates, S., Candès, E. J., Jordan, M. I., and Lei, L. (2021). Learn then test: Calibrating predictive algorithms to achieve risk control. arXiv preprint arXiv:2110.01052.

Arlot, S., Blanchard, G., and Roquain, E. (2010). Some nonasymptotic results on resampling in high dimension, i: Confidence regions. The Annals of Statistics, 38(1):51–82.

Arora, S. and Kannan, R. (2005). Learning mixtures of separated nonspherical Gaussians. The Annals of Applied Probability, 15(1A):69 – 92.

Arthur, D. and Vassilvitskii, S. (2006). k-means++: The advantages of careful seeding. Technical report, Stanford.

Audibert, J.-Y. and Tsybakov, A. B. (2007). Fast learning rates for plug-in classifiers. The Annals of statistics, 35(2):608–633.

Azizyan, M., Singh, A., and Wasserman, L. (2013). Minimax theory for high-dimensional gaussian mixtures with sparse mean separation. In Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, page 2139–2147, Red Hook, NY, USA. Curran Associates Inc.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. The Annals of Statistics, 45(1):77 – 120.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). Conformal prediction for reliable machine learning: theory, adaptations and applications. Newnes.

Baraud, Y. (2016). Bounding the expectation of the supremum of an empirical process over a (weak) vc-major class. Electronic journal of statistics, 10(2):1709–1728.

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

Barber, R. F. and Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. The Annals of Statistics, 47(5):2504 – 2537.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+. The Annals of Statistics, 49(1):486–507.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2022). Conformal prediction beyond exchangeability. arXiv preprint arXiv:2202.13415.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2020). Robust inference with knockoffs. The Annals of Statistics, 48(3):1409 – 1431.

Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning. fairmlbook. org.

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. (2019). Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. The Journal of Machine Learning Research, 20(1):2285–2301.

Bartlett, P. L. and Wegkamp, M. H. (2008). Classification with a reject option using a hinge loss. Journal of Machine Learning Research, 9(59):1823–1840.

Bashari, M., Epstein, A., Romano, Y., and Sesia, M. (2023). Derandomized novelty detection with fdr control via conformal e-values. arXiv preprint arXiv:2302.07294.

Bates, S., Candès, E., Janson, L., and Wang, W. (2021). Metropolized knockoff sampling. Journal of the American Statistical Association, 116(535):1413–1427.

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.

Bekker, J. and Davis, J. (2020). Learning from positive and unlabeled data: A survey. Machine Learning, 109(4):719–760.

Ben-Hur, A. and Noble, W. S. (2005). Kernel methods for predicting protein-protein interactions. In Proceedings Thirteenth International Conference on Intelligent Systems for Molecular Biology 2005, Detroit, MI, USA, 25-29 June 2005, pages 38–46.

Benjamini, Y. (2010). Discovering the false discovery rate. Journal of the Royal Statistical Society: series B (statistical methodology), 72(4):405–416.

Benjamini, Y., Hechtlinger, Y., and Stark, P. B. (2019). Confidence intervals for selected parameters. arXiv preprint arXiv:1906.00505.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Statist. Soc. Ser. B, 57(1):289–300.

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. Ann. Statist., 29(4):1165–1188.

Blanchard, G., Lee, G., and Scott, C. (2010a). Semi-supervised novelty detection. The Journal of Machine Learning Research, 11:2973–3009.

Blanchard, G., Lee, G., and Scott, C. (2010b). Semi-supervised novelty detection. J. Mach. Learn. Res., 11:2973–3009.

Blanchard, G., Neuvial, P., and Roquain, E. (2021). On agnostic post-hoc approaches to false positive control. In Handbook of Multiple Comparisons, pages 211–232. Chapman and Hall/CRC.

Blanchard, G. and Roquain, E. (2009). Adaptive false discovery rate control under independence and dependence. J. Mach. Learn. Res., 10:2837–2871.

Bleakley, K., Biau, G., and Vert, J.-P. (2007). Supervised reconstruction of biological networks with local models. Bioinformatics, 23(13):i57–i65.

Bogdan, M., van den Berg, E., Sabatti, C., Su, W., and Candès, E. J. (2015). SLOPE—adaptive variable selection via convex optimization. Ann. Appl. Stat., 9(3):1103–1140.

Bos, T. and Schmidt-Hieber, J. (2021). Convergence rates of deep relu networks for multiclass classification.

Cai, T., Sun, W., and Wang, W. (2019). Covariate-assisted ranking and screening for large-scale two-sample inference. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 81(2):187–234.

Cai, T. T. and Jin, J. (2010). Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. The Annals of Statistics, 38(1):100 – 145.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: finding needles in multiple haystacks. J. Amer. Statist. Assoc., 104(488):1467–1481.

Calvo, B., Larranaga, P., and Lozano, J. A. (2007). Learning bayesian classifiers from positive and unlabeled examples. Pattern Recognition Letters, 28(16):2375–2384.

Cannon, A., Howse, J., Hush, D., and Scovel, C. (2002). Learning with the neyman-pearson and min-max criteria. Los Alamos National Laboratory, Tech. Rep. LA-UR, pages 02–2951.

Chow, C. (1970). On optimum recognition error and reject tradeoff. IEEE Transactions on Information Theory, 16(1):41–46.

Chretien, S., Dombry, C., and Faivre, A. (2019). The Guedon-Vershynin Semi-Definite Programming approach to low dimensional embedding for unsupervised clustering. Frontiers in Applied Mathematics and Statistics.

Christian, R. R. and Luczkovich, J. J. (1999). Organizing and understanding a winter's seagrass foodweb network through effective trophic levels. Ecological Modelling, 117(1):99–124.

Clauset, A., Moore, C., and Newman, M. E. (2008). Hierarchical structure and the prediction of missing links in networks. Nature, 453(7191):98–101.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

Denis, C. and Hebiri, M. (2020). Consistency of plug-in confidence sets for classification in semi-supervised learning. Journal of Nonparametric Statistics, 32(1):42–72.

Dickhaus, T. (2014). Simultaneous statistical inference: with applications in the life sciences. Springer Science & Business Media.

Du Plessis, M. C., Niu, G., and Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. Advances in neural information processing systems, 27:703–711.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. J. Am. Stat. Assoc., 99(465):96–104.

Efron, B. (2007). Doing thousands of hypothesis tests at the same time. Metron - International Journal of Statistics, LXV(1):3–21.

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

Efron, B. (2009). Empirical Bayes estimates for large-scale prediction problems. J. Am. Stat. Assoc., 104(487):1015–1028.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. J. Amer. Statist. Assoc., 96(456):1151–1160.

Errica, F., Podda, M., Bacciu, D., and Micheli, A. (2020). A fair comparison of graph neural networks for graph classification. In Proceedings of the 8th International Conference on Learning Representations (ICLR).

Ferreira, J. A. and Zwinderman, A. H. (2006). On the Benjamini-Hochberg method. Ann. Statist., 34(4):1827–1849.

Fisher, A. (2021). Saffron and lord ensure online control of the false discovery rate under positive dependence. arXiv preprint arXiv:2110.08161.

Fithian, W. and Lei, L. (2020). Conditional calibration for false discovery rate control under dependence.

Foygel Barber, R. and Ramdas, A. (2015). The p-filter: multi-layer fdr control for grouped hypotheses. ArXiv e-prints, pages arXiv–1512.

Friedman, J. H. (2003). On multivariate goodness–of–fit and two–sample testing. Statistical Problems in Particle Physics, Astrophysics, and Cosmology, 1:311.

Gao, Z. and Zhao, Q. (2023). Simultaneous hypothesis testing using internal negative controls with an application to proteomics.

Gaucher, S., Klopp, O., and Robin, G. (2021). Outlier detection in networks with missing links. Computational Statistics & Data Analysis, 164:107308.

Geifman, Y. and El-Yaniv, R. (2017). Selective classification for deep neural networks. In Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, page 4885–4894, Red Hook, NY, USA. Curran Associates Inc.

Genovese, C. and Wasserman, L. (2002). Operating characteristics and extensions of the false discovery rate procedure. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):499–517.

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

Giraud, C. (2022). Introduction to high-dimensional statistics, volume 168. Boca Raton, FL: CRC Press.

Giraud, C. and Verzelen, N. (2018). Partial recovery bounds for clustering with the relaxed K-means. Mathematical Statistics and Learning, 1(3):317–374.

Goeman, J. J. and Solari, A. (2011). Multiple Testing for Exploratory Research. Statistical Science, 26(4):584 – 597.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep Learning. MIT Press. `http://www.deeplearningbook.org`.

Grandvalet, Y., Rakotomamonjy, A., Keshet, J., and Canu, S. (2008). Support vector machines with a reject option. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008, pages 537–544. Curran Associates, Inc.

Grün, B. (2019). Model-based clustering. In Handbook of mixture analysis, pages 157–192. Chapman and Hall/CRC.

Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of the 34th International Conference on Machine Learning - Volume 70, ICML'17, page 1321–1330. JMLR.org.

Guo, T., Xu, C., Huang, J., Wang, Y., Shi, B., Xu, C., and Tao, D. (2020). On positive-unlabeled classification in gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Haroush, M., Frostig, T., Heller, R., and Soudry, D. (2022). A statistical framework for efficient out of distribution detection in deep neural networks. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net.

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction, volume 2. Springer.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778.

Heller, R., Bogomolov, M., and Benjamini, Y. (2014). Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study. Proceedings of the National Academy of Sciences, 111(46):16262–16267.

Heller, R. and Rosset, S. (2021). Optimal control of false discovery criteria in the two-group model. Journal of the Royal Statistical Society Series B: Statistical Methodology, 83(1):133–155.

Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. The Annals of Applied Statistics, 8(1):481 – 498.

Herbei, R. and Wegkamp, M. H. (2006). Classification with reject option. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, 34(4):709–721.

Ho, N. and Nguyen, X. (2016). On strong identifiability and convergence rates of parameter estimation in finite mixtures. Electronic Journal of Statistics, 10(1):271 – 307.

Huang, K., Jin, Y., Candes, E., and Leskovec, J. (2023). Uncertainty quantification over graph with conformalized graph neural networks. arXiv preprint arXiv:2305.14535.

Ioannidis, J. P. (2005). Why most published research findings are false. PLoS medicine, 2(8):e124.

Ivanov, D. (2020). Dedpul: Difference-of-estimated-densities-based positive-unlabeled learning. In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pages 782–790. IEEE.

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., and Gray, A. (2019). Statistics, data mining, and machine learning in astronomy: A practical python guide for the analysis of survey data. Princeton University Press.

Ivezić, Ž., Vivas, A. K., Lupton, R. H., and Zinn, R. (2005). The selection of RR lyrae stars using single-epoch data. The Astronomical Journal, 129(2):1096.

Janson, L. and Su, W. (2016). Familywise error rate control via knockoffs. Electronic Journal of Statistics, 10(1):960 – 975.

Javanmard, A. and Javadi, H. (2019). False discovery rate control via debiased lasso. Electronic Journal of Statistics, 13(1):1212–1253.

Jin, J. and Cai, T. T. (2007). Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. Journal of the American Statistical Association, 102(478):495–506.

Jin, Y. and Candès, E. J. (2022). Selection by prediction with conformal p-values.

Jordan, M. I. and Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245):255–260.

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.

Kipf, T. N. and Welling, M. (2016a). Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Kipf, T. N. and Welling, M. (2016b). Variational graph auto-encoders. NIPS Workshop on Bayesian Deep Learning.

Korn, E. L., Troendle, J. F., McShane, L. M., and Simon, R. (2004). Controlling the number of false discoveries: application to high-dimensional genomic data. J. Statist. Plann. Inference, 124(2):379–398.

Kovács, I. A., Luck, K., Spirohn, K., Wang, Y., Pollis, C., Schlabach, S., Bian, W., Kim, D.-K., Kishore, N., Hao, T., Calderwood, M. A., Vidal, M., and Barabási, A.-L. (2019). Network-based prediction of protein interactions. Nature Communications, 10(1):1240.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems, 30.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324.

LeCun, Y. and Cortes, C. (2010). MNIST handwritten digit database.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44(3):907 – 927.

Lee, K., Lee, H., Lee, K., and Shin, J. (2018). Training confidence-calibrated classifiers for detecting out-of-distribution samples. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net.

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

Lei, J. and Rinaldo, A. (2015). Consistency of spectral clustering in stochastic block models. The Annals of Statistics, 43(1):215 – 237.

Lei, L., D'Amour, A., Ding, P., Feller, A., and Sekhon, J. (2021). Distribution-free assessment of population overlap in observational studies. Technical report.

Lei, L. and Fithian, W. (2018). Adapt: an interactive procedure for multiple testing with side information. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 80(4):649–679.

Li, T., Wu, Y.-J., Levina, E., and Zhu, J. (2023). Link prediction for egocentrically sampled networks. Journal of Computational and Graphical Statistics, 0(0):1–24.

Li, X. and Chen, H. (2013). Recommendation as link prediction in bipartite graphs: A graph kernel-based machine learning approach. Decision Support Systems, 54(2):880–890.

Liang, Z., Sesia, M., and Sun, W. (2022). Integrative conformal p-values for powerful out-of-distribution testing with labeled outliers.

Liang, Z., Zhou, Y., and Sesia, M. (2023). Conformal inference is (almost) free for neural networks trained with early stopping. arXiv preprint arXiv:2301.11556.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In 2008 eighth ieee international conference on data mining, pages 413–422. IEEE.

Loper, J. H., Lei, L., Fithian, W., and Tansey, W. (2019). Smoothed nested testing on directed acyclic graphs. arXiv preprint arXiv:1911.09200.

Lu, Y. and Zhou, H. H. (2016). Statistical and computational guarantees of lloyd's algorithm and its variants. arXiv preprint arXiv:1612.02099.

Lunde, R., Levina, E., and Zhu, J. (2023). Conformal prediction for network-assisted regression. arXiv preprint arXiv:2302.10095.

Luo, R., Nettasinghe, B., and Krishnamurthy, V. (2021). Anomalous edge detection in edge exchangeable social network models. arXiv preprint arXiv:2109.12727.

Ma, R., Tony Cai, T., and Li, H. (2021a). Global and simultaneous hypothesis testing for high-dimensional logistic regression models. Journal of the American Statistical Association, 116(534):984–998.

Ma, X., Wu, J., Xue, S., Yang, J., Zhou, C., Sheng, Q. Z., Xiong, H., and Akoglu, L. (2021b). A comprehensive survey on graph anomaly detection with deep learning. IEEE Transactions on Knowledge and Data Engineering.

Marandon, A., Lei, L., Mary, D., and Roquain, E. (2022). Adaptive novelty detection with false discovery rate guarantee.

Marandon, A., Rebafka, T., Roquain, E., and Sokolovska, N. (2023). False membership rate control in mixture models.

Mary, D. and Roquain, E. (2022). Semi-supervised multiple testing. Electronic Journal of Statistics, 16(2):4926–4981.

Mary-Huard, T., Perduca, V., Blanchard, G., and Marie-Laure, M.-M. (2021). Error rate control for classification rules in multiclass mixture models.

Massart, P. (2007). Concentration Inequalities and Model Selection Ecole d'Eté de Probabilités de Saint-Flour XXXIII - 2003. École d'Été de Probabilités de Saint-Flour, 1896. Springer Berlin Heidelberg, Berlin, Heidelberg, 1st ed. 2007. edition.

Melnykov, V. (2013). On the distribution of posterior probabilities in finite mixture models with application in clustering. Journal of Multivariate Analysis, 122:175–189.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). Foundations of Machine Learning. The MIT Press.

Mukherjee, S. S. and Chakrabarti, S. (2019). Graphon estimation from partially observed network data. CoRR, abs/1906.00494.

Najafi, A., Motahari, S. A., and Rabiee, H. R. (2020). Reliable clustering of Bernoulli mixture models. Bernoulli, 26(2):1535 – 1559.

Nguyen, T.-B., Chevalier, J.-A., Thirion, B., and Arlot, S. (2020). Aggregation of multiple knockoffs. In International Conference on Machine Learning, pages 7283–7293. PMLR.

Patcha, A. and Park, J.-M. (2007). An overview of anomaly detection techniques: Existing solutions and latest technological trends. Computer networks, 51(12):3448–3470.

Peel, D. and McLachlan, G. J. (2000). Robust mixture modelling using the t distribution. Statistics and Computing, 10(4):339–348.

Podkopaev, A. and Ramdas, A. (2021). Distribution-free uncertainty quantification for classification under label shift. In Uncertainty in Artificial Intelligence, pages 844–853. PMLR.

Ramdas, A., Chen, J., Wainwright, M. J., and Jordan, M. I. (2019a). A sequential algorithm for false discovery rate control on directed acyclic graphs. Biometrika, 106(1):69–86.

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

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. Journal of the American Statistical association, 66(336):846–850.

Rava, B., Sun, W., James, G. M., and Tong, X. (2021). A burden shared is a burden halved: A fairness-adjusted approach to classification. arXiv preprint arXiv:2110.05720.

Rebafka, T., Roquain, É., and Villers, F. (2022). Powerful multiple testing of paired null hypotheses using a latent graph model. Electronic Journal of Statistics, 16(1):2796 – 2858.

Regev, O. and Vijayaraghavan, A. (2017). On learning mixtures of well-separated gaussians. In 2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS), pages 85–96.

Rigollet, P. and Tong, X. (2011). Neyman-pearson classification, convexity and stochastic constraints. Journal of Machine Learning Research.

Romano, J. P. and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. J. Amer. Statist. Assoc., 100(469):94–108.

Roquain, E. and Verzelen, N. (2022). False discovery rate control with unknown null distribution: Is it possible to mimic the oracle? The Annals of Statistics, 50(2):1095–1123.

Roquain, E. and Villers, F. (2011). Exact calculations for false discovery proportion with application to least favorable configurations. Ann. Statist., 39(1):584–612.

Rosset, S., Heller, R., Painsky, A., and Aharoni, E. (2022). Optimal and maximin procedures for multiple testing problems. Journal of the Royal Statistical Society Series B: Statistical Methodology, 84(4):1105–1128.

Ruff, L., Vandermeulen, R. A., Görnitz, N., Deecke, L., Siddiqui, S. A., Binder, A., Müller, E., and Kloft, M. (2018). Deep one-class classification. In Proceedings of the 35th International Conference on Machine Learning, volume 80, pages 4393–4402.

Sarkar, S. K. (2008). On methods controlling the false discovery rate. Sankhya, Ser. A, 70:135–168.

Sarkar, S. K. and Tang, C. Y. (2022). Adjusting the benjamini–hochberg method for controlling the false discovery rate in knockoff-assisted variable selection. Biometrika, 109(4):1149–1155.

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., and Williamson, R. C. (2001). Estimating the support of a high-dimensional distribution. Neural computation, 13(7):1443–1471.

Scott, C. and Nowak, R. (2005). A neyman-pearson approach to statistical learning. IEEE Transactions on Information Theory, 51(11):3806–3819.

Sen, B. (2018). A gentle introduction to empirical process theory and applications. Lecture Notes, Columbia University, 11:28–29.

Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., and Eliassi-Rad, T. (2008). Collective classification in network data. AI magazine, 29(3):93–93.

Seo, E. and Hutchinson, R. (2018). Predicting links in plant-pollinator interaction networks using latent factor models with implicit feedback. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).

Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, USA.

Shenhav, L., Heller, R., and Benjamini, Y. (2015). Quantifying replicability in systematic reviews: the r-value. arXiv preprint arXiv:1502.00088.

Shervashidze, N., Schweitzer, P., van Leeuwen, E. J., Mehlhorn, K., and Borgwardt, K. M. (2011). Weisfeiler-lehman graph kernels. J. Mach. Learn. Res., 12:2539–2561.

Storey, J. D. (2002). A direct approach to false discovery rates. J. R. Stat. Soc. Ser. B Stat. Methodol., 64(3):479–498.

Storey, J. D. (2003). The positive false discovery rate: a bayesian interpretation and the q-value. The Annals of Statistics, 31(6):2013–2035.

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. J. R. Stat. Soc. Ser. B Stat. Methodol., 66(1):187–205.

Sugiyama, M., Suzuki, T., and Kanamori, T. (2012). Density ratio estimation in machine learning. Cambridge University Press.

Sun, W. and Cai, T. T. (2007). Oracle and adaptive compound decision rules for false discovery rate control. J. Am. Stat. Assoc., 102(479):901–912.

Sun, W. and Cai, T. T. (2009). Large-scale multiple testing under dependence. J. R. Stat. Soc. Ser. B Stat. Methodol., 71(2):393–424.

Tabouy, T., Barbillon, P., and Chiquet, J. (2020). Variational inference for stochastic block models from sampled data. Journal of the American Statistical Association, 115(529):455–466.

Tarassenko, L., Hayton, P., Cerneaz, N., and Brady, M. (1995). IET Conference Proceedings, pages 442–447(5).

Terry, J. C. D. and Lewis, O. T. (2020). Finding missing links in interaction networks. Ecology, 101(7):e03047.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series B (Methodological), 58(1):267–288.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. Advances in neural information processing systems, 32.

Vapnik, V. N. (1998). Statistical learning theory. Chichester: Wiley.

Vempala, S. and Wang, G. (2004). A spectral algorithm for learning mixture models. Journal of Computer and System Sciences, 68(4):841–860. Special Issue on FOCS 2002.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. Nature, 417(6887):399–403.

Vovk, V. (2015). Cross-conformal predictors. Annals of Mathematics and Artificial Intelligence, 74(1):9–28.

Vovk, V., Gammerman, A., and Shafer, G. (2005). Algorithmic learning in a random world, volume 29. Springer.

Wang, Y., Kaji, T., and Rockova, V. (2022). Approximate bayesian computation via classification. Journal of Machine Learning Research, 23(350):1–49.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world'networks. Nature, 393(6684):440–442.

Weber Zendrera, A., Sokolovska, N., and Soula, H. A. (2019). Robust structure measures of metabolic networks that predict prokaryotic optimal growth temperature. BMC Bioinformatics, 20(1):499.

Weber Zendrera, A., Sokolovska, N., and Soula, H. A. (2021). Functional prediction of environmental variables using metabolic networks. Scientific Reports, 11(1):12192.

Wegkamp, M. and Yuan, M. (2011). Support vector machines with a reject option. Bernoulli, 17(4):1368–1385.

Weinstein, A. (2021). On permutation invariant problems in large-scale inference. arXiv preprint arXiv:2110.06250.

Weinstein, A., Barber, R., and Candès, E. (2017). A power and prediction analysis for knockoffs with lasso statistics.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., and Philip, S. Y. (2020). A comprehensive survey on graph neural networks. IEEE transactions on neural networks and learning systems, 32(1):4–24.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. (2018). How powerful are graph neural networks? arXiv preprint arXiv:1810.00826.

Yang, C.-Y., Lei, L., Ho, N., and Fithian, W. (2021). Bonus: Multiple multivariate testing with a data-adaptivetest statistic.

Zaidi, S., Zela, A., Elsken, T., Holmes, C. C., Hutter, F., and Teh, Y. (2021). Neural ensemble search for uncertainty estimation and dataset shift. Advances in Neural Information Processing Systems, 34:7898–7911.

Zeevi, Y., Astashenko, S., and Benjamini, Y. (2020). Ignored evident multiplicity harms replicability–adjusting for it offers a remedy. arXiv preprint arXiv:2006.11585.

Zhang, M. and Chen, Y. (2018). Link prediction based on graph neural networks. In Advances in Neural Information Processing Systems, pages 5165–5175.

Zhao, Y., Wu, Y.-J., Levina, E., and Zhu, J. (2017). Link prediction for partially observed networks. Journal of Computational and Graphical Statistics, 26(3):725–733.

Zrnic, T., Ramdas, A., and Jordan, M. I. (2021). Asynchronous online testing of multiple hypotheses. J. Mach. Learn. Res., 22:33–1.

Appendix A

Supplementary material of Chapter 2

A.1 Proofs and results for Section 3

A.1.1 Proof of Lemma 2

Let

$$\begin{aligned} U = (U_1, \dots, U_{n+m_0}) &= (Y_1, \dots, Y_n, X_i, i \in \mathcal{H}_0); \\ V = (V_1, \dots, V_{m_1}) &= (X_i, i \in \mathcal{H}_1); \\ W = h(U, V) &= ((Z_1, \dots, Z_k), \{Z_{k+1}, \dots, Z_{n+m}\}); \\ S_i &= g(U_i, W), \ \ i \in \{1, \dots, n + m_0\}, \end{aligned}$$

for given measurable function g that satisfies the condition (2.8). Then, for any permutation π of $\{1, \dots, n+m_0\}$ that do not permute $\{1, \dots, k\}$, Assumption 1 implies that $U \,|\, V \sim U^\pi \,|\, V$ and thus $(U, V) \sim (U^\pi, V)$. This entails $(U, V, W) \sim (U^\pi, V, h(U^\pi, V)) = (U^\pi, V, h(U, V)) = (U^\pi, V, W)$. Hence,

$$(g(U_1, W), \dots, g(U_{n+m_0}, W)) \,|\, V, W \sim (g(U_{\pi(1)}, W), \dots, g(U_{\pi(n+m_0)}, W)) \,|\, V, W$$

Since $\pi(i) = i$ for all $i \in \{1, \dots, k\}$, we obtain that

$$\begin{aligned} &(g(U_{k+1}, W), \dots, g(U_{n+m_0}, W)) \,|\, (g(V_1, W), \dots, g(V_{m_1}, W)) \\ &\sim (g(U_{\pi(k+1)}, W), \dots, g(U_{\pi(n+m_0)}, W)) \,|\, (g(V_1, W), \dots, g(V_{m_1}, W)), \end{aligned}$$

which completes the proof.

A.1.2 Key properties

In this section, we present key properties of empirical p-values derived from exchangeable scores. The first result provides a representation that characterizes the dependence structure of the empirical p-values that is a key step for the proof of FDR control. It generalizes the representation by Bates et al. (2023) for independent scores. The proof is deferred to Section A.1.7.

Theorem 18. *Consider any family of scores $(S_{k+1}, \dots, S_{n+m})$ that satisfy Assumptions 2 and 3 and the corresponding family of empirical p-values* (2.10)*. For any $i \in \mathcal{H}_0$, define*

$$W_i = \big(\{S_{k+1}, \dots, S_n, S_{n+i}\}, (S_{n+j}, j \in \mathcal{H}_0, j \neq i), (S_{n+j}, j \in \mathcal{H}_1)\big) \tag{A.1}$$

and, for any $j \in \{1, \dots, m\} \backslash \{i\}$,

$$C_{i,j} = (\ell + 1)^{-1} \sum_{s \in \{S_{k+1}, \dots, S_n, S_{n+i}\}} \mathbb{1}\, s > S_{n+j}. \tag{A.2}$$

Further, let $S_{(1)} > \cdots > S_{(\ell+1)}$ *be the order statistics of* $\{S_{k+1}, \dots, S_n, S_{n+i}\}$. *Then the following holds:*

(i) The sequences $(S_{n+j})_{j \in \{1,\dots,m\} \backslash \{i\}}$ *and* $(C_{i,j})_{j \in \{1,\dots,m\} \backslash \{i\}}$ *are both measurable with respect to* W_i.

(ii) For all $j \in \{1, \dots, m\} \backslash \{i\}$,

$$p_j = C_{i,j} + \mathbb{1}\, S_{n+i} \leq S_{n+j}/(\ell + 1) = C_{i,j} + \mathbb{1}\, S_{((\ell+1)p_i)} \leq S_{n+j}/(\ell + 1), \tag{A.3}$$

which is a nondecreasing function of p_i *for any given* W_i.

(iii) p_i *is independent of* W_i.

(iv) $(\ell + 1)p_i$ *is uniform distributed on* $\{1, \dots, \ell + 1\}$.

The second result characterizes the distribution of other null p-values conditional on a given null p-value as well as the ordered scores. We defer the proof to Section A.1.8.

Theorem 19. *In the setting of Theorem 18, the distribution of the family* $(p_j,\ j \in \mathcal{H}_0)$ *satisfies*

$$\begin{aligned}
&(p_j,\ j \in \mathcal{H}_0 \backslash \{i\}) \,|\, p_i = 1/(\ell + 1), \{S_{(1)}, \dots\ S_{(\ell+1)}\} \\
&\sim (C_{i,j} + \mathbb{1}\, S_{(1)} \leq S_{n+j}/(\ell + 1),\ j \in \mathcal{H}_0 \backslash \{i\}) \,|\, \{S_{(1)}, \dots\ S_{(\ell+1)}\} \\
&\sim (p'_j,\ j \in \mathcal{H}_0 \backslash \{i\}) \,|\, \{U_{(1)}, \dots\ U_{(\ell+1)}\}
\end{aligned}$$

for which $U_{(1)} > \cdots > U_{(\ell+1)}$ *denote the order statistics of* $\ell+1$ *i.i.d.* $U(0, 1)$ *random variables* $U_1, \dots, U_{\ell+1}$, *and where* p'_j, $j \in \mathcal{H}_0 \backslash \{i\}$, *are conditionally on* $\{U_{(1)}, \dots\ U_{(\ell+1)}\}$, *i.i.d. with common distribution* $F(x) = (1 - U_{(\lfloor x(\ell+1) \rfloor + 1)})\, \mathbb{1}\, 1/(\ell + 1) \leq x < 1 + \mathbb{1}\, x \geq 1$, $x \in \mathbb{R}$.

Noting that $U_{(b)}$ follows a beta distribution, Theorem 19 can be used to compute the distribution of $(p_j,\ j \in \mathcal{H}_0 \backslash \{i\})$ conditionally on $p_i = 1/(\ell + 1)$ by a simple integration. This is used in the proof of FDR control for π_0-adaptive AdaDetect (Theorem 6 and Corollary 7).

A.1.3 Proof of Theorem 3

By Theorem 18 (iii), p_i is independent of W_i. Hence, by integration, it is sufficient to establish that for any nondecreasing measurable set $D \subset [0, 1]^{m-1}$, the function

$$r \in \{1, \dots, \ell + 1\} \mapsto \mathbb{P}((p_j, j \neq i) \in D \,|\, p_i = r/(\ell + 1), W_i) \tag{A.4}$$

is nondecreasing. By Theorem 18 (ii), we have that $(p_j, j \neq i)$ is a deterministic function of p_i and W_i, which is nondecreasing in p_i. This gives (A.4) and proves Theorem 3.

A.1.4 Proof of Theorem 4

This proof combines Theorem 18 with Lemma 25, a property of the BH algorithm that slightly extends the classical result.

Recall that the AdaDetect procedure is the BH algorithm applied to the empirical p-values given by (2.10). We apply Lemma 25 with the empirical p-values $(p_j, 1 \le j \le m)$ being the empirical p-values, any $i \in \mathcal{H}_0$, $p'_i = 1/(\ell+1)$, and $p'_j = C_{i,j}$ defined in (A.2) for $j \neq i$. By definition, the condition (A.23) holds. Moreover, if $p_j > p_i$, we have $S_{n+i} > S_{n+j}$ in which case $\mathbb{1}\, S_{n+i} \le S_{n+j} = 0$, implying that $p'_j = p_j$. Letting $\widehat{k} = \widehat{k}(p_i, 1 \le i \le m)$ and $\widehat{k}'_i = 1 \vee \widehat{k}(p'_i, 1 \le i \le m)$, Lemma 25 entails that

$$\{p_i \le \alpha \widehat{k}/m\} = \{p_i \le \alpha \widehat{k}'_i/m\} \subset \{\widehat{k} = \widehat{k}'_i\}.$$

Let W_i be defined in (A.1). Then

$$\begin{aligned}
\mathrm{FDR}(\mathrm{AdaDetect}_\alpha) &= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\, p_i \le \alpha \hat{k}/m}{\hat{k} \vee 1}\right] \\
&= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\, p_i \le \alpha \hat{k}'_i/m}{\hat{k}'_i}\right] \\
&= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\, p_i \le \alpha \hat{k}'_i/m}{\hat{k}'_i} \,\middle|\, W_i\right]\right] \\
&= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{P}\left((\ell+1)p_i \le \alpha(\ell+1)\hat{k}'_i/m \,\middle|\, W_i\right)}{\hat{k}'_i}\right],
\end{aligned}$$

where the last line is due to that $\hat{k}'_i$ is measurable with respect to W_i, which is implied by Theorem 18 (i). Then Theorem 18 (iii) and (iv) implies

$$\mathrm{FDR}(\mathrm{AdaDetect}_\alpha) = \sum_{i \in \mathcal{H}_0} \mathbb{E}\left(\frac{\lfloor \alpha(\ell+1)\hat{k}'_i/m \rfloor}{(\ell+1)\hat{k}'_i}\right).$$

The result is then proved by letting $K_i = \hat{k}'_i$.

A.1.5 Proof of Theorem 6

Letting $\widehat{k} = \widehat{k}(p_i, 1 \le i \le m)$ the number of rejections of $\mathrm{AdaDetect}_{\alpha m/G(p)}$, we have

$$\begin{aligned}
\mathrm{FDR}(\mathrm{AdaDetect}_{\alpha m/G(p)}) &= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\, p_i \le \alpha \hat{k}(p)/G(p)}{\hat{k}(p) \vee 1}\right] \\
&= \sum_{i \in \mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{1}\, p_i \le \alpha (\hat{k}(p) \vee 1)/G(p)}{\hat{k}(p) \vee 1}\right]
\end{aligned}$$

By Theorem 18 (ii), we can write $(p_j, j \neq i)$ as $\Psi_1(W_i, p_i)$ for some deterministic function Ψ_1 that is nondecreasing in p_i where W_i is defined in (A.1). As a result, we can write $\hat{k}(p) \vee 1$ (resp. $1/G(p)$) as $\Psi_2(W_i, p_i)$ (resp. $\Psi_3(W_i, p_i)$) for some deterministic functions Ψ_2, Ψ_3 that are nonincreasing in p_i since $\hat{k}$ and $1/G$ are both coordinate-wise nonincreasing. Let

$$\begin{aligned}
&c^*(W_i) = \max \mathcal{N}(W_i) \\
&\mathcal{N}(W_i) = \{j/(\ell+1) \,:\, 1 \le j \le \ell+1, j/(\ell+1) \le \alpha \Psi_2(W_i, j/(\ell+1))\Psi_3(W_i, j/(\ell+1))\}.
\end{aligned}$$

Above, we define $c^*(W_i) = 1/(\ell+1)$ if $\mathcal{N}(W_i) = \emptyset$. By definition, $\mathcal{N}(W_i)$ is thus the set of all values that the empirical p-value p_i can take if it is rejected and $c^*(W_i)$ is the largest possible value. Thus, $\{p_i \leq \alpha\hat{k}(p)/G(p)\} = \{p_i \in \mathcal{N}(W_i)\} = \{p_i \leq c^*(W_i), \mathcal{N}(W_i) \neq \emptyset\}$. This entails

$$\begin{aligned}
\mathrm{FDR}(\mathrm{AdaDetect}_{\alpha m/G(p)}) &\leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\, p_i \leq c^*(W_i)}{\hat{k}(p)\vee 1}\,\mathbb{1}\,\mathcal{N}(W_i)\neq\emptyset \mid W_i\right]\right] \\
&\leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}\, p_i \leq c^*(W_i)}{\Psi_2(W_i, c^*(W_i))}\,\mathbb{1}\,\mathcal{N}(W_i)\neq\emptyset \mid W_i\right]\right] \\
&= \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{\mathbb{P}\left[p_i \leq c^*(W_i) \mid W_i\right]}{\Psi_2(W_i, c^*(W_i))}\,\mathbb{1}\,\mathcal{N}(W_i)\neq\emptyset\right] \\
&\leq \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\frac{c^*(W_i)}{\Psi_2(W_i, c^*(W_i))}\,\mathbb{1}\,\mathcal{N}(W_i)\neq\emptyset\right],
\end{aligned}$$

where the last two lines use Theorem 18 (iii) and (iv), respectively. By definition of $c^*(W_i)$, we obtain

$$\begin{aligned}
\mathrm{FDR}(\mathrm{AdaDetect}_{\alpha m/G(p)}) &\leq \alpha \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\Psi_3(W_i, c^*(W_i))\right] \\
&\leq \alpha \sum_{i\in\mathcal{H}_0} \mathbb{E}\left[\Psi_3(W_i, 1/(\ell+1))\right] \\
&\leq \alpha \sum_{i\in\mathcal{H}_0} \mathbb{E}\left(\frac{1}{G(q'_j, 1\leq j\leq m)}\right),
\end{aligned}$$

where $q'_i = 1/(\ell+1)$, $q'_j = 0$ for $j \in \mathcal{H}_1$ and $q'_j = C_{i,j} + \mathbb{1}\, S_{(1)} \leq S_{n+j}/(\ell+1)$ for $j \in \mathcal{H}_0\backslash\{i\}$. The proof is completed by applying Theorem 19.

A.1.6 Proof of Corollary 7

By using (2.14) in Theorem 6, the result is established if we prove in each case

$$\sum_{i\in\mathcal{H}_0} \mathbb{E}\left(\frac{1}{G(p'_j, 1\leq j\leq m)}\right) \leq 1, \tag{A.5}$$

with $p' = (p'_j, 1 \leq j \leq m) \sim \mathcal{D}_i$ defined in (2.13).

Proof for the Storey estimator In this case,

$$G(p') = \frac{1 + \sum_{j=1}^m \mathbb{1}\, p'_j \geq \lambda}{1-\lambda} = \frac{1 + \sum_{j\in\mathcal{H}_0\backslash\{i\}} \mathbb{1}\, p'_j \geq \lambda}{1-\lambda},$$

with $\lambda = K/(\ell+1)$ for $K \in \{2, \ldots, \ell\}$. Recall that, conditionally on $\{U_{(1)}, \ldots, U_{(\ell+1)}\}$, $(p'_j : j \in \mathcal{H}_0\backslash\{i\})$ are i.i.d. with the c.d.f. $F(x) = (1 - U_{(\lfloor x(\ell+1)\rfloor+1)})\,\mathbb{1}\, 1/(\ell+1) \leq x < 1 + \mathbb{1}\, x \geq 1$, $x \in \mathbb{R}$. Therefore, we have for $j \in \mathcal{H}_0\backslash\{i\}$,

$$\begin{aligned}
\mathbb{P}(p'_j \geq \lambda \mid \{U_{(1)}, \ldots, U_{(\ell+1)}\}) &= \mathbb{P}(p'_j \geq K/(\ell+1) \mid \{U_{(1)}, \ldots, U_{(\ell+1)}\}) \\
&= \mathbb{P}(p'_j > (K-1)/(\ell+1) \mid \{U_{(1)}, \ldots, U_{(\ell+1)}\}) \\
&= 1 - (1 - U_{(K)}) = U_{(K)}.
\end{aligned}$$

Thus, it is enough to prove

$$\mathbb{E}\left(\frac{1}{1+\mathcal{B}(m_0-1,U_{(K)})}\right) \leq \frac{1}{m_0(1-\lambda)}, \tag{A.6}$$

where $\mathcal{B}(m_0-1,U_{(K)})$ denotes a Binomial random variable with parameters m_0-1 and $U_{(K)}$. By Lemma 1 in Benjamini et al. (2006),

$$\mathbb{E}\left(\frac{1}{1+\mathcal{B}(m_0-1,U_{(K)})} \mid \{U_{(1)},\dots,U_{(\ell+1)}\}\right) \leq \frac{1}{m_0 U_{(K)}}.$$

Hence, the LHS of (A.6) is bounded by $\mathbb{E}\left(\frac{1}{m_0 U_{(K)}}\right)$. It is well-known that $U_{(K)} \sim \beta(\ell+2-K,K)$ (note that $U_{(1)},\dots,U_{(\ell+1)}$ is a decreasing sequence) and the expectation of the inverse of a Beta random variable with scale parameters a and b is $(a+b-1)/(a-1)$. Hence,

$$\mathbb{E}\left(\frac{1}{m_0 U_{(K)}}\right) = \frac{\ell+1}{m_0(\ell+1-K)} = \frac{1}{m_0(1-\lambda)}.$$

Proof for the Quantile estimator In that case,

$$G(p') = \frac{m-k_0+1}{1-p'_{(k_0)}},$$

where $p'_{(k_0)}$ denotes the k_0-smallest element of $(p'_j, 1\leq j\leq m)$. If $k_0 \leq m-m_0+1$, we have $G(p') \geq m_0$ and (A.5) holds. Thus, we assume $k_0 \geq m-m_0+2$, in which case $k_0+m_0-m-1\geq 1$. Let $j_0 = k_0+m_0-m-1 \in [1,m_0]$. For the rest of the proof, we will fix i and write $\mathbb{E}[\cdot]$ for $\mathbb{E}_{p'\sim\mathcal{D}_i}[\cdot]$ in short. Then

$$\mathbb{E}\left(\frac{1}{G(p'_j, 1\leq j\leq m)}\right) = \mathbb{E}\left(\frac{1-p'_{(j_0:\mathcal{H}_0\setminus\{i\})}}{m-k_0+1}\right),$$

where $p'_{(j_0:\mathcal{H}_0\setminus\{i\})}$ denotes the j_0-smallest element of $(p'_j, j\in\mathcal{H}_0\setminus\{i\})$. To prove (A.5), it is left to prove

$$\mathbb{E}\left(1-p'_{(j_0:\mathcal{H}_0\setminus\{i\})}\right) \leq \frac{m-k_0+1}{m_0} \iff \mathbb{E}(p'_{(j_0:\mathcal{H}_0\setminus\{i\})}) \geq j_0/m_0. \tag{A.7}$$

By definition of $\mathcal{D}_i$,

$$\begin{aligned}
\mathbb{E}(p'_{(j_0:\mathcal{H}_0\setminus\{i\})} \mid \{U_{(1)},\dots,U_{(\ell+1)}\}) &= \int_0^\infty \mathbb{P}(p'_{(j_0:\mathcal{H}_0\setminus\{i\})} > x \mid \{U_{(1)},\dots,U_{(\ell+1)}\})dx \\
&= \int_0^\infty \mathbb{P}\left(\sum_{j\in\mathcal{H}_0\setminus\{i\}} \mathbb{1}\, p'_j \leq x < j_0 \mid \{U_{(1)},\dots,U_{(\ell+1)}\}\right) dx \\
&= \int_0^\infty \mathbb{P}\left(\mathcal{B}(m_0-1,F(x)) < j_0\right) dx,
\end{aligned}$$

where $\mathcal{B}(m_0-1,F(x))$ denotes a Binomial random variable with parameters m_0-1 and $F(x)$, where $F(x) = (1-U_{(\lfloor x(\ell+1)\rfloor+1)})\,\mathbb{1}\,1/(\ell+1)\leq x < 1+\mathbb{1}\,x\geq 1$, $x\in\mathbb{R}$. Hence, the last display

is equal to

$$
\begin{aligned}
&(\ell+1)^{-1} + \sum_{b=2}^{\ell+1} \int_0^1 \mathbb{1}\left\lfloor x(\ell+1) \right\rfloor + 1 = b\mathbb{P}\left(\mathcal{B}(m_0-1, 1-U_{(b)}) < j_0\right) dx \\
&= (\ell+1)^{-1} + (\ell+1)^{-1} \sum_{b=2}^{\ell+1} \mathbb{P}\left(\mathcal{B}(m_0-1, 1-U_{(b)}) < j_0\right) \\
&\geq (\ell+1)^{-1} \sum_{b=1}^{\ell+1} \mathbb{P}\left(\mathcal{B}(m_0-1, 1-U_{(b)}) < j_0\right).
\end{aligned}
$$

Hence,

$$
\mathbb{E}(p'_{(j_0:\mathcal{H}_0\setminus\{i\})}) \geq \sum_{k=0}^{j_0-1} \binom{m_0-1}{k} (\ell+1)^{-1} \sum_{b=1}^{\ell+1} \mathbb{E}[(1-U_{(b)})^k U_{(b)}^{m_0-1-k}].
$$

Since $U_{(b)} \sim \beta(\ell+2-b, b)$ (recall that $U_{(1)}, \ldots, U_{(\ell+1)}$ being a decreasing sequence), we have $1-U_{(b)} \sim \beta(b, \ell+2-b)$. Hence,

$$
\begin{aligned}
\mathbb{E}[(1-U_{(b)})^k U_{(b)}^{m_0-1-k}] &= \frac{(\ell+1)!}{(b-1)!(\ell+1-b)!} \int_{\mathbb{R}} x^k (1-x)^{m_0-1-k} x^{b-1} (1-x)^{\ell+2-b-1} dx \\
&= \frac{(\ell+1)!}{(b-1)!(\ell+1-b)!} \frac{(k+b-1)!(m_0+\ell-(k+b))!}{(m_0+\ell)!},
\end{aligned}
$$

where the last line uses the fact that the integrand is proportional to the density of $\beta(k+b, m_0-k+\ell+1-b)$. As a result, we get

$$
\begin{aligned}
&\binom{m_0-1}{k} (\ell+1)^{-1} \sum_{b=1}^{\ell+1} \mathbb{E}[(1-U_{(b)})^k U_{(b)}^{m_0-1-k}] \\
&= m_0^{-1} \binom{m_0+\ell}{m_0}^{-1} \sum_{b=1}^{\ell+1} \binom{k+b-1}{k} \binom{m_0-k-1+\ell-b+1}{m_0-k-1} \\
&= m_0^{-1} \binom{m_0+\ell}{m_0}^{-1} \sum_{b'=0}^{\ell} \binom{k+\ell-b'}{k} \binom{m_0-k-1+b'}{m_0-k-1}.
\end{aligned}
$$

By applying Lemma 21 with $j = m_0-k-1$, $u = k$, $v = \ell$ (hence $j+u = m_0-1$ and $j+u+v = \ell+m_0-1$), the RHS is equal to $1/m_0$. This proves (A.7) and hence the theorem..

A.1.7 Proof of Theorem 18

By Assumption 3, we can assume $(S_{k+1}, \ldots, S_{n+m})$ has no ties throughout the proof. The result (i) is obvious. For (ii), note that $(\ell+1)p_i = 1 + \sum_{s\in\{S_{k+1},\ldots,S_n\}} \mathbb{1}\, s > S_{n+i}$ is the rank of S_{n+i} within the set $\{S_{k+1}, \ldots, S_n, S_{n+i}\}$. Hence, $S_{((\ell+1)p_i)} = S_{n+i}$ and, for all $j \in \{1, \ldots, m\}\setminus\{i\}$,

$$
\begin{aligned}
p_j &= (\ell+1)^{-1} \left(1 + \sum_{\substack{s\in\{S_{k+1},\ldots,S_n,S_{n+i}\} \\ s\neq S_{((\ell+1)p_i)}}} \mathbb{1}\, s > S_{n+j}\right) \\
&= (\ell+1)^{-1} \left(\sum_{s\in\{S_{k+1},\ldots,S_n,S_{n+i}\}} \mathbb{1}\, s > S_{n+j} + \mathbb{1}\, S_{((\ell+1)p_i)} \leq S_{n+j}\right).
\end{aligned}
$$

This proves (ii). By Assumption 2, for any permutation σ of $\{k+1,\dots,n,n+i\}$, we have

$$((S_{k+1},\dots,S_n,S_{n+i}),W_i) \sim ((S_{\sigma(k+1)},\dots,S_{\sigma(n)},S_{\sigma(n+i)}),W_i^\sigma) = ((S_{\sigma(k+1)},\dots,S_{\sigma(n)},S_{\sigma(n+i)}),W_i),$$

where

$$W_i^\sigma = \big(\{S_{\sigma(k+1)},\dots,S_{\sigma(n)},S_{\sigma(n+i)}\},(S_{n+j},j\in\mathcal{H}_0,j\neq i),(S_{n+j},j\in\mathcal{H}_1)\big) = W_i.$$

This implies that $(S_{k+1},\dots,S_n,S_{n+i})$ is exchangeable conditionally on W_i. Now let $R_1,\dots,R_{\ell+1}$ be the ranks of $(S_{k+1},\dots,S_n,S_{n+i})$ within the same set. Then $S_{j+k}=S_{(R_j)}$ $(j=1,\dots,\ell)$ and $S_{n+i}=S_{(R_{\ell+1})}$, where $S_{(1)}>S_{(2)}>\cdots>S_{(\ell+1)}$ are the order statistics. Since $(S_{k+1},\dots,S_n,S_{n+i})$ are exchangeable conditionally on W_i and almost surely mutually distinct, we have that

$$(R_1,\dots,R_{\ell+1}) \perp\!\!\!\perp W_i, \quad \text{and} \quad (R_1,\dots,R_{\ell+1}) \sim \text{Unif}(\mathfrak{S}(\{1,\dots,n-k+1\})),$$

where $\mathfrak{S}(\{1,\dots,\ell+1\})$ denotes the set of permutations of $\{1,\dots,\ell+1\}$. The results (iii) and (iv) then follow from the fact that $p_i = R_{\ell+1}/(\ell+1)$.

A.1.8 Proof of Theorem 19

By (A.3), p_j is a function of p_i and W_i for all $j\in\mathcal{H}_0\backslash\{i\}$. Replacing p_i by $1/(\ell+1)$ in that expression, we get

$$\begin{aligned}
p_j &= (\ell+1)^{-1} \sum_{s\in\{S_{k+1},\dots,S_n,S_{n+i}\}} \mathbb{1}\, s > S_{n+j} + \mathbb{1}\, S_{(1)} \le S_{n+j}/(\ell+1) \\
&= (\ell+1)^{-1} \sum_{q=1}^{\ell+1} \mathbb{1}\, S_{(q)} > S_{n+j} + \mathbb{1}\, S_{(1)} \le S_{n+j}/(\ell+1) \\
&= (\ell+1)^{-1} \left(1 + \sum_{q=2}^{\ell+1} \mathbb{1}\, S_{(q)} > S_{n+j}\right).
\end{aligned}$$

By Lemma 23, we have

$$\begin{aligned}
&(p_j,\ j\in\mathcal{H}_0\backslash\{i\}) \,|\, p_i = 1/(\ell+1), \{S_{(1)},\dots\, S_{(\ell+1)}\} \\
&\sim (p'_j,\ j\in\mathcal{H}_0\backslash\{i\}) \,|\, p'_i = 1/(\ell+1), \{U_{(1)},\dots\, U_{(\ell+1)}\} \\
&\sim (p'_j,\ j\in\mathcal{H}_0\backslash\{i\}) \,|\, \{U_{(1)},\dots\, U_{(\ell+1)}\}
\end{aligned}$$

where $p'_j = \frac{1+\sum_{q=2}^{\ell+1}\mathbb{1}\, U_{(q)}>V_j}{\ell+1}$, $U_1,\dots,U_{\ell+1},V_j (j\in\mathcal{H}_0)$ are i.i.d. from $U(0,1)$, and $U_{(1)}>\cdots>U_{(\ell+1)}$ denote the order statistics of $U_1,\dots,U_{\ell+1}$. As a result, conditional on $\{U_{(1)},\dots\, U_{(\ell+1)}\}$, $(p'_j,\ j\in\mathcal{H}_0\backslash\{i\})$ are i.i.d. with a c.d.f.

$$\begin{aligned}
F(x) &= \mathbb{P}(p'_j \le x \,|\, U_{(1)},\dots,U_{(\ell+1)}) \\
&= \mathbb{P}\left(\sum_{q=2}^{\ell+1} \mathbb{1}\, U_{(q)} > V_j \le \lfloor x(\ell+1)\rfloor - 1 \,|\, U_{(1)},\dots,U_{(\ell+1)}\right) \\
&= \mathbb{P}(U_{(\lfloor x(\ell+1)\rfloor+1)} \le V_j) = 1 - U_{(\lfloor x(\ell+1)\rfloor+1)}.
\end{aligned}$$

because $\sum_{q=2}^{\ell+1}\mathbb{1}\, U_{(q)} > v \ge \lfloor x(\ell+1)\rfloor$ is equivalent to $U_{(\lfloor x(\ell+1)\rfloor+1)} > v$. This completes the proof.

A.2 Proofs for Section 4

A.2.1 Proof of Theorem 8

Let $T(x) = 1 - r(x) = \pi_0 f_0(x)/f(x)$ and $t(\alpha) = 1 - c(\alpha) \in (0,1)$. Then $R = \{i : T(X_i) \le t(\alpha)\}$. Consider any procedure $R' = \{i : T'(X_i) \le t'\}$ with $\mathrm{mFDR}(R') \le \alpha$. Since $\mathrm{mFDR}(R) = \alpha$, we have both

$$0 = \int \mathbb{1}\, T(x) \le t(\alpha)(T(x) - \alpha) f(x) d\nu(x)$$
$$0 \ge \int \mathbb{1}\, T'(x) \le t'(T(x) - \alpha) f(x) d\nu(x).$$

The first equality implies $t(\alpha) \ge \alpha$. If $t(\alpha) = \alpha$, then $T(X) = \alpha$ almost surely under f. This implies that all hypotheses are rejected with probability 1 and hence R is never less powerful than R'.

Assume $t(\alpha) > \alpha$. Then the two equalities imply

$$\int (\mathbb{1}\, T(x) \le t(\alpha) - \mathbb{1}\, T'(x) \le t')(T(x) - \alpha) f(x) d\nu(x) \ge 0. \tag{A.8}$$

Since $T(x) \le t(\alpha)$ is equivalent to $\frac{T(x)-\alpha}{1-T(x)} \le \frac{t(\alpha)-\alpha}{1-t(\alpha)}$ (even when $T(x) = 1$), we obtain

$$\frac{t(\alpha) - \alpha}{1 - t(\alpha)} \int (\mathbb{1}\, T(x) \le t(\alpha) - \mathbb{1}\, T'(x) \le t')(1 - T(x)) f(x) d\nu(x) \ge 0.$$

Since $t(\alpha) > \alpha$,

$$\int (\mathbb{1}\, T(x) \le t(\alpha) - \mathbb{1}\, T'(x) \le t') \bar{f}_1(x) d\nu(x) \ge 0. \tag{A.9}$$

A.2.2 Proof of Lemma 9

For case (i), (2.22) can be expressed as

$$2J_\lambda(g) = \int \Big\{ k(1 + g(x))_+ + \lambda(\ell + m)(1 - g(x))_+ f_\gamma(x)/f_0(x) \Big\} f_0(x) d\nu(x).$$

For any $u, v > 0$ and $a \in \mathbb{R}$,

$$u(1+a)_+ + v(1-a)_+ = v(1-a)\,\mathbb{1}\, a < -1 + (u+v+a(u-v))\,\mathbb{1}\, -1 \le a \le 1 + u(1+a)\,\mathbb{1}\, a > 1.$$

As a function of a, it is continuous and piecewise linear with two turning points $(-1, 2v)$ and $(1, 2u)$. When $u \ne v$, the unique minimum is attained at $a = \mathrm{sign}(v/u - 1)$. The proof is then completed by setting $u = k$ and $v = \lambda(\ell + m) f_\gamma(x)/f_0(x)$ and the assumption that $\mathbb{P}(f_\gamma(X)/f_0(X) = k/\lambda(\ell + m)) = 0$.

For case (ii), (2.22) is given by

$$J_\lambda(g) = \int \Big\{ -k \log(1 - g(x)) - \lambda(\ell + m) \log(g(x)) f_\gamma(x)/f_0(x) \Big\} f_0(x) d\nu(x).$$

For any $u > 0$, $v \ge 0$, the map $a \in [0,1] \mapsto u \log(1-a) + v \log(a)$ has a unique maximizer at $a = v/(u+v)$.

A.3 Proofs for Section 5

A.3.1 Proof of Theorem 10

For all $g \in \mathcal{G}$, let

$$\tilde{R}_0(g) = \ell^{-1} \sum_{i=k+1}^{n} \mathbb{1}\, g(Z_i) \geq 0; \tag{A.10}$$

$$\hat{R}_{\gamma,0}(g) = (m_0+\ell)^{-1} \left(\sum_{i=k+1}^{n} \mathbb{1}\, g(Z_i) < 0 + \sum_{i \in \mathcal{H}_0} \mathbb{1}\, g(Z_{n+i}) < 0 \right); \tag{A.11}$$

$$\hat{R}_{\gamma,1}(g) = m_1^{-1} \sum_{i \in \mathcal{H}_1} \mathbb{1}\, g(Z_{n+i}) < 0, \tag{A.12}$$

so that $\hat{R}_\gamma(g) = (m+\ell)^{-1} \sum_{i=k+1}^{n+m} \mathbb{1}\, g(Z_i) < 0 = (1-\gamma)\hat{R}_{\gamma,0}(g) + \gamma \hat{R}_{\gamma,1}(g)$. For notational convenience, for any $i \geq 1$, let

$$e_i = \sqrt{\frac{V(\mathcal{G}) + \log(1/\delta)}{i}}.$$

Define the following events:

$$\Omega_0 = \left\{ \sup_{g \in \mathcal{G}} |\hat{R}_0(g) - R_0(g)| \leq b e_k \right\};$$

$$\tilde{\Omega}_0 = \left\{ \sup_{g \in \mathcal{G}} |\tilde{R}_0(g) - R_0(g)| \leq b e_\ell \right\};$$

$$\Omega_{\gamma,0} = \left\{ \sup_{g \in \mathcal{G}} |\hat{R}_{\gamma,0}(g) - (1 - R_0(g))| \leq b e_{m_0+\ell} \right\};$$

$$\Omega_{\gamma,1} = \left\{ \sup_{g \in \mathcal{G}} |\hat{R}_{\gamma,1}(g) - R_1(g)| \leq b e_{m_1} \right\}.$$

We choose b such that

$$\mathbb{P}(\Omega_0 \cap \tilde{\Omega}_0 \cap \Omega_{\gamma,0} \cap \Omega_{\gamma,1}) \geq 1 - \delta. \tag{A.13}$$

The well-known result for empirical processes on finite VC classes (e.g., Example 7.10 of Sen (2018)) implies that b only depends on δ. Throughout the rest of the proof, we choose

$$C = b, \quad C' = 30b, \tag{A.14}$$

where C and C' are the constants in expressions of ϵ_0 and Δ, respectively.

Note that on $\Omega_{\gamma,0} \cap \Omega_{\gamma,1}$, we have

$$\begin{aligned} &\sup_{g \in \mathcal{G}} |\hat{R}_\gamma(g) - R_\gamma(g)| \leq b(1-\gamma) e_{m_0+\ell} + b\gamma e_{m_1} \\ &= b \left(\sqrt{\frac{m_0+\ell}{m+\ell}} + \sqrt{\frac{m_1}{m+\ell}} \right) e_{m+\ell} \leq 2b e_{m+\ell}. \end{aligned} \tag{A.15}$$

On Ω_0, (2.27) implies that

$$R_0(\hat{g}) = \hat{R}_0(\hat{g}) + R_0(\hat{g}) - \hat{R}_0(\hat{g}) \leq \beta + \epsilon_0 + b e_k = \beta + 2b e_k.$$

Clearly, $2be_k \le 30\gamma^{-1}be_k \le \Delta$. This proves the first claim of (i). Moreover, on $\Omega_0 \cap \tilde{\Omega}_0$, we have

$$\tilde{R}_0(\hat{g}) = R_0(\hat{g}) + \tilde{R}_0(\hat{g}) - R_0(\hat{g}) \le \beta + 3b(e_k \vee e_\ell) \le \beta + 0.1\gamma\Delta.$$

Equivalently,

$$\sum_{i=k+1}^{n} \mathbb{1}\,\hat{g}(Z_i) \ge 0 \le \ell(\beta + 0.1\gamma\Delta). \tag{A.16}$$

By the assumption that $R_0(g_{\mathcal{G}}^\sharp) = \beta$, on the event Ω_0, $\hat{R}_0(g_{\mathcal{G}}^\sharp) \le R_0(g_{\mathcal{G}}^\sharp) + be_k = \beta + \epsilon_0$. By definition (2.27) of $\hat{g}$, we have

$$\hat{R}_\gamma(\hat{g}) \le \hat{R}_\gamma(g_{\mathcal{G}}^\sharp). \tag{A.17}$$

By (A.15) and (A.17), on $\Omega_0 \cap \Omega_{\gamma,0} \cap \Omega_{\gamma,1}$,

$$R_\gamma(\hat{g}) = \hat{R}_\gamma(\hat{g}) + R_\gamma(\hat{g}) - \hat{R}_\gamma(\hat{g}) \le \hat{R}_\gamma(g_{\mathcal{G}}^\sharp) + 2be_{m+\ell} \le R_\gamma(g_{\mathcal{G}}^\sharp) + 4be_{m+\ell}.$$

Together with (2.26), this implies

$$(1-\gamma)(1 - R_0(\hat{g})) + \gamma R_1(\hat{g}) \le (1-\gamma)(1 - R_0(g_{\mathcal{G}}^\sharp)) + \gamma R_1(g_{\mathcal{G}}^\sharp) + 4be_{m+\ell}$$

and thus

$$R_1(\hat{g}) \le R_1(g_{\mathcal{G}}^\sharp) + \gamma^{-1}(R_0(\hat{g}) - R_0(g_{\mathcal{G}}^\sharp) + 4be_{m+\ell}).$$

By definition, $R_0(\hat{g}) \le \beta + \epsilon_0 = R_0(g_{\mathcal{G}}^\sharp) + \epsilon_0 = R_0(g_{\mathcal{G}}^\sharp) + be_k$. Thus,

$$R_1(\hat{g}) \le R_1(g_{\mathcal{G}}^\sharp) + \gamma^{-1}b[4e_{m+\ell} + e_k] \le R_1(g_{\mathcal{G}}^\sharp) + 5\gamma^{-1}be_k.$$

Clearly, $5\gamma^{-1}be_k \le 30\gamma^{-1}be_k \le \Delta$. This proves the second claim of (i). Then, on the event $\Omega_{\gamma,1}$, we have

$$\hat{R}_{\gamma,1}(\hat{g}) \le R_1(\hat{g}) + be_{m_1} \le R_1(g_{\mathcal{G}}^\sharp) + be_{m_1} + 5\gamma^{-1}be_k.$$

Since $e_{m_1} = \gamma^{-1/2}e_{m+\ell} \le \gamma^{-1}e_{m+\ell} \le \gamma^{-1}e_k$, we have

$$\hat{R}_{\gamma,1}(\hat{g}) \le R_1(g_{\mathcal{G}}^\sharp) + 6\gamma^{-1}be_k.$$

Since $C' \ge 6b$,

$$\Delta = C'\gamma^{-1}(e_k \vee e_\ell) \ge 6\gamma^{-1}be_k.$$

Thus, on $\Omega_0 \cap \Omega_{\gamma,0} \cap \Omega_{\gamma,1}$,

$$\frac{1}{m_1}\sum_{i\in\mathcal{H}_1} \mathbb{1}\,\hat{g}(Z_{n+i}) < 0 = \hat{R}_{\gamma,1}(\hat{g}) \le R_1(g_{\mathcal{G}}^\sharp) + \Delta.$$

Equivalently, recalling that $X_i = Z_{n+i}$,

$$\sum_{i\in\mathcal{H}_1} \mathbb{1}\,\hat{g}(X_i) \ge 0 \ge M. \tag{A.18}$$

Let $\eta = 1/\ell + \beta + 0.1\gamma\Delta$. Then (A.16) implies that

$$1 + \sum_{i=k+1}^{n} \mathbb{1}\,\hat{g}(Z_i) \ge 0 \le \eta\ell.$$

To validate the conditions in Lemma 24, we only need to show $\eta \le \alpha M/m$. Since $1-R_1(g_{\mathcal{G}}^\sharp) \ge (1+\alpha^{-1})\Delta$, we have

$$\Delta \le \alpha(1 - R_1(g_{\mathcal{G}}^\sharp) - \Delta) \le \frac{\alpha M}{m_1} \le \frac{\alpha M}{\gamma m}.$$

By the assumptions that $\ell \ge 2m/(\alpha M)$ and $\beta \le 0.4\alpha M/m$, we have

$$\eta \le \frac{\alpha M}{2m} + \frac{0.4\alpha M}{m} + \frac{0.1\alpha M}{m} = \frac{\alpha M}{m}.$$

By Lemma 24, on the event $\Omega_0 \cap \tilde{\Omega}_0 \cap \Omega_{\gamma,0} \cap \Omega_{\gamma,1}$,

$$\mathrm{AdaDetect}_\alpha \supset \{i \in \{1,\dots,m\} : \hat{g}(X_i) \ge 0\}.$$

By (A.18),

$$|\mathrm{AdaDetect}_\alpha \cap \mathcal{H}_1|/m_1 \ge M/m_1 \ge 1 - R_1(g_{\mathcal{G}}^\sharp) - \Delta.$$

By (A.13), this occurs with probability at least $1-\delta$.

A.3.2 Proof of Theorem 11

We first prove the result for BH_α^*, i.e.,

$$\mathbb{P}\left(\mathcal{R}' \cap \{\mathrm{BH}_\alpha^* \subset \mathrm{AdaDetect}_{\alpha(1+\delta)(1+\zeta_r(\eta))}\}^c\right) \le \mathbb{P}\left(\hat{\eta} > \eta\right) - me^{-(3/28)(\ell+1)\delta^2\alpha(r\vee 1)/m}, \quad \text{(A.19)}$$

where $\mathcal{R}' = \{|\mathrm{BH}_\alpha^*| \ge r\}$. First we note that, while $\hat{g}(Y_i)$ are dependent through the score function $\hat{g}(\cdot)$, the $g^*(Y_i)$ are i.i.d., allowing us to apply concentration inequalities. For $s \in \mathbb{R}$, define

$$\widehat{G}_0(s) = (\ell+1)^{-1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\,\hat{g}(Y_i) \ge s\right);$$
$$\widehat{G}_0^*(s) = (\ell+1)^{-1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\,g^*(Y_i) \ge s\right).$$

For notational convenience, let $\tilde{\alpha} = \alpha(r\vee 1)/m$. Consider in addition the following events:

$$\Omega_1 = \left\{\max_{k+1\le i\le n+m} |\hat{g}(Z_i) - g^*(Z_i)| \le \eta\right\};$$
$$\Omega_2 = \left\{\sup_{1\le i\le m}\left(\frac{\widehat{G}_0^*(g^*(X_i)-2\eta) - \overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}}{\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}}\right) \le \delta\right\}.$$

On $\Omega_1\cap\Omega_2$, we have for all $i\in\{1,\dots,m\}$,

$$\widehat{G}_0(\hat{g}(X_i)) \le \widehat{G}_0^*(\hat{g}(X_i)-\eta) \le \widehat{G}_0^*(g^*(X_i)-2\eta) \le (\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha})(1+\delta)$$
$$= (\overline{G}_0(g^*(X_i))\vee\tilde{\alpha})\left(1 + \frac{\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha} - \overline{G}_0(g^*(X_i))\vee\tilde{\alpha}}{\overline{G}_0(g^*(X_i))\vee\tilde{\alpha}}\right)(1+\delta).$$

Let $u = \overline{G}_0(g^*(X_i))\vee\tilde{\alpha}$. Since $\overline{G}_0$ is nonincreasing, $u = \overline{G}_0(g^*(X_i)\wedge\overline{G}_0^{-1}(\tilde{\alpha}))$. Thus,

$$\frac{\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha} - \overline{G}_0(g^*(X_i))\vee\tilde{\alpha}}{\overline{G}_0(g^*(X_i))\vee\tilde{\alpha}}$$
$$= \frac{\overline{G}_0((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha})) - u}{u}$$
$$\le \frac{\overline{G}_0(\overline{G}_0^{-1}(u)-2\eta) - u}{u}.$$

Since $u \geq \tilde{\alpha}$, the LHS is bounded by $\zeta_r(\eta)$. As a result, on $\Omega_1 \cap \Omega_2$,

$$\widehat{G}_0(\hat{g}(X_i)) \leq (\overline{G}_0(g^*(X_i)) \vee \tilde{\alpha})(1+\zeta_r(\eta))(1+\delta). \tag{A.20}$$

Then, for all $t \in \{\alpha k/m, r \vee 1 \leq k \leq m\}$,

$$\mathbb{1}\,\overline{G}_0(g^*(X_i)) \leq t = \mathbb{1}\,\overline{G}_0(g^*(X_i)) \vee \tilde{\alpha} \leq t \leq \mathbb{1}\,\widehat{G}_0(\hat{g}(X_i)) \leq t(1+\delta)(1+\zeta_r(\eta)).$$

By applying Lemma 20 with $p_i = \overline{G}_0(g^*(X_i))$, $p_i' = \widehat{G}_0(\hat{g}(X_i))$, $\beta = \alpha$, and $\beta' = \alpha(1+\zeta_r(\eta))(1+\delta)$, we obtain that $\mathrm{BH}^*_\alpha \subset \mathrm{AdaDetect}_{\alpha(1+\delta)(1+\zeta_r(\eta))}$ on $\Omega_1 \cap \Omega_2 \cap \mathcal{R}'$. We are left to show that

$$\mathbb{P}(\Omega_2^c) \leq m \exp(-(3/28)(\ell+1)\delta^2\tilde{\alpha}).$$

Since X_i's and Y_i's are independent, the union bound implies

$$\begin{aligned}
\mathbb{P}(\Omega_2^c) &\leq \mathbb{P}\left(\sup_{1\leq i\leq m}\left(\frac{\widehat{G}_0^*(g^*(X_i)-2\eta) - \overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}}{\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}}\right) \geq \delta\right) \\
&\leq \sum_{i=1}^m \mathbb{P}\left(\frac{\widehat{G}_0^*(g^*(X_i)-2\eta) - \overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}}{\overline{G}_0(g^*(X_i)-2\eta)\vee\tilde{\alpha}} \geq \delta\right) \\
&\leq \sum_{i=1}^m \mathbb{P}\left(\frac{\widehat{G}_0^*(g^*(X_i)-2\eta) - \overline{G}_0((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha}))}{\overline{G}_0((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha}))} \geq \delta\right) \\
&\leq \sum_{i=1}^m \mathbb{P}\left(\frac{\widehat{G}_0^*((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha})) - \overline{G}_0((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha}))}{\overline{G}_0((g^*(X_i)-2\eta)\wedge\overline{G}_0^{-1}(\tilde{\alpha}))} \geq \delta\right) \\
&\leq m \sup_{s\leq\overline{G}_0^{-1}(\tilde{\alpha})} \mathbb{P}\left(\frac{\widehat{G}_0^*(s) - \overline{G}_0(s)}{\overline{G}_0(s)} \geq \delta\right).
\end{aligned}$$

For all $s \leq \overline{G}_0^{-1}\tilde{\alpha}$,

$$\begin{aligned}
&\mathbb{P}\left(\widehat{G}_0^*(s) - \overline{G}_0(s) \geq \delta\overline{G}_0(s)\right) \\
&\leq \mathbb{P}\left(\sum_{i=k+1}^n (\mathbb{1}\,g^*(Y_i) \geq s - \overline{G}_0(s)) \geq (\ell+1)\delta\overline{G}_0(s) - 1\right) \\
&\leq \mathbb{P}\left(\sum_{i=k+1}^n (\mathbb{1}\,g^*(Y_i) \geq s - \overline{G}_0(s)) \geq 0.5(\ell+1)\delta\overline{G}_0(s)\right),
\end{aligned}$$

where the last line uses the fact that $(\ell+1)\delta\overline{G}_0(s) \geq (\ell+1)\delta\tilde{\alpha} \geq 2$. Let $A = 0.5(\ell+1)\delta\overline{G}_0(s)$. Since the $g^*(Y_i)$'s are independent, By Bernstein's inequality,

$$\begin{aligned}
\mathbb{P}(\Omega_2^c) &\leq m \sup_{s\leq\overline{G}_0^{-1}(\tilde{\alpha})} \exp\left(-\frac{A^2}{2(\ell+1)\overline{G}_0(s) + 2A/3}\right) \\
&= m \sup_{s\leq\overline{G}_0^{-1}(\tilde{\alpha})} \exp\left(-0.5\frac{A^2}{4A/\delta + 2A/3}\right) \\
&\leq m \sup_{s\leq\overline{G}_0^{-1}(\tilde{\alpha})} \exp\left(-0.5\frac{A^2}{4A/\delta + 2A/3\delta}\right) \\
&= m \sup_{s\leq\overline{G}_0^{-1}(\tilde{\alpha})} \exp\left(-\frac{3(\ell+1)\delta^2\overline{G}_0(s)}{28}\right) \\
&\leq m\exp(-(3/28)(\ell+1)\delta^2\tilde{\alpha}).
\end{aligned} \tag{A.21}$$

The proof of (A.19) is then completed.

Next, we prove (2.37), i.e., the result for AdaDetect$_\alpha$. Recall that $\mathcal{R} = \{|\text{AdaDetect}^*_\alpha| \geq r\}$. Similar to Ω_2, we define

$$\begin{aligned}\Omega_3 &= \left\{ \sup_{1\leq i\leq m} \left(\frac{\overline{G}_0(g^*(X_i)) \vee \tilde{\alpha} - \widehat{G}^*_0(g^*(X_i)) \vee \tilde{\alpha}}{\widehat{G}^*_0(g^*(X_i)) \vee \tilde{\alpha}} \right) \leq \delta \right\} \\ &= \left\{ \sup_{1\leq i\leq m} \left(\frac{\overline{G}_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha})) - \widehat{G}^*_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha}))}{\widehat{G}^*_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha}))} \right) \leq \delta \right\}.\end{aligned}$$

By (A.20), on $\Omega_1 \cap \Omega_2 \cap \Omega_3$, for all $i \in \{1, \ldots, m\}$ and $t \in \{\alpha k/m, r \vee 1 \leq k \leq m\}$,

$$\begin{aligned}\widehat{G}_0(\hat{g}(X_i)) &\leq (\overline{G}_0(g^*(X_i)) \vee \tilde{\alpha})(1 + \zeta_r(\eta))(1 + \delta) \\ &\leq (\widehat{G}^*_0(g^*(X_i)) \vee \tilde{\alpha})(1 + \zeta_r(\eta))(1 + \delta)^2 \\ &\leq (\widehat{G}^*_0(g^*(X_i)) \vee \tilde{\alpha})(1 + \zeta_r(\eta))(1 + 3\delta).\end{aligned}$$

Thus, on $\Omega_1 \cap \Omega_2 \cap \Omega_3$,

$$\mathbb{1}\, \widehat{G}^*_0(g^*(X_i)) \vee \tilde{\alpha} \leq t \leq \mathbb{1}\, \widehat{G}_0(g(X_i)) \leq t(1 + 3\delta)(1 + \zeta_r(\eta)).$$

Applying Lemma 20 with $p_i = \widehat{G}^*_0(g^*(X_i))$, $p'_i = \widehat{G}_0(g(Z_i))$, $\beta = \alpha$, and $\beta' = \alpha(1 + \zeta_r(\eta))(1 + 3\delta))$, we obtain that AdaDetect$^*_\alpha \subset$ AdaDetect$_{\alpha(1+3\delta)(1+\zeta_r(\eta))}$ on $\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \mathcal{R}$. It remains to prove that

$$\mathbb{P}(\Omega_3^c) \leq m \exp(-(3/28)(\ell + 1)\delta^2\tilde{\alpha}).$$

Similar to $\mathbb{P}(\Omega_2^c)$, we have

$$\begin{aligned}\mathbb{P}(\Omega_3^c) &\leq \mathbb{P}\left(\sup_{1\leq i\leq m} \left(\frac{\overline{G}_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha})) - \widehat{G}^*_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha}))}{\widehat{G}^*_0(g^*(X_i) \wedge \overline{G}_0^{-1}(\tilde{\alpha}))} \right) \geq \delta \right) \\ &\leq m \sup_{s\leq \overline{G}_0^{-1}(\tilde{\alpha})} \mathbb{P}\left(\frac{\overline{G}_0(s) - \widehat{G}^*_0(s)}{\widehat{G}^*_0(s)} \geq \delta \right) \\ &\leq m \sup_{s\leq \overline{G}_0^{-1}(\tilde{\alpha})} \mathbb{P}\left(\widehat{G}^*_0(s) - \overline{G}_0(s) \leq -0.5\delta\overline{G}_0(s) \right),\end{aligned}$$

where the last line uses the fact that $(1 + \delta)^{-1} \leq 1 - 0.5\delta$ for any $\delta \in [0, 1]$. Now, for for all $s \leq \overline{G}_0^{-1}(\tilde{\alpha})$,

$$\begin{aligned}&\mathbb{P}\left(\widehat{G}^*_0(s) - \overline{G}_0(s) \leq -0.5\delta\overline{G}_0(s) \right) \\ &= \mathbb{P}\left(\sum_{i=k+1}^{n} (\mathbb{1}\, g^*(Y_i) \geq s - \overline{G}_0(s)) \leq -0.5(\ell + 1)\delta\overline{G}_0(s) - 1 + \overline{G}_0(s) \right) \\ &\leq \mathbb{P}\left(\sum_{i=k+1}^{n} (\mathbb{1}\, g^*(Y_i) \geq s - \overline{G}_0(s)) \leq -0.5(\ell + 1)\delta\overline{G}_0(s) \right).\end{aligned}$$

Applying the Bernstein inequality as in (A.21), we conclude that $\mathbb{P}(\Omega_3^c) \leq m \exp(-(3/28)(\ell + 1)\delta^2\tilde{\alpha})$.

Remark 16. The inequality (A.19) generalizes the power oracle inequality of Mary and Roquain (2022) for a fixed score function and $r = 0$.

A.3.3 Proof of Corollary 12

Letting $\mathcal{R} = \{|\text{AdaDetect}^*_\alpha| \geq m_1\epsilon\}$, we have

$$\begin{aligned}\text{TDR}(\text{AdaDetect}^*_\alpha) &\leq \mathbb{E}[\text{TDP}(\text{AdaDetect}^*_\alpha)] \\ &\leq \epsilon + \int_\epsilon^1 \mathbb{P}(\text{TDP}(\text{AdaDetect}^*_\alpha) \geq u)du \\ &= \epsilon + \int_\epsilon^1 \mathbb{P}(\mathcal{R} \cap \{\text{TDP}(\text{AdaDetect}^*_\alpha) \geq u\})du\end{aligned}$$

Now applying Theorem 11 with $r = \lceil m_1\epsilon \rceil$, we obtain

$$\begin{aligned}\text{TDR}(\text{AdaDetect}^*_\alpha) &\leq \epsilon + \int_\epsilon^1 \mathbb{P}(\text{TDP}(\text{AdaDetect}_{\alpha'}) \geq u)du + \mathbb{P}\left(\hat{\eta} > \eta\right) + 2me^{-(3/28)(\ell+1)\delta^2\alpha\lceil m_1\epsilon\rceil/m} \\ &\leq \text{TDR}(\text{AdaDetect}_{\alpha'}) + \epsilon + \mathbb{P}\left(\hat{\eta} > \eta\right) + 2me^{-(3/28)(\ell+1)\delta^2\alpha\lceil m_1\epsilon\rceil/m}\end{aligned}$$

which gives (2.38).

A.4 Useful lemmas

Lemma 20. *Let $(p_i, 1 \leq i \leq m)$ and $(p'_i, 1 \leq i \leq m)$ be two sets of p-values and $\beta, \beta' \in (0,1)$, $r \in \{0, \ldots, m\}$. Assume that for all $t \in \{\beta k/m, r \vee 1 \leq k \leq m\}$,*

$$\mathbb{1}\, p_i \leq t \leq \mathbb{1}\, p'_i \leq t\beta'/\beta. \tag{A.22}$$

Then, on the event where the BH algorithm applied to p-values $(p_i, 1 \leq i \leq m)$ at level β has at least r rejections, all these rejections would be rejected by the BH algorithm applied to p-values $(p'_i, 1 \leq i \leq m)$ at level β'.

Proof. Recall that BH algorithm applied to p-values $(p_i, 1 \leq i \leq m)$ at level β rejects the i-th hypothesis iff $p_i \leq \beta\hat{k}/m$ where

$$\hat{k} = \max\left\{k \in \{0, \ldots, m\} \,:\, \sum_{i=1}^m \mathbb{1}\, p_i \leq \beta k/m \geq k\right\}.$$

Similarly, the BH algorithm at level β' rejects i-th hypothesis iff $p'_i \leq \beta\hat{k}'/m$ where

$$\hat{k}' = \max\left\{k \in \{0, \ldots, m\} \,:\, \sum_{i=1}^m \mathbb{1}\, p'_i \leq \beta k/m \geq k\right\}.$$

When $\hat{k} = 0$, the conclusion is trivial. Now assume $\hat{k} \geq r \vee 1$. Setting $t = \beta\hat{k}/m$ in (A.22), we obtain that

$$\mathbb{1}\, p_i \leq \beta\hat{k}/m \leq \mathbb{1}\, p'_i \leq \beta'\hat{k}/m, \quad 1 \leq i \leq m.$$

This implies $\hat{k}' \geq \hat{k} \geq r \vee 1$. Hence, for all i,

$$\mathbb{1}\, p_i \leq \beta\hat{k}/m \leq \mathbb{1}\, p'_i \leq \beta'\hat{k}/m \leq \mathbb{1}\, p'_i \leq \beta'\hat{k}'/m.$$

□

Lemma 21 (Vandermonde's equality)**.** *For all $0 \le j \le k \le n$, we have*

$$\binom{n+1}{k+1} = \sum_{m=j}^{n-k+j} \binom{m}{j}\binom{n-m}{k-j}.$$

Equivalently, for all $j, u, v \ge 0$

$$\binom{j+u+v+1}{j+u+1} = \sum_{b=0}^{v} \binom{j+b}{j}\binom{u+v-b}{u}.$$

Lemma 22. *Consider any set of scores $(S_{k+1}, \dots, S_{n+m})$ and the corresponding empirical p-values $(p_1, \dots, p_m)$ defined in (2.10). Then on the event where the scores $(S_j, j \in \{k+1, \dots, n\} \cup (n + \mathcal{H}_0))$ are mutually distinct, the null p-values $(p_j, j \in \mathcal{H}_0)$ are mesurable with respect to the ranks of $(S_j, j \in \{k+1, \dots, n\} \cup (n + \mathcal{H}_0))$.*

Proof. Let $R_1, \dots, R_{n-k+m_0}$ and $S_{(1)}, \dots, S_{(n-k+m_0)}$ be the ranks and the order statistics of $(S_j, j \in \{k+1, \dots, n\} \cup (n+\mathcal{H}_0))$, respectively. Since the score $(S_j, j \in \{k+1, \dots, n\} \cup (n+\mathcal{H}_0))$ are mutually distinct, we have that $S_i > S_j$ iff $R_i < R_j$. Hence, by (2.10), we obtain for all $j \in \mathcal{H}_0$,

$$p_j = \frac{1}{n-k+1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\, S_i > S_{n+j}\right) = \frac{1}{n-k+1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\, R_i < R_{n+j}\right).$$

□

Lemma 22 implies the following equivalent representation of the distribution of p-values.

Lemma 23. *For any set of scores $(S_{k+1}, \dots, S_{n+m})$ satisfying Assumptions 2 and 3, the joint distribution of the null empirical p-values $(p_j, j \in \mathcal{H}_0)$ does not depend on the joint distribution of scores. In particular, it can be characterized by generating $(S_j, j \in \{k+1, \dots, n\} \cup (n+\mathcal{H}_0))$ i.i.d. from $U(0,1)$.*

Proof. By Lemma 22, almost surely, $(p_j, j \in \mathcal{H}_0)$ is a function of the ranks of $(S_j, j \in \{k+1, \dots, n\} \cup (n+\mathcal{H}_0))$, which is uniformly distributed on the permutations of $\{1, \dots, n-k+m_0\}$ according to Assumption 2. The result follows. □

Lemma 24. *Fix $M \in \{1, \dots, m\}$ and $\eta \in (0,1)$. Assume that the scores $S_{k+1}, \dots, S_n$ in the NTS satisfy $1 + \sum_{i=k+1}^{n} \mathbb{1}\, S_i \ge 0 \le \eta\ell$ and the scores $S_{n+1}, \dots, S_{n+m}$ in the test sample satisfy $\sum_{j=1}^{m} \mathbb{1}\, S_{n+j} \ge 0 \ge M$. Then, if $\eta \le \alpha M/m$, AdaDetect$_\alpha$ would reject all hypotheses with nonnegative scores, i.e.,*

$$\{j \in \{1, \dots, m\} \,:\, S_{n+j} \ge 0\} \subset \mathit{AdaDetect}_\alpha.$$

Proof. By definition, for each $j \in \{1, \dots, m\}$, the empirical p-value p_j (2.10) satisfies

$$\begin{aligned} p_j &= \frac{1}{\ell+1}\left(1 + \sum_{i=k+1}^{n} \mathbb{1}\, S_i > S_{n+j}\right) \\ &\le \frac{1}{\ell+1}\left(1 + \sum_{i=k+1}^{n} (\mathbb{1}\, S_i \ge 0 + \mathbb{1}\, S_{n+j} < 0)\right) \\ &\le \eta + \mathbb{1}\, S_{n+j} < 0, \end{aligned}$$

by the assumptions. Hence, letting $M' = \sum_{j=1}^{m} \mathbb{1}\, S_{n+j} \geq 0 \geq M$, we have

$$\sum_{j=1}^{m} \mathbb{1}\, p_j \leq \alpha M'/m \geq \sum_{j=1}^{m} \mathbb{1}\, p_j \leq \alpha M/m \geq \sum_{j=1}^{m} \mathbb{1}\, p_j \leq \eta \geq \sum_{j=1}^{m} \mathbb{1}\, S_{n+j} \geq 0 = M'$$

Since AdaDetect_α is the BH algorithm applied to the empirical p-values, the result follows. □

Lemma 25. *Write the number of rejections $\widehat{k} = \widehat{k}(p_i, 1 \leq i \leq m)$ given by (2.4) a function of p-values. Fix any $i \in \{1, \dots, m\}$ and consider two sets of p-values $(p_j, 1 \leq j \leq m)$ and $(p'_j, 1 \leq j \leq m)$ which satisfy almost surely that*

$$\forall j \in \{1, \dots, m\}, \begin{cases} p'_j \leq p_j & \text{if } p_j \leq p_i \\ p'_j = p_j & \text{if } p_j > p_i \end{cases} \tag{A.23}$$

Let $\widehat{k} = \widehat{k}(p_i, 1 \leq i \leq m)$ and $\widehat{k}' = 1 \vee \widehat{k}(p'_i, 1 \leq i \leq m)$. Then

$$\{p_i \leq \alpha\widehat{k}/m\} = \{p_i \leq \alpha\widehat{k}'/m\} \subset \{\widehat{k} = \widehat{k}'\}.$$

This lemma is closely related to many previous results on the structure of the BH algorithm; see, e.g., Ferreira and Zwinderman (2006); Roquain and Villers (2011); Ramdas et al. (2019b). It states that the rejected p-values can be made arbitrarily smaller without changing the number of rejections.

Proof. First, since $p'_j \leq p_j$ for all $j \in \{1, \dots, m\}$, we clearly have $\widehat{k} \leq \widehat{k}'$. Now we prove the equivalence

$$p_i \leq \alpha\widehat{k}/m \iff p_i \leq \alpha\widehat{k}'/m. \tag{A.24}$$

Clearly, $p_i \leq \alpha\widehat{k}/m$ implies $p_i \leq \alpha\widehat{k}'/m$. Now we prove the other direction that $p_i \leq \alpha\widehat{k}'/m$ implies $\widehat{k} \geq \widehat{k}'$. Note that

$$\begin{aligned}
\sum_{j=1}^{m} \mathbb{1}\, p_j \leq \alpha\widehat{k}'/m &= \sum_{j=1}^{m} \mathbb{1}\, p_j \leq p_i\, \mathbb{1}\, p_j \leq \alpha\widehat{k}'/m + \sum_{j=1}^{m} \mathbb{1}\, p_j > p_i\, \mathbb{1}\, p_j \leq \alpha\widehat{k}'/m \\
&= \sum_{j=1}^{m} \mathbb{1}\, p_j \leq p_i + \sum_{j=1}^{m} \mathbb{1}\, p_j > p_i\, \mathbb{1}\, p'_j \leq \alpha\widehat{k}'/m \\
&= \sum_{j=1}^{m} \mathbb{1}\, p_j \leq p_i\, \mathbb{1}\, p'_j \leq \alpha\widehat{k}'/m + \sum_{j=1}^{m} \mathbb{1}\, p_j > p_i\, \mathbb{1}\, p'_j \leq \alpha\widehat{k}'/m \\
&= \sum_{j=1}^{m} \mathbb{1}\, p'_j \leq \alpha\widehat{k}'/m \geq 1 \vee \widehat{k}(p'_i, 1 \leq i \leq m) = \widehat{k}',
\end{aligned}$$

where the second and third lines is due to (A.23) and that $p_i \leq \alpha\widehat{k}'/m$. The fourth line uses the definition of $\widehat{k}(p'_i, 1 \leq i \leq m)$ and that $p_i \leq \alpha\widehat{k}'/m$). By definition of $\widehat{k}$, the above inequality implies $\widehat{k} \geq \widehat{k}'$. Thus, we must have $\widehat{k} = \widehat{k}'$ and the result follows. □

A.5 Auxiliary results for Section 5

A.5.1 Bounding $\zeta_r(\cdot)$

In this section, we provide explicit bounds on $\zeta_r(\cdot)$ for any given $r \in \{0, \dots, m\}$ and sample size m in the two following cases.

Lemma 26. *Assume $P_0 = U([0,1]^d)$ and $P_i = U([0,1]^{d-1}) \otimes Q$ for any $i \in \mathcal{H}_1$ where Q is a distribution supported on $[0,1]$ with a strictly decreasing and differentiable density function $h(x)$ on $(0,1)$. Assume $h(0) > 0$ and $c = \inf_{x\in(0,1)} |h'(x)| > 0$. For any $\alpha \in (0,1)$, $m \geq 1$, and $r \in \{0,\dots,m\}$, the following results hold.*

(i) If $g^*(x) = A + Bf_1(x)/f_0(x)$ for some $A \in \mathbb{R}$ and $B > 0$, then, for any $\eta < (B/2) \cdot (h(\alpha) - h(1))$,

$$\zeta_r(\eta) \leq \frac{m}{\alpha(r \vee 1)} \frac{2}{Bc} \eta. \tag{A.25}$$

(ii) If $g^*(x) = \Psi(f_1(x)/f_0(x))$ where $\Psi(y) = 1/(1 + (A + By)^{-1})$ for some $A, B > 0$, then, for any $\eta < (1/2)(\Psi \circ h(\alpha) - \Psi \circ h(1))$,

$$\zeta_r(\eta) \leq \frac{m}{\alpha(r \vee 1)} \frac{2(A + Bh(0) + 1)^2}{Bc} \eta. \tag{A.26}$$

Proof of Lemma 26. Under the assumptions, $f_1(x)/f_0(x) = h(x_d)$ where x_d is the d-th coordinate of x. For case (i), $g^*(x) = A + Bf_1(x)/f_0(x) = A + Bh(x_d)$. Since h is strictly decreasing, for any $s \in [A + Bh(1), A + Bh(0)]$,

$$\overline{G}_0(s) = \mathbb{P}_{X_d \sim U([0,1])}(A + Bh(X_d) \geq s) = \mathbb{P}(X_d \leq h^{-1}((s-A)/B)) = h^{-1}((s-A)/B).$$

Thus, $\overline{G}_0^{-1}(u) = A + Bh(u)$ for all $u \in [0,1]$. Fix any $\eta < (B/2)(h(\alpha) - h(1))$. Then for any $u \in [0, \alpha]$,

$$A+Bh(0) \geq \overline{G}_0^{-1}(u) \geq \overline{G}_0^{-1}(u) - 2\eta \geq \overline{G}^{-1}(\alpha) - 2\eta > A + Bh(\alpha) - B(h(\alpha) - h(1)) = A + Bh(1).$$

Note that both $\overline{G}_0$ and $\overline{G}_0^{-1}$ are decreasing,

$$\begin{aligned}
\zeta_r(\eta) &= \max_{u \in [\alpha(r\vee 1)/m, \alpha]} \left\{ \frac{\overline{G}_0(\overline{G}_0^{-1}(u) - 2\eta) - u}{u} \right\} \\
&\leq \frac{m}{\alpha(r \vee 1)} \max_{u \in [\alpha(r\vee 1)/m, \alpha]} (h^{-1}(h(u) - 2\eta/B) - h^{-1}(h(u))) \\
&\leq \frac{m}{\alpha(r \vee 1)} \frac{2\eta}{B} \max_{u \in [\alpha(r\vee 1)/m, \alpha]} |(h^{-1})'(u)|.
\end{aligned}$$

The result is then proved by noting that $(h^{-1})'(u) = 1/h'(h^{-1}(u))$ and the assumption that $|h'(h^{-1}(u))| > c$.

For case (ii), $g^*(x) = 1/(1 + (A + Bh(x_d))^{-1}) = \Psi \circ h(x_1)$, where $\Psi(u) = (A + Bu)/(A + Bu + 1)$ for any $u \in [h(1), h(0)]$. Note that $\Psi'(u) = B/(A + Bu + 1)^2$. Then for any s in $[1/(1 + (A + Bh(1))^{-1}), 1/(1 + (A + Bh(0))^{-1})]$, the image of Ψ,

$$\overline{G}_0(s) = \mathbb{P}_{X_d \sim U([0,1])}(\Psi(h(X_d)) \geq s) = (\Psi \circ h)^{-1}(s).$$

Hence, for all $u \in [0,1]$, $\overline{G}_0^{-1}(u) = \Psi \circ h(u)$. Fix any $\eta < (1/2)(\Psi \circ h(\alpha) - \Psi \circ h(1))$. Then for any $u \in [\alpha(r \vee 1)/m, \alpha]$,

$$\Psi \circ h(0) \geq \overline{G}_0^{-1}(u) \geq \overline{G}_0^{-1}(u) - 2\eta \geq \overline{G}_0^{-1}(\alpha) - 2\eta > \Psi \circ h(1).$$

Note that both $\overline{G}_0$ and $\overline{G}_0^{-1}$ are decreasing,

$$\begin{aligned}
\zeta_r(\eta) &= \max_{u\in[\alpha(r\vee 1)/m,\alpha]} \left\{ \frac{\overline{G}_0(\overline{G}_0^{-1}(u) - 2\eta) - u}{u} \right\} \\
&\leq \frac{m}{\alpha(r\vee 1)} \max_{u\in[\alpha(r\vee 1)/m,\alpha]} \left((\Psi\circ h)^{-1}(\Psi\circ h(u) - 2\eta) - u \right) \\
&\leq \frac{m}{\alpha(r\vee 1)} \max_{u\in[\alpha(r\vee 1)/m,\alpha]} \left((\Psi\circ h)^{-1}(\Psi\circ h(u) - 2\eta) - (\Psi\circ h)^{-1}(\Psi\circ h(u)) \right) \\
&\leq \frac{m}{\alpha(r\vee 1)}(2\eta) \max_{u\in[\alpha(r\vee 1)/m,\alpha]} |((\Psi\circ h)^{-1})'(u)| \\
&\leq \frac{m}{\alpha(r\vee 1)}(2\eta) \frac{1}{\inf_{x\in(0,1)}|h'(x)| \times \inf_{y\in[h(1),h(0)]}|\Psi'(y)|}.
\end{aligned}$$

The result is proved by noting that $\inf_{y\in[h(1),h(0)]}|\Psi'(y)| = B/(A + Bh(0) + 1)^2$. □

Lemma 27. *Assume $P_0 = \mathcal{N}(\mu_0, I_d)$ and $P_i = \mathcal{N}(\mu, I_d)$ for any $i \in \mathcal{H}_1$, where $\mu_0 \neq \mu_1$ are the null and alternative mean vectors, respectively. For any $\alpha \in (0,1)$, $m \geq 1$, and $r \in \{0,\ldots,m\}$, the following results hold.*

(i) If $g^*(x) = A + Bf_1(x)/f_0(x)$ for some $A \in \mathbb{R}$ and $B > 0$, then for any $\alpha \in (0, \overline{\Phi}(1))$ and $\eta \in [0,1]$ with $4\eta \leq (eb^2B') \wedge B'$,

$$\zeta_r(\eta) \leq C\eta, \tag{A.27}$$

where $C = C(B,\mu,\mu_0) = \frac{8}{b^2B'}$, $B' = Be^{-b^2/2}$ and $b = \|\mu - \mu_0\|$.

(ii) If $g^*(x) = 1/(A + Bf_0(x)/f_1(x))$, $A > 0$, $B > 0$, and $\alpha \in (0, \overline{\Phi}(1))$, $\eta \in [0,1]$ with $4\eta(A + B''e^{-b\overline{\Phi}^{-1}(\alpha)}) \leq 1$ and $\eta Ce^{(b+1)\sqrt{2\log(m/(\alpha(r\vee 1)))}}/(2e) \leq 1$,

$$\zeta_r(\eta) \leq C\eta e^{(b+1)\sqrt{2\log(m/(\alpha(r\vee 1)))}}, \tag{A.28}$$

for $C = C(A,B,\mu,\mu_0) = 8e(A + B'')^2/(bB'')$ with $B'' = Be^{b^2/2}$ and $b = \|\mu - \mu_0\|$.

(iii) For $g^*(x) = 1/(1 + (A + Bf_1(x)/f_0(x))^{-1})$, $A > 0$, $B > 0$, and $\alpha \in (0, \overline{\Phi}(1))$, $\eta \in [0,1]$ with $16\eta\left((A + B'e^{b\sqrt{2\log(m/(\alpha(r\vee 1)))}})^2 \vee 1\right)/B' \leq 1 \wedge b$, then

$$\zeta_r(\eta) \leq C\eta\left((A + B'e^{b\sqrt{2\log(m/(\alpha(r\vee 1)))}})^2 \vee 1\right), \tag{A.29}$$

for $C = C(B,\mu,\mu_0) = 64e/(bB')$, $B' = Be^{-b^2/2}$ and $b = \|\mu - \mu_0\|$.

Proof. Let us first consider the case of $g^*(x) = A + Bf_1(x)/f_0(x)$. We thus have

$$\begin{aligned}
g^*(x) &= A + B\exp\left\{(x-\mu_0)^T(\mu-\mu_0) - (1/2)\|\mu-\mu_0\|^2\right\} \\
&= \Psi((x-\mu_0)^T(\mu-\mu_0)/\|\mu-\mu_0\|)
\end{aligned}$$

where we let $\Psi(t) = A + B'\exp(bt)$, $t \in \mathbb{R}$, for $B' = Be^{-(1/2)\|\mu-\mu_0\|^2}$ and $b = \|\mu - \mu_0\|$. Hence $\Psi^{-1}(v) = b^{-1}\log((v-A)/B')$ for $v > A$. In that case, we have

$$\begin{aligned}
\Psi^{-1}(\Psi(t) - 2\eta) = \Psi^{-1}(A + B'e^{bt} - 2\eta) &= b^{-1}\log\left(e^{bt} - 2\eta/B'\right) \\
&= t + b^{-1}\log\left(1 - 2\eta e^{-bt}/B'\right)
\end{aligned}$$

Since $\log(1-x) \geq -2x$ for all $x \in [0, 1/2]$, we have $\log\left(1 - 2\eta e^{-bt}/B'\right) \geq -4\eta e^{-bt}/B'$ because $4\eta e^{-bt}/B' \leq 1$. This entails that for $4\eta e^{-b\overline{\Phi}^{-1}(\alpha)}/B' \leq 1$, for $u \in [\alpha(r \vee 1)/m, \alpha]$, (by taking $t = \overline{\Phi}^{-1}(u)$ in the above relations)

$$\overline{\Phi} \circ \Psi^{-1}(\Psi \circ \overline{\Phi}^{-1}(u) - 2\eta) - u \leq \overline{\Phi}\left(\overline{\Phi}^{-1}(u) - 4\eta e^{-b\overline{\Phi}^{-1}(u)}/(bB')\right) - u.$$

Now, we can use (A.30) in Lemma 28 with $y = 4\eta e^{-b\overline{\Phi}^{-1}(u)}/(bB')$ (checking that $u \leq \alpha \leq \overline{\Phi}(1)$) to obtain that

$$\begin{aligned}\overline{\Phi} \circ \Psi^{-1}(\Psi \circ \overline{\Phi}^{-1}(u) - 2\eta) &\leq 8u\eta e^{-b\overline{\Phi}^{-1}(u)}(bB')^{-1}\overline{\Phi}^{-1}(u)\exp(4\eta e^{-b\overline{\Phi}^{-1}(u)}(bB')^{-1}\overline{\Phi}^{-1}(u)) \\ &\leq \frac{8u}{eb^2B'}\eta\exp(4\eta/(eb^2B')),\end{aligned}$$

because $\forall x \geq 1$, we have $xe^{-xb} \leq 1/(eb)$. This gives (A.27).

Let us now turn to prove (A.28) by considering $g^*(x) = 1/(A + Bf_0(x)/f_1(x))$, $A > 0$, $B > 0$. Similarly to above, we have

$$g^*(x) = \Psi((x - \mu_0)^T(\mu - \mu_0)/\|\mu - \mu_0\|)$$

where we let $\Psi(t) = 1/(A + B'' \exp(-bt))$, $t \in \mathbb{R}$, for $B'' = Be^{(1/2)\|\mu-\mu_0\|^2}$ and $b = \|\mu - \mu_0\|$. Hence $\Psi^{-1}(v) = -b^{-1}\log\left((1/v - A)/B''\right)$ for $v < 1/A$. In that case, we have

$$\begin{aligned}\Psi^{-1}(\Psi(t) - 2\eta) &= \Psi^{-1}(1/(A + B''e^{-bt})) - 2\eta) \\ &= -b^{-1}\log\left(\left(\frac{A + B''e^{-bt}}{1 - 2\eta(A + B''e^{-bt})} - A\right)/B''\right).\end{aligned}$$

Now using that $1/(1-x) \leq 1 + 2x$ for all $x \in [0, 1/2]$, we have that for $4\eta(A + B''e^{-bt}) \leq 1$,

$$\left(\frac{A + B''e^{-bt}}{1 - 2\eta(A + B''e^{-bt})} - A\right)/B'' \leq e^{-bt} + 4\eta(A + B''e^{-bt})^2/B''$$

This entails

$$\begin{aligned}\Psi^{-1}(\Psi(t) - 2\eta) &\geq -b^{-1}\log\left(e^{-bt} + 4\eta(A + B''e^{-bt})^2/B''\right) \\ &= t - b^{-1}\log\left(1 + 4\eta(A + B''e^{-bt})^2e^{bt}/B''\right) \\ &\geq t - 4\eta(A + B''e^{-bt})^2e^{bt}/(bB'') \\ &\geq t - 4\eta(A + B'')^2e^{bt}/(bB''),\end{aligned}$$

because $\log(1+x) \leq x$ for all $x \geq 0$. Hence for $4\eta(A+B''e^{-b\overline{\Phi}^{-1}(\alpha)}) \leq 1$, for $u \in [\alpha(r\vee1)/m, \alpha]$,

$$\overline{\Phi} \circ \Psi^{-1}(\Psi \circ \overline{\Phi}^{-1}(u) - 2\eta) - u \leq \overline{\Phi}\left(\overline{\Phi}^{-1}(u) - 4\eta(A + B'')^2e^{b\overline{\Phi}^{-1}(u)}/(bB'')\right) - u.$$

Now, we can use (A.30) in Lemma 28 with $y = 4\eta(A + B'')^2e^{b\overline{\Phi}^{-1}(u)}/(bB'')$ to obtain

$$\begin{aligned}\overline{\Phi} \circ \Psi^{-1}(\Psi \circ \overline{\Phi}^{-1}(u) - 2\eta) &\leq 2uh\left(4(A + B'')^2\eta e^{(b+1)\overline{\Phi}^{-1}(u)}(bB'')^{-1}\right) \\ &\leq 2uh\left(4(A + B'')^2\eta e^{(b+1)\sqrt{2\log(m/(\alpha(r\vee1)))}}(bB'')^{-1}\right),\end{aligned}$$

for $h(x) = xe^x$ and because $\overline{\Phi}^{-1}(u) \leq \overline{\Phi}^{-1}(\alpha(r \vee 1)/m) \leq \sqrt{2\log(m/(\alpha(r \vee 1)))}$ (and using that $\forall x \geq 0$, we have $x \leq e^x$). This gives (A.28) because $xe^x \leq ex$ when $x \leq 1$.

Let us now turn to prove (A.29) by considering $g^*(x) = 1/(1 + (A + Bf_1(x)/f_0(x))^{-1})$, $A > 0$, $B > 0$. Similarly to above, we have

$$g^*(x) = \Psi((x - \mu_0)^T(\mu - \mu_0)/\|\mu - \mu_0\|)$$

where $\Psi(t) = 1/(1 + (A + B'e^{bt})^{-1})$, $t \in \mathbb{R}$, for $B' = Be^{-(1/2)\|\mu-\mu_0\|^2}$ and $b = \|\mu - \mu_0\|$. Hence $\Psi^{-1}(v) = b^{-1}\log\left(((1/v - 1)^{-1} - A)/B'\right)$ for $v \in (0, 1)$. In that case, we have for $4\eta\Psi(t)^{-1} \leq 1$,

$$\begin{aligned}
\frac{1}{\Psi(t) - 2\eta} - 1 &= \Psi(t)^{-1}\frac{1}{1 - 2\eta\Psi(t)^{-1}} - 1 \\
&\leq \Psi(t)^{-1}(1 + 4\eta\Psi(t)^{-1}) - 1 \\
&= (1 + (A + B'e^{bt})^{-1})(1 + 4\eta\Psi(t)^{-1}) - 1 \\
&= 4\eta\Psi(t)^{-1} + (A + B'e^{bt})^{-1}(1 + 4\eta\Psi(t)^{-1}),
\end{aligned}$$

by using $1/(1 - x) \leq 1 + 2x$, $x \in [0, 1/2]$. Hence,

$$\begin{aligned}
\left(\frac{1}{\Psi(t) - 2\eta} - 1\right)^{-1} &\geq \frac{A + B'e^{bt}}{1 + 4\eta\Psi(t)^{-1}}\frac{1}{1 + 4\eta\Psi(t)^{-1}\frac{A+B'e^{bt}}{1+4\eta\Psi(t)^{-1}}} \\
&\geq \frac{A + B'e^{bt}}{1 + 4\eta\Psi(t)^{-1}}\frac{1}{1 + 4\eta\Psi(t)^{-1}(A + B'e^{bt})} \\
&\geq (A + B'e^{bt})(1 - 4\eta\Psi(t)^{-1})\left(1 - 4\eta\Psi(t)^{-1}(A + B'e^{bt})\right) \\
&\geq (A + B'e^{bt})\left(1 - 8\eta\Psi(t)^{-1}\left((A + B'e^{bt}) \vee 1\right)\right) \\
&\geq (A + B'e^{bt})\left(1 - 16\eta\frac{(A + B'e^{bt})^2 \vee 1}{A + B'e^{bt}}\right),
\end{aligned}$$

by using $1/(1 + x) \geq 1 - x$ and $(1 - x)^2 \geq 1 - 2x$, $x \in [0, 1]$, $\Psi(t)^{-1}\left((A + B'e^{bt}) \vee 1\right) \leq 2\frac{(A+B'e^{bt})^2\vee 1}{A+B'e^{bt}}$, and provided that $16\eta\frac{(A+B'e^{bt})^2\vee 1}{A+B'e^{bt}} \leq 1$. This entails

$$\left(\left(\frac{1}{\Psi(t) - 2\eta} - 1\right)^{-1} - A\right)/B' \geq e^{bt} - 16\eta\left((A + B'e^{bt})^2 \vee 1\right)/B'.$$

Thus, we have

$$\begin{aligned}
\Psi^{-1}(\Psi(t) - 2\eta) &\geq b^{-1}\log\left(e^{bt} - 16\eta\left((A + B'e^{bt})^2 \vee 1\right)/B'\right) \\
&= t + b^{-1}\log\left(1 - 16\eta\left((A + B'e^{bt})^2 \vee 1\right)e^{-bt}/B'\right) \\
&\geq t + b^{-1}\log\left(1 - 16\eta\left((A + B'e^{bt})^2 \vee 1\right)/B'\right) \\
&\geq t - 32\eta\left((A + B'e^{bt})^2 \vee 1\right)/(bB'),
\end{aligned}$$

because $\log(1 - x) \geq -2x$ for all $x \in [0, 1/2]$ and provided that $16\eta\left((A + B'e^{bt})^2 \vee 1\right)/B' \leq 1/2$. (Also note that the latter condition implies both previous conditions $4\eta\Psi(t)^{-1} = 4\eta(1 + (A + B'e^{bt})^{-1}) \leq 1$ and $16\eta\frac{(A+B'e^{bt})^2\vee 1}{A+B'e^{bt}} \leq 1$).

Hence for $16\eta\left((A+B'e^{b\overline{\Phi}^{-1}(\alpha/m)})^2\vee 1\right)/B'\le 1/2$ and $u\in[\alpha(r\vee 1)/m,\alpha]$,

$$\overline{\Phi}\circ\Psi^{-1}(\Psi\circ\overline{\Phi}^{-1}(u)-2\eta)-u\le\overline{\Phi}\left(\overline{\Phi}^{-1}(u)-32\eta\left((A+B'e^{b\overline{\Phi}^{-1}(u)})^2\vee 1\right)/(bB')\right)-u.$$

Now, we can use (A.30) in Lemma 28 with $y=32\eta\left((A+B'e^{b\overline{\Phi}^{-1}(u)})^2\vee 1\right)/(bB')$ to obtain

$$\begin{aligned}\overline{\Phi}\circ\Psi^{-1}(\Psi\circ\overline{\Phi}^{-1}(u)-2\eta)&\le 2uh\left(32\eta\left((A+B'e^{b\overline{\Phi}^{-1}(u)})^2\vee 1\right)/(bB')\right)\\&\le 2uh\left(32\eta\left((A+B'e^{b\sqrt{2\log(m/(\alpha(r\vee 1)))}})^2\vee 1\right)/(bB')\right),\end{aligned}$$

for $h(x)=xe^x$ and because $\overline{\Phi}^{-1}(u)\le\overline{\Phi}^{-1}(\alpha(r\vee 1)/m)\le\sqrt{2\log(m/(\alpha(r\vee 1)))}$ (and using that $\forall x\ge 0$, we have $x\le e^x$). This gives (A.29). □

Lemma 28. *For all $y\ge 0$ and $u\in(0,\overline{\Phi}(1)]$, we have*

$$\overline{\Phi}(\overline{\Phi}^{-1}(u)-y)-u\le 2uy\overline{\Phi}^{-1}(u)\exp(y\overline{\Phi}^{-1}(u)). \tag{A.30}$$

Proof. By using the classical relations on upper-tail distribution of standard Gaussian, we have

$$\begin{aligned}\overline{\Phi}(\overline{\Phi}^{-1}(u)-y)-u&\le y\left(\phi(\overline{\Phi}^{-1}(u)-y)\vee\phi(\overline{\Phi}^{-1}(u))\right)\\&=y\phi(\overline{\Phi}^{-1}(u))\left(1\vee\exp(y\overline{\Phi}^{-1}(u)-y^2/2)\right)\\&\le 2uy\overline{\Phi}^{-1}(u)\exp(y\overline{\Phi}^{-1}(u)),\end{aligned}$$

since $\overline{\Phi}^{-1}(u)\ge 1$, because $\phi(x)\le 2x\overline{\Phi}(x)$ for all $x\ge 1$. □

A.5.2 Case of density estimation

Consistency

Consider g^* given by (2.21), that is $g^*(x)=f_\gamma(x)/f_0(x)$. Assuming for simplicity that f_0 is known (hence $k=0$ and $n=\ell$ here), we propose the estimator $g(x)=\hat{f}_\gamma(x)/f_0(x)$, where $\hat{f}_\gamma$ is the histogram estimator of f_γ given by

$$\hat{f}_\gamma(x)=M^d\sum_{j=1}^{M^d}(n+m)^{-1}\sum_{i=1}^{n+m}\mathbb{1}\,Z_i\in\mathcal{D}_j\,\mathbb{1}\,x\in\mathcal{D}_j,\ \ x\in[0,1]^d, \tag{A.31}$$

where $\{\mathcal{D}_1,\dots,\mathcal{D}_{M^d}\}$ is a regular partition of $[0,1]^d$ formed by M^d d-dimensional cubes of side size $1/M$ and Lebesgue measure $|\mathcal{D}_1|=1/M^d$, with $M=\lceil(n+m)^{1/(2+d)}\rceil$.

Remember that the corresponding AdaDetect procedure controls the FDR even if the estimation quality of $\hat{f}_\gamma$ is poor. In addition, we show in this section that, in a suitable case where the estimation quality is good enough, the power of AdaDetect consistently converges to that of the oracle, that is, (2.40) holds. We use for this Corollary 12.

Assume $m\asymp\ell=n\gg m/m_1$, let Assumptions 4 and 5 be true and consider the uniformly bounded case described in Section A.5.1. Then we have both (2.41) with $\kappa\in(0,1/(2+d))$ (Lemma 29, see next section) and $\zeta_{\lceil m_1\epsilon\rceil}(\eta)\lesssim\epsilon^{-1}\frac{m}{\alpha m_1}\eta/\gamma\asymp\epsilon^{-1}\eta/(m_1/m)^2$ for η small enough (observe that $\gamma\sim m_1/m$). Hence, in the not too sparse scenario

$$m_1/m\gg m^{-\kappa/2}$$

(including the dense case $m_1\asymp m$), choosing $\eta\asymp m^{-\kappa}$, $\delta=\epsilon=1/\log((m_1/m)^2/m^{-\kappa})$, we have $\ell\delta^2\epsilon m_1/m=m^{1-\kappa/2}\frac{(m_1/m)/m^{-\kappa/2}}{\log^3((m_1/m)^2/m^{-\kappa})}\gg 1$, which ensures (2.39) and thus proves the consistency (2.40).

Bounding $\hat{\eta}$ in case of density estimation

Lemma 29. *Let Assumptions 4 and 5 be true and consider the case where $P_0 = U([0,1]^d)$ and P_i, $i \in \mathcal{H}_1$ have a common distribution, supported on $[0,1]^d$, with a density f_1 which is L-Lipschitz and uniformly upper-bounded by C^d. Then the estimator $g(x) = \hat{f}_\gamma(x)/f_0(x)$ of g^* given above satisfies that for $n+m \geq N(C,d)$,*

$$\mathbb{P}\left(\hat{\eta} \geq c_0(n+m)^{-1/(2+d)}\sqrt{\log(n+m)}\right) \leq 2/(n+m), \tag{A.32}$$

where $\hat{\eta}$ is given by (2.36) and $c_0(d,C,L) = 2L + 8(2C)^{d/2}$.

Proof. First note that when f_1 is L-Lipschitz and upper bounded by C^d, then f_γ is also L-Lipschitz and upper bounded by C^d (because $\gamma \leq 1$). Hence, by Lemma 30 we have for some constant $c_0 = c(d,C,L) > 0$ and $N = N(C)$, for all $m \geq N$, for all $x \in [0,1]^d$,

$$\mathbb{P}\left(|\hat{f}_\gamma(x) - f_\gamma(x)| \geq c_0(n+m)^{-1/(2+d)}\sqrt{\log(n+m)}\right) \leq 2/(n+m)^2. \tag{A.33}$$

Now, we have for all $\eta > 0$,

$$\mathbb{P}(\hat{\eta} \geq \eta) = \mathbb{P}\left(\max_{1\leq i\leq n+m} |g^*(Z_i) - g(Z_i)| \geq \eta\right) \leq \sum_{i=1}^{n+m} \mathbb{P}\left(|\hat{f}_\gamma(Z_i) - f_\gamma(Z_i)| \geq \eta\right),$$

because f_0 is the density of the uniform distribution on $[0,1]^d$. A technical point here is that Z_i also appears in $\hat{f}_\gamma$ (at one place), nevertheless, denoting $\hat{f}'_\gamma$ the quantities (A.34) for which Z_i has been changed to an independent copy Z'_i (at that place), we have for all $x \in [0,1]^d$,

$$|\hat{f}_\gamma(x) - \hat{f}'_\gamma(x)| \leq M^d(n+m)^{-1} \leq 2^d(n+m)^{-2/(2+d)}.$$

Combined with (A.33), this gives

$$\mathbb{P}(\hat{\eta} \geq \eta) \leq \sum_{i=1}^{n+m} \mathbb{P}\left(|\hat{f}'_\gamma(Z_i) - f_\gamma(Z_i)| \geq \eta - 2^d(n+m)^{-2/(2+d)}\right) \leq 2/(n+m),$$

by choosing η such that $\eta \geq 2^d(n+m)^{-2/(2+d)} + c_0(n+m)^{-1/(2+d)}\sqrt{\log(n+m)}$. We have established Lemma 29.

□

Lemma 30. *[Histogram density estimator, non i.i.d. version] Consider $Z_1, \ldots, Z_n$ independent random variables take values in $[0,1]^d$ where Z_i has for density f_i, $1 \leq i \leq n$. We assume that all the f_i's are L-Lipschitz and pointwise upper bounded by C^d (for some constant value $C \in (0,1)$), where L, C does not depend on i. Let $f(x) = n^{-1}\sum_{i=1}^n f_i(x)$ for $x \in [0,1]^d$. We consider the histogram estimator of f given by*

$$\hat{f}_n(x) = M^d \sum_{j=1}^{M^d} n^{-1} \sum_{i=1}^{n} \mathbb{1}\, Z_i \in \mathcal{D}_j\, \mathbb{1}\, x \in \mathcal{D}_j, \;\; x \in [0,1]^d, \tag{A.34}$$

where $\{\mathcal{D}_1, \ldots, \mathcal{D}_{M^d}\}$ is a regular partition of $[0,1]^d$ formed by M^d d-dimensional cubes of side size $1/M$ and Lebesgue measure $|\mathcal{D}_1| = 1/M^d$. Then, choosing $M = \lceil n^{1/(2+d)} \rceil$, for $n \geq N(C)$, we have for all $x \in [0,1]^d$,

$$\mathbb{P}\left(|\hat{f}_n(x) - f(x)| \geq c_0(d,C,L)n^{-1/(2+d)}\sqrt{\log n}\right) \leq 2/n^2, \tag{A.35}$$

where $c_0(d,C,L) = L + 4(2C)^{d/2}$.

Proof. We have for all $x \in [0,1]^d$,

$$|\hat{f}_n(x) - f(x)| \le |\hat{f}_n(x) - \mathbb{E}\,\hat{f}_n(x)| + |\,\mathbb{E}\,\hat{f}_n(x) - f(x)|,$$

which contains a bias and a variance term. For the bias, we have

$$|\,\mathbb{E}\,\hat{f}_n(x) - f(x)| \le \sum_{j=1}^{M^d} \mathbb{1}\, x \in \mathcal{D}_j |\mathcal{D}_j|^{-1} \int_{\mathcal{D}_j} n^{-1} \sum_{i=1}^{n} |f_i(y) - f_i(x)| dy \le L/M,$$

because $\sup_{1\le k\le d} |x_k - y_k| \le 1/M$ when x, y belongs to the same $\mathcal{D}_j$.

For the variance term, by denoting j_x the only j such that $x \in \mathcal{D}_j$ and $p_{i,x} = \mathbb{P}(Z_i \in \mathcal{D}_{j_x}) \in [0, (C/M)^d]$, we have

$$\begin{aligned}\mathbb{P}(|\hat{f}_n(x) - \mathbb{E}\,\hat{f}_n(x)| \ge \delta) &\le \mathbb{P}\left(\left|\sum_{i=1}^{n} (\mathbb{1}\, Z_i \in \mathcal{D}_{j_x} - p_{i,x})\right| \ge \delta n M^{-d}\right) \\ &\le 2\exp\left(-\frac{1}{2}\frac{A^2}{\sum_{i=1}^{n} p_{i,x} + A/3}\right) \le 2\exp\left(-\frac{1}{2}\frac{A^2}{n(C/M)^d + A/3}\right),\end{aligned}$$

by letting $A = \delta n M^{-d}$ and applying Bernstein's inequality. Choosing δ such that $A \le n(C/M)^d$, that is, $\delta \le C^d$, we obtain

$$\mathbb{P}(|\hat{f}_n(x) - \mathbb{E}\,\hat{f}_n(x)| \ge \delta) \le 2\exp\left(-(3/8)\frac{A^2}{n(C/M)^d}\right) = 2\exp\left(-(3/8)n\delta^2 M^{-d} C^{-d}\right),$$

by choosing δ such that $(3/8)n\delta^2 M^{-d} C^{-d} = 2\log n$, that is, $\delta = (4/\sqrt{3})(MC)^{d/2}\sqrt{(\log n)/n}$ gives

$$\mathbb{P}(|\hat{f}_n(x) - \mathbb{E}\,\hat{f}_n(x)| \ge \delta) \le 2/n^2,$$

provided that $\delta \le C^d$. Now, we choose $M = \lceil n^{1/(2+d)} \rceil$, so that for $n \ge N(C)$, $\delta = 4(2C)^{d/2} n^{-1/(2+d)}\sqrt{\log n}$ is a valid choice. This gives the bound (A.35). □

A.5.3 From the two-sample setting to classical two-group setting

Existing results of the form (2.41) typically assume that the observations are i.i.d. draws from a two-group mixture model under which the class labels are random. By contrast, we focus on a two-sample setting with fixed labels in which case the observations are *non-identically distributed* because the assumptions are weaker than the two-group setting. Nevertheless, we can easily adapt our theory in the two-group setting, in which case a plethora of existing results can be applied to understand the scale of $\zeta_r(\eta)$ and $\hat{\eta}$. The two-group setting can be formalized by the following assumptions.

Assumption 11. *The sample $(Z_1, \ldots, Z_{n+m})$ is obtained in the following way:*

- $(Z_1, \ldots, Z_k) = (W_i, 1 \le i \le n+m \ : \ A_i = 0, B_i = 0, C_i = 1)$;
- $(Z_{k+1}, \ldots, Z_n) = (W_i, 1 \le i \le n+m \ : \ A_i = 0, B_i = 0, C_i = 0)$;
- $(X_i, i \in \mathcal{H}_0) = (W_i, 1 \le i \le n+m \ : \ A_i = 0, B_i = 1, C_i = 0)$;
- $(X_i, i \in \mathcal{H}_1) = (W_i, 1 \le i \le n+m \ : \ A_i = 1, B_i = 1, C_i = 0)$,

where (A_i, B_i, C_i, W_i), $1 \leq i \leq n+m$, *are i.i.d. with* $A_i \sim \mathcal{B}(\pi_1)$ *(indicator of being a novelty),* $B_i \sim \mathcal{B}(\pi_B)$ *(indicator of being in the test sample),* $C_i \sim \mathcal{B}(\pi_C)$ *(indicator of being in the first NTS), for some proportions* $\pi_A, \pi_B, \pi_C \in (0,1)$ *and* $W_i \mid A_i \sim f_0$ *if* $A_i = 0$ *and* $W_i \mid A_i = 1 \sim f_1$, *where* f_0 *is the density of* P_0 *and* f_1 *is the common density of all* P_i, $i \in \mathcal{H}_1$.

Under Assumption 11, the sample size $n+m$ (number of trials) is fixed while the sample sizes k, ℓ, m and m_1 are random. Also, we easily see that, conditionally on A, B, C, the sample $(Z_1, \ldots, Z_{n+m})$ satisfies Assumptions 4 and 5 with $f_i = f_1$ for $i \in \mathcal{H}_1$.

As a result, under Assumption 11, and letting $L_i = 1 - (1 - A_i)(1 - B_i)C_i$ we have $(Z_1, \ldots, Z_k) = (W_i, 1 \leq i \leq n+m \,:\, L_i = 0)$, and $(Z_{k+1}, \ldots, Z_{n+m}) = (W_i, 1 \leq i \leq n+m \,:\, L_i = 1)$ with $(W_i, L_i)_{1 \leq i \leq n+m}$ i.i.d., $L_i \sim \mathcal{B}(1 - (1 - \pi_A)(1 - \pi_B)\pi_C)$ and $W_i \mid L_i = 0 \sim f_0$ and $W_i \mid L_i = 1 \sim (1 - \pi) f_0 + \pi f_1$, for some $\pi \in (0,1)$. Also, the knowledge of the samples $(Z_1, \ldots, Z_k)$ and $(Z_{k+1}, \ldots, Z_{n+m})$ is equivalent to the knowledge of $(W_i, L_i)_{1 \leq i \leq n+m}$. This means that, under Assumption 11, the score function (2.7) is based on $(W_i, L_i)_{1 \leq i \leq n+m}$, which is a standard classification setting where the covariates and labels are jointly i.i.d..

A.6 Details of simulation studies in Section 6

A.6.1 Methods

First, we describe the score functions used in each version of AdaDetect.

- `AdaDetect oracle`: the oracle score function r defined in (2.19).
- `AdaDetect parametric` and `AdaDetect KDE`: $g(x) = \widehat{f_\gamma}(x)/\widehat{f_0}(x)$, where $\widehat{f_\gamma}$ is a density estimator of f_γ (2.18) computed on mixed sample $(Z_{k+1}, \ldots, Z_{n+m})$ and $\widehat{f_0}$ is a density estimator of f_0 based on $(Z_1, \ldots, Z_k)$. For `AdaDetect parametric`, $\hat{f}_0$ is estimated by the Ledoit-Wolf method and $\hat{f}_\gamma$ is estimated by a two-component mixture of Gaussians via an expectation-maximization (EM) algorithm with 100 random restarts. For `AdaDetect KDE`, they are given by non-parametric Gaussian kernel density estimators (KDE).
- `AdaDetect SVM`: $\widehat{g}$ is obtained by minimizing the empirical risk (2.24) with the hinge loss, $\lambda = 1$ and a suitable regularization with the cost parameter C set to 1; see Hastie et al. (2009).
- `AdaDetect RF`: $\widehat{g}$ is obtained by random forest with the maximum depth 10;
- `AdaDetect NN`: $\widehat{g}$ is obtained by minimizing the cross entropy loss (2.24) with $\lambda = 1$ and the NN function class with 1 hidden layer, 100 neurons, and the ReLU activation function.
- `AdaDetect NN cv`: AdaDetect NN with the number of hidden layers and the number of neurons per layer chosen by the cross-validation procedure described in Section 2.4.5.

They are compared to several existing methods:

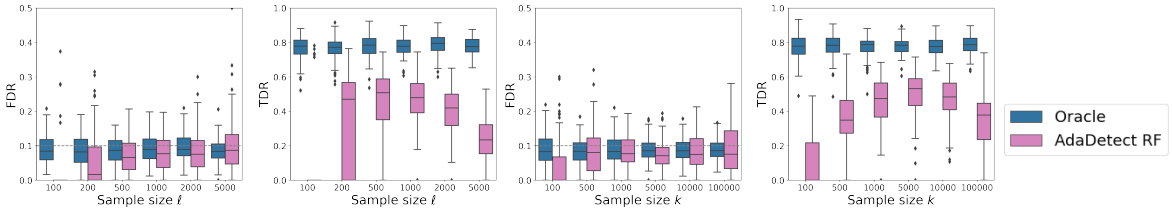
- `SC parametric` and `SC KDE`: the procedure of Sun and Cai (2007) based on local FDR estimates $\ell_i = \hat{\pi}_0 \hat{f}_0(X_i)/\hat{f}(X_i)$, where the densities are estimated by the same methods for `AdaDetect parametric` and `SC KDE`, respectively, except that $\hat{f}_0$ is based on the whole NTS $Y = (Y_1, \ldots, Y_n)$ and $\hat{f}$ is only based on the test sample $X = (X_1, \ldots, X_m)$. We set $\hat{\pi}_0 = 1$ for a fair comparison with the non-adaptive versions of AdaDetect;
- `CAD SVM` and `CAD IForest`: the conformal anomaly detection procedure proposed by Bates et al. (2023) based on one-class SVM and Isolation Forest, respectively.

A.6.2 Additional experiments with varying k and ℓ

To examine the effect of k and ℓ, we study the performance of `AdaDetect RF` with `AdaDetect oracle` in the simple setting where $P_0 = \mathcal{N}(0, I_d)$ and $P_1 = \mathcal{N}(\mu, I_d)$ with $d = 4$ and $\mu = (\sqrt{2}, \ldots, \sqrt{2})$.

Figure A.1a presents the results for $k = m = 1000$ and varying ℓ. Interestingly, the TDR is not monotone in ℓ. This is because a small ℓ makes p-value inaccurate while a large ℓ dilutes the signal in the mixed sample $(X_{k+1}, \ldots, X_{m+n})$ and degrades the quality of classification-based score function.

For $\ell = m = 1000$ and varying values of k, a similar pattern is observed in Figure A.1b. This suggests that an imbalanced classification problem may arise due to the presence of extreme values of k.



(a) (b)

Figure A.1: FDR and TDR as a function of ℓ (panel (a)) and k (panel (b)). The dashed line indicates the nominal level.

A.6.3 Additional experiments with varying n, m, and m_1

In this section, we report results for additional experiments in more challenging settings for AdaDetect.

- Small sample sizes: $m = 200$, with $k = 4m$ and $\ell = m$ as before. The results are reported in Table A.1.
- Small null sample sizes with $n < m$: $n = m/2$, with $k = m/4$ and $\ell = m/4$. The results are reported in Table A.2.
- Sparse novelties: we set $m_1/m = 2\%$, with $m = 1000$, $k = 4m$ and $\ell = m$ as before. The results are reported in Table A.3.

As in Section 2.6.2, we set the target FDR level $\alpha = 0.1$ and report the mean value and the standard deviation (in brackets) over 100 runs. We only highlight in bold the best-performing method if its FDR is below α.

Table A.1: Same as Table 2.3 for $m = 200$ ($k = 4m$, $\ell = m$).

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
			FDR			
CAD SVM	0.04 (0.09)	0.00 (0.00)	0.00 (0.00)	0.02 (0.09)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.07 (0.10)	0.05 (0.08)	0.07 (0.11)	0.01 (0.06)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.03 (0.09)	0.00 (0.00)	0.00 (0.00)	0.01 (0.08)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.05 (0.10)	0.01 (0.05)	0.00 (0.00)	0.02 (0.06)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.04 (0.10)	0.02 (0.07)	0.01 (0.04)	0.01 (0.06)	0.00 (0.00)	0.01 (0.03)
AdaDetect RF	0.07 (0.09)	0.07 (0.09)	0.08 (0.09)	0.05 (0.11)	0.00 (0.00)	0.02 (0.12)
AdaDetect NN	0.05 (0.09)	0.06 (0.09)	0.06 (0.14)	0.03 (0.09)	0.01 (0.10)	0.02 (0.10)
AdaDetect cv NN	0.05 (0.08)	0.06 (0.08)	0.05 (0.11)	0.04 (0.10)	0.01 (0.10)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.01 (0.07)
AdaDetect CNN	-	-	-	-	-	0.01 (0.10)
			TDR			
CAD SVM	0.12 (0.22)	0.00 (0.00)	0.00 (0.00)	0.03 (0.10)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.28 (0.28)	0.25 (0.32)	0.42 (0.47)	0.02 (0.11)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.13 (0.25)	0.00 (0.00)	0.00 (0.00)	0.01 (0.07)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.22 (0.33)	0.01 (0.07)	0.00 (0.00)	0.06 (0.17)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.16 (0.26)	0.03 (0.11)	0.06 (0.21)	0.01 (0.05)	0.00 (0.00)	0.02 (0.11)
AdaDetect RF	**0.57 (0.30)**	**0.71 (0.21)**	**0.97 (0.04)**	**0.11 (0.21)**	0.00 (0.00)	0.00 (0.01)
AdaDetect NN	0.31 (0.35)	0.46 (0.34)	0.28 (0.40)	0.07 (0.19)	0.00 (0.00)	0.01(0.05)
AdaDetect cv NN	0.28 (0.36)	0.47 (0.33)	0.31 (0.39)	0.10 (0.22)	0.00 (0.00)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.00 (0.02)
AdaDetect CNN	-	-	-	-	-	0.02 (0.10)

Table A.2: Same as Table 2.3 for $n = m/2$ ($k = m/4$, $\ell = m/4$).

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
			FDR			
CAD SVM	0.03 (0.06)	0.00 (0.00)	0.00 (0.00)	0.02 (0.08)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.05 (0.07)	0.04 (0.07)	0.04 (0.07)	0.01 (0.06)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.03 (0.05)	0.00 (0.00)	0.00 (0.00)	0.23 (0.39)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.05 (0.08)	0.01 (0.07)	0.00 (0.00)	0.04 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.06 (0.07)	0.03 (0.06)	0.00 (0.04)	0.00 (0.04)	0.00 (0.00)	0.00 (0.00)
AdaDetect RF	0.07 (0.06)	0.07 (0.06)	0.08 (0.06)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect NN	0.05 (0.06)	0.06 (0.07)	0.01 (0.05)	0.05 (0.08)	0.01 (0.10)	0.00 (0.00)
AdaDetect cv NN	0.05 (0.07)	0.04 (0.08)	0.02 (0.06)	0.04 (0.08)	0.02 (0.14)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.03 (0.03)
AdaDetect CNN	-	-	-	-	-	0.03 (0.09)
			TDR			
CAD SVM	0.07 (0.16)	0.00 (0.00)	0.00 (0.00)	0.02 (0.08)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.22 (0.24)	0.20 (0.29)	0.34 (0.46)	0.01 (0.07)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.12 (0.23)	0.00 (0.00)	0.00 (0.00)	0.05 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.19 (0.30)	0.01 (0.06)	0.00 (0.00)	0.07 (0.16)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.47 (0.29)	0.20 (0.32)	0.00 (0.02)	0.00 (0.03)	0.00 (0.00)	0.00 (0.00)
AdaDetect RF	**0.66 (0.20)**	**0.72 (0.16)**	**0.98 (0.02)**	0.11 (0.18)	0.00 (0.00)	0.00 (0.00)
AdaDetect NN	0.30 (0.36)	0.39 (0.35)	0.07 (0.24)	**0.16 (0.22)**	0.00 (0.00)	0.00 (0.00)
AdaDetect cv NN	0.30 (0.36)	0.21 (0.31)	0.09 (0.27)	0.13 (0.21)	0.00 (0.00)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.03 (0.03)
AdaDetect CNN	-	-	-	-	-	0.06 (0.17)

Table A.3: Same as Table 2.3 for $m_1/m = 2\%$ ($k = 4m$, $\ell = m$).

	Shuttle	Credit card	KDDCup99	Mammography	Musk	MNIST
			FDR			
CAD SVM	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.05 (0.11)	0.03 (0.11)	0.03 (0.08)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.03 (0.09)	0.03 (0.08)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect RF	0.09 (0.10)	0.07 (0.09)	0.08 (0.08)	0.03 (0.11)	0.00 (0.00)	0.00 (0.00)
AdaDetect NN	0.06 (0.12)	0.06 (0.10)	0.03 (0.07)	0.03 (0.09)	0.00 (0.00)	0.00 (0.00)
AdaDetect cv NN	0.04 (0.10)	0.05 (0.09)	0.02 (0.05)	0.02 (0.09)	0.00 (0.00)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.01 (0.10)
AdaDetect CNN	-	-	-	-	-	0.02 (0.14)
			TDR			
CAD SVM	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
CAD IForest	0.10 (0.20)	0.04 (0.13)	0.13 (0.31)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect parametric	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect KDE	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect SVM	0.10 (0.27)	0.11 (0.26)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
AdaDetect RF	**0.60 (0.29)**	**0.64 (0.27)**	**0.98 (0.03)**	0.05 (0.14)	0.00 (0.00)	0.00 (0.00)
AdaDetect NN	0.12 (0.24)	0.38 (0.34)	0.22 (0.38)	0.04 (0.13)	0.00 (0.00)	0.00 (0.00)
AdaDetect cv NN	0.08 (0.21)	0.37 (0.35)	0.18 (0.36)	0.03 (0.11)	0.00 (0.00)	0.00 (0.00)
CAD SVDD + CNN	-	-	-	-	-	0.01 (0.09)
AdaDetect CNN	-	-	-	-	-	0.01 (0.07)

A.7 Additional experimental results for the astronomy application

In this section, we provide more results for more settings. Recall that, for each experiment, we sample n nonvariable stars from as the NTS along with m_1 variable stars and $m_0 = m - m_1$ additional nonvariable stars as the test sample. For all experiments, we set $m = 100$.

First, we show how TDR varies with sparsity measured by m_1. Figure A.2 presents the TDR for $m_1 \in \{5, 15, 40, 90\}$ with different target FDR levels shown in the title of each panel. When α is low, the left two panels show that AdaDetect RF substantially outperforms the other methods. When the novelties are sparse, the panels in the middle column shows that AdaDetect RF still performs well when $\alpha = 0.2$ but underperforms when $\alpha = 0.5$, though $\alpha = 0.5$ is arguably less relevant in practice. When the novelties are dense, the right panels show that AdaDetect with adaptive scores outperform AdaDetect with non-adaptive scores when $\alpha = 0.05$ and they are nearly indistinguishable when $\alpha = 0.5$.

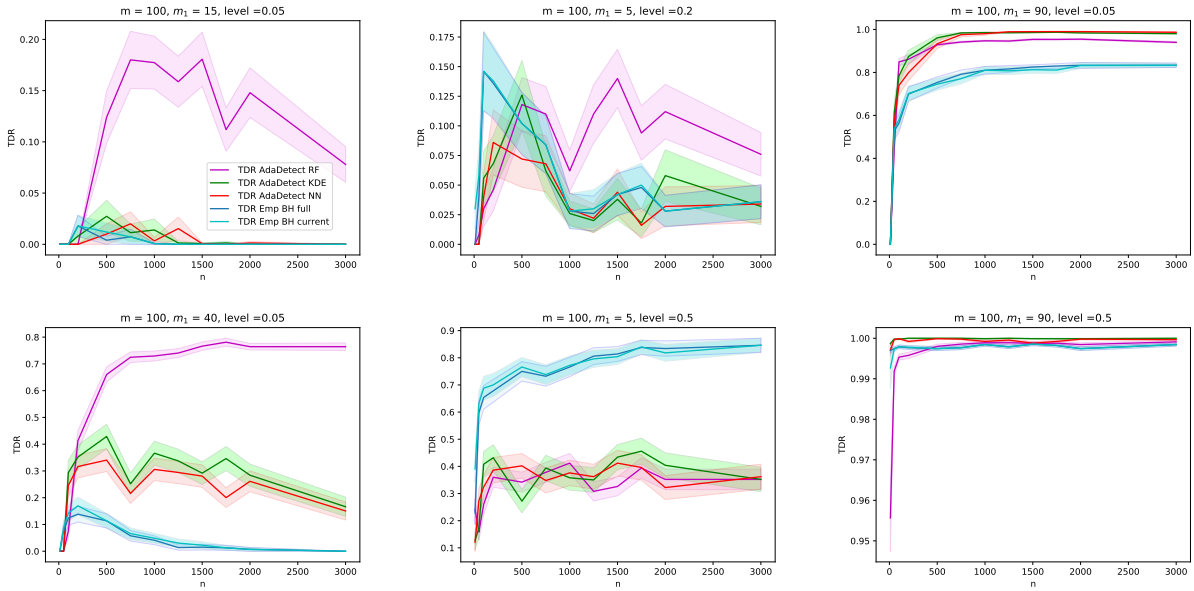


Figure A.2: TDR for different sparsity regimes: see m_1 and α in the titles. In all plots $m = 100$.

Next, we compare the FDR and TDR of AdaDetect KDE and Empirical BH with both m_1 and n varying in Figure A.3. For the purpose of visualization, we only show the point estimates of FDR and TDR without uncertainty measures. The left column shows that both methods control the FDR, though AdaDetect KDE is generally less conservative. The right column shows that AdaDetect KDE almost always has a higher power and it starts to reject at a higher sparsity level. An interesting observation is that the power of AdaDetect KDE is decreasing in n when the novelties are sparse (e.g., $m_1 \approx 40$). This can be explained by the fact that the 'contamination' of the NTS increases with increasing n, leading to a more noisy estimation of the test statistic.

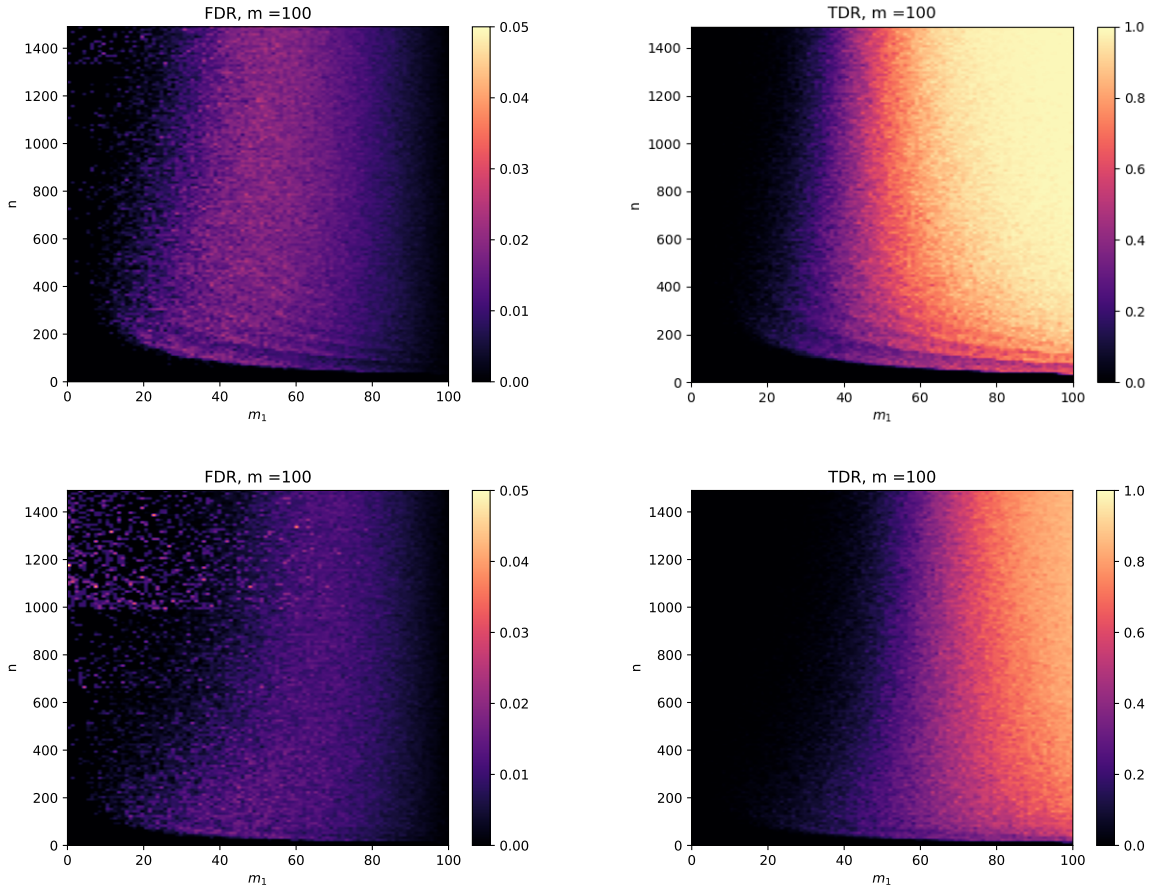


Figure A.3: FDR (left column) and TDR (right column) for AdaDetect KDE (top row) and Empirical BH (bottom row). In all plots $m = 100$ and $\alpha = 0.05$.

Appendix B

Supplementary material of Chapter 3

B.1 Proof of Theorems 16 and 17

B.1.1 A general result

In this section, we establish a general result, from which Theorems 16 and 17 can be deduced. It provides non-asymptotic bounds on the mFMR and the FMR of the plug-in procedure and on its average selection number, by relying only on Assumption 6. To state the result, we introduce some additional quantities measuring the regularity of the model which will appear in our remainder terms. Recall definitions (3.4), (3.5) and (3.13) of $\ell_q(X,\theta)$, $T(X,\theta)$ and $t^*(\alpha)$ respectively, and let for $\epsilon, \delta, v > 0$,

$$\mathcal{W}_\ell(\epsilon) = \sup\left\{ \sup_{x\in\mathbb{R}^d} \left[\max_{1\le q\le Q} |\ell_q(x,\theta^*) - \ell_q(x,\theta)| \right] ,\ \|\theta-\theta^*\|_2 \le \epsilon,\ \theta\in\Theta \right\}; \tag{B.1}$$

$$\mathcal{W}_T(\delta) = \sup\{ |\, \mathbb{P}_{\theta^*}(T(X,\theta^*) < t') - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \,, \tag{B.2}$$

$$t, t' \in [0,1], |t'-t| \le \delta\}; \tag{B.3}$$

$$\Psi(\epsilon) = \mathcal{W}_T(\mathcal{W}_\ell(\epsilon)^{1/2}) + \mathcal{W}_\ell(\epsilon)^{1/2}; \tag{B.4}$$

$$\mathcal{W}_{t^*,\alpha}(v) = \sup\left\{ |t^*(\alpha+\beta) - t^*(\alpha)| \,,\ |\beta| \le v \right\}. \tag{B.5}$$

Theorem 31. *Let Assumption 6 be true. For any $\alpha \in (\alpha_c, \bar{\alpha})$ and constants $s^* = s^*(\alpha,\theta^*) \in (0,1)$ and $e^* = e(\alpha,\theta^*) > 0$ depending only on α and θ^*, the following holds. Consider the plug-in procedure $\widehat{\mathcal{C}}_\alpha^{PI} = (\widehat{\mathbf{Z}}^{PI}, \widehat{S}_\alpha^{PI})$ introduced in Algorithm 4 and based on an estimator $\hat{\theta}$ satisfying Assumption 10, with $\eta(\epsilon,\theta^*)$ defined by (3.18). Then for $\epsilon \le e^*$ and $n \ge (2e)^3$, letting*

$$\Delta_n(\epsilon) = 2\left(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta_n + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta_n\right),$$

for $\delta_n = C\sqrt{(\log n)/n}/s^*$ where $C = 2 + 56Q\sqrt{\mathscr{V}} + 28Q^2\sqrt{\mathscr{V}_-}$ and with the quantities $\mathcal{W}_T$, $\mathcal{W}_\ell$, Ψ, $\mathcal{W}_{t^*,\alpha}$ defined by (B.3), (B.1), (B.4), (B.5), respectively, it holds:

- *The procedure $\widehat{\mathcal{C}}_\alpha^{PI}$ controls both the FMR and the mFMR at level close to α in the following sense:*

$$\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{PI}) \le \alpha + \Delta_n(\epsilon)/s^* + 5/n^2 + \eta(\epsilon,\theta^*);$$
$$\mathrm{mFMR}(\widehat{\mathcal{C}}_\alpha^{PI}) \le \alpha + \Delta_n(\epsilon)/s^* + s^{*-1}\left[50/n^2 + 10\eta(\epsilon,\theta^*)\right].$$

- *The procedure $\widehat{\mathcal{C}}_\alpha^{PI}$ is nearly optimal in the following sense: for any other procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ that controls the mFMR at level α,*

$$n^{-1}\,\mathbb{E}_{\theta^*}(|\widehat{S}_\alpha^{PI}|) \geq n^{-1}\,\mathbb{E}_{\theta^*}(|S|) - \Delta_n(\epsilon).$$

Before proving this result (which will be done in the next subsections), let us first show that Theorem 31 implies Theorems 16 and 17.

Proof of Theorem 16 By Lemma 32 below, $\Delta_n(\epsilon)$ tends to 0 when n tends to infinity and ϵ tends to 0. Moreover, by consistency of $\hat{\theta}$, $\eta(\epsilon, \theta^*)$ tends to 0 for all $\epsilon > 0$. This implies the result.

Lemma 32. *Under Assumption 6, we have $\lim_{\delta\to 0} \mathcal{W}_T(\delta) = 0$, $\lim_{v\to 0} \mathcal{W}_{t^*,\alpha}(v) = 0$. Under Assumption 7, we have $\lim_{\epsilon\to 0} \mathcal{W}_\ell(\epsilon) = 0$. Under both assumptions, we have $\lim_{\epsilon\to 0} \Psi(\epsilon) = 0$.*

Proof. The only non-trivial fact is for $\mathcal{W}_{t^*,\alpha}(v)$. Assumption 6 and Lemma 36 provide that $t \mapsto \mathrm{mFMR}_t^*$ is a one-to-one continuous increasing map from $(t^*(\alpha_c), t^*(\bar{\alpha}))$ to $(\alpha_c, \bar{\alpha})$. Hence, for $\alpha \in (\alpha_c, \bar{\alpha})$, $\beta \mapsto t^*(\alpha + \beta)$ is continuous in 0 and $\lim_{v\to 0} \mathcal{W}_{t^*,\alpha}(v) = 0$. □

Proof of Theorem 17 By using Assumption 8 (with the notation therein) and Lemma 33 below, we have

$$\begin{aligned}
\Delta_n(\epsilon) &= 2\left(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta_n + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta_n\right) \\
&\leq 2C_2C_3\left(2\delta_n + (8/s^*)\sqrt{C_1}(C_2+1)\sqrt{\epsilon}\right) + 8\sqrt{C_1}(C_2+1)\sqrt{\epsilon} + 4\delta_n \\
&= 8\sqrt{C_1}(C_2+1)(1+2C_2C_3)\sqrt{\epsilon}/s^* + 4(C_2C_3+1)C\sqrt{\log n/n}/s^*,
\end{aligned}$$

because $s^* \leq 1$ and by definition of δ_n. This gives (3.19) and (3.20) with $A = 8\sqrt{C_1}(C_2 + 1)(1 + 2C_2C_3)/s^{*2}$ and $B = 4(C_2C_3 + 1)C/s^{*2}$.

Lemma 33. *Under Assumption 8, we have $\mathcal{W}_\ell(\epsilon) \leq C_1\epsilon$, $\mathcal{W}_T(\delta) \leq C_2\delta$, $\mathcal{W}_{t^*,\alpha}(v) \leq C_3 v$ and $\Psi(\epsilon) \leq \sqrt{C_1}(C_2+1)\sqrt{\epsilon}$ for ϵ, δ, v small enough.*

B.1.2 An optimal procedure

We consider in this section the procedure that serves as an optimal procedure in our theory. For $t \in [0,1]$, let $\mathcal{C}_t^* = (\widehat{\mathbf{Z}}^*, S_t^*)$ be the procedure using the Bayes clustering $\widehat{\mathbf{Z}}^*$ (3.6) and the selection rule $S_t^* = \{i \in \{1,\dots,n\} : T_i^* < t\}$. Let us consider the map $t \in [0,1] \mapsto \mathrm{mFMR}(\mathcal{C}_t^*)$ and note that $\mathrm{mFMR}(\mathcal{C}_t^*) = \mathrm{mFMR}_t^*$ as defined by (3.12). Lemma 36 below provides the key properties for this function.

Definition 6. *The optimal procedure at level α is defined by $\mathcal{C}_{t^*(\alpha)}^*$ where $t^*(\alpha)$ is defined by* (3.13).

Note that the optimal procedure is not the same as the oracle procedure defined in Section 3.3.1, although these two procedures are expected to behave roughly in the same way (at least for a large n).

Under Assumption 6, Lemma 36 entails that, for $\alpha > \alpha_c$, $\mathrm{mFMR}(\mathcal{C}_{t^*(\alpha)}^*) \leq \alpha$. Hence, $\mathcal{C}_{t^*(\alpha)}^*$ controls the mFMR at level α. In addition, it is optimal in the following sense: any other mFMR controlling procedure should select less items than $\mathcal{C}_{t^*(\alpha)}^*$.

Lemma 34 (Optimality of $\mathcal{C}_{t^*(\alpha)}^*$)**.** *Let Assumption 6 be true and choose $\alpha \in (\alpha_c, \bar{\alpha}]$. Then the oracle procedure $\mathcal{C}_{t^*(\alpha)}^* = (\widehat{\mathbf{Z}}^*, S_{t^*(\alpha)}^*)$ satisfies the following:*

(i) $\mathrm{mFMR}(\mathcal{C}^*_{t^*(\alpha)}) = \alpha$*;*

(ii) for any procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ *such that* $\mathrm{mFMR}(\mathcal{C}) \leq \alpha$*, we have* $\mathbb{E}_{\theta^*}(|S|) \leq \mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|)$.

B.1.3 Preliminary steps for proving Theorem 31

To keep the main proof concise, we need to define several additional notation. Let for $t \in [0,1]$ and $\theta \in \Theta$ (recall (3.5))

$$\widehat{\mathbf{L}}_0(\theta, t) = \frac{1}{n} \sum_{i=1}^{n} T(X_i, \theta)\, \mathbb{1}_{T(X_i,\theta)<t}; \tag{B.6}$$

$$\widehat{\mathbf{L}}_1(\theta, t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{T(X_i,\theta)<t}\,. \tag{B.7}$$

Denote $\widehat{\mathbf{L}} = \widehat{\mathbf{L}}_0/\widehat{\mathbf{L}}_1$, $\mathbf{L}_0 = \mathbb{E}_{\theta^*}\widehat{\mathbf{L}}_0$, $\mathbf{L}_1 = \mathbb{E}_{\theta^*}\widehat{\mathbf{L}}_1$, $\mathbf{L} = \mathbf{L}_0/\mathbf{L}_1$ (with the convention $0/0 = 0$). Note that for any $\alpha > \alpha_c$, the mFMR of the optimal procedure $\mathcal{C}^*_{t^*(\alpha)}$ defined in Section B.1.2 is given by $\mathrm{mFMR}(\mathcal{C}^*_{t^*(\alpha)}) = \mathbf{L}(\theta^*, t^*(\alpha)) = \alpha$.

Also, we denote from now on $\ell^*_{i,q} = \mathbb{P}_{\theta^*}(Z_i = q|X_i)$ for short and introduce for any parameter $\theta \in \Theta$ (recall (3.4) and (3.5))

$$\bar{q}(X_i, \theta) \in \operatorname*{argmax}_{q\in\{1,\dots,Q\}} \ell_q(X_i, \theta), \quad 1 \leq i \leq n; \tag{B.8}$$

$$U(X_i, \theta) = 1 - \ell^*_{i,\bar{q}(X_i,\theta)}, \quad 1 \leq i \leq n; \tag{B.9}$$

$$\widehat{\mathbf{M}}_0(\theta, t) = \frac{1}{n} \sum_{i=1}^{n} U(X_i, \theta)\, \mathbb{1}_{T(X_i,\theta)<t}, \quad t \in [0,1], \tag{B.10}$$

Note that $\widehat{\mathbf{M}}_0(\theta^*, t) = \widehat{\mathbf{L}}_0(\theta^*, t)$ but in general $\widehat{\mathbf{M}}_0(\theta, t)$ is different from $\widehat{\mathbf{L}}_0(\theta, t)$. We denote $\widehat{\mathbf{M}} = \widehat{\mathbf{M}}_0/\widehat{\mathbf{L}}_1$, $\mathbf{M}_0 = \mathbb{E}_{\theta^*}\widehat{\mathbf{M}}_0$ and $\mathbf{M} = \mathbf{M}_0/\mathbf{L}_1$ (with the convention $0/0 = 0$).

When $\alpha \in (\alpha_c, \bar{\alpha}]$ (recall (3.14) and (3.15)), we also let

$$s^* = s^*(\alpha, \theta^*) = n^{-1}\, \mathbb{E}_{\theta^*}\left(|S^*_{t^*(\frac{\alpha+\alpha_c}{2})}|\right) = \mathbf{L}_1(\theta^*, t^*((\alpha+\alpha_c)/2)) > 0. \tag{B.11}$$

We easily see that the latter is positive: if it was zero then $S^*_{t^*((\alpha+\alpha_c)/2))}$ would be empty which would entails that $\mathrm{mFMR}(\mathcal{C}^*_{t^*((\alpha+\alpha_c)/2)})$ is zero. This is excluded by definition (3.14) of α_c because $(\alpha+\alpha_c)/2 > \alpha_c$.

Also, we are going to extensively use the event

$$\Omega_\epsilon = \left\{ \min_{\sigma\in[Q]} \|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon \right\}.$$

On this event, we fix any permutation $\sigma \in [Q]$ (possibly depending on X) such that $\|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon$. Now using Lemma 37, the plug-in selection rule can be rewritten as $\widehat{S}^{\mathrm{PI}}_\alpha = \{i \in \{1,\dots,n\} \,:\, \hat{T}_i < \hat{t}(\alpha)\}$ (denoted by $\widehat{S}$ in the sequel for short), where

$$\hat{t}(\alpha) = \sup\{t \in [0,1] \,:\, \widehat{\mathbf{L}}(\hat{\theta}, t) \leq \alpha\}. \tag{B.12}$$

With the above notation, we can upper bound what is inside the brackets of $\mathrm{FMR}(\widehat{\mathcal{C}}^{\mathrm{PI}})$ and $\mathrm{mFMR}(\widehat{\mathcal{C}}^{\mathrm{PI}})$ as follows.

Lemma 35. *For the permutation σ in Ω_ϵ realizing $\|\hat{\theta}^\sigma - \theta^*\|_2 < \epsilon$, we have on the event Ω_ϵ the following relations:*

$$|\widehat{S}| = \widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, \hat{t}(\alpha));$$
$$\min_{\sigma' \in [Q]} \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X} \right) \leq \widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, \hat{t}(\alpha));$$
$$\min_{\sigma' \in [Q]} \mathbb{E}_{\theta^*}\left(\frac{\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}), \mathbf{Z})}{\max(|\widehat{S}|, 1)} \,\middle|\, \mathbf{X} \right) \leq \widehat{\mathbf{M}}(\hat{\theta}^\sigma, \hat{t}(\alpha)).$$

Finally, we make use of the concentration of the empirical processes $\widehat{\mathbf{L}}_0(\theta, t)$, $\widehat{\mathbf{L}}_1(\theta, t)$, and $\widehat{\mathbf{M}}_0(\theta, t)$, uniformly with respect to $\theta \in \mathcal{D}$ (where $\mathcal{D}$ is defined in Assumption 10). Thus, we define the following events, for $\delta > 0$ (recall s^* defined by (B.11)):

$$\Gamma_{0,\delta,t} = \left\{ \sup_{\theta \in \mathcal{D}} \left| \widehat{\mathbf{L}}_0(\theta, t) - \mathbf{L}_0(\theta, t) \right| \leq \delta \right\};$$
$$\Gamma_{1,\delta,t} = \left\{ \sup_{\theta \in \mathcal{D}} \left| \widehat{\mathbf{L}}_1(\theta, t) - \mathbf{L}_1(\theta, t) \right| \leq \delta \right\};$$
$$\Gamma_{\delta,t} = \left\{ \sup_{\substack{\theta \in \mathcal{D}, \\ \mathbf{L}_1(\theta,t) \geq s^*}} \left| \widehat{\mathbf{L}}(\theta, t) - \mathbf{L}(\theta, t) \right| \leq \delta \right\};$$
$$\Upsilon_{0,\delta,t} = \left\{ \sup_{\theta \in \mathcal{D}} \left| \widehat{\mathbf{M}}_0(\theta, t) - \mathbf{M}_0(\theta, t) \right| \leq \delta \right\}.$$

Note that the following holds:

$$\Gamma_{0,\delta s^*/2,t} \cap \Gamma_{1,\delta s^*/2,t} \subset \Gamma_{\delta,t}. \tag{B.13}$$

Indeed, on the event $\Gamma_{0,\delta s^*/2,t} \cap \Gamma_{1,\delta s^*/2,t}$, provided that $\mathbf{L}_1(\theta, t) \geq s^*$, we have

$$\begin{aligned}
&\left| \frac{\widehat{\mathbf{L}}_0(\theta, t)}{\widehat{\mathbf{L}}_1(\theta, t)} - \frac{\mathbf{L}_0(\theta, t)}{\mathbf{L}_1(\theta, t)} \right| \\
&\leq \left| \frac{\mathbf{L}_0(\theta, t) - \widehat{\mathbf{L}}_0(\theta, t)}{\mathbf{L}_1(\theta, t)} \right| + \widehat{\mathbf{L}}_0(\theta, t) \left| \frac{1}{\widehat{\mathbf{L}}_1(\theta, t)} - \frac{1}{\mathbf{L}_1(\theta, t)} \right| \\
&\leq (\delta s^*/2)/s^* + (\delta s^*/2)/s^* = \delta,
\end{aligned}$$

because $\widehat{\mathbf{L}}_0(\theta, t) \leq \widehat{\mathbf{L}}_1(\theta, t)$. This proves the desired inclusion.

B.1.4 Proof of Theorem 31

Let us now provide a proof for Theorem 31.

Step 1: bounding $\hat{t}(\alpha)$ w.r.t. $t^*(\alpha)$ Recall (3.13), (B.12) and (B.11). In this part, we only consider realizations on the event Ω_ϵ. Let $\beta \in [\frac{2\alpha + \alpha_c}{3}, \frac{\alpha + \bar{\alpha}}{2}]$. By Lemma 38, we have

$$\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\beta)) \geq \mathbf{L}_1(\theta^*, t^*(\beta)) - \Psi(\|\hat{\theta}^\sigma - \theta^*\|_2) \geq \mathbf{L}_1(\theta^*, t^*((2\alpha + \alpha_c)/3)) - \Psi(\epsilon),$$

because $t^*(\beta) \geq t^*(\frac{2\alpha + \alpha_c}{3})$ since $t^*(\cdot)$ is non decreasing by Lemma 36. Hence $\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\beta)) \geq s^*$ for ϵ smaller than a threshold only depending on θ^* and α. Hence, we have on $\Gamma_{\delta, t^*(\beta)}$ that

$$\mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)) - \delta \leq \widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\beta)) \leq \delta + \mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)).$$

By using again Lemma 38, we have

$$\mathbf{L}(\theta^*, t^*(\beta)) - 3\Psi(\epsilon)/s^* \leq \mathbf{L}(\hat{\theta}^\sigma, t^*(\beta)) \leq \mathbf{L}(\theta^*, t^*(\beta)) + 3\Psi(\epsilon)/s^*.$$

Given that $\mathbf{L}(\theta^*, t^*(\beta)) = \mathrm{mFMR}(\mathcal{C}^*_{t^*(\beta)}) = \beta$ (see Lemma 34 (i)), it follows that for $\gamma = \gamma(\epsilon, \delta) = \delta + 4\Psi(\epsilon)/s^*$, on the event $\Gamma_{\delta, t^*(\alpha-\gamma)} \cap \Gamma_{\delta, t^*(\alpha+\gamma)}$,

$$\widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) \leq \alpha, \quad \widehat{\mathbf{L}}(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) > \alpha,$$

where we indeed check that $\alpha - \gamma \geq \frac{2\alpha+\alpha_c}{3}$ and $\alpha + \gamma \leq \frac{\alpha+\bar{\alpha}}{2}$ for δ and ϵ smaller than some threshold only depending on θ^* and α. In a nutshell, we have established

$$\Gamma_{\delta, t^*(\alpha-\gamma)} \cap \Gamma_{\delta, t^*(\alpha+\gamma)} \cap \Omega_\epsilon \subset \left\{ t^*(\alpha-\gamma) \leq \hat{t}(\alpha) \leq t^*(\alpha+\gamma) \right\}. \tag{B.14}$$

Step 2: upper-bounding the FMR Let us consider the event

$$\begin{aligned}\Lambda_{\alpha,\delta,\epsilon} := \Gamma_{0,\delta s^*/2, t^*(\alpha-\gamma)} \cap \Gamma_{1,\delta s^*/2, t^*(\alpha-\gamma)} \cap \Gamma_{0,\delta s^*/2, t^*(\alpha+\gamma)} \\ \cap \Gamma_{1,\delta s^*/2, t^*(\alpha+\gamma)} \cap \Upsilon_{0,\delta, t^*(\alpha+\gamma)} \cap \Omega_\epsilon,\end{aligned}$$

where the different events have been defined in the previous section.

Let us prove (3.19). By using Lemma 35 and (B.14),

$$\begin{aligned}\mathrm{FMR}(\hat{\mathcal{C}}) &\leq \mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}(\hat{\theta}^\sigma, \hat{t}(\alpha))\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\leq \mathbb{E}_{\theta^*}\left[\frac{\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma))}{\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma))}\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}} \right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c).\end{aligned}$$

Now using a concentration argument on the event $\Lambda_{\alpha,\delta,\epsilon} \subset \Gamma_{1,\delta,t^*(\alpha-\gamma)} \cap \Upsilon_{0,\delta,t^*(\alpha+\gamma)}$, we have

$$\begin{aligned}\mathrm{FMR}(\hat{\mathcal{C}}) &\leq \mathbb{E}_{\theta^*}\left[\frac{\mathbf{M}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) + \delta}{\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) - \delta}\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}} \right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\leq \frac{\mathbf{M}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &= \frac{\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),\end{aligned} \tag{B.15}$$

by using Lemma 38 and that $\mathbf{M}_0(\theta^*, t) = \mathbf{L}_0(\theta^*, t)$ for all t by definition. Now, using again Lemma 38, we have

$$\begin{aligned}\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) &\leq \mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(t^*(\alpha+\gamma) - t^*(\alpha-\gamma)) \\ &\leq \mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma))\end{aligned}$$

This entails

$$\begin{aligned}\mathrm{FMR}(\hat{\mathcal{C}}) &\leq \frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) + \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma)) + 3\Psi(\epsilon) + \delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\leq \frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma))}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} \\ &\quad + (s^*/2)^{-1} \left(\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\gamma)) + 3\Psi(\epsilon) + \delta \right) + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),\end{aligned}$$

by choosing ϵ, δ smaller than a threshold (only depending on θ^* and α) so that $\mathbf{L}_1(\theta^*, t^*(\alpha - \gamma)) - \Psi(\epsilon) - \delta \geq s^*/2$. Now using $\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma)) = (\alpha-\gamma)\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma))$, we have

$$\begin{aligned}\frac{\mathbf{L}_0(\theta^*, t^*(\alpha-\gamma))}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta} &= (\alpha-\gamma)\left(1 + \frac{\Psi(\epsilon)+\delta}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta}\right) \\ &\leq \alpha\left(1 + (s^*/2)^{-1}(\Psi(\epsilon)+\delta)\right).\end{aligned}$$

This leads to

$$\mathrm{FMR}(\hat{\mathcal{C}}) \leq \alpha + (2/s^*)\left[\mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(2\delta + 8\Psi(\epsilon)/s^*)) + 4\Psi(\epsilon) + 2\delta\right] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),$$

which holds true for δ, ϵ smaller than a threshold only depending on θ^* and α.

Step 3: upper-bounding the mFMR We apply a similar technique as for step 2. By using Lemma 35 and (B.14),

$$\begin{aligned}\mathrm{mFMR}(\hat{\mathcal{C}}) &\leq \frac{\mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, \hat{t}(\alpha))\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, \hat{t}(\alpha))\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]} \\ &\leq \frac{\mathbb{E}_{\theta^*}[\widehat{\mathbf{M}}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma))\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[\widehat{\mathbf{L}}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma))\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]}.\end{aligned}$$

Now using a concentration argument on $\Lambda_{\alpha,\delta,\epsilon} \subset \Gamma_{1,\delta,t^*(\alpha-\gamma)} \cap \Upsilon_{0,\delta,t^*(\alpha+\gamma)}$, we have

$$\begin{aligned}\mathrm{mFMR}(\hat{\mathcal{C}}) &\leq \frac{\mathbb{E}_{\theta^*}[(\mathbf{M}_0(\hat{\theta}^\sigma, t^*(\alpha+\gamma)) + \delta)\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}] + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbb{E}_{\theta^*}[(\mathbf{L}_1(\hat{\theta}^\sigma, t^*(\alpha-\gamma)) - \delta)\, \mathbb{1}_{\Lambda_{\alpha,\delta,\epsilon}}]} \\ &\leq \frac{\mathbf{M}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)} \\ &= \frac{\mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta + \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)}{\mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)},\end{aligned}$$

by using Lemma 38 and that $\mathbf{M}_0(\theta^*, t) = \mathbf{L}_0(\theta^*, t)$ by definition. Letting $x = \mathbf{L}_0(\theta^*, t^*(\alpha+\gamma)) + 3\Psi(\epsilon) + \delta$, $y = \mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta$ and $u = \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$, we have obtained the bound $(x+u)/(y-u)$, which has to be compared with the FMR bound (B.15), which reads $x/y + u$. Now, when $y \in [0,1]$, $x \geq 0$, $x/y \leq 2$, $u/y \leq 1/2$, $y - u \geq s^*/2$, we have

$$(x+u)/(y-u) \leq \frac{x/y}{1-u/y} + (2/s^*)u \leq x/y(1+2u/y) + (2/s^*)u \leq x/y + (10/s^*)u.$$

As a result, for ϵ, δ small enough, and $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \leq s^*/4$, we obtain the same bound as for the FMR, with $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$ replaced by $(10/s^*)\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$.

Step 4: lower-bounding the selection rate In Step 3, when bounding the mFMR, we derived a lower bound for the denominator of the mFMR, that is, $\mathbb{E}_{\theta^*}(|\hat{S}|)$. It reads

$$\begin{aligned}n^{-1}\,\mathbb{E}_{\theta^*}(|\hat{S}|) &\geq \mathbf{L}_1(\theta^*, t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\geq \mathbf{L}_1(\theta^*, t^*(\alpha)) - \mathcal{W}_T(t^*(\alpha) - t^*(\alpha-\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \\ &\geq n^{-1}\,\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) - \mathcal{W}_T(\mathcal{W}_{t^*,\alpha}(\gamma)) - \Psi(\epsilon) - \delta - \mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c),\end{aligned}$$

by using (B.3) and (B.5). Now consider another procedure $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ that controls the mFMR at level α, that is, $\mathrm{mFMR}(\mathcal{C}) \leq \alpha$. By Lemma 34, we then have $\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) \geq \mathbb{E}_{\theta^*}(|S|)$.

Step 5: concentration Finally, we bound $\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c)$ by using Lemma 39 with $x = (1 + 2c)\sqrt{\frac{\log n}{n}}$ (with c defined in Lemma 39). This gives for $\delta = 2x/s^*$, and $n \geq (2e)^3$

$$\mathbb{P}((\Lambda_{\alpha,\delta,\epsilon})^c) \leq 5/n^2 + \mathbb{P}(\Omega_\epsilon^c).$$

B.2 Proofs of lemmas

Proof of Lemma 13 The clustering risk of $\widehat{\mathbf{Z}}$ is given by

$$\begin{aligned} R(\widehat{\mathbf{Z}}) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(n^{-1}\sum_{i=1}^n \mathbb{1}\{Z_i \neq \sigma(\hat{Z}_i)\} \,\middle|\, \mathbf{X}\right)\right) \\ &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \sigma(\hat{Z}_i) \mid \mathbf{X})\right) \\ &\geq \mathbb{E}_{\theta^*}\left(\min_{\widehat{\mathbf{Z}}} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i \mid \mathbf{X})\right), \end{aligned}$$

where, by independence, the minimum in the lower bound is achieved for the Bayes clustering. Thus, $R(\widehat{\mathbf{Z}}) \geq n^{-1}\sum_{i=1}^n \mathbb{E}_{\theta^*}(T_i^*)$. Moreover, $n^{-1}\sum_{i=1}^n \mathbb{E}_{\theta^*}(T_i^*) \geq R(\widehat{\mathbf{Z}}^*)$, since

$$\begin{aligned} R(\widehat{\mathbf{Z}}^*) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \sigma(\hat{Z}_i^*) \mid \mathbf{X})\right) \\ &\leq \mathbb{E}_{\theta^*}\left(n^{-1}\sum_{i=1}^n \mathbb{P}_{\theta^*}(Z_i \neq \hat{Z}_i^* \mid \mathbf{X})\right). \end{aligned}$$

Thus, $\min_{\widehat{\mathbf{Z}}} R(\widehat{\mathbf{Z}}) = R(\widehat{\mathbf{Z}}^*)$ and the proof is completed.

Proof of Lemma 14 Following the reasoning of the proof of Lemma 13, we have

$$\begin{aligned} \mathrm{FMR}_{\theta^*}(\mathcal{C}) &= \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\frac{\sum_{i\in S}\mathbb{1}\{Z_i \neq \sigma(\hat{Z}_i^*)\}}{\max(|S|,1)} \,\middle|\, \mathbf{X}\right)\right) \\ &= \mathbb{E}_{\theta^*}\left(\frac{\sum_{i\in S} T_i^*}{\max(|S|,1)}\right). \end{aligned}$$

Proof of Lemma 15 By definition, we have

$$\mathrm{FMR}(\widehat{\mathcal{C}}_\alpha^{\mathrm{PI}}) = \mathbb{E}_{\theta^*}\left(\min_{\sigma\in[Q]} \mathbb{E}_{\theta^*}\left(\frac{\sum_{i=1}^n \mathbb{1}_{Z_i\neq\sigma(\hat{Z}_i^{\mathrm{PI}}(\mathbf{X}))}\,\mathbb{1}\{i \in \hat{S}^{\mathrm{PI}}(\mathbf{X})\}}{\max(|\hat{S}^{\mathrm{PI}}(\mathbf{X})|,1)} \,\middle|\, \mathbf{X}\right)\right),$$

so that (3.9) follows by a direct integration w.r.t. the latent variable Z.

Proof of Lemma 34 By Lemma 36, we have that $\mathrm{mFMR}(\mathcal{C}_t^*)$ is monotonous in t and continuous w.r.t. t on $(t^*(\alpha_c), 1]$, thus for $\alpha \in (\alpha_c, \bar{\alpha}]$, $\mathrm{mFMR}(\mathcal{C}_{t^*(\alpha)}^*) = \alpha$ which gives (i). For (ii), let $\mathcal{C} = (\widehat{\mathbf{Z}}, S)$ be a procedure such that $\mathrm{mFMR}(\mathcal{C}) \leq \alpha$. Let us consider the procedure $\mathcal{C}'$ with the Bayes clustering $\widehat{\mathbf{Z}}^*$ and the same selection rule S. Since $\mathcal{C}'$ is based on a Bayes

clustering, by the same reasoning leading to $R(\widehat{\mathbf{Z}}^*) \leq R(\widehat{\mathbf{Z}})$ in Section 3.3.1, we have that $\mathrm{mFMR}(\mathcal{C}') \leq \mathrm{mFMR}(\mathcal{C}) \leq \alpha$ with

$$\mathrm{mFMR}(\mathcal{C}') = \frac{\mathbb{E}_{\theta^*}\left(\sum_{i\in S} T_i^*\right)}{\mathbb{E}_{\theta^*}(|S|)}.$$

Hence,

$$\mathbb{E}_{\theta^*}\left(\sum_{i\in S} T_i^*\right) \leq \alpha\, \mathbb{E}_{\theta^*}(|S|). \tag{B.16}$$

Now we use an argument similar to the proof of Theorem 1 in Cai et al. (2019). By definition of $S^*_{t^*(\alpha)}$, we have that

$$\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - t^*(\alpha)) \leq 0$$

which we can rewrite as

$$\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - t^*(\alpha) + \alpha - \alpha) \leq 0$$

and so

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - \alpha)\right) \\
&\leq (t^*(\alpha) - \alpha)\, \mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)\right) \\
&= (t^*(\alpha) - \alpha)(\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) - \mathbb{E}_{\theta^*}(|S|)).
\end{aligned}$$

On the other hand, $\mathrm{mFMR}(\mathcal{C}^*_{t^*(\alpha)}) = \alpha$ together with (B.16) implies that

$$\begin{aligned}
&\mathbb{E}_{\theta^*}\left(\sum_{i=1}^{n} \left(\mathbb{1}_{i\in S^*_{t^*(\alpha)}}(\mathbf{X}) - \mathbb{1}_{i\in S(\mathbf{X})}\right)(T_i^* - \alpha)\right) \\
&= \mathbb{E}_{\theta^*}\left(\sum_{i\in S^*_{t^*(\alpha)}} T_i^* - \alpha|S^*_{t^*(\alpha)}| - \sum_{i\in S} T_i^* + \alpha|S|\right) \geq 0.
\end{aligned}$$

Combining, the relations above provides

$$(t^*(\alpha) - \alpha)(\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) - \mathbb{E}_{\theta^*}(|S|)) \geq 0.$$

Finally, noting that $t^*(\alpha) - \alpha > 0$ since $\alpha = \mathrm{mFMR}(\mathcal{C}^*_{t^*(\alpha)}) < t^*(\alpha)$ by (ii) Lemma 36, this gives $\mathbb{E}_{\theta^*}(|S^*_{t^*(\alpha)}|) - \mathbb{E}_{\theta^*}(|S|) \geq 0$ and concludes the proof.

Proof of Lemma 35 First, we have by definition $\ell_q(X_i, \theta^\sigma) = \ell_{\sigma(q)}(X_i, \theta)$ and thus $T(X_i, \hat{\theta}) = T(X_i, \hat{\theta}^\sigma)$ by taking the maximum over q. This gives $\widehat{S}^\sigma = \widehat{S}$ and yields the first equality. Next, we have on Ω_ϵ,

$$\begin{aligned}
\min_{\sigma'\in[Q]} \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma'(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right) &\leq \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}}(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right) \\
&\leq \mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}^\sigma}(\sigma(\widehat{\mathbf{Z}}), \mathbf{Z}) \,\middle|\, \mathbf{X}\right),
\end{aligned}$$

still because $\widehat{S}^\sigma = \widehat{S}$. Now observe that,

$$\begin{aligned}\mathbb{E}_{\theta^*}\left(\varepsilon_{\widehat{S}^\sigma}(\sigma(\widehat{\mathbf{Z}}),\mathbf{Z})\,\middle|\,\mathbf{X}\right) &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{P}_{\theta^*}(Z_i \neq \sigma(\bar{q}(X_i,\hat{\theta}))\,|\,\mathbf{X})\,\mathbb{1}_{T(X_i,\hat{\theta}^\sigma)<\hat{t}(\alpha)}\\ &= \frac{1}{n}\sum_{i=1}^{n}(1-\ell^*_{i,\sigma(\bar{q}(X_i,\hat{\theta}))})\,\mathbb{1}_{T(X_i,\hat{\theta}^\sigma)<\hat{t}(\alpha)}\\ &= \widehat{\mathbf{M}}_0(\hat{\theta}^\sigma,\hat{t}(\alpha)),\end{aligned}$$

because $\sigma(\bar{q}(X_i,\hat{\theta})) = \bar{q}(X_i,\hat{\theta}^\sigma)$. This proves the result.

B.3 Auxiliary results

Lemma 36. *Let us consider the procedure $\mathcal{C}^*_t$ defined in Section B.1.2 and the functional* mFMR^*_t *defined by* (3.12). *Then we have*

$$\mathrm{mFMR}(\mathcal{C}^*_t) = \frac{\mathbb{E}_{\theta^*}\left(\sum_{i=1}^n T^*_i\,\mathbb{1}_{T^*_i<t}\right)}{\mathbb{E}_{\theta^*}\left(\sum_{i=1}^n \mathbb{1}_{T^*_i<t}\right)} = \mathrm{mFMR}^*_t, \quad t\in[0,1]. \tag{B.17}$$

Moreover, the following properties for the function $t\in[0,1]\mapsto \mathrm{mFMR}(\mathcal{C}^*_t)$:

(i) $\mathrm{mFMR}(\mathcal{C}^*_t)$ *is non-decreasing in $t\in[0,1]$ and, under Assumption 6, it is increasing in $t\in(t^*(\alpha_c),t^*(\bar{\alpha}))$;*

(ii) $\mathrm{mFMR}(\mathcal{C}^*_t) < t$ *for $t\in(0,1]$;*

(iii) Under Assumption 6, $\mathrm{mFMR}(\mathcal{C}^*_t)$ *is continuous w.r.t. t on $(t^*(\alpha_c),1]$, where $t^*(\alpha_c)$ is given by* (3.14).

Proof. First, (B.17) is obtained similarly than (3.8). For proving (i), let $t_1,t_2\in[0,1]$ such that $t_1<t_2$. We show that $\mathrm{mFMR}(\mathcal{C}^*_{t_1}) \leq \mathrm{mFMR}(\mathcal{C}^*_{t_2})$. Remember here the convention $0/0=0$ and that $\mathrm{mFMR}(\mathcal{C}^*_t) = \mathbb{E}_{\theta^*}\left(T(X,\theta^*)\,|\,T(X,\theta^*)<t\right)$. First, if $\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)=0$ then the result is immediate. Otherwise, we have that

$$\begin{aligned}&\mathrm{mFMR}(\mathcal{C}^*_{t_1}) - \mathrm{mFMR}(\mathcal{C}^*_{t_2})\\ &= \left(\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)\right)^{-1}\\ &\cdot \mathbb{E}_{\theta^*}\left(T(X,\theta^*)\left\{\mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)}{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_2\right)}\,\mathbb{1}_{T(X,\theta^*)<t_2}\right\}\right),\end{aligned}$$

where, given that $t_1<t_2$, the quantity in the brackets is positive when $T(X,\theta^*)<t_1$ and is negative or zero otherwise. Hence,

$$\begin{aligned}&T(X,\theta^*)\left\{\mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)}{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_2\right)}\,\mathbb{1}_{T(X,\theta^*)<t_2}\right\}\\ &\leq t_1\left\{\mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)}{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_2\right)}\,\mathbb{1}_{T(X,\theta^*)<t_2}\right\}.\end{aligned}$$

Taking the expectation makes the right-hand-side equal to zero, from which the result follows. Now, to show the increasingness, if $\mathrm{mFMR}(\mathcal{C}^*_{t_1}) = \mathrm{mFMR}(\mathcal{C}^*_{t_2})$ for $t^*(\alpha_c)<t_1<t_2<t^*(\bar{\alpha})$, then the above reasoning shows that

$$\left(T(X,\theta^*)-t_1\right)\left\{\mathbb{1}_{T(X,\theta^*)<t_1} - \frac{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_1\right)}{\mathbb{P}_{\theta^*}\left(T(X,\theta^*)<t_2\right)}\,\mathbb{1}_{T(X,\theta^*)<t_2}\right\} \leq 0$$

and has an expectation equal to 0. Hence, given that $T(X,\theta^*)$ is continuous, we derive that almost surely

$$\mathbb{P}_{\theta^*}\left(T(X,\theta^*) < t_2\right) \mathbb{1}_{T(X,\theta^*)<t_1} = \mathbb{P}_{\theta^*}\left(T(X,\theta^*) < t_1\right) \mathbb{1}_{T(X,\theta^*)<t_2},$$

that is, $\mathbb{P}_{\theta^*}(t_1 \leq T_i^* < t_2) = 0$, which is excluded by Assumption 6. This entails $\mathrm{mFMR}(\mathcal{C}^*_{t_1}) < \mathrm{mFMR}(\mathcal{C}^*_{t_2})$.

For proving (ii), let $t > 0$. If $\mathbb{P}_{\theta^*}\left(T(X,\theta^*) < t\right) = 0$ then the result is immediate. Otherwise, we have that $\mathrm{mFMR}(\mathcal{C}^*_t) - t = \left(\mathbb{P}_{\theta^*}\left(T(X,\theta^*) < t\right)\right)^{-1} \mathbb{E}_{\theta^*}((T(X,\theta^*) - t)\, \mathbb{1}\{T(X,\theta^*) < t\})$. The latter is clearly not positive, and is moreover negative because $(T(X,\theta^*)-t)\, \mathbb{1}\{T(X,\theta^*) < t\} \leq 0$ and $\mathbb{P}_{\theta^*}(T(X,\theta^*) = t) = 0$ by Assumption 6.

For proving (iii), let $\psi_0(t) = \mathbb{E}_{\theta^*}(T(X,\theta^*)\, \mathbb{1}\{T(X,\theta^*) < t\})$ and $\psi_1(t) = \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)$, the numerator and denominator of $\mathrm{mFMR}(\mathcal{C}^*_t) = \mathrm{mFMR}^*_t$, respectively. $\psi_1(t)$ is non-decreasing in t, with $\psi_1(0) = 0$ and $\psi_1(1) > 0$. Moreover, ψ_0 and ψ_1 are both continuous under Assumption 6. Then denote by t_c the largest t s.t. $\psi_1(t) = 0$. ψ_1 is zero on $[0, t_c]$ then strictly positive and non-decreasing on $(t_c, 1]$, and we have that $t_c = t^*(\alpha_c)$. Hence, $\mathrm{mFMR}(\mathcal{C}^*_t)$ is zero on $[0, t_c]$ then strictly positive and continuous on $(t_c, 1]$. □

Remark 17. With the notation of the above proof, $t \mapsto \mathrm{mFMR}(\mathcal{C}^*_t)$ may have a discontinuity point at t_c since for $t_n \underset{t_n > t_c}{\to} t_c$, as $\psi_1(t_n) \to 0$, one does not necessarily have that $\mathrm{mFMR}(\mathcal{C}^*_t) \to 0$.

Lemma 37 (Expression of plug-in procedure as a thresholding rule)**.** *For any $\alpha \in (0,1)$, let us consider the plug-in procedure $\widehat{\mathcal{C}}^{PI}_\alpha = (\widehat{\mathbf{Z}}^{PI}, \widehat{S}^{PI}_\alpha)$ defined by Algorithm 4 and denote $K = |\widehat{S}^{PI}_\alpha|$ the maximum of the $k \in \{0,\dots,n\}$ such that $\max(k,1)^{-1} \sum_{j=1}^k \hat{T}_{(j)} \leq \alpha$ for $\hat{T}_i = 1 - \max_q \ell_q(X_i, \hat{\theta})$, $1 \leq i \leq n$. Consider also $\hat{t}(\alpha)$ defined by* (B.12)*. Let Assumption 6 be true and consider an estimator $\hat{\theta}$ satisfying Assumption 10. Then it holds that $\hat{t}(\alpha) = \hat{T}_{(K+1)}$ and*

$$\widehat{S}^{PI}_\alpha = \{i \in \{1,\dots,n\} \,:\, \hat{T}_i < \hat{t}(\alpha)\}.$$

Proof. If $\hat{T}_{(K)} < \hat{T}_{(K+1)}$ then the result is immediate. Thus it suffices to show that $\hat{T}_{(K)} = \hat{T}_{(K+1)}$ occurs with probability 0. From Assumption 10 (with the countable set $\mathcal{D}$ defined therein), we have

$$\mathbb{P}_{\theta^*}(\hat{T}_{(K)} = \hat{T}_{(K+1)}) \leq \mathbb{P}_{\theta^*}\left(\bigcup_{i\neq j} \{\hat{T}_i = \hat{T}_j\}\right) \leq \mathbb{P}_{\theta^*}\left(\bigcup_{\theta\in\mathcal{D}} \bigcup_{i\neq j} \{T(X_i,\theta) = T(X_j,\theta)\}\right).$$

Now, the right term is a countable union of events which are all of null probability under Assumption 6. The result follows. □

Lemma 38. *We have for all $\theta \in \Theta$,*

$$\sup_{t\in[0,1]} |\mathbf{L}_1(\theta,t) - \mathbf{L}_1(\theta^*,t)| \leq \Psi(\|\theta^* - \theta\|); \tag{B.18}$$

$$\sup_{t\in[0,1]} |\mathbf{L}_0(\theta,t) - \mathbf{L}_0(\theta^*,t)| \leq 2\Psi(\|\theta^* - \theta\|); \tag{B.19}$$

$$\sup_{t\in[t^*((\alpha+\alpha_c)/2),1]} |\mathbf{L}(\theta,t) - \mathbf{L}(\theta^*,t)| \leq 3\Psi(\|\theta^* - \theta\|)/s^*; \tag{B.20}$$

$$\sup_{t\in[0,1]} |\mathbf{M}_0(\theta,t) - \mathbf{M}_0(\theta^*,t)| \leq 3\Psi(\|\theta^* - \theta\|); \tag{B.21}$$

$$\sup_{t\in[t^*((\alpha+\alpha_c)/2),1]} |\mathbf{M}(\theta,t) - \mathbf{M}(\theta^*,t)| \leq 4\Psi(\|\theta^* - \theta\|)/s^*; \tag{B.22}$$

where $\alpha \in (\alpha_c, \bar{\alpha}]$ *and* $s^* > 0$ *is given by* (B.11). *In addition, for all* $\theta \in \Theta$ *and* $t, t' \in [0,1]$,

$$|\mathbf{L}_0(\theta, t) - \mathbf{L}_0(\theta, t')| \leq 4\Psi(\|\theta^* - \theta\|) + \mathcal{W}_T(|t - t'|). \tag{B.23}$$

Proof. Fix $\theta \in \Theta$ and $t \in [0,1]$. We have for any $\delta > 0$,

$$\begin{aligned}
&|\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \\
&\leq (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t + \delta) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)) \vee (\mathbb{P}_{\theta^*}(T(X,\theta^*) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t - \delta)) \\
&\quad + \mathbb{P}_{\theta^*}(|T(X,\theta^*) - T(X,\theta)| > \delta) \\
&\leq \mathcal{W}_T(\delta) + \mathbb{E}_{\theta^*}(|T(X,\theta^*) - T(X,\theta)|)/\delta.
\end{aligned}$$

In addition, by definition (3.5),

$$\begin{aligned}
|T(X,\theta^*) - T(X,\theta)| &\leq |\max_{1\leq q\leq Q} \ell_q(X,\theta^*) - \max_{1\leq q\leq Q} \ell_q(X,\theta)| \\
&\leq \max_{1\leq q\leq Q} |\ell_q(X,\theta^*) - \ell_q(X,\theta)|.
\end{aligned}$$

Hence,

$$|\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \leq \inf_{\delta\in(0,1)} \{\mathcal{W}_T(\delta) + \mathcal{W}_\ell(\|\theta^* - \theta\|)/\delta\} \leq \Psi(\|\theta^* - \theta\|),$$

which establishes (B.18).

Next, we have

$$\begin{aligned}
&\mathbf{L}_0(\theta, t) - \mathbf{L}_0(\theta^*, t) \\
&= \mathbb{E}_{\theta^*}[T(X,\theta)(\mathbb{1}_{T(X,\theta)<t} - \mathbb{1}_{T(X,\theta^*)<t}) + \mathbb{1}_{T(X,\theta^*)<t}(T(X,\theta) - T(X,\theta^*))] \\
&\leq t|\,\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| \\
&\quad + |\,\mathbb{E}_{\theta^*}[\mathbb{1}_{T(X,\theta^*)<t}(T(X,\theta) - T(X,\theta^*))]| \\
&\leq |\,\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| + \mathbb{E}_{\theta^*}|T(X,\theta) - T(X,\theta^*)| \\
&\leq 2\Psi(\|\theta^* - \theta\|)
\end{aligned}$$

By exchanging the role of θ and θ^* in the above reasoning, the same bound holds for $\mathbf{L}_0(\theta^*, t) - \mathbf{L}_0(\theta, t)$, which gives (B.19). To prove (B.20), we use for any $t \in [t^*(\frac{\alpha+\alpha_c}{2}), 1]$,

$$\begin{aligned}
&\left|\frac{\mathbf{L}_0(\theta,t)}{\mathbf{L}_1(\theta,t)} - \frac{\mathbf{L}_0(\theta^*,t)}{\mathbf{L}_1(\theta^*,t)}\right| \\
&\leq \left|\frac{\mathbf{L}_0(\theta,t) - \mathbf{L}_0(\theta^*,t)}{\mathbf{L}_1(\theta^*,t)}\right| + \mathbf{L}_0(\theta,t)\left|\frac{1}{\mathbf{L}_1(\theta^*,t)} - \frac{1}{\mathbf{L}_1(\theta,t)}\right| \\
&\leq 2\Psi(\|\theta^* - \theta\|)/s^* + \frac{1}{\mathbf{L}_1(\theta^*,t)}\frac{\mathbf{L}_0(\theta,t)}{\mathbf{L}_1(\theta,t)}\,|\mathbb{P}_{\theta^*}(T(X,\theta^*) < t) - \mathbb{P}_{\theta^*}(T(X,\theta) < t)| \\
&\leq 3\Psi(\|\theta^* - \theta\|)/s^*,
\end{aligned}$$

because $\mathbf{L}_0(\theta,t) \leq \mathbf{L}_1(\theta,t)$ and $\mathbf{L}_1(\theta^*,t) \geq s^*$ by monotonicity. Similarly to the bound on $\mathbf{L}_0$, we derive

$$\begin{aligned}
&|\mathbf{M}_0(\theta,t) - \mathbf{M}_0(\theta^*,t)| \\
&\leq |\,\mathbb{P}_{\theta^*}(T(X,\theta) < t) - \mathbb{P}_{\theta^*}(T(X,\theta^*) < t)| + \mathbb{E}_{\theta^*}|U(X,\theta) - U(X,\theta^*)|.
\end{aligned}$$

Define $\bar{q}(X,\theta) \in \underset{q\in\{1,\dots,Q\}}{\operatorname{argmax}} \ell_q(X,\theta)$. Now, since $U(X,\theta^*) \leq U(X,\theta)$ by definition (B.9), we have

$$\begin{aligned}
\mathbb{E}_{\theta^*}|U(X,\theta)-U(X,\theta^*)| &= \mathbb{E}_{\theta^*}[U(X,\theta)-U(X,\theta^*)] \\
&= \mathbb{E}_{\theta^*}[\ell_{\bar{q}(X,\theta)}(X,\theta^*)-\ell_{\bar{q}(X,\theta^*)}(X,\theta^*)] \\
&= \mathbb{E}_{\theta^*}[\ell_{\bar{q}(X,\theta)}(X,\theta^*)-\ell_{\bar{q}(X,\theta)}(X,\theta) \\
&\quad + \ell_{\bar{q}(X,\theta)}(X,\theta)-\ell_{\bar{q}(X,\theta^*)}(X,\theta^*)] \\
&\leq \mathbb{E}_{\theta^*}[\max_{1\leq q\leq Q}|\ell_q(X,\theta^*)-\ell_q(X,\theta)|] \\
&\quad + \mathbb{E}_{\theta^*}[\max_{1\leq q\leq Q}\ell_q(X,\theta)-\max_{1\leq q\leq Q}\ell_q(X,\theta^*)] \\
&\leq 2\,\mathbb{E}_{\theta^*}[\max_{1\leq q\leq Q}|\ell_q(X,\theta^*)-\ell_q(X,\theta)|] \leq 2\Psi(\|\theta^*-\theta\|).
\end{aligned}$$

This proves (B.21) and leads to (B.22) by following the reasoning that provided (B.20).

Next, we have for $0 \leq t' \leq t \leq 1$, by (B.19),

$$|\mathbf{L}_0(\theta,t)-\mathbf{L}_0(\theta,t')| \leq |\mathbf{L}_0(\theta^*,t)-\mathbf{L}_0(\theta^*,t')| + 4\Psi(\|\theta^*-\theta\|).$$

Moreover,

$$\begin{aligned}
|\mathbf{L}_0(\theta^*,t)-\mathbf{L}_0(\theta^*,t')| &= \mathbf{L}_0(\theta^*,t)-\mathbf{L}_0(\theta^*,t') = \mathbb{E}_{\theta^*}[T(X,\theta^*)\mathbb{1}_{t'\leq T(X,\theta^*)<t}] \\
&\leq \mathbb{E}_{\theta^*}[\mathbb{1}_{t'\leq T(X,\theta^*)<t}] \\
&= \mathbb{P}_{\theta^*}(T(X,\theta^*)<t)-\mathbb{P}_{\theta^*}(T(X,\theta^*)<t'),
\end{aligned}$$

which is below $\mathcal{W}_T(t-t')$ by (B.3). This leads to (B.23). □

Lemma 39 (Concentration of $\widehat{\mathbf{L}}_\mathbf{0}$ (B.6), $\widehat{\mathbf{L}}_\mathbf{1}$ (B.7), and $\widehat{\mathbf{M}}_\mathbf{0}$ (B.10))**.** *Let Assumption 6 be true. Recall $\mathscr{V}, \mathscr{V}_-$ defined by* (3.16), (3.17) *respectively, set $c := 14Q\sqrt{\mathscr{V}} + 7Q^2\sqrt{\mathscr{V}_-}$ and consider any countable set $\mathcal{D} \subset \Theta$. For all $t \in (0,1]$ and for $n \geq (2e)^3$, we have*

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{L}}_0(\theta,t)-\mathbf{L}_0(\theta,t)\right| > x\right) \leq n^{-2}; \tag{B.24}$$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{L}}_1(\theta,t)-\mathbf{L}_1(\theta,t)\right| > x\right) \leq n^{-2}; \tag{B.25}$$

$$\mathbb{P}_{\theta^*}\left(\sup_{\theta\in\mathcal{D}}\left|\widehat{\mathbf{M}}_0(\theta,t)-\mathbf{M}_0(\theta,t)\right| > x\right) \leq n^{-2}, \tag{B.26}$$

for any $x \geq (1+2c)\sqrt{\frac{\log n}{n}}$ and provided that $(1+2c)\sqrt{\frac{\log n}{n}} \leq 1$.

Proof. For a fixed $t \in (0,1]$, let $\mathscr{F}_{L_0} = \{T(.,\theta)\,\mathbb{1}\{T(.,\theta)\leq t\}, \theta\in\mathcal{D}\}$, $\mathscr{F}_{L_1} = \{\mathbb{1}\{T(.,\theta)\leq t\}, \theta\in\mathcal{D}\}$, and $\mathscr{F}_{M_0} = \{U(.,\theta)\,\mathbb{1}\{T(.,\theta)\leq t\}, \theta\in\mathcal{D}\}$. We apply Lemma 42 and Lemma 43 for $\xi_i = X_i$, $1\leq i\leq n$, $b=1$, $a=0$ and for each $\mathscr{F} \in \{\mathscr{F}_{L_0}, \mathscr{F}_{L_1}, \mathscr{F}_{M_0}\}$ to get that the corresponding probability in (B.24)-(B.25)-(B.26) is at most n^{-2} by taking

$$x \geq \sqrt{\frac{\log n}{n}} + 2\,\mathbb{E}\,\mathfrak{R}_n(\mathscr{F}),$$

where $\mathfrak{R}_n(\mathscr{F})$ denotes the Rademacher complexity of $\mathscr{F}$, see (B.30). We now bound each $\mathfrak{R}_n(\mathscr{F})$ by using Lemma 40:

$$\begin{aligned}\mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_0}) &\leq \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_1}) + \mathbb{E}\,\mathfrak{R}_n(\{T(.,\theta), \theta\in\Theta\}) \\ &\leq \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_1}) + \sum_{q=1}^{Q} \mathbb{E}\,\mathfrak{R}_n(\{\ell_q(.,\theta), \theta\in\Theta\}); \qquad \text{(B.27)}\\ \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_1}) &\leq \sum_{q=1}^{Q} \mathbb{E}\,\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) < 1-t\}, \theta\in\Theta\}); \qquad \text{(B.28)}\\ \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{M_0}) &\leq \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_1}) + \mathbb{E}\,\mathfrak{R}_n(\{U(.,\theta), \theta\in\Theta\}),\end{aligned}$$

where for (B.27) and (B.28), we used that $T(.,\theta) = 1 - \max_q \ell_q(.,\theta)$ and $\mathbb{1}\{T(.,\theta)\leq t\} = 1 - \prod_{q=1}^{Q}\mathbb{1}\{\ell_q(.,\theta) < 1-t\}$ and the fact that the variables $\ell_q(X_i,\theta)$ are continuous by Assumption 6. Similarly, we have $U(.,\theta) = \sum_{q=1}^{Q}\ell_q(.,\theta^*)\prod_{k\neq q}\mathbb{1}\{\ell_q(.,\theta)\geq \ell_k(.,\theta)\}$. Hence, Lemma 40 once again entails that

$$\begin{aligned}\mathbb{E}\,\mathfrak{R}_n(\{U(.,\theta), \theta\in\Theta\}) \leq &\sum_{q=1}^{Q}\sum_{k=1,k\neq q}^{Q} \mathbb{E}\,\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) - \ell_k(.,\theta)\geq 0\}, \theta\in\Theta\}) \\ &+ \sum_{q=1}^{Q}\mathbb{E}\,\mathfrak{R}_n(\{\ell_q(.,\theta), \theta\in\Theta\}). \qquad \text{(B.29)}\end{aligned}$$

To bound both $\mathbb{E}\,\mathfrak{R}_n(\{\ell_q(.,\theta), \theta\in\Theta\})$ and $\mathbb{E}\,\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) < 1-t\}, \theta\in\Theta\})$, we use the results of Baraud (2016) (more specifically the proof of Theorem 1 therein), to obtain that they are bounded by

$$\sqrt{\mathscr{V}\log\frac{2en}{\mathscr{V}}}\frac{\sqrt{2}}{\sqrt{n}} + 4\,\mathscr{V}\log\frac{2en}{\mathscr{V}}\frac{1}{n} \leq 7\sqrt{\mathscr{V}\frac{\log n}{n}},$$

provided that $\mathscr{V}(\log n)/n \leq 1$ and for $n \geq (2e)^3$. Similarly, $\mathbb{E}\,\mathfrak{R}_n(\{\mathbb{1}\{\ell_q(.,\theta) - \ell_k(.,\theta)\geq 0\}, \theta\in\Theta\})$ is bounded by

$$\sqrt{\mathscr{V}_-\log\frac{2en}{\mathscr{V}_-}}\frac{\sqrt{2}}{\sqrt{n}} + 4\,\mathscr{V}_-\log\frac{2en}{\mathscr{V}_-}\frac{1}{n} \leq 7\sqrt{\mathscr{V}_-\frac{\log n}{n}},$$

$\mathscr{V}_-(\log n)/n \leq 1$ and for $n \geq (2e)^3$. Combining this with what is above entails

$$\begin{aligned}\mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_1}) &\leq 7Q\sqrt{\mathscr{V}\frac{\log n}{n}} \\ \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{L_0}) &\leq 14Q\sqrt{\mathscr{V}\frac{\log n}{n}} \\ \mathbb{E}\,\mathfrak{R}_n(\mathscr{F}_{M_0}) &\leq 14Q\sqrt{\mathscr{V}\frac{\log n}{n}} + 7Q^2\sqrt{\mathscr{V}_-\frac{\log n}{n}}.\end{aligned}$$

In particular, all expectations are upper-bounded by $c\sqrt{\frac{\log n}{n}}$, which leads to the result. □

Lemma 40. *If $\mathcal{F}$ is a class of indicator functions and $\mathcal{G}$ is a class of functions from $\mathbb{R}^d$ to $[0,1]$, we have*

$$\begin{aligned}\mathbb{E}\,\mathfrak{R}_n(\mathcal{F}\cdot\mathcal{G}) &\leq \mathbb{E}\,\mathfrak{R}_n(\mathcal{F}) + \mathbb{E}\,\mathfrak{R}_n(\mathcal{G}) \\ \mathbb{E}\,\mathfrak{R}_n(\max(\mathcal{F},\mathcal{G})) &\leq \mathbb{E}\,\mathfrak{R}_n(\mathcal{F}) + \mathbb{E}\,\mathfrak{R}_n(\mathcal{G}),\end{aligned}$$

where we denoted $\mathcal{F}\cdot\mathcal{G} = \{fg, f\in\mathcal{F}, g\in\mathcal{G}\}$ and $\max(\mathcal{F},\mathcal{G}) = \{f\vee g, f\in\mathcal{F}, g\in\mathcal{G}\}$.

Proof. We have

$$\begin{aligned}\mathbb{E}\,\mathfrak{R}_n(\mathcal{F}\cdot\mathcal{G}) &= \mathbb{E}\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i f.g(X_i)\right|\right)\\ &\le \mathbb{E}\left(\sup_{f\in\mathcal{F},g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i (f(X_i)+g(X_i))\right|\right)\\ &\le \mathfrak{R}_n(\mathcal{F}+\mathcal{G}),\end{aligned}$$

because $fg = (f+g-1)_+ = 0.5(f+g-1+|f+g-1|)$ and by applying the contraction lemma of Talagrand (see e.g. Lemma 5.7 in Mohri et al. (2012)) with $x \mapsto 0.5(x-1+|x-1|)$ which is 1-Lipchitz. Then we conclude by using the triangular inequality. For the max we use $\max(f,g) = 0.5(f+g+|f-g|)$. □

Lemma 41. *Consider the case where $Q=2$ and $\{F_u, u\in\mathcal{U}\}$ is an exponential family, i.e. there exists some functions A,B,C,D such that $f(x,u) = \exp\big(A(u)^tB(x) - C(u) + D(x)\big)$. Let k be the dimension of the sufficient statistic vector $B(x)$. If $k\ge 3$, then $\mathscr{V},\mathscr{V}_-$ defined by* (3.16), (3.17) *satisfy $\mathscr{V},\mathscr{V}_- \le Qk(k+1)\,[3\log(k(k+1)) + 2(Q-1)]$. In addition, this bound still holds for $\mathscr{V}_-$ in the case $Q\ge 3$.*

Proof. Let us first bound $\mathscr{V}$. Given that, for $Q=2$, $\theta=(\pi_1,\pi_2,\phi_1,\phi_2)$, $\ell_1(x,\theta)\ge t$ is equivalent to $\pi_1 f(x,\phi_1)/\pi_2 f(x,\phi_2) \ge g(t)$ for some function g, we get that $\ell_1(x,\theta)\ge t$ if and only if $a(\theta)^tB(x) - b(\theta) \ge h(t)$ for some functions a,b,h. The set family is a subset of $\{\{x\in\mathbb{R}^d, a^tB(x)+b\ge 0\}, a\in\mathbb{R}^k, b\in\mathbb{R}\}$, whose VC dimension is bounded by $k(k+1)\,[3\log(k(k+1))+2]$ for $k\ge 3$, see Lemma 10.3 in Shalev-Shwartz and Ben-David (2014). By symmetry, this bound also holds for the VC dimension of $\{\ell_2(\cdot,\theta),\theta\in\Theta\}$. It follows that $\mathscr{V} \le Qk(k+1)\,[3\log(k(k+1))+2] + 2(Q-1)$ (see, e.g., Exercice 3.24 in Mohri et al. (2012) on the VC dimension of the union of two classes with bounded VC dimension).

For $\mathscr{V}_-$, we have that for any $q\ne q'\in\{1,\dots,Q\}$, $\ell_q(x,\theta) - \ell_{q'}(x,\theta)\ge 0$ is equivalent to $\pi_q f(x,\phi_q)/\pi_{q'} f(x,\phi_{q'}) \ge 1$. The rest of the proof follows similarly as for $\mathscr{V}$. □

Lemma 42 (Talagrand's inequality, Theorem 5.3. in Massart (2007))**.** *Let $\xi_1,\dots,\xi_n$ independent r.v., $\mathscr{F}$ a countable class of measurable functions s.t. $a\le f\le b$ for every $f\in\mathscr{F}$ for some real numbers $a\le b$, and $W = \sup_{f\in\mathscr{F}}|\sum_{i=1}^n f(\xi_i) - \mathbb{E}(f(\xi_i))|$. Then, for any $x>0$,*

$$\mathbb{P}(W - \mathbb{E}(W)\ge x) \le e^{-\frac{2x^2}{n(b-a)^2}}.$$

Lemma 43 (Rademacher complexity bound, see, e.g., Lemma 1 in Baraud (2016))**.** *In the setting of Lemma 42 (and with the notation therein), we have*

$$\mathbb{E}(W) \le 2\,\mathfrak{R}_n(\mathscr{F}),$$

where

$$\mathfrak{R}_n(\mathscr{F}) = \sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(\xi_i)\right| \tag{B.30}$$

is the Rademacher complexity of the class $\mathscr{F}$ (with $\varepsilon_1,\dots,\varepsilon_n$ being i.i.d. random signs).

B.4 Auxiliary results for the Gaussian case

B.4.1 Convergence rate for parameter estimation

The following result presents two situations where the parameter of a Gaussian mixture model can be consistently estimated, with an explicit rate.

Proposition 44. *Consider the mixture model (Section 3.2.1) in the d-multivariate Gaussian case with true parameter $\theta^* = (\pi^*, \phi^*)$, where $\phi_q^* = (\mu_q^*, \Sigma_q^*)$, $1 \leq q \leq Q$. Then $\eta(\epsilon, \theta^*)$ defined by (3.18) is such that $\eta(\epsilon_n, \theta^*) \leq 1/n$ for $\epsilon_n \geq C\sqrt{\log n/n}$, where $C > 0$ is a sufficiently large constant, in two following situations:*

(i) $\hat{\theta}$ is the constrained MLE, that is, computed for $\phi_q = (\mu_q, \Sigma_q) \in \mathcal{U}$ with constrained parameter space $\mathcal{U} = [-a_n, a_n]^d \times \{\Sigma \in S_d^{++}, \underline{\lambda} \leq \lambda_1(\Sigma) \leq \lambda_d(\Sigma) \leq \bar{\lambda}\}$[1] where $a_n \leq L(\log n)^\gamma$ for some $L, \gamma > 0$ and S_d^{++} denotes the space of positive definite matrices, with $\underline{\lambda}, \bar{\lambda} > 0$. In that case, C only depends on θ^* and $L, \gamma, \underline{\lambda}, \bar{\lambda}$.

(ii) $\hat{\theta}$ is the estimator coming from EM algorithm (when the iteration number is infinite) for an initialization $\mu_1^{(0)}, \mu_2^{(0)}$ such that $\|(\mu_1^{(0)} - \mu_2^{(0)}) - (\mu_1 - \mu_2)\| \leq \Delta/4$, where $\Delta = \|\mu_1 - \mu_2\|_2$ is the separation between the true means. Here, we consider an homoscedastic model with $\Sigma_1 = \Sigma_2 = \Sigma = \nu I_d$ with known ν. The conclusion applies if the signal-to-noise ratio Δ/ν is large enough, and for a constant C of the form $c(\nu, \Delta)\sqrt{d}$.

Proof. Since case (ii) is a direct application of Balakrishnan et al. (2017), we focus in what follows on proving case (i), by revisiting the result of Ho and Nguyen (2016). First, in the considered model, any mixture can be defined in terms of $\{f_u, u \in \mathcal{U}\}$ and a discrete mixing measure $G = \sum_{q=1}^Q \pi_q \delta_{\phi_q}$ with Q support points, as $\sum_{q=1}^Q \pi_q f_{\phi_q} = \int f_u(x) dG(u)$. As shown by Ho and Nguyen (2016), the convergence of mixture model parameters can be measured in terms of a Wasserstein distance on the space of mixing measures. Let $G_1 = \sum_{q=1}^Q \pi_q^1 \delta_{\phi_q^1}$ and $G_2 = \sum_{q=1}^Q \pi_q^2 \delta_{\phi_q^2}$ be two discrete probability measures on some parameter space, which is equipped with metric $\|.\|$. The Wasserstein distance of order 1 between G_1 and G_2 is given by

$$W_1(G_1, G_2) = \inf_p \sum_{q,l} p_{q,l} \|\phi_q^1 - \phi_l^2\|$$

where the infimum is over all couplings $(p_{q,l})_{1 \leq q,l \leq Q} \in [0,1]^{Q \times Q}$ such that $\sum_l p_{q,l} = \pi_q^1$ and $\sum_q p_{q,l} = \pi_l^2$. Let $G^*, \hat{G}_n$ denote the true mixing measure and the mixing measure that corresponds to the restricted MLE considered here, respectively. Theorem 4.2. in Ho and Nguyen (2016) implies that, with the notation of Ho and Nguyen (2016), for any $\epsilon_n \geq (\sqrt{C_1}/c)\delta_n$, and $\delta_n \leq C\sqrt{\log n/n}$, we have $\mathbb{P}_{\theta^*}(W_1(\hat{G}_n, G^*) \geq (c/C_1)\epsilon_n) \leq ce^{-n\epsilon_n^2}$. We apply this relation for $\epsilon_n = \max((\sqrt{C_1}/c)\delta_n, \sqrt{\log(cn)/n})$. In that case, we have still ϵ_n of order $\sqrt{\log n/n}$ and the upper-bound is at most $1/n$. On the other hand, if we have a convergence rate in terms of W_1, then we have convergence of the mixture model parameters in terms of $\|.\|$ at the same rate, see Lemma 45. This concludes the proof. □

Lemma 45. *Let $G_n = \sum_{q=1}^Q \pi_q^n \delta_{\phi_q^n}$ be a sequence of discrete probability measures on $\mathcal{U}$, and let G^*, W_1 be defined as in the proof of Proposition 44. There exists a constant C only depending on G^* such that if $W_1(G_n, G^*) \to 0$, then for sufficiently large n,*

$$W_1(G_n, G^*) \geq C \min_{\sigma \in [Q]} \|\theta_n^\sigma - \theta^*\|.$$

[1] Here, $\lambda_1(\Sigma)$ (resp. $\lambda_d(\Sigma)$) denotes the smallest (resp. largest) eigenvalue of Σ.

Proof. In what follows, we let $\{p^n_{q,l}\}$ denote the corresponding probabilities of the optimal coupling for the pair (G_n, G^*). We start by showing that $(\phi^n_q)_q \to (\phi^*_q)_q$ in $\|.\|$ up to a permutation of the labels. Let σ^n the permutation of the labels such that $\|\phi^n_q - \phi^*_l\| \geq \|\phi^n_{\sigma^n(l)} - \phi^*_l\|$ for all $q, l \in \{1, ..., Q\}$. Then, by definition,

$$\begin{aligned} W_1(G_n, G^*) &\geq \sum_{1\leq q,l\leq Q} p^n_{q,l} \|\phi^n_{\sigma(l)} - \phi^*_l\| \\ &= \sum_l \pi^*_l \|\phi^n_{\sigma^n(l)} - \phi^*_l\|. \end{aligned}$$

It follows that each $\|\phi^n_{\sigma^n(l)} - \phi^*_l\|$ must converge to zero. Since $(\phi^n_q)_q \to (\phi^*_q)_q$ up to a permutation of the labels, without loss of generality we can assume that $\phi^n_q \to \phi^*_q$ for all q. Let $\Delta\phi^n_q := \phi^n_q - \phi^*_q$ and $\Delta\pi^n_q := \pi^n_q - \pi^*_q$. Write $W_1(G_n, G^*)$ as

$$W_1(G_n, G^*) = \sum_q p^n_{qq} \|\Delta\phi^n_q\| + \sum_{q\neq l} p^n_{ql} \|\phi^n_q - \phi^*_l\|$$

Define $C_{ql} = \|\phi^*_q - \phi^*_l\|$ and $C = \min_{q\neq l} C_{ql} > 0$. It follows from the convergence of ϕ^n that for $q \neq l$, $\|\phi^n_q - \phi^*_l\| \geq C/2$ for sufficiently large n. Thus,

$$W_1(G_n, G^*) \geq \frac{C}{2} \sum_{q\neq l} p^n_{ql}$$

We deduce that $\sum_{q\neq l} p^n_{ql} \to 0$. As a result, $p^n_{qq} = \pi^*_q - \sum_{l\neq q} p^n_{lq} \to \pi^*_q$, and so, $p^n_{qq} \geq (1/2)\pi^*_{\min} := \min_l \pi^*_l$ for sufficiently large n. On the other hand, $\sum_{q\neq l} p^n_{ql} = \sum_q \pi^n_q - p^n_{qq} = \sum_q \pi^*_q - p^n_{qq}$ where $p^n_{qq} \leq \min(\pi^n_q, \pi^*_q)$. Thus, $\sum_{q\neq l} p^n_{ql} \geq \sum_q \pi^n_q - \min(\pi^n_q, \pi^*_q) = \sum_{q,\pi^n_q \geq \pi^*_q} \pi^n_q - \pi^*_q = \sum_{q,\pi^n_q\geq\pi^*_q} |\pi^n_q - \pi^*_q|$ and similarly we have that $\sum_{q\neq l} p^n_{ql} \geq \sum_{q,\pi^*_q\geq\pi^n_q} |\pi^n_q - \pi^*_q|$. It follows that $2\sum_{q\neq l} p^n_{ql} \geq \sum_q |\pi^n_q - \pi^*_q|$. Therefore, for sufficiently large n,

$$W_1(G_n, G^*) \geq \frac{1}{2}\pi^*_{\min} \sum_q \|\Delta\phi^n_q\| + \frac{C}{4} \sum_q |\Delta\pi^n_q|.$$

This gives the result. □

B.4.2 Gaussian computations

The following lemma holds.

Lemma 46. *Let us consider the multivariate Gaussian case where $\phi_q = (\mu_q, \Sigma_q)$, $1 \leq q \leq Q$, with $Q = 2$, $\Sigma_1 = \Sigma_2$ is an invertible covariance matrix and μ_1 and μ_2 are two different vectors of $\mathbb{R}^d$. Then Assumptions 6, 7 and 8 hold true for $\alpha_c = 0$ and for a level $\alpha \in (0, \bar{\alpha})\backslash\mathcal{E}$ for $\mathcal{E}$ a set of Lebesgue measure 0.*

Proof. Let us first prove that $\ell_q(X, \theta)$ is a continuous random variable under $\mathbb{P}_{\theta^*}$ (this is established below without assuming $\Sigma_1 = \Sigma_2$ for the sake of generality). We have

$$\begin{aligned} &\mathbb{P}_{\theta^*}(\ell_1(X, \theta) = t) \\ &= \mathbb{P}_{\theta^*}(f_{\phi_1}(X)/f_{\phi_2}(X) = t\pi_2/\pi_1) \\ &= \mathbb{P}_{\theta^*}\left((X-\mu_1)^t\Sigma_1^{-1}(X-\mu_1) - (X-\mu_2)^t\Sigma_2^{-1}(X-\mu_2) = -2\log(t\pi_2/\pi_1) - \log(|\Sigma_1|/|\Sigma_2|)\right). \end{aligned}$$

Now,

$$\begin{aligned}
&(X-\mu_1)^t\Sigma_1^{-1}(X-\mu_1) - (X-\mu_2)^t\Sigma_2^{-1}(X-\mu_2)\\
&= (X-\mu_1)^t\Sigma_1^{-1}(X-\mu_1) - (X-\mu_1)^t\Sigma_2^{-1}(X-\mu_2) - (\mu_1-\mu_2)^t\Sigma_2^{-1}(X-\mu_2)\\
&= (X-\mu_1)^t(\Sigma_1^{-1}-\Sigma_2^{-1})(X-\mu_1) - (X-\mu_1)^t\Sigma_2^{-1}(\mu_1-\mu_2) - (\mu_1-\mu_2)^t\Sigma_2^{-1}(X-\mu_2)\\
&= (X-\mu_1)^t(\Sigma_1^{-1}-\Sigma_2^{-1})(X-\mu_1) - (\mu_1-\mu_2)^t\Sigma_2^{-1}(2X-\mu_2-\mu_1).
\end{aligned}$$

Since the real matrix $\Sigma_1^{-1}-\Sigma_2^{-1}$ is symmetric, we can diagonalize it and we end up with a subset of $\mathbb{R}^d$ of the form

$$\left\{ y \in \mathbb{R}^d \,:\, \sum_{j=1}^{d} \left(\alpha_j y_j^2 + \beta_j y_j\right) + \gamma = 0 \right\},$$

for some real parameters $\alpha_j, \beta_j, \gamma$. The result follows because this set has a Lebesgue measure equal to 0 in any case.

Now, since $\Sigma_1 = \Sigma_2 = \Sigma$, we have for all $t \in (0,1)$,

$$\begin{aligned}
\{T(X,\theta) > t\} &= \left\{ \forall q \in \{1,\dots,Q\}, \pi_q f_{\phi_q}(X) < (1-t)\sum_{\ell=1}^{Q} \pi_\ell f_{\phi_\ell}(X) \right\}\\
&= \{\pi_1 f_{\phi_1}(X) < (1/t-1)\pi_2 f_{\phi_2}(X)\} \cap \{\pi_2 f_{\phi_2}(X) < (1/t-1)\pi_1 f_{\phi_1}(X)\}\\
&= \left\{ (1/t-1)^{-1} < \frac{\pi_1 f_{\phi_1}(X)}{\pi_2 f_{\phi_2}(X)} < (1/t-1) \right\}.
\end{aligned}$$

Applying $2\log(\cdot)$ on each part of the relation, we obtain

$$\{T(X,\theta) > t\} = \left\{ -2\log(1/t-1) < a^t X + b < 2\log(1/t-1) \right\},$$

for

$$\begin{aligned}
a = a(\theta) &= 2\Sigma^{-1}(\mu_1-\mu_2) \in \mathbb{R}^d \backslash \{0\}\\
b = b(\theta) &= -(\mu_1-\mu_2)^t\Sigma^{-1}(\mu_1+\mu_2) + 2\log(\pi_1/\pi_2) \in \mathbb{R}^d.
\end{aligned}$$

Since under P_{θ^*} we have $X \sim \pi_1^*\mathcal{N}(\mu_1^*, \Sigma^*) + \pi_2^*\mathcal{N}(\mu_2^*, \Sigma^*)$, we have $a^tX + b \sim \pi_1^*\mathcal{N}(a^t\mu_1^* + b, a^t\Sigma^* a) + \pi_2^*\mathcal{N}(a^t\mu_2^* + b, a^t\Sigma^* a)$. This yields for all $t \in (0,1)$,

$$\begin{aligned}
\mathbb{P}_{\theta^*}(T(X,\theta) > t) =& \pi_1 \left[\Phi\left(\frac{2\log(1/t-1) - a^t\mu_1^* - b}{(a^t\Sigma^* a)^{1/2}} \right) \right.\\
&\left. - \Phi\left(\frac{-2\log(1/t-1) - a^t\mu_1^* - b}{(a^t\Sigma^* a)^{1/2}} \right) \right]\\
&+ \pi_2 \left[\Phi\left(\frac{2\log(1/t-1) - a^t\mu_2^* - b}{(a^t\Sigma^* a)^{1/2}} \right) \right.\\
&\left. - \Phi\left(\frac{-2\log(1/t-1) - a^t\mu_2^* - b}{(a^t\Sigma^* a)^{1/2}} \right) \right]. \qquad \text{(B.31)}
\end{aligned}$$

A direct consequence is that for all $t \in (0,1)$, we have $\mathbb{P}_{\theta^*}(T(X,\theta) > t) < 1$, that is, $\mathbb{P}_{\theta^*}(T(X,\theta) \le t) = \mathbb{P}_{\theta^*}(T(X,\theta) < t) > 0$. Hence, α_c defined in (3.14) is equal to zero. Moreover, from (B.31), we clearly have that $t \in (0,1) \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) > t)$ is decreasing, so that $t \in (0,1) \mapsto \mathbb{P}_{\theta^*}(T(X,\theta) \le t)$ is increasing. This proves that Assumption 6 holds in that

case.

Let us now check Assumptions 7 and 8. Assumptions 7 and 8 (i) follow from Result 2.1 in Melnykov (2013).

As for Assumption 8 (ii), from (B.31), we only have to show that the function $t \in (0,1) \mapsto \frac{\partial}{\partial t}\Phi\left(\frac{\log(1/t-1)-\alpha^*}{\beta^*}\right)$ is uniformly bounded by some constant $C = C(\alpha^*, \beta^*)$, for any $\alpha^* \in \mathbb{R}$ and $\beta^* > 0$. A straightforward calculation leads to the following: for all $t \in (0,1)$,

$$\left|\frac{\partial}{\partial t}\Phi\left(\frac{\log(1/t-1)-\alpha^*}{\beta^*}\right)\right| = \frac{e^{-(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2/2}}{\beta^*\sqrt{2\pi}}\frac{1}{t(1-t)}. \tag{B.32}$$

Consider now $t_0 = t_0(\alpha^*, \beta^*) \in (0, 1/2)$ such that $(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2 \geq 2\log(1/t)$ for all $t \in (0, t_0)$. It is clear that the right-hand-side of (B.32) is upper-bounded by $\frac{1}{\beta^*\sqrt{2\pi}(1-t_0)}$ on $t \in (0, t_0)$. Similarly, let $t_1 = t_1(\alpha^*, \beta^*) \in (1/2, 1)$ such that $(\frac{\log(1/t-1)-\alpha^*}{\beta^*})^2 \geq 2\log(1/(1-t))$ for all $t \in (t_1, 1)$. It is clear that the right-hand-side of (B.32) is upper-bounded by $\frac{1}{\beta^*\sqrt{2\pi}t_1}$ on $t \in (t_1, 1)$. Finally, for $t \in [t_0, t_1]$, the upper-bound $\frac{1}{\beta^*\sqrt{2\pi}t_0(1-t_1)}$ is valid. This proves that Assumption 8 (ii) holds.

Let us now finally turn to Assumption 8 (iii). Lemma 36 ensures that $t \in (0, t^*(\bar{\alpha})) \mapsto \mathrm{mFMR}_t^*$ is continuous increasing. Hence, $t^* : \beta \in (0, \bar{\alpha}) \mapsto t^*(\beta)$ defined in (3.13) is the inverse of this function and is also continuous increasing. It is therefore differentiable almost everywhere in $(0, \bar{\alpha})$, so everywhere in $(0, \bar{\alpha})\backslash\mathcal{E}$ where $\mathcal{E}$ is a set of Lebesgue measure 0. By taking α in $(0, \bar{\alpha})\backslash\mathcal{E}$, this ensures that t^* is differentiable in α and thus that Assumption 8 (iii) holds. □

Lemma 47. *In the multivariate gaussian case with $Q = 2$ and $\Sigma_1 = \Sigma_2$, we have that $\mathscr{V} \leq 2d + 4$ and $\mathscr{V}_- \leq 2d + 4$.*

Proof. In that case, we have that (see the proof of Lemma 46)

$$\{\ell_q(x, \theta) \leq u, x \in \mathbb{R}^d\} = \{a_\theta^t x + b_\theta \geq g(u), x \in \mathbb{R}^d\}.$$

Since the VC dimension of the vector space of real-valued affine functions is bounded by $d+1$ (see, e.g., Exercice 3.19 in Mohri et al. (2012)). We obtain the result by applying the usual bound on the VC dimension of the union of two classes with bounded VC dimension (see, e.g., Exercice 3.24 in Mohri et al. (2012)). □

Appendix C

Supplementary material of Chapter 4

C.1 Additional experimental details

Datasets `Cora` (Sen et al., 2008) is a citation network. Each publication in the dataset is described by a 0/1-valued vector indicating the absence/presence of a each word of a dictionary that consists of 1433 unique words. `Yeast` (Von Mering et al., 2002) is a protein-protein interaction network of *S. cerevisiae*. `T. Albus` (Weber Zendrera et al., 2021) is a metabolic network of *Thermocrinis Albus*. `C. ele` (Watts and Strogatz, 1998) is a neural network of *C. elegans*. `Florida Food web` (Christian and Luczkovich, 1999) is a food web network downloaded from the Web of Life Repository (`https://www.web-of-life.es/`).

Hyper-parameters for SEAL SEAL (Zhang and Chen, 2018) is used with a hop number of 2, for the GNN we use GIN (Xu et al., 2018) with 3 layers and 32 neurons, and we train for 10 epochs with a learning rate of $0,001$.

C.2 Exchangeability condition for FDR control

Recent works (Bates et al., 2023; Mary and Roquain, 2022; Marandon et al., 2022) proved finite-sample FDR control guarantees for the use of conformal p-values in the context of novelty detection. Specifically, given a data sample $Z_1, \dots, Z_{n+m}$ where $Z_1, \dots, Z_n$ are marginally distributed as some unknown P_0, the aim is to test the null hypotheses $H_{0j} : Z_{n+j} \sim P_0$ simultaneously for all $1 \leq j \leq m$. Consider $\mathcal{H}_0 \subset \{1, \dots, m\}$ the set of nominals and $\mathcal{H}_1 \subset \{1, \dots, m\}$ the set of novelties in the test sample, and for $j \in \{1, \dots, m\}$ the score $S_j = \hat{g}(Z_j)$ with $\hat{g}$ is some real-valued function that may depend on the data. We have the following result (Weinstein et al., 2017; Mary and Roquain, 2022; Marandon et al., 2022): if the null scores $(S_i)_{i \in \{1,\dots,n\} \cup \mathcal{H}_0}$ are exchangeable i.e. if, for all permutation π of $\{1, \dots, n+m\}$ that lets invariant $\mathcal{H}_1$, we have that $(S_{\pi(i)})_{1 \leq i \leq n+m} \sim (S_i)_{1 \leq i \leq n+m}$, then the Counting Knock-off algorithm (Weinstein et al., 2017) (i.e. Algorithm 6 with $\mathcal{D}_{\text{test}} = \{n+1, \dots, n+m\}$ and $\mathcal{D}_{\text{cal}} = \{1, \dots, n\}$) controls the FDR at level α.

In our set-up, we have that $\mathcal{H}_1$, $\mathcal{H}_0$, $\mathcal{D}_{\text{test}} = \mathcal{H}_0 \cup \mathcal{H}_1$ and $\mathcal{D}_{\text{cal}}$ are all random. In that case, the afore-mentioned exchangeability condition can be extended to the following: conditionally on $(\mathcal{D}_{\text{cal}}, \mathcal{H}_0, \mathcal{H}_1)$, for $\pi \sim \text{Unif}[\Pi]$ where Π is the set of permutations of $\mathcal{D}_{\text{cal}} \cup \mathcal{D}_{\text{test}}$ that lets invariant the set $\mathcal{H}_1$,

$$(S_{\pi(i,j)})_{1 \leq i,j \leq n} \sim (S_{i,j})_{1 \leq i,j \leq n}. \tag{C.1}$$

However this is not a standard exchangeability assumption, because the permutation π in (C.1) is random due to $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$ being random in our set up.