Some Contributions to Modern Multiple Hypothesis Testing in High-dimension

Binh Tuan Nguyen

université PARIS-SACLAY

Some Contributions to Modern Multiple Hypothesis Testing in High-dimension

Quelques contributions aux tests d'hypothèses multiples en grande dimension

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574 Mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques. Référent : Faculté des sciences d'Orsay

Thèse préparée au **Laboratoire de mathématiques d'Orsay** (Université Paris-Saclay, CNRS) et à l'**Inria Saclay-Île-de-France** (Université Paris-Saclay, Inria), sous la direction de **Sylvain ARLOT**, Professeur, et la co-direction de **Bertrand THIRION**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 10 Décembre 2021, par

Binh Tuan NGUYEN

Composition du jury

Christophe GIRAUD Professeur, Université Paris-Saclay	Président
Etienne ROQUAIN Maître de Conférences (HDR), Sorbonne Université	Rapporteur & Examinateur
Jelle GOEMAN Professeur, Leiden University	Rapporteur & Examinateur
Jeanette MUMFORD Maître de Conférences, Stanford University	Examinatrice
Claire BOYER Maître de Conférences, Sorbonne Université	Examinatrice
Sylvain ARLOT Professeur, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT
NNT : 2021UPASM047

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Titre : Quelques contributions aux tests d'hypothèses multiples en grande dimension
Mots clés : inférence statistique, inférence multivariée, tests multiples,

Résumé : Cette thèse traite des problèmes de tests multiples en grande dimension, un régime qui est devenu populaire dans l'inférence statistique moderne. Son objectif principal est de fournir des algorithmes efficaces et fiables pour l'inférence multivariée, un problème difficile qui souffre du fléau de la dimension. Nos solutions améliorent les méthodes de l'état de l'art, les rendent plus stables et efficaces tout en conservant leurs garanties théoriques sur le contrôle des métriques de tests multiples. De plus, nous montrons que nos contributions sont raisonnablement performantes par rapport à l'état de l'art dans des applications concrètes, à savoir des problèmes issus des sciences de la vie, comme les neurosciences, l'imagerie médicale et la génomique. En particulier, nous étudions les propriétés des filtres "knockoff", une méthode de contrôle du taux de fausses découvertes (False Discovery Rate – FDR), qui nécessite peu d'hypothèses sur la loi des données. Nous proposons ensuite des méthodes d'agrégation de plusieurs échantillonnages pour traiter le caractère aléatoire du filtre knockoff, et prouvons des résultats théoriques non asymptotiques sur le knockoff agrégé, en particulier une garantie de contrôle du FDR, qui repose sur certaines inégalités de concentration. En outre, nous étendons la méthode, en fournissant une version qui peut s'adapter à un régime de dimension extrêmement élevée. L'une des étapes clés est l'utilisation d'un regroupement (clustering) aléatoire des covariables, afin d'éviter le fléau de la dimension, puis l'assemblage de plusieurs exécutions afin de limiter les biais causés par l'utilisation d'un seul regroupement. Afin de prendre en compte la compression des données qui résulte de l'étape de clustering, nous introduisons une relaxation spatiale du taux de fausses découvertes. Enfin, nous considérons le problème de construction de p-valeurs pour l'inférence conditionnelle avec la régression logistique en grande dimension. Cette méthode est une variante du test de randomisation conditionnel, avec un schéma de décorrélation supplémentaire qui donne des statistiques de test plus précises et plus puissantes que les estimateurs précédents. Nous concluons la thèse par une discussion sur certaines questions ouvertes, qui nous semblent importantes et peuvent servir de directions de travail pour améliorer les méthodes d'inférence en grande dimension et leur application à des domaines tels que la génomique ou l'imagerie cérébrale.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Title : Some Contributions to Modern Multiple Hypothesis Testing in High-dimension
Keywords : statistical inference, multivariate inference, multiple testing

Abstract : This thesis deals with multiple testing problems in high-dimension, a regime which has become increasingly common nowadays in statistical inference. Its main goal is to provide efficient and reliable algorithms for multivariate inference, a hard problem that suffers from the curse of dimensionality. Our solutions improve on state-of-the-art methods, make them more stable and efficient while still maintaining their theoretical guarantees on controlling multiple testing metrics. Moreover, we show that our contributions perform reasonably well compared to the state-of-the-art in practical applications, considering analysis problems from life-sciences, such as neuroscience, medical imaging and genomics. In particular, we study the properties of knockoff filters, a method for controlling False Discovery Rate (FDR) that requires limited distribution assumptions. We then propose methods for aggregating several samplings to address knockoff filter's randomness, and prove non-asymptotic theoretical results on the aggregated knockoff, specifically guaranteed FDR control, which relies on some concentration inequalities. Furthermore, we extend the method, providing a version that can scale to extremely high dimensional regime. One of the key steps is to use randomized clustering to reduce the dimension to avoid the curse of dimensionality, and then to ensemble several runs to tame the bias from the selection of a fixed clustering. In order to take into account the compression of data that results from the clustering step, we introduce a spatial relaxation of the False Discovery Rate. Finally, we consider the problem of outputting p-values in the conditional inference for high-dimensional logistic regression. This method is a variant of the Conditional Randomization Test, with an additional decorrelation scheme that yields more accurate test statistics and more powerful than previous estimators. We conclude the thesis with a discussion about some open questions, which we believe are important and can serve as potential directions to further improve high-dimensional inference methods and their application to fields such as genomics or brain imaging.

Contents

Abstract

This thesis deals with multiple testing problems in high-dimension, a regime which has become increasingly common nowadays in statistical inference. Its main goal is to provide efficient and reliable algorithms for multivariate inference, a hard problem that suffers from the curse of dimensionality. Our solutions improve on state-of-the-art methods, make them more stable and efficient while still maintaining their theoretical guarantees on controlling multiple testing metrics. Moreover, we show that our contributions perform reasonably well compared to the state-of-the-art in practical applications, considering analysis problems from life-sciences, such as neuroscience, medical imaging and genomics. In particular, we study the properties of knockoff filters, a method for controlling False Discovery Rate (FDR) that requires limited distribution assumptions. We then propose methods for aggregating several samplings to address knockoff filter's randomness, and prove non-asymptotic theoretical results on the aggregated knockoff, specifically guaranteed FDR control, which relies on some concentration inequalities. Furthermore, we extend the method, providing a version that can scale to extremely high dimensional regime. One of the key steps is to use randomized clustering to reduce the dimension to avoid the curse of dimensionality, and then to ensemble several runs to tame the bias from the selection of a fixed clustering. In order to take into account the compression of data that results from the clustering step, we introduce a spatial relaxation of the False Discovery Rate. Finally, we consider the problem of outputting p-values in the conditional inference for high-dimensional logistic regression. This method is a variant of the Conditional Randomization Test, with an additional decorrelation scheme that yields more accurate test statistics and more powerful than previous estimators. We conclude the thesis with a discussion about some open questions, which we believe are important and can serve as potential directions to further improve high-dimensional inference methods and their application to fields such as genomics or brain imaging.

Résumé

Cette thèse traite des problèmes de tests multiples en grande dimension, un régime qui est devenu populaire dans l'inférence statistique moderne. Son objectif principal est de fournir des algorithmes efficaces et fiables pour l'inférence multivariée, un problème difficile qui souffre du fléau de la dimension. Nos solutions améliorent les méthodes de l'état de l'art, les rendent plus stables et efficaces tout en conservant leurs garanties théoriques sur le contrôle des métriques de tests multiples. De plus, nous montrons que nos contributions sont raisonnablement performantes par rapport à l'état de l'art dans des applications concrètes, à savoir des problèmes issus des sciences de la vie, comme les neurosciences, l'imagerie médicale et la génomique. En particulier, nous étudions les propriétés des filtres "knockoff", une méthode de contrôle du taux de fausses découvertes (False Discovery Rate – FDR), qui nécessite peu d'hypothèses sur la loi des données. Nous proposons ensuite des méthodes d'agrégation de plusieurs échantillonnages pour traiter le caractère aléatoire du filtre knockoff, et prouvons des résultats théoriques non asymptotiques sur le knockoff agrégé, en particulier une garantie de contrôle du FDR, qui repose sur certaines inégalités de concentration. En outre, nous étendons la méthode, en fournissant une version qui peut s'adapter à un régime de dimension extrêmement élevée. L'une des étapes clés est l'utilisation d'un regroupement (clustering) aléatoire des covariables, afin d'éviter le fléau de la dimension, puis l'assemblage de plusieurs exécutions afin de limiter les biais causés par l'utilisation d'un seul regroupement. Afin de prendre en compte la compression des données qui résulte de l'étape de clustering, nous introduisons une relaxation spatiale du taux de fausses découvertes. Enfin, nous considérons le problème de construction de p-valeurs pour l'inférence conditionnelle avec la régression logistique en grande dimension. Cette méthode est une variante du test de randomisation conditionnel, avec un schéma de décorrélation supplémentaire qui donne des statistiques de test plus précises et plus puissantes que les estimateurs précédents. Nous concluons la thèse par une discussion sur certaines questions ouvertes, qui nous semblent importantes et peuvent servir de directions de travail pour améliorer les méthodes d'inférence en grande dimension et leur application à des domaines tels que la génomique ou l'imagerie cérébrale.

Acknowledgements

Writing a doctoral thesis is by no means an easy task, and in the process of doing so, I have been incredibly fortunate to know and work with an enormous amount of talented people.

First and foremost, I would like to thank my two supervisors, Bertrand and Sylvain. Bertrand – thank you for giving me a chance to work with you, to start everything written in this document: from the day my email randomly appeared in your inbox, asking for a research internship, until the day of my thesis defense. You are always supportive, yet also efficient, a skill I always admire and want to be better at. Sylvain – thank you for your constant encouragement, your scientific thoroughness, and your huge collection of knowledge on mathematical statistics. You are a role model for me, the type of scientist I look forward to become. To both of you – I hope that the end of this doctoral study is just a beginning of our collaboration.

A warm thank you have to go to the jury members of my thesis defense, especially Etienne Roquain and Jelle Goeman, for without their thoughtful comments, the writing of this manuscript would be incomplete.

I spent a large part of my time working on this thesis inside INRIA Parietal team, and as such, would like to thank all the members of the team for our time together. I consider myself to be in a very special intersection of two Parietal's generations, which hardly overlap: each member of one would likely not even know ones from the other. The older generation – I would like to thank Jérôme-Alexis (for our close collaboration and numerous fun times spent together), Hugo (for a friendship I neglected in the beginning, but have been lucky enough to cherish later on), Kamalaker, Jérôme, Thomas Bazeille, Ana-Luisa, Antonia, Pierre, Omar, Arthur, Carole, Loubna, Hamza, Patricio, Maëliss, Zaccharie, Hicham, Valentin, Thomas Moreau, and Marine. The newer generation, by chance (and COVID-19) I am glad to have come to know – I would like to thank two Alexis (Thual and Cvetkov-Iliev), Thomas Chapalain, Raphaël, Joseph, Julia, Charlotte, Apolline, Lilian, Cédric, Benoît, two Matthieu, Bénédicte, Judith, and Léo. Thanks to all the PIs in the team: Alexandre, Gaël, Demian, Denis and Philippe.

Not least importantly, I want to express the deepest gratitude to my parents. I know having me to earn a doctorate degree would be a dream come true for both of you, and I am grateful to make you a very proud parents. Dad – the first person to teach me mathematics, along with various scientific subjects, and logic/strategy through chess. Mom – for always showing your unconditional love, despite us living too far apart is something you do not wish.

Finally, thank you, Hoa and Chi, two of the most important women in my life. Thank you Hoa, for even though not being able to be here, you are always a source of inspiration and constant support. Thank you for having been my best friend for such a long time, and I hope it will remain, no matter what would happen. Thank you Chi, for being my hope and future.

Chapter 1

Overview

This thesis is divided into two parts: the first part introduces necessary background information and motivation for the second part, which mainly deals with new hypothesis testing/variable selection procedures.

In particular, for the background part, we start with an introduction to hypothesis testing in Chapter 2. The reader will be familiarized with the system of notations, problem of multiple testing and the hardness of doing multivariate inference in high-dimension. This relies on a key distinction between conditional and marginal inference. Specifically, we focus on popular metrics of multiple testing, such as Family-Wise Error Rate or False Discovery Rates and procedures to control these metrics. We then move to recent advances in high-dimensional multiple testing, which are knockoff filters and conditional randomization test. We conclude this chapter with aggregation techniques that help stabilize the inference results of multiple testing procedures. Chapter 3 introduces functional Magnetic Resonance Imaging (fMRI), one of the major techniques for neuro-imaging. FMRI data analysis makes it possible to understand how our brain works, and will be one of the main application of the variable selection procedures that this thesis presented. We will briefly review the pipeline for fMRI data-analysis: from acquisition to processing, modeling and finally statistical inference methods on fMRI data.

The contribution part of this thesis is organized around three major directions, which are presented as follows.

1.1 Stabilizing Inference Results Of Knockoff Filter

Knockoff filter [BC15, CFJL18] is one of the recently introduced methods for multiple testing, with the aim to control False Discovery Rate. Despite being relatively new to the literature, this method enjoys a great popularity thanks to its flexibility in model settings and works well in mildly high-dimension settings. However, a main drawback of knockoff filter is its instability: the method's approach relies on sampling of new noisy copies of original variables, therefore the inference results are inherently random. In Chapter 4, we introduce a methodology to fix the this randomness issue. Our algorithm, `AKO`, makes the variable selection more stable by combining multiple runs of knockoff inference using p-values aggregation technique of [MMB09]. We provide a theoretical analysis of `AKO`, most importantly proves the guarantee for controlling FDR under predefined level. Finally, we perform empirical validation of `AKO` on numerical, neuro-imaging and genetics dataset.

Published work Nguyen, T., Chevalier, J., Thirion, B. & Arlot, S.. (2020). Aggregation of Multiple Knockoffs. *In Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. Available from `https://proceedings.mlr.press/v119/nguyen20a.html`.

1.2 Multiple Testing In Extremely High-Dimension: A Combination Of Knockoff Filter And Randomized Clustering

In fMRI data-analysis, the number of variables (brain voxels) can be over hundred of thousands, while only a small number of observations of at most several thousands is available. In such setting, finding explanatory variables that are truly related to the response while controlling the rate of false discoveries is a difficult problem, both in computational and statistical senses [SP20]. In Chapter 5, we introduce an algorithm that combines randomized clustering and knockoff filter. The first step of this algorithm, randomized hierarchical clustering, is crucial: it acts as a dimension reduction step, making the inference task tractable in a lower resolution for the fMRI data. We then perform knockoff filter to select clusters (groups) of brain voxels which are significant, while aiming to control FDR. Again, to make the solution more stable, we perform multiple runs of clustering and knockoff inference, then combine it using p-value aggregation. We called the algorithm Ensemble of Clustered Knockoffs, or `ECKO`. More importantly, to take into account the lower resolution when perform clustering, we introduce a new notion of the FDR with a spatial tolerance $\delta > 0$. In short, non-activated voxels (null variables) selected at a distance closer than δ from activated voxels (significant variables) given a stimuli would not be considered as false positives. We call this metric FDR^{δ}, and introduce the necessary definition of the concept. With various benchmarks on both simulated and realistic neuroimaging dataset, we demonstrate that `ECKO` performs well compared to contemporary methods.

Published work T.-B. Nguyen, J.-A Chevalier, & B. Thirion (2019), `ECKO`: Ensemble of Clustered Knockoffs for Robust Multivariate Inference on fMRI Data. *In International Conference on Information Processing in Medical Imaging (pp. 454-466). Springer, Cham.*

1.3 A Conditional Randomization Test For Logistic Regression In High-Dimension

In Chapter 6, we introduce a conditional independence test procedure that works with non-linear relationships between covariates and responses in high-dimension, in particular for logistic regression. In high-dimension setting, testing for conditional independence is highly non-trivial, as conditioning on a large number of variables is intrinsically difficult [SP20]. Conditional randomization test (CRT) is another recent attempt to tackle this problem, besides knockoff filters. The concept of CRT can be understood as a randomization procedure to sample the empirical distribution of the test statistics for each variable j, in order to test the independence of j with the response *conditionally* to all other variables. It was discussed originally in the knockoff paper [CFJL18] as an alternative procedure, should the statistician wants to output p-values with knockoff variables. Despite that, CRT comes with a prohibitive computational cost, due to the multiple sampling of noisy knockoff variables for each test statistic computed. The recent work of [LKJR20] introduced a distillation operation of CRT, called dCRT, that shows promise in fixing this crucial disadvantage. However, our empirical observations show that dCRT encounters statistical issue when running in non-linear regime, in particular logistic regression. Despite its popularity, there is currently no good solution to allow accurate inference on logistic regression's coefficients in the high-dimensional regime. To tackle this issue, we combine the distillation operation from [LKJR20] with a decorrelation step that adapts this operation to non-linear regime, and call this method `CRT-Logit`. Additionally, we present an algorithm that combines clustering and p-values aggregation for logistic

regression, similar to the idea of `ECKO`. Results from synthetic, neuro-imaging and genomics datasets suggest that `CRT-logit` performs well compared to competitive method.

Preprint T.-B. Nguyen, S. Arlot, & B. Thirion (2021+), Upscaling the Conditional Randomization Test to Extremely High-Dimensional Logistic Regression. *To be submitted.*

1.4 Conclusion

Finally, we conclude the thesis in Chapter 7 with a summary of contributions and a wider perspectives about open questions that can be further addressed.

1.5 Other Works

There are additional preprints/publications that we list as follows, but will not present in this thesis manuscript.

- J.-A. Chevalier, T.-B. Nguyen, B. Thirion J. Salmon, Spatially relaxed inference on high-dimensional linear models (2021+). *In submission.*
- J.-A. Chevalier, T.-B. Nguyen, J. Salmon, G. Varoquaux, B. Thirion, Decoding with confidence: Statistical control on decoder maps. *In NeuroImage, Volume 234, 2021, 117921, ISSN 1053-8119.*

1.6 Software

To foster scientific reproducibility, we have developed an open-source implementation of a part of the procedures presented in this thesis. The software is written in Python and can be found on: https://ja-che.github.io/hidimstat/

Notations & Abbreviations

Notations	
n	number of observations (sample size)
p	number of variables/hypotheses to be tested
$[k]$	set of natural numbers from 1 to k
$\mathcal{H}_0, \mathcal{H}_a$	the null and alternative hypotheses to be tested
$\mathcal{H}_0^1, \ldots, \mathcal{H}_0^p$	family of p hypotheses to be tested
α	significance level used in hypothesis testing
$x, \mathbf{x}, \mathbf{X}, \mathcal{X}$	scalar, vector (bold lowercase), matrix (bold uppercase), set (calligraphic)
$\|\mathbf{x}\|_q$	ℓ_q norm of vector $\mathbf{x} \in \mathbb{R}^p := \left(\sum_{i=1}^p \|x_i\|^q\right)^{1/q}$
$E^C, \mathbb{1}_{\mathcal{X}}, \text{card}\,(\mathcal{X})$	complementary set, indicator function, cardinality of set $\mathcal{X}$
$\mathcal{S}$	support set
$\mathbb{N}, \mathbb{R}$	set of natural numbers, real numbers
$\mathcal{N}(\mu, \mathbf{\Sigma})$	Gaussian distribution with mean μ and covarianace $\mathbf{\Sigma}$
Abbreviations	
i.i.d.	independent and identically distributed (random variables)
i.e.	*id est* (Latin) – in other words
e.g.	*exempli gratia* (Latin) – for example

Part I

Background

Chapter 2

Hypothesis Testing in High-dimension

Summary. In this chapter, we introduce the motivation of this thesis's main theme: the problem of hypothesis testing for multivariate inference in high-dimension. In particular, we start from the classical problem of single hypothesis test, then move to the approaches of multiple hypothesis test, which has found popularity in modern data analysis. Some popular metrics for multiple hypothesis testing are also presented, along with the corresponding procedures to control them. We then move on to present the difficulty of doing statistical inference in high-dimensional regime, where the number of observations is far less than the number of variables. Next, we discuss the differences between univariate and multivariate approaches in this regime, and stress the importance of the latter for applications. The knockoff filter and conditional randomization test, two of the popular methods recently introduced for high-dimensional multivariate inference, are then presented. As discussed in Chapter 8, these two methods serve as a backbone for the main direction of this thesis. We end the chapter with the description of two important ingredients of our approach: a randomized clustering and p-value aggregation technique.

2.1 Determining The Importance Of A Single Variable *w.r.t.* The Response: A Classical Problem

Historically, the concept of hypothesis testing was introduced almost a century ago by the works of Ronald Fisher [Fis25, Fis36]. Hypothesis testing is a procedure that uses observed data to take decisions regarding the properties of the unknown data generating model. More formally, the definition of a hypothesis test is as follows.

Definition 2.1 (Hypothesis test, [CB02])**.** *A hypothesis test is a rule that specifies:*

- *For which sample values the decision is made to accept $\mathcal{H}_0$, the* null hypothesis*.*
- *For which sample values $\mathcal{H}_0$ is rejected and $\mathcal{H}_a$, the* alternative hypothesis*, is accepted.*

Example 2.1. One of the earliest recorded events that involve hypothesis testing is described in Ronald Fisher's book *Design of Experiments* in 1936 [Fis36]. The famed statistician wanted to test the claim of a female colleague that she could distinguish whether the milk or the tea was first placed in the cup just by tasting. Fisher then designed an experiment to give his colleague eight cups of tea, four of each variety (putting milk before tea and vice versa), in random order. One could then ask what the probability was for her getting the number she got correct, but just by chance.

The null hypothesis $\mathcal{H}_0$ was that the lady had no such ability, and the alternative hypothesis $\mathcal{H}_a$ is she could indeed classify the order of preparing a tea cup.

Suppose we have a sample of n *i.i.d.* observations $X = X_1, \ldots, X_n$ from a data distribution P_X with mean $\mu \in \mathbb{R}$, *i.e.* $X_i \sim P_X$. The problem could be testing whether

$$\mathcal{H}_0 : \mu \leq 0 \quad \text{vs.} \quad \mathcal{H}_a : \mu > 0.$$

Typically, to perform hypothesis testing, we specify a *test statistic*, denoted $T(X_1, \ldots, X_n)$.

Definition 2.2 (Test statistic)**.** *A test statistic is a function that captures an aspect of the sample of observations.*

Example 2.2. An example of a test statistic is the sample mean: $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i/n$.

After calculating the test statistic, we have to make a decision to declare whether the result of the hypothesis test is *statistically significant* or not. One way is to compare the test statistic with a certain quantile of its null distribution; this quantile is called *significance level* – denoted α – the probability that we reject the null hypothesis $\mathcal{H}_0$ given that the $\mathcal{H}_0$ is assumed to be true. Alternatively, one can report the result of a hypothesis test based on a *p-value*, defined as follows.

Definition 2.3 (p-value)**.** *A* p-value $p(X)$ *is a test statistic satisfying the following properties:*

1. $p(X) \in [0, 1]$.
2. *(Stochastic Domination) When* P_X *satisfies* $\mathcal{H}_0$*, for all* $t \in [0, 1]$*,*

$$\mathbb{P}_{X \sim P_X}(p(X) \leq t) \leq t.$$

Property 2 is important in the sense that it makes a p-value *valid*. Moreover, a consequence of this property is that if $p(X)$ is a valid p-value, it is possible to construct a test statistic at the significance level α based on $p(X)$, for any $\alpha \in (0, 1)$. Intuitively, we can see that the smaller the p-value $p(X)$, the more confidently the statistician can reject $\mathcal{H}_0$.

Example 2.3 (Two-sided normal p-value [CB02])**.** Perhaps one of the most common test that statisticians perform is testing the mean of a random sample $X_1, \ldots, X_n$ from a $\mathcal{N}(\mu, 1)$ distribution. We want to check $\mathcal{H}_0 : \mu = \mu_0$ versus $\mathcal{H}_a : \mu \neq \mu_0$. The test statistic is $T(X) = (\bar{X} - \mu_0)/\sqrt{n}$ where $\bar{X} = \sum_{i=1}^{n} X_i/n$. Under the null hypothesis, this test statistic follows a Student's t distribution with $n - 1$ degrees of freedom. Therefore, we obtain the two-sided p-value with a formula:

$$p(X) = 2\mathbb{P}(t_{n-1} \geq T(X)),$$

where t_{n-1} follows Student's t distribution with $n - 1$ degrees of freedom.

In any testing problem, two types of errors can be made. We say we make a *false positive*, or *type I error*, whenever we reject a true null hypothesis. On the other hand, a *false negative*, or *type II error* occurs when we fail to reject a false null hypothesis. Ideally, one would like to simultaneously minimize both the frequency of type I and type II errors. Unfortunately, this is not feasible and one has to find a trade-off between these two types of errors. This trade-off typically involves the minimization of type II errors while subjecting to a type I error constraint. Minimizing the number of false negatives can also be understood as maximizing the quantity of true discoveries, which is the *statistical power* of the hypothesis test.

Definition 2.4 (Statistical power)**.** *The statistical power of a single hypothesis test is the probability that the test correctly rejects the null hypothesis* $\mathcal{H}_0$ *when the alternative hypothesis* $\mathcal{H}_a$ *is true.*

2.2 Multiple Hypothesis Testing

While univariate testing still remains popular and has been widely studied theoretically, nowadays, with a given phenomenon, statisticians rarely ask only one question in the inference task. This gives rise to the problem of assessing statistical significance for multiple features simultaneously, or *multiple hypothesis testing*. More concretely, consider the general problem of simultaneously testing m hypotheses $\{\mathcal{H}_0^i\}_{i=1}^m$. If we assume that statistical tests are available for each individual hypothesis, the problem becomes how to combine them into a simultaneous test procedure. A naive approach would be to ignore the multiplicity and simply test each hypothesis at level α, just like in the univariate case. However, with such a procedure the probability of one or more false positives rapidly increases when m becomes large, since

$$\begin{aligned}&\mathbb{P}\left(\text{make at least one false positive when testing m hypotheses}\right)\\ &\quad=\mathbb{P}\left(\bigcup_{i=1}^{m}\text{false positive on i-th null hypothesis}\right).\end{aligned}\tag{2.1}$$

In other words, we can make more erroneous rejections when the number of tests grows large, which corresponds to the high-dimensional data setting. The following example demonstrates this issue in brain image analysis, but it also occurs commonly in other applications, *e.g.* genomics or health data, where there are abundance of claims in scientific studies that cannot be reproduced due to potentially incorrect usage of hypothesis testing procedures. [Ioa05, BIM+13, ENK16].

Example 2.4 (Naive approach of univariate testing in brain imaging)**.** A typical example in brain-imaging analysis: suppose $p = 100000$ of brain voxels, of which only 2000 are important. If a statistician naively performs the test of importance for each variable with the standard false positive rate of 5%, then even if he discovers all the important variables, on average there are still around 5000 $(5\% \times 98,000)$ unimportant variables that are falsely detected. The result is a list of discovered (supposedly important) variables of which about 70% are actually unimportant. It means the claim that the procedure controls the probability of false rejections at level 5% is misleading. Figure 2.1 illustrates this problem.

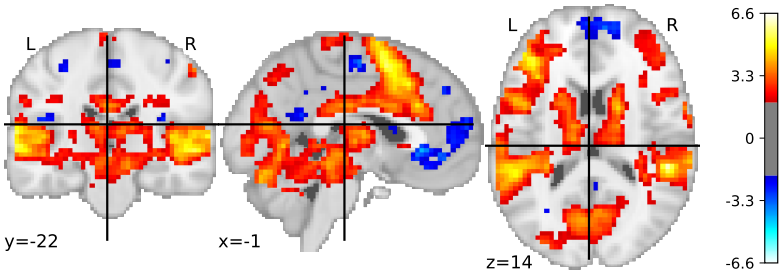


Figure 2.1: P-values (unadjusted for multiple testing) associated with each brain voxels at significance level 5%. The hypothesis to test here is which brain voxels are activated when the human subjects are presented with visual stimulation. The selected active parts of the brain are far more than the general consensus in neuroscience. More specifically, some activated areas of the brain belong to the auditory cortex, which is not related to vision tasks.

Example 2.4 shows that in order to perform multiple testing correctly, we have to reduce the significance level for each individual test. Naturally, the more hypotheses to test simultaneously, the more stringent the significance level. This procedure is called *multiple testing correction*, and with it comes several metrics that are more suitable. We can use p-value correction procedures to adjust the significance level to

these metrics. In the next section, we will introduce two of the most popular of such metrics: Family-Wise Error Rate (FWER) and False Discovery Rate (FDR).

	$\mathcal{H}_0$ true	$\mathcal{H}_a$ true	Total
$\mathcal{H}_0$ rejected	V (false positives, type-I error)	S	R
$\mathcal{H}_0$ not rejected	U	T (false negatives, type-II error)	$m - R$
Total	m_0	$m - m_0$	m

Table 2.1: Notation of possible outcomes from testing m hypotheses

Remark 2.1. *With notations in Table 2.1, a formula for statistical power in multiple testing is*

$$Power(R) = \mathbb{E}\left[\frac{S}{m - m_0}\right]. \tag{2.2}$$

Remark 2.2. *The reader can refer to [Roq15] as an extensive survey for the theory of multiple testing, including some recent results in the literature.*

2.2.1 Family-Wise Error Rate

The Family-Wise Error Rate is the probability of making one or more false positives when performing multiple tests, where the term "family" refers to the collection of hypotheses for simultaneous testing. Readers perhaps can recognize from this definition that it is the probability on either side of Eq. (2.1). FWER is a popular replacement for asserting significance in multiple testing. Instead of trying to control the probability of making type I error on one test not exceeding a given statistical significance level α, we focus on controlling the FWER below that level:

$$\text{FWER} \leq \alpha.$$

Controlling FWER is hence much stronger than normal type-I control of each single hypothesis test, since we control the probability of making a type I error in *any* tests, when the null is true for all of them, to be below α. Methods that control this metric are often obtained by individual test's p-value correction.

Definition 2.1 (Bonferroni's procedure)**.** *Let $p_1, \ldots, p_m$ be p-values associated with the family of hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$. The Bonferroni's correction rejects the null hypothesis for each $\mathcal{H}_0^i$ when:*

$$p_i \leq \frac{\alpha}{m}.$$

Theorem 2.1. Bonferroni's procedure controls FWER at significance level α for the family of hypotheses H_i, for all $i = 1, \ldots, m$.

Proof. Following notations in Table 2.1, let m_0 be the total number of null hypotheses that are true. From union bound and the definition of p-value we have, for Bonferroni's procedure, we have

$$\text{FWER} = \mathbb{P}\left[\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right] \leq \sum_{i=1}^{m_0}\mathbb{P}\left[p_i \leq \frac{\alpha}{m}\right] \leq m_0\frac{\alpha}{m} \leq \alpha.$$

□

The naming of this method follows Bonferroni inequality [Bon36], a generalized version of the union bound we use to prove the control of FWER. This method is a type of *single-step* procedure, *i.e.* we reject a hypothesis if its corresponding p-value is less than a threshold (which in the Bonferroni case is α/m). Another method

to control FWER, Holm's procedure, belongs to another class of multiple testing called *step-down* procedures. These types of procedure rely on thresholding the p-values based on a set of critical values $\alpha_1 \leq \cdots \leq \alpha_m$ with $\alpha_i \in (0,1)$, then start with comparing the smallest p-value with the smallest α-value α_1, and so on. An advantage of this procedure compared with Bonferroni's is that in general, Holm's procedure return higher statistical power. It can be described as follows.

Definition 2.2 (Holm's procedure)**.** *Let $p_1, \ldots, p_m$ be the (uncorrected) p-values associated with the family of hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$.*

1. *Order the m p-values ascendingly, denoted by $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ associated with $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}$.*
2. *For $i = 1, \ldots, m$:*
   - *If $p_{(i)} < \alpha/(m-i+1)$, reject $\mathcal{H}_0^{(i)}$.*
   - *Else if $p_{(i)} \geq \alpha/(m-i+1)$, stop the process and accept (not reject) $\mathcal{H}_0^{(i)}, \ldots, \mathcal{H}_0^{(m)}$.*

Theorem 2.2. Holm's procedure provides a control of FWER under significance level α.

Proof. Define $\mathcal{S}^c = \{i : \mathcal{H}_0^i \text{ is true null}\}$, so $\text{card}\,(\mathcal{S}^c) = m_0$ with our system of notation. Let j be the smallest index satisfying $p_{(j)} = \min_{i \in \mathcal{S}^c} p_i$. By definition, the Holm procedure makes a false positive only if

$$p_{(j)} \leq \frac{\alpha}{m-j+1},$$

or

$$\min_{i \in \mathcal{S}^c} p_i \leq \frac{\alpha}{m-j+1} \leq \frac{\alpha}{m_0},$$

since given our definition of p_j, we have $j \leq m - m_0 + 1$. Using Bonferroni inequality:

$$\text{FWER} \leq \mathbb{P}\left[\min_{i \in \mathcal{S}^c} p_i \leq \frac{\alpha}{m_0}\right] \leq \sum_{i \in \mathcal{S}^c} \mathbb{P}\left[p_i \leq \frac{\alpha}{m_0}\right] \leq \alpha.$$

□

One advantage of Bonferroni's and Holm's correction is that no assumption on the correlation structure of the p-values is needed, as we only rely on a union bound to prove the theoretical guarantee for controlling FWER. However, controlling FWER comes with a cost of being extremely conservative when the set of hypotheses is large. This may defeat the purpose of practitioners, which is to discover features that show an effect of interest. We will go through a different type of error rate that attempts to fix this problem in the next section.

2.2.2 False Discovery Rate

The seminal paper of Benjamini and Hochberg [BH95] introduced a new multiple testing criterion that takes into account the proportion of false positives, or false discoveries amongst all rejected hypotheses instead of the probability of making false positives as with FWER. Ever since its introduction, False Discovery Rate has become a popular metric for multiple testing, thanks to the better tradeoff between controlling type-I and type-II error than that of FWER. This is partly motivated by the increasing availability of many "wide" datasets, in which we have few observations compared with a large number of variables (or test hypotheses m), *e.g.* in genomics or brain-imaging data. In other words, when the number of hypotheses m is large, for instance when considering high-resolution datasets, controlling FWER

would provide a strong control on false positives, but can make too few of variables to reach statistical significance (figure 3.2d). The FDR controlling procedures help finding new discoveries in a less restrictive fashion, while still maintaining some level of error control, hence avoiding a too large proportion of spurious detections. First, we define the False Discovery Proportion based on the notation introduced in table 2.1.

Definition 2.3 (False Discovery Proportion & False Discovery Rate [BH95])**.**

(a) False Discovery Proportion: $FDP = \dfrac{V}{R \vee 1} = \dfrac{\textit{number of false discoveries}}{\textit{number of discoveries}}$

(b) False Discovery Rate: $FDR = \mathbb{E}\left[FDP\right] = \mathbb{E}\left[\dfrac{V}{R \vee 1}\right]$.

We have to take expectation on the False Discovery Proportion based on the fact that testing procedure is done on a sample of observations, therefore FDP is random. In other words, the False Discovery Rate is the *expected* proportion of discoveries that are false. In [BH95], a procedure to control the FDR was also proposed. This procedure belongs to the family of *step-up procedures*. Similar to step-down procedures, *e.g.* the Holm's procedure stated above, step-up procedures also begin with ordering p-values ascendingly, and comparing them with a set of critical values $\alpha_1 \leq \cdots \leq \alpha_m$, $\alpha_i \in (0, 1)$. However, step-up procedures start the comparison using the *least* significant p-value with the *largest* α-value, which is demonstrated with the Benjamini-Hochberg (BH) procedure as follows.

Definition 2.4 (Benjamini-Hochberg procedure)**.** *Let $p_1, \ldots, p_m$ be the associated p-values with the family of hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$. We reorder them ascendingly, denoted by $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, which associated with $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}$.*

1. *Find $\widehat{k}_{BH}$ such that*

$$\widehat{k}_{BH} \in \max\left\{k = [m] \mid p_{(k)} \leq \alpha_k\right\}, \quad \textit{where } \alpha_k \overset{def.}{=} \frac{k\alpha}{m},$$

2. *If $\widehat{k}_{BH}$ exists, reject $\mathcal{H}_0^{(1)}, \ldots \mathcal{H}_0^{(\widehat{k}_{BH})}$, otherwise accept all hypotheses.*

Theorem 2.3. When the p-values $p_1, \ldots, p_m$ are independent, the BH procedure controls FDR at level α.

As the reader might notice, the BH procedure relies on a crucial assumption that the test statistics used to derive p-values be *independent*, which is rather problematic in practical settings. The study of [BY01] provided an in-depth discussion of this problem, and relaxed this assumption to milder case. In particular, they proved that FDR is still controlled when the joint distribution of p-values satisfies a property called *positive regression dependency*, besides independence case. More importantly, the work also provides a procedure which has guaranteed control for FDR under *arbitrary dependence*, *i.e.* in any possible correlation structure, which is called Benjamini-Yekutieli (BY) procedure. We describe this procedure as follows.

Definition 2.5 (Benjamini-Yekutieli procedure)**.** *Let $p_1, \ldots, p_m$ be the associated p-values with the family of hypotheses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$. We reorder them ascendingly, denoted by $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ and $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}$.*

1. *Find $\widehat{k}_{BY}$ such that*

$$\widehat{k}_{BY} = \max\left\{k \in [m] \mid p_{(k)} \leq \alpha_k, \right\} \quad \textit{where } \alpha_k \overset{def.}{=} \frac{k\alpha}{m\sum_{i=1}^{m} 1/i}.$$

2. *If* $\widehat{k}_{BY}$ *exists, reject* $\mathcal{H}_0^{(1)}, \dots \mathcal{H}_0^{(\widehat{k}_{BY})}$, *otherwise accept all hypotheses.*

Theorem 2.4. Under any arbitrary dependence structure of p-values $p_1, \dots, p_m$, the BY procedure controls FDR at level α.

Remark 2.3. *The proof for FDR controlling of BH procedure can be found in [BH95] using a proof by induction. Another one with a different approach that uses martingale theory is presented in [Sto02]. For proving FDR controlling of BY procedure, readers can refer to [BY01].*

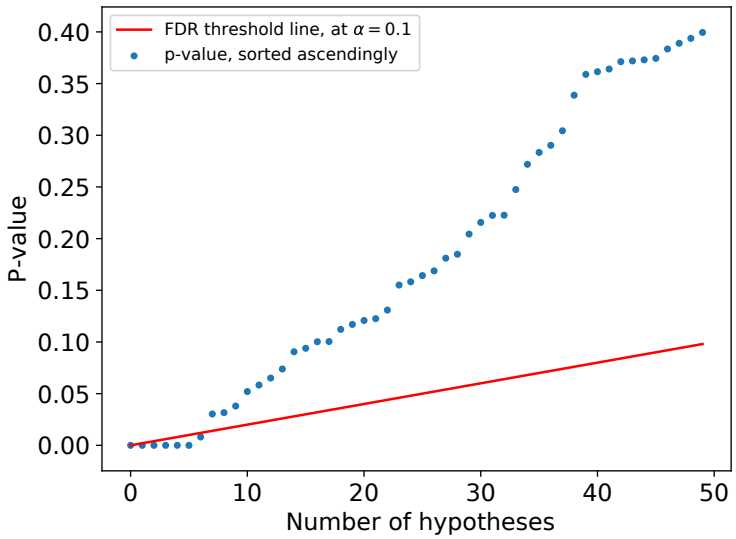


Figure 2.2: Benjamini-Hochberg FDR thresholding procedure, with the redline illustrating the $(k\alpha)/m$ function, for $k = 1, \dots, m$. We reject all hypotheses k that have p-values $p_{(k)} \leq \dfrac{k\alpha}{m}$, corresponding to dots lie under the redline.

Potential problems of FDR as a multiple testing metric. As mentioned above, while we have the control of the FDR, the FDP is still a random variable [Roq15]. This means that a procedure that guarantees FDR $\leq \alpha$ can still fail to achieve FDP below the same significance level α for different samples of observations. This is problematic in many real life applications, for example in medical imaging where avoiding making a wrong call on whether the patient has lung disease or not is a must. A natural solution to that consists in building a confidence interval for the FDP. There are several studies that develop this direction, for example the early work of [GW06], or more recent work of [GS11, GMKS19, BNR20] and [KR20] that introduces a *post hoc* multiple-testing procedure. In such a procedure, the statistician chooses the collection of rejected hypotheses freely, and the multiple testing procedure returns the associated quality criterion. This is in contrast with traditional FWER and FDR controlling procedures, which return the collection of rejected hypotheses based on a *pre-specified* level.

Limitations of BH and BY procedure Besides the problem with FDR, there are also limitations with the two classic procedures to control this metric. Benjamini-Hochberg's procedure relies on independence or positive dependence of the test-statistics, which may not hold in some scenarii. On the other hand, while one can rely on Benjamini-Yekutieli procedure on such arbitrary dependence cases, the reader should note that the factor $\sum_{i=1}^{m} 1/i$ in Definition 2.5 of BY procedure approaches

$\log m$ when m goes large, which makes each p-value threshold much smaller. This means although we have strong control of FDR in any dependence structure, BY is inherently much more conservative than BH procedure, much similar to the conservativeness of the Bonferroni's procedure that requires no further assumptions on the test-statistic distribution to control the FWER.

2.3 Statistical Inference in High-dimension

Problem setting Suppose we have a sample of observations with n i.i.d. random variables, each with the form $\mathbf{x} = (x_1, \dots, x_p) \in \mathbb{R}^p$ with the response $y \in \mathbb{R}$. Furthermore, they follow the linear relationship

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon}. \tag{2.3}$$

with $\mathbf{X} \in \mathbb{R}^{n \times p}$ is the design matrix that contains n rows of $\mathbf{x}$, $\mathbf{y}$ the vector of responses, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ the unknown vector of regression coefficients, $\boldsymbol{\varepsilon}$ the Gaussian noise vector with noise magnitude σ, *i.e.* we have $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. We will further assume that the distribution of $\mathbf{X}$ is multivariate normal with mean vector zero and population covariance $\boldsymbol{\Sigma}$.

Conditional Independence Testing Besides multiple hypothesis testing, another important issue the statistician has to face is the choice of using a *univariate* versus *multivariate* approach for statistical inference. The testing problem stated in Section 2.1 is one example of univariate approach, or marginal inference. As the name suggested, in this approach, we only care about testing the marginal relationship of one variable $\mathbf{x}_j$ versus the response $\mathbf{y}$ without taking into the interaction of j and other variables $\mathbf{X}_{-j} \overset{\text{def.}}{=} \{\mathbf{x}_1, \dots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \dots, \mathbf{x}_p\}$. This, however, is unrealistic in most practical applications, especially nowadays that datasets are becoming evermore complex thanks to the advances in acquisition technologies. For instance, in neuro-imaging, a common understanding is that there is a positive correlation between neighboring brain voxels. In particular, given a mental simulation like watching a short video, we usually observe activation from a region of multiple brain voxels. Therefore, it is more likely to phrase the question in the sense of conditional relationship between $\mathbf{y}$ amd $\mathbf{x}_j$: is the brain voxel $\mathbf{x}_j$ activated when presenting with the mental stimuli $\mathbf{y}$, given its interaction with other brain voxels $\mathbf{X}_{-j}$? This is called multivariate inference, or testing for conditional independence. Figure 2.3 illustrates this idea in the neuroscience setting.

We define formally the conditional hypothesis testing problem as follows. Let $\mathcal{H}_0^1, \dots \mathcal{H}_0^p$ be the family of p hypotheses to test for. With each variable j, we want to test its independence with the response $\mathbf{y}$, conditionally to other variables, or in general:

$$\mathcal{H}_0^j : \mathbf{x}_j \perp y_j \mid \mathbf{x}_{-j} \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_a^j : \mathbf{x}_j \not\perp y_j \mid \mathbf{x}_{-j}, \tag{2.4}$$

for each feature $j \in [p]$. Equivalently, in linear relationship of Eq. 2.3, we have, for each $j \in [p]$:

$$\text{(null) } \mathcal{H}_0^j : \beta_j^0 = 0 \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_a^j : \beta_j^0 \neq 0.$$

A classical approach is first finding the estimator $\hat{\boldsymbol{\beta}}$ by solving the least-squares estimation problem

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{2.5}$$

where $\|\cdot\|_2$ is the Euclidean norm. This problem is easily solvable when $n > p$, or the "low-dimension" regime. More specifically, we have the closed form formula for $\hat{\boldsymbol{\beta}}$ when taking the first order condition of the optimization problem:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{2.6}$$

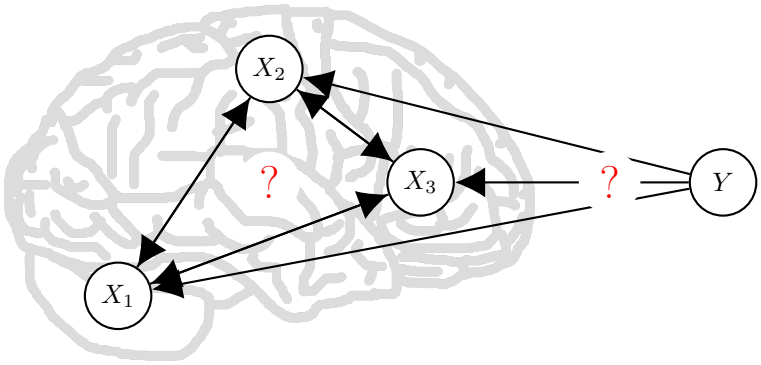


Figure 2.3: Illustration of interaction between 3 brain voxels in cognitive neuroscience and a behavior variable Y (that is typically elicited by a controled stimulus). The goal is to investigate whether X_j is activated given the mental condition Y, conditionally to the interaction with the other two brain voxels X_{-j}. Figure is an adaptation from [WMO+15, WP20].

where we assume that the matrix $\mathbf{X}^T\mathbf{X}$ is invertible (feasible in low-dimension). Calculation of the test statistics $T(\mathbf{x}_j, y)$ for each variable j, which often involves $\hat{\beta}_j$, therefore comes naturally. With this we can have p-values and doing multiple testing correction comes naturally with the procedures mentioned in Section 2.2. However, when $n < p$, or in high-dimension regime, this matrix is not invertible, and there will be multiple solutions to the optimization problem in eq. (2.5). Furthermore, the iteration cost of approximating the pseudo-inverse of $\mathbf{X}^T\mathbf{X}$ is cubic in p, which means the computation becomes prohibitive when p grows very large. We therefore have to think of different approaches to adapt to this regime, which will be presented in the next section.

2.4 Some Approaches for Multiple Hypothesis Testing in High-dimension

2.4.1 Penalized Regression: Lasso and Ridge Estimator

As mentioned in Section 2.3, the ordinary least squares estimator behaves poorly in high-dimension, as when $n < p$ the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible. A common solution is to regularize the least-squares objective.

Least-squares in high-dimension One popular solution is to rely on penalized regression by adding a penalty term in the form of either ℓ_2-norm or ℓ_1-norm of the coefficient to the optimization problem in equation (2.5). This is called *ridge* and *lasso* regression, respectively.

Definition 2.1 (Ridge least-squares regression)**.** *Let $\lambda > 0$ be a regularization parameter, the ridge regression estimator $\hat{\boldsymbol{\beta}}^{RIDGE}$ is defined as*

$$\hat{\boldsymbol{\beta}}^{RIDGE}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2. \tag{2.7}$$

Definition 2.2 (Lasso least-squares regression)**.** *Let $\lambda > 0$ be a regularization parameter, the lasso regression estimator $\hat{\boldsymbol{\beta}}^{RIDGE}$ is defined as*

$$\hat{\boldsymbol{\beta}}^{LASSO}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{2.8}$$

For the ridge regression, we do actually have a closed form formula for the solution

$$\hat{\boldsymbol{\beta}}^{\text{RIDGE}}(\lambda) = \frac{1}{n}\left(\mathbf{X}^T\mathbf{X}/n + \lambda\mathbf{I}\right)^{-1}\mathbf{X}^T\mathbf{y},$$

where $\mathbf{I}$ is the identity matrix. This formula is obtained by taking the first order derivative of the objective function, then set it to zero and solving for $\hat{\boldsymbol{\beta}}^{\text{RIDGE}}(\lambda)$. However, for the lasso, we have to resort to using optimization schemes as ℓ_1-norm is non-smooth. The two popular solutions are iterative hard-thresholding [BD08] and coordinate descent [FHT10]. Since it is not the main topic of this thesis, we will not provide any further analysis for these algorithms. A geometric demonstration of regularized least-squares estimation can be found in Figure 2.4.

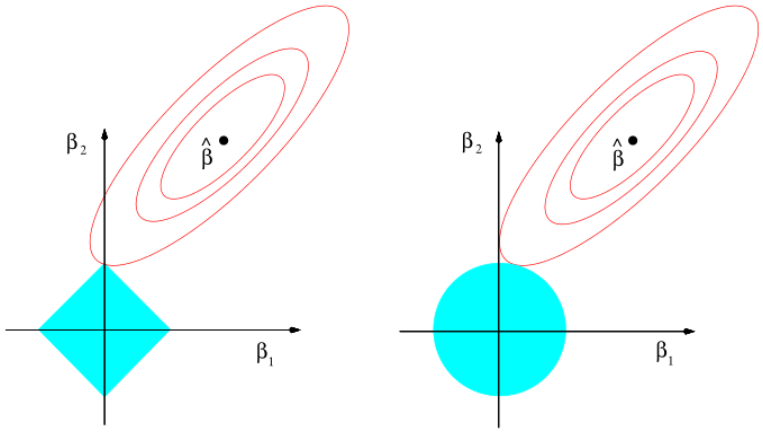


Figure 2.4: A simple visualization of lasso and ridge regression. The red contours are of least-squares loss function, while the solid blue areas are the constraint regions for ℓ_1 (left) and ℓ_2 (right) penalty term. Graphic taken from [HTF09, chapter 3].

The regularization parameter λ is called a hyper-parameter, which should be adapted to the data. Note that in practice, cross-validation is usually employed to choose this parameter. Readers can refer to [AC10] for a detailed survey on this topic. A desired property of penalized regression, especially with the lasso, is that they shrink the estimated coefficients. Depending on the magnitude of the regularization parameter λ, we can allow a small proportion (if λ is small) or a large proportion of the regression coefficients $\hat{\boldsymbol{\beta}}$ to be zero. We can see this phenomenon on the left side of Figure 2.4: with lasso regularization, when the contour of the loss function touches the corner of the constraint region, one of the coefficients will be zero. Because of this property, the lasso is also a very popular estimator in compressive sensing literature, where one needs to estimate/recover a sparse signal from the data. The induced sparsity can also be understood as a variable selection step. However, even with only a small subset of variables being greater than zero, we cannot provide conclusions that these variables are significant. This is because generally, in standard ridge and lasso regression, there are still no way to calculate test-statistics that follow a known distribution under the null hypothesis, hence p-values also cannot be calculated for each variable.

Remark 2.1. *There is a recent work that suggests a method for multiple testing directly on the estimated coefficients of Least Angle Regression (LARS), a close variant of Lasso. See the work of [ADC21] for more details.*

In the recent statistics literature, there exists two lines of methods that attempt to fix the issue of getting p-values in high-dimension. One is based on the technique called *post-selective inference*. As the name suggested, this procedure uses only non-zero fitted coefficients to calculate test-statistics, which is possible since the model

is now in lower-dimension. This technique will be discussed in Section 2.4.2. A different line of works [vdGBRD14, JM14, ZZ14] provides a valid way to construct p-values in high-dimension for the lasso estimator without refitting the estimated coefficients in lower-dimension. These works point out that $\hat{\boldsymbol{\beta}}^{\text{LASSO}}(\lambda)$ is in fact biased, depending on the choice of λ. Then, their idea is to correct the bias and derive an unbiased estimator that follows a standard normal distribution under the null hypothesis. However, their assumptions are strong and might be difficult to verify.

Debiased Lasso The following formula is taken following [vdGBRD14], but readers can also refer to the parallel works of [JM14, ZZ14], which are all based on the same principle. First, let $\boldsymbol{\Theta}$ be the precision matrix, or inverse of the true covariance matrix $\boldsymbol{\Sigma}$. The idea is to correct the $\hat{\boldsymbol{\beta}}^{\text{LASSO}}(\lambda)$ with a quantity proportional to the regression residuals and the diagonal elements of the estimated precision matrix $\hat{\boldsymbol{\Theta}}$. Estimating $\hat{\boldsymbol{\Theta}}$ is therefore the most crucial step of whole procedure. This is done via an operation called nodewise lasso regression. For each $j = 1, \ldots, p$, let $\mathbf{X}_{-j}$ be the design matrix with the jth column removed. We then solve the lasso for each regression problem $\mathbf{x}_j$ versus $\mathbf{X}_{-j}$:

$$\hat{\boldsymbol{\gamma}}_j = \text{argmin}_{\boldsymbol{\gamma} \in \mathbb{R}^{p-1}} \left\{ \frac{1}{n} \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\gamma}\|_2^2 + 2\lambda_j \|\boldsymbol{\gamma}\|_1 \right\}. \tag{2.9}$$

We then define matrices $\hat{\mathbf{C}}$ and $\hat{\mathbf{T}}$ as

$$\begin{aligned} \hat{\tau}_j^2 &= \frac{1}{n}(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\gamma}}_j)^T \mathbf{x}_j, \\ \hat{\mathbf{C}} &= \begin{pmatrix} 1 & -\hat{\gamma}_{1,2} & \ldots & -\hat{\gamma}_{1,p} \\ -\hat{\gamma}_{2,1} & 1 & \ldots & -\hat{\gamma}_{2,p} \\ \vdots & \vdots & \vdots & \vdots \\ -\hat{\gamma}_{p,1} & -\hat{\gamma}_{p,2} & \ldots & 1 \end{pmatrix} \\ \hat{\mathbf{T}}^2 &= \text{diag}(\hat{\tau}_1^2, \ldots, \hat{\tau}_p^2), \end{aligned} \tag{2.10}$$

where $\hat{\boldsymbol{\gamma}}_j = \{\hat{\gamma}_{j,k}; k \neq j\}$. We use these quantities to estimate the precision matrix:

$$\hat{\boldsymbol{\Theta}} = \hat{\mathbf{T}}^{-2}\hat{\mathbf{C}}. \tag{2.11}$$

Finally, the debiased estimated coefficient is defined by

$$\hat{b}^{\text{LASSO}} = \hat{\boldsymbol{\beta}}^{\text{LASSO}} + \frac{1}{n}\hat{\boldsymbol{\Theta}}\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{\text{LASSO}}). \tag{2.12}$$

The same work [vdGBRD14, Theorem 3.1] also established a theoretical result regarding the asymptotic distribution of $\hat{b}^{\text{LASSO}}$ as follows. Under assumptions regarding the eigenvalues of the population covariance matrix, we have, for each $j = 1, \ldots, p$

$$\frac{\sqrt{n}(\hat{b}_j^{\text{LASSO}} - \beta_j^0)}{\hat{\sigma}} \xrightarrow[n\to\infty]{\text{d}} \mathcal{N}(0, 1), \tag{2.13}$$

where $\hat{\sigma}$ is the estimated noise magnitude. We can then use the standardized debiased coefficient as a test statistic with standard normal distribution under the null hypothesis, and output the corresponding p-values, even when we step into high-dimensional regime of $n < p$. The debiased lasso is a nice development in the multiple testing literature in recent years, and is still actively developed with the work of [JJ19] that adapts this procedure for FDR controlling, [BZ21] adding the degree-of-freedom adjustment for the debiasing step, and [MM18, CMW20] providing extended theoretical analysis with more relaxed assumptions than the work of [vdGBRD14].

As we shall see in the chapters that present this thesis' contributions, debiased lasso would always appear in the experimental results section as one of the alternative methods for comparison.

Remark 2.2. *While it is not presented here, there also exists a debiasing approach with the same principle for ridge estimator, presented in [Bü13].*

2.4.2 Post-Selective Inference

As mentioned briefly in the previous section, post-selective inference usually contains two steps. The first step is a variable screening/model selection step in high-dimension problems, that reduces the model to a low-dimensional one. The second step is to perform inference only on the variables selected in the first step. One of the early works that introduce this type of solution is [WR09]. They proposed a method called single data-splitting: the data is split into two parts: the first part is used to reduce the number of dimension from p to $k \ll p$, and the second part is used to calculate p-values by ordinary least squares estimation on these k variables. An example could be using lasso for variable screening step, and fitting the OLS on the reduced model with inference step. What makes data-splitting work is that variable screening step is performed on data that are independent from the one used for inference. Hence, the inference is automatically valid conditional on the screening step, assuming the true support is a subset of screened variables. However, because only one half of the sample of observations is used for inference, splitting data often results in a loss of statistical power. This gives rise to a new line of studies, for example of [BBB+13], which performed statistical inference conditionally to the selection of the model based on the conditional coverage; [LTTT14] proposed using lasso-path for calculation of the test statistics; [LT15, RTT15] consider the procedure for cluster of variables instead of selecting them individually. Post-selective inference has also been extended beyond the linear model setting with [YUFT18, SCAV19] using a kernel method. For a more detailed introduction to this procedure, readers can refer to the short survey of [TT15]. One of the main drawbacks for post-selective inference is that the screening step still suffers from the curse of dimensionality, as we have to estimate the regression coefficients $\hat{\boldsymbol{\beta}}$ for all variables. We therefore do not consider further such procedures, and proceed to the next section with alternative approaches.

2.5 Knockoff Filter: A Modern Approach for Controlling False Discovery Rate in high-dimension

Introduced in the seminal work of [BC15] and further developed by [CFJL18], the knockoff filter is one of the recent breakthroughs in the multiple testing literature. The most distinctive feature in this method's approach is the creation of noisy copies of the original variables, which helps with the calculation of test statistics. These noisy copied variables are called knockoffs. More formally, suppose that we continue with the setting in previous section, *i.e.* linear relationship in Eq. (2.3) on n i.i.d. random variables, each with the form $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ with the response $y \in \mathbb{R}$. We are looking to test the conditional independence of variable x_j with y conditionally to all other variables $\mathbf{x}_{-j}$, similar to the formal setting of Eq. (2.4). The definition of knockoffs variable is as follows.

Definition 2.1 (Knockoff variables, [BC15, CFJL18])**.** *$\tilde{\mathbf{x}}$ is a knockoff of a family of random variables $\mathbf{x} = (x_1, \ldots, x_p)$ if $\tilde{\mathbf{x}}$ satisfies two following properties:*

1. *For any subset $\mathcal{K} \subset \{1, \ldots, p\}$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(\mathcal{K})} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$, where the vector $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(\mathcal{K})}$ denotes the swap of entries x_j and $\tilde{x}_j$ for all $j \in \mathcal{K}$, and $\overset{d}{=}$ denotes equality in distribution.*
2. *$\tilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}$.*

The first property in Def. 2.1 is called exchangability. Figure 2.5 illustrates an example of knockoff variable with $\mathcal{K} = j, k$ with this swap property. To sample knockoffs, [CFJL18] relies on the assumption that the distribution of the data P_X is known. They called this second-order knockoff sampling, which, as the name suggested, uses the mean (first moment) and the variance (second moment) of $\mathbf{X}$ as input:

$$\mathbb{E}[\tilde{\mathbf{X}}] = \mathbb{E}[\mathbf{X}], \quad \mathbb{E}[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}] = \boldsymbol{\Sigma} \quad \text{and} \quad \mathbb{E}[\tilde{\mathbf{X}}^T\mathbf{X}] = \boldsymbol{\Sigma} - \text{diag}(\{)\mathbf{s}\}, \tag{2.14}$$

where $\boldsymbol{\Sigma}$ is the population covariance, $\text{diag}(\{)\mathbf{s}\}$ is a diagonal matrix acting as a perturbation that makes the correlation matrix between $\mathbf{X}$ and $\tilde{\mathbf{X}}$ a non-trivial one, *i.e.* different from $\boldsymbol{\Sigma}$. It can be calculated by either solving a semi-definite optimization program, or using a closed-form formula. If the known distribution P_X is Gaussian, we can then sample $\tilde{\mathbf{X}}$ satisfying the above condition by

$$\tilde{\mathbf{x}}_j \mid \mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}), \tag{2.15}$$

where

$$\boldsymbol{\mu} = \mathbf{X} - \mathbf{X}\boldsymbol{\Sigma}^{-1}\text{diag}(\{)\mathbf{s}\} \tag{2.16}$$

$$\text{and } \mathbf{V} = 2\text{diag}(\{)\mathbf{s}\} - \text{diag}(\{)\mathbf{s}\}\Sigma^{-1}\text{diag}(\{)\mathbf{s}\}. \tag{2.17}$$

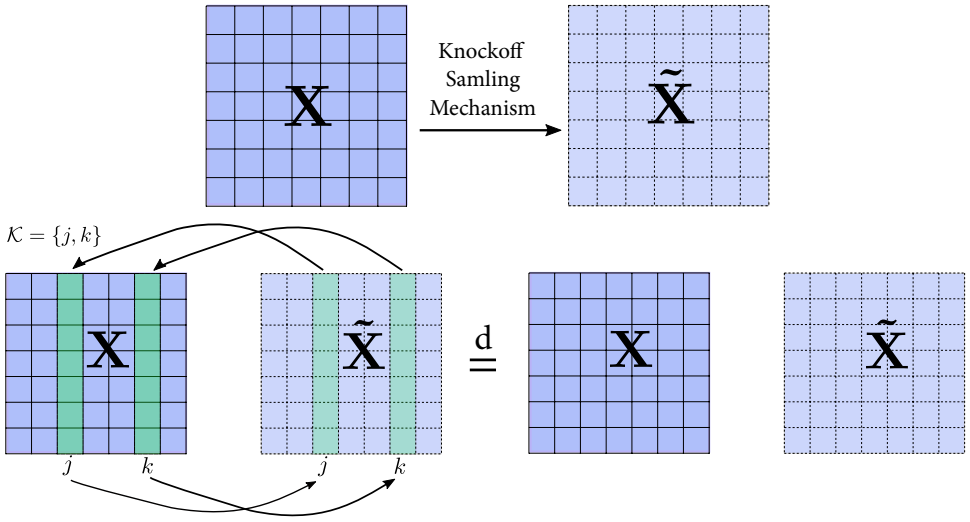


Figure 2.5: A visualization of knockoff covariates sampling (creating noisy copies of original variables) and swapping property of knockoff variables. Some of the knockoff sampling methods include: Second-order knockoff [CFJL18], Hidden Markov Knockoff [SSC18], Metropolis-Hasting Knockoff [BCJW20], Deep Knockoff variants: Autoencoder and Generative Adversarial Networks [RSC18, JY19]

After obtaining knockoff variables, it is crucial for the calculation of knockoff statistics (or knockoff importance score), defined as follows.

Definition 2.2 (Knockoff statistic, [BC15, CFJL18])**.** *A knockoff statistic* $\mathbf{W} = \{W_j\}_{j=1}^p$ *is a measure of feature importance that satisfies the two following properties:*

1. $\mathbf{W}$ *is the function of* $\mathbf{X}, \tilde{\mathbf{X}}$ *and* $\mathbf{y}$ *only*

$$\mathbf{W} = f(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}).$$

2. *Swapping the original variable column* $\mathbf{x}_j$ *and its knockoff column* $\tilde{\mathbf{x}}_j$ *switches the sign of* W_j*, for any* $\mathcal{K} \subseteq 1, \ldots, p$

$$W_j\left([\mathbf{X}, \tilde{\mathbf{X}}]_{swap(\mathcal{K})}, y\right) = \begin{cases} W_j\left([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}\right) & \text{if } j \in \mathcal{K}^c \\ -W_j\left([\mathbf{X}, \tilde{\mathbf{X}}], y\right) & \text{if } j \in \mathcal{K}, \end{cases}$$

The idea is to take W_j such that a large positive value of W_j provides evidence against the null hypothesis $H_0^j : \boldsymbol{\beta}_j^0 = 0$. In particular, if we denote $\mathcal{S}$ the true support set, defined as $\mathcal{S} \overset{\text{def.}}{=} \{j \in [p] : \beta_j^0 \neq 0\}$ where β_j^0 the true regression coefficient, the previous sentence means that the probability $\mathbb{P}(W_j \geq k)$ with $k > 0$ should be large when $j \in \mathcal{S}$. From the swapping property, it is easy to verify that the null distribution of knockoff statistic W_j is symmetric around 0. This property will be useful for calculating the FDR controlling threshold later. There are several ways to define a knockoff statistic that satisfies the definition, *e.g.*

(1) $W_j = |\mathbf{x}_j^T y| - |\tilde{\mathbf{x}}_j^T y|$.

(2) $W_j = |\hat{\beta}^{\text{LS}}|_j - |\hat{\beta}^{\text{LS}}|_{j+p}$ in which $\hat{\boldsymbol{\beta}}^{\text{LS}} = \left(\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^T \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] \right)^{-1} \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^T \mathbf{y}$ is the classical least-square solution on the augmented matrix $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$.

However, note that these suggestions only work when $n > 2p$, or the setting of the original *fixed-X knockoff*, proposed in [BC15]. In order to calculate knockoff statistics in high-dimension, [CFJL18] suggested to first run the Lasso program on $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2p}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{y} - \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{2.18}$$

Then, we can define W_j based on the estimated Lasso coefficients:

(1) Take $z_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$ for $j = 1, \ldots, 2p$, then take $W_j = (z_j \vee z_{j+p}) \times \operatorname{sign}(z_j - z_{j+p})$. This is called *Lasso-max statistic*.

(2) $W_j = |\widehat{\beta}_j(\lambda)| - |\widehat{\beta}_{j+p}(\lambda)|$, the *Lasso coefficient-difference (LCD)*.

In this thesis, we choose to work exclusively with the LCD statistics and model-X knockoffs, since this statistic provides flexibility and has shown good performance versus other choices of knockoff statistics [CFJL18, KLM20].

After acquiring knockoff statistics, the final step is to calculate a multiple testing threshold with the aim to control FDR at significance level $\alpha \in [0, 1]$.

$$\tau = \min \left\{ t > 0 : \widehat{\text{FDP}}(t) \leq \alpha \right\}, \quad \text{where } \widehat{\text{FDP}}(t) \overset{\text{def.}}{=} \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \tag{2.19}$$

and the set of selected variables is $\hat{\mathcal{S}} = \{j \in [p] : W_j \geq \tau\}$.

Remark 2.1. *The threshold τ is in fact called* `knockoff+` *threshold in the original literature [BC15, CFJL18]. The name is inspired from the $+1$ quantity in the numerator of the quantity $\widehat{FDP}(t)$ in Eq. (2.19) to distinguish with the one without this quantity. We omit the need to introduce two different notions since only the* `knockoff+` *threshold has theoretical guarantee for controlling FDR under significance level α.*

Let us get some intuition about the estimation of $\widehat{\text{FDP}}(t)$ in Eq. (2.19). Figure 2.6 shows the histogram of one sample of knockoff statistics W_j for each variable $j = 1, \ldots, p$. For any $t > 0$, the red area denote the number of the test statistic W_j that is less than or equal to $-t$. This quantity could be understood as an estimation of the number of falsely rejected null hypothesis. Indeed, because of the symmetry of knockoff statistics null distribution, this is equal to the number of $W_j \geq t$, or the number of H_j that we falsely rejected, *i.e.* the false positives. On the other hands, the blue area denotes the estimation of the support $\hat{\mathcal{S}}$, the set of rejected hypotheses. It is also the denominator of the formula of $\widehat{\text{FDP}}(t)$. Hence, by definition, if we divide the red area by the blue area we would have an estimation of False Discovery Proportion, for any given $t > 0$. We also note that if $j \in \mathcal{S}$, W_j will be positive with high probability due to the property of the knockoff statistic. This ensures that we have a good estimation of the $\widehat{\text{FDP}}(t)$ with the proposed formula.

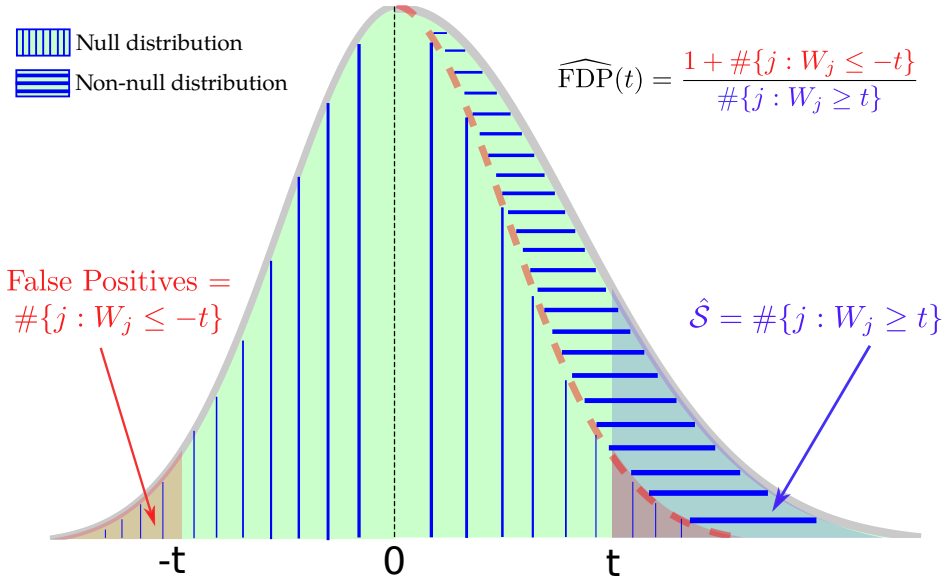


Figure 2.6: Illustration of False Discovery Proportion estimation procedure with Knockoff Statistics: the figure shows a histogram of one sample of knockoff statistics. The FDP is the ratio of the red over the blue area on the right tail of the distribution. While the red area on the right tail (which shows the number of false positives) cannot be calculated in reality, it can be estimated by the red area on the left tail of the distribution due to symmetry property of knockoff statistics.

Advantages of Knockoff Filter A strong advantage of knockoff filter is its flexibility: the methodology relies on few assumptions. As a result, we can use knockoff on a large class of problems and test statistics. For example, statistician can select different types of knockoff statistics, some of which are listed above, and one can switch from solving standard Lasso program with the generalized linear model with ℓ_1 regularization. Moreover, the flexibility could also come from the knockoff construction method. One could choose second-order knockoff as suggested in original knockoff paper [CFJL18], or from the extensions of [RSC18, JY19]. The knockoff construction proposed in the latter two works could scale better in very high-dimension thanks to modern deep network architectures. To take into account of interaction between latent variables, which is especially true in the genomics application, [SSC18] proposed the Hidden Markov Knockoff , and [BCJW20] extended the idea with fast Metropolis-Hasting Knockoff. Beyond the model-X regime, [BCS18] introduces a more robust method of knockoff sampling without assumption that P_X is known. Deep neural network could also be used to calculate knockoff statistic in non-linear relationship [LFLN18]. We also refer the readers to the work of [KLM20] for a generalization of knockoff filter as an FDR control method. Another advantage of knockoff filter is relatively low computation cost. Although some knockoff sampling methods can be costly, the original proposal of second-order knockoff is cheap to compute. The method also requires only one solution of the optimization problem for test statistic calculation. Compared to permutation test, for example, we would need to have multiple runs to output a meaningful empirical p-values with much higher costs, although technically knockoff filter does not provide p-values for each variable.

Some drawbacks of Knockoff Filter While knockoff inference is in general a flexible and fast method, it suffers from a crucial problem. The introduction of noisy copies $\tilde{\mathbf{X}}$ means there is an inherent randomness in the method. We illustrate this

idea in Figure 2.7 with a synthetic data experiment. First, we generate a synthetic dataset of 400 samples of $\mathbf{x}_j$ follows a multivariate normal distribution with dimension 600. We generate the response y according to linear relationship of Eq. (2.3) with $\boldsymbol{\beta}^0$ sparse. We then make 2500 runs of knockoff filter on this same dataset, and plot histogram of the resulting FDP and statistical power. Although the FDR, *i.e.* average of FDP, is controlled under level $\alpha = 0.1$, we observe a great variability in both metrics. This suggests the knockoff filter is highly unstable in inference results. Stabilizing the method, therefore, is a non-trivial question, and this thesis provides one of the solution for this problem. This new algorithm is based on a p-value aggregation technique, and will be presented in Chapter 4.

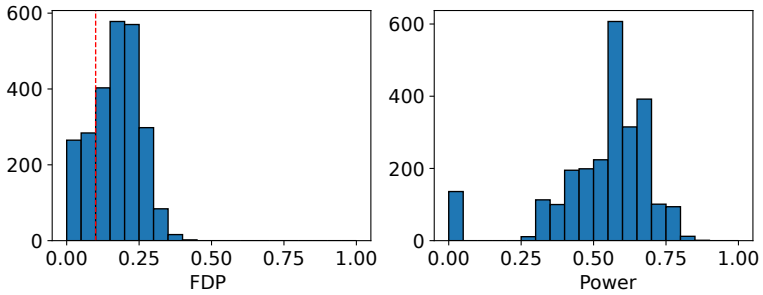


Figure 2.7: Histogram of FDP and statistical power of 2500 runs model-X knockoff on the *same* synthetic dataset; $n = 400, p = 600$. FDR is controlled at level $\alpha = 0.1$. As stated above, the FDP varies substantially and can be much higher that α. But the power also varies dramatically, and is even stays at zero in a fraction of the simulations.

2.6 Outputting valid p-values in high-dimension problem via Conditional Randomization Testing

Introduced in the same work that presented model-X knockoffs [CFJL18], the *Conditional Randomization Test (CRT)* is an alternative approach for multiple testing that makes use of the knockoff variable. The idea is to have a randomization test that first draws a new sample $\tilde{\mathbf{x}}_j$ of each variable $\mathbf{x}_j$ from the conditional distribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$. Then, let $T_j \overset{\text{def.}}{=} T(\mathbf{x}_j, \mathbf{y})$ be a test statistic calculated using $\mathbf{x}_j$ and $\mathbf{y}$. The next step is to find this statistic, usually by solving an optimization problem with regularization. In fact, these first two steps are similar to those of knockoff inference: sampling noisy copies of $\mathbf{x}_j$ and calculating knockoff statistics with the Lasso program. However, contrary to knockoff filter, the purpose of CRT as a type of randomization test is to output an empirical p-value by running these two steps multiple times. We then count the number of noisy test scores $\tilde{T}_j \overset{\text{def.}}{=} T(\tilde{\mathbf{x}}_j, \mathbf{y})$ that are greater than the original test score T_j. The idea is that under the null hypothesis of conditional independence between $\mathbf{y}$ and $\mathbf{x}_j$ given $\mathbf{x}_{-j}$, T_j and $\tilde{T}_j$ are identically distributed, therefore, by taking

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B} \tag{2.20}$$

where B is the number of sampling, we can have a good measure of evidence against the null if $\hat{p}_j$ is small, which translates to a significantly larger value of T_j against $\tilde{T}_j$, conditionally to $\mathbf{x}_{-j}$. Note that this empirical p-values formula is based on the same principle of p-values of the permutation test. A full description of CRT is presented in Algorithm 2.1.

Algorithm 2.1: Conditional Randomization Test [CFJL18]

INPUT dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^p$, number of sampling run B, test statistic T_j, conditional ditribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$ for each $j = 1, \ldots, p$;

OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;

for $j = 1, 2, \ldots, p$ **do**

  for $b = 1, 2, \ldots, B$ **do**

    1. Generate $\tilde{\mathbf{x}}_j^{(b)}$, a knockoff sample from $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$;

    2. Compute test statistics for original variable T_j and for knockoff variables $\tilde{T}_j$;

  end

  Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

The CRT is a powerful method in the sense that we can sample from any type of conditional distribution of any test statistic, a flexibility that one also has with knockoff filter. However, as with most types of randomization tests, CRT is costly to perform. It requires computing randomized p-values for all variables and for a large enough number B of samplings. Moreover, for each sampling and calculation of test statistics T_j, we would have to take into account the full dimensionality of the data, which means a time complexity of $\mathcal{O}(p^3)$ is required. Therefore, the time complexity of CRT is $\mathcal{O}(Bp^4)$, which is prohibitive. Reducing the computation cost, hence, is one of the main goals for some extensions of the CRT. We will briefly reviewing three of the variants of CRT, called Holdout Randomization Test, Conditional Permutation Test and Distilled Conditional Randomization Test.

Conditional Permutation Test [BWBS19] Perhaps the closest cousin, the Conditional Permutation Test (CPT) consists almost the same number of steps as CRT, with the exception that instead of using the conditional distribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$, we apply a permutation function to original variables to create $\tilde{\mathbf{x}}_j$. To ensure the flip-sign property of the test statistic T_j and $\tilde{T}_j$, this permutation function must satisfy some specific properties, defined in the same work. The authors also provided arguments that CPT is more robust than CRT in practice, and the permutation scheme in some cases can be faster than sampling from the conditional distribution of the variables.

Holdout Randomization Test [TVZ+18]. The basic idea behind Holdout Randomization Test (HRT) is to choose the test statistic $\tilde{T}_j$s in a way that minimizes the calculation cost for each of the sampling run. More specifically, the difference is that we first split the dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ into two parts: train data $\mathcal{D}_{train}$ and test data $\mathcal{D}_{test}$. The train dataset will be used to find an estimation $\hat{\boldsymbol{\beta}}$ of true parameter $\boldsymbol{\beta}^0$. We then use $\mathcal{D}_{test}$ to evaluate the empirical risk function, denoted $L(\mathcal{D}_{test}, \hat{\boldsymbol{\beta}})$, which will also be used as the test statistic T_j to measure variable importance. The calculation of empirical p-values is similar to CRT: for each variable j, we run knockoff sampling for B times, using the conditional distribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$. Then in each run, we replace samples of $\mathbf{x}_j$ in $\mathcal{D}_{test}$ with that of $\tilde{\mathbf{x}}_j$, and recalculate the test statistic, or empirical risk $L(\tilde{\mathcal{D}}_{test}, \hat{\boldsymbol{\beta}})$. A similar formula for empirical p-values with the CRT is then computed. The pseudo-code of HRT can be found in Algorithm 2.2.

Algorithm 2.2: Holdout Randomization Test [TVZ+18]

INPUT dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, a empirical risk function $L(\cdot)$, number of sampling run B, test statistic T_j, conditional ditribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$ for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\boldsymbol{\pi} = \{\hat{p}_j\}_{j=1}^p$;
Split $\mathcal{D}$ into two parts: $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$;
$\hat{\boldsymbol{\beta}} \leftarrow \texttt{fit}(\mathcal{D}_{train})$;
Compute empirical risk, used as test statistic: $T_j \overset{\text{def.}}{=} L(\mathcal{D}_{test}, \hat{\boldsymbol{\beta}})$;
for $j = 1, 2, \ldots, p$ **do**
  for $b = 1, 2, \ldots, B$ **do**
    1. Generate $\tilde{\mathbf{x}}_j^{(b)}$, a knockoff sample from $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$;
    2. Replace samples of $\mathbf{x}_j$ in $\mathcal{D}_{test}$ with $\tilde{\mathbf{x}}_j^{(b)}$ to form $\tilde{\mathcal{D}}_{test}$;
    3. $\tilde{T}_j \overset{\text{def.}}{=} L(\tilde{\mathcal{D}}_{test}, \hat{\boldsymbol{\beta}})$;
  end
  Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j \geq T_j}}{1 + B}$$

end

The HRT, while improving computation time by eliminating the estimator fitting part for each sample, still requires sampling new noisy copies for each variable j in each sampling run b. In large scale datasets, this could still be very costly. Moreover, HRT poses a different problem that CRT does not have: the split of the dataset into two parts can potentially make HRT lose a fair amount of statistical power.

Distilled Conditional Randomization Test [LKJR20] The Distilled CRT (dCRT) aims to fix the prohibitive computation cost of vanilla CRT. This is done through a process called *distillation operator*. The key idea behind distillation is that important information of a variable $\mathbf{x}_j$ is "distilled" back to the response $\mathbf{y}$ and to the remaining $j-1$ variables, i.e. the data matrix $\mathbf{X}_{-j}$. More formally, for each variable j, we perform the distillation by first solving two lasso problems:

$$\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{2.21}$$

and

$$\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \tag{2.22}$$

Then, a test statistic can be calculated by the formula

$$T_j \overset{\text{def.}}{=} T(x_j, y) = \left[(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T (\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})\right]^2 . \tag{2.23}$$

A nice additional property of this procedure is that it can directly output an analytical p-value, since feature-importance statistic T_j is assumed to follow Gaussian distribution, conditionally to $\mathbf{y}$ and $\mathbf{X}_{-j}$ [LKJR20]:

$$(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T (\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}}) \sim \mathcal{N}(0, \sigma^2 \left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|^2). \tag{2.24}$$

It follows that we can output the exact p-value for each variable j by taking

$$p_j = 2\left[1 - \Phi\left(\frac{\sqrt{T_j}}{\hat{\sigma}\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|}\right)\right], \tag{2.25}$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF) and $\hat{\sigma}$ is the estimated noise magnitude.

Although [LKJR20] framed dCRT as a general framework for outputting p-values with flexible choice of distillation operator, what we presented above is a variant called the Lasso Distillation CRT (Lasso-dCRT), which the authors of the work have mostly focused on in both theoretical and empirical arguments. This means that we work under the linear-relationship of Eq.(2.3). Lasso-dCRT gains on vanilla CRT by removing multiple sampling steps for each variable. This leads to a reasonable reduction of the computation cost. However, the reader can still notice that the runtime complexity is still a polynomial of order 4 of p, or $\mathcal{O}(p^4)$. To deal with this problem, [LKJR20] proposed an initial screening step with the assumption that all relevant variables are selected during this step, and set all p-values of unselected variables to 1. The resulting runtime is reduced to $\mathcal{O}(kp^3)$, where k is the total number of selected variables post-screening and is often much smaller than p in sparsity regime. More formally, the screening step returns the estimation of the support $\hat{\mathcal{S}} := \{j \in [p] : \hat{\beta}_j \neq 0\}$, and k is the cardinality of this set. A full algorithm block is given in Algorithm 2.3.

Algorithm 2.3: Distilled Conditional Randomization Test (dCRT) [LKJR20]

INPUT dataset $\mathcal{D} = \{\mathbf{x_i}, y_i\}_{i=1}^n$, test statistic T_j for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{p_j\}_{j=1}^p$;
$\hat{\mathcal{S}}^{\text{SCREENING}} = \{j : j \in [p], \hat{\beta}_j^{\text{LASSO}} \neq 0\}$;
for $j \in \hat{\mathcal{S}}^{SCREENING}$ **do**
    1. Distill information of $\mathbf{X}_{-j}$ to $\mathbf{x}_j$ and to $\mathbf{y}$ by finding:
    - $\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$
    - $\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$

    2. Obtain test statistic:

$$T_j = \left[(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})\right]^2$$

    3. Compute (two-sided) p-value

$$p_j = 2\left[1 - \Phi\left(\frac{\sqrt{T_j}}{\hat{\sigma}\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|}\right)\right]$$

end

Advantages and drawbacks of dCRT To our knowledge, among all the conditional randomization test variants, dCRT is the only one that introduces an analytical step to calculate p-values instead of having to running multiple sampling/permutation. Therefore, the computational time is reduced significantly. However, as we have already noted, the distillation formula proposed in Eq. (2.22) and Eq. (2.21) only works in the linear-relationship setting of Eq (2.3). Moreover, the analytical formula of p-values relies on the assumption that the test-statistics T_j to be follow standard normal distribution. This assumption is rather strong and a deviation from it, which is often encountered in real-life applications, would make the procedure incorrect. In Chapter 6, we investigate further this phenomenon in high-dimensional logistic regression. The idea is to decorrelate the formula of the dCRT test-statistic to take

into account the non-linear relationship.

2.7 Multiple Random-sampling Variable Selection

2.7.1 Data Multi-Split and P-Value Aggregation

As discussed in Section 2.4.2, the data single-split procedure faces a problem of low statistical power if there are not enough samples. Moreover, we have to rely on the assumption that the screening procedure using one half of the data would output a superset of the support $\mathcal{S}$, or the set of all truly predictive variables. At the same time, when performing a single-split of the data, we rely on pure chance for this to happen in realistic scenario: different data splits might produce different inference results, a problem similar to the one we have discussed with the knockoff filter in section 2.5. To stabilize the inference result, the seminal work of [MMB09] suggested performing multiple splits of data instead of a single one, and having an aggregation of several p-values for each variable. Assuming for each variable $j = 1, \dots, p$, we have B p-values $p_j^{(b)}$ corresponding to B runs of inference, then [MMB09] argued a suitable function for aggregating these p-values relies on a quantile:

$$\bar{p}_j = \min \left\{ 1, \frac{q_\gamma\big(\{p_j^{(b)} : b \in [B]\}\big)}{\gamma} \right\}, \tag{2.26}$$

whereby $q_\gamma(\cdot)$ is the empirical γ-quantile function with fixed $\gamma \in (0, 1)$. Since choosing a reasonable γ might be difficult, [MMB09] also proposed an adaptive aggregation version:

$$\bar{p}_{j,\text{adapt}} = \min \left\{ 1, (1 - \log \gamma_{min}) \inf_{\gamma \in (\gamma_{min}, 1)} \bar{p}_j \right\}, \tag{2.27}$$

with $\gamma_{min} \in (0, 1)$ is the lower bound for γ. The full procedure is summarized in Algorithm 2.4.

Algorithm 2.4: Multi-split inference [MMB09]

INPUT dataset $\mathcal{D} = \{\mathbf{x_i}, y_i\}_{i=1}^n$, parameter $\gamma > 0$ or $\gamma_{min} > 0$;
OUTPUT vector of p-values $\{\bar{p}_j\}_{j=1}^p$;
for $b = 1, \dots, B$ **do**
- Randomly split $\mathcal{D}$ into two parts of equal size $\mathcal{D}_{in}^{(b)}$ and $\mathcal{D}_{out}^{(b)}$;
- Use $\mathcal{D}_{in}^{(b)}$ to find $\hat{\mathcal{S}}^{(b)}$, an estimation of true support $\mathcal{S}$;
- Use $\mathcal{D}_{out}^{(b)}$, fit variables in $\hat{\mathcal{S}}^{(b)}$ with OLS regression, output p-values $\{p_j^{(b)}\}_{j \in \hat{\mathcal{S}}^{(b)}}$, set $\{p_j^{(b)}\}_{j \notin \hat{\mathcal{S}}^{(b)}} = 1$;

end
$\bar{p}_j \leftarrow$ `aggregate`$\{p_j^{(b)}\}_{b \in [B]}$ for each $j = 1, \dots, p$, with either eq. (2.26) or eq. (2.27)

Although the multi-split procedure still relies on the assumption that the screening step with $\mathcal{D}_{in}^{(b)}$ produces $\hat{\mathcal{S}}^{(b)} \supseteq \mathcal{S}$, running B splits and aggregating results later produces a more stable solution. In other words, aggregation step makes the assumption on screening set less problematic in real settings. The authors also proved theoretical guarantees of the aggregated p-values usage in FWER and FDR controlling procedure. We shall see in later Chapter 4 and Chapter 5 that the stabilization of knockoff inference is heavily influenced by this p-value aggregation technique.

Remark 2.1. *For an excellent review of the topic on data-splitting and p-value based aggregation techniques, the reader can refer to the work of [RD19a].*

2.7.2 Ensemble of Randomized Clusters

In some of the modern applications of multiple hypothesis testing, the number of hypotheses needed to test simultaneously is so large that doing any type of estimation, even with regularization, is both prohibitive computationally and statistically powerless. For example, when doing brain-imaging data analysis, the typical dimension is over 100,000 of brain-voxels, while the number of samples can only be around 100-times less than that due to constraints in data-acquisition step. One of the popular solutions is to reduce the dimension of the dataset by clustering, *i.e.* grouping correlated neighboring variables. The relevant literature on this topic includes [BRvdGZ13], which provides a combination of statistical inference and fixed clustering; [VGT12] suggests clustering on a subsample the data to create randomized clusters, then do variable selection. Taking inspiration from the latter work and from the p-value aggregation technique presented in the previous section 2.7.1, a more recent line of works of [CST18, CNS+21, CNTS21] proposes an ensemble of clusters inference approach for a more stable solution. In particular, this algorithm combines several inference results on spatially constrained clustering, then aggregates on these cluster-level solutions. A pseudo-code for this algorithm is presented in Algorithm 2.5.

Algorithm 2.5: Mulitple testing with ensemble of randomized clusters [CNTS21]

INPUT dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^p$, parameter B the number of subsampling runs, γ or γ_{min} the choice of quantile;

OUTPUT vector of p-values $\{\bar{p}_j\}_{j=1}^p$;

for $b = 1, \ldots, B$ **do**

- $\tilde{\mathcal{D}}^{(b)} \leftarrow \texttt{subsampling}(\mathcal{D})$ // $\tilde{\mathcal{D}}^{(b)} = \{\mathbf{x}_i, y_i\}_{i \in \mathcal{K}}$ where $\mathcal{K} \subset [n]$
- $\mathcal{Z}^{(b)} = \{\mathbf{z}_i, y_i\}_{i \in \mathcal{K}} \leftarrow \texttt{clustering}(\tilde{\mathcal{D}}^{(b)})$ // $\mathbf{z}_i \in \mathbb{R}^k$ where $k \ll p$
- $\{q_j^{(b)}\}_{j \in [k]} \leftarrow \texttt{inference}(\mathcal{Z}^{(b)})$ // cluster-wise p-values
- $\{p_j^{(b)}\}_{j=1}^p \leftarrow \texttt{broadcast}\left(\{q_j^{(b)}\}_{j \in [k]}, \mathcal{Z}^{(b)}, \tilde{\mathcal{D}}^{(b)}\right)$ // variables of same cluster get same p-value

end

$\bar{p}_j \leftarrow \texttt{aggregate}\{p_j^{(b)}\}_{b \in [B]}$ for each $j = 1, \ldots, p$, with either eq. (2.26) or eq. (2.27)

There are several ways to group variables into lower-resolution clusters, but we stick with hierarchical agglomerative clustering with Ward's linkage, a popular choice for clustering in both brain-imaging and genomic applications [MGV+12, VGT12, RVN21]. Note that when we compress the data by clustering (replacing initial features by the cluster representatives), statistical inference at the initial feature-level resolution will be difficult. Therefore, in order to maintain error control, we have to add a spatial tolerance, denoted $\delta > 0$. This means that null variables rejected at a distance closer than δ from the truly significant variables would not be considered as false positives. The error control metrics FDR and FWER are also be reformulated to take into account of with this quantity: the work of [CST18, CNTS21] dealing with δ-FWER; [NCT19, GZ19a] dealing with FDR$^{\delta}$. More specifically, in chapters 5 and 6, we introduce the extensions of this algorithm for knockoff filter and for conditional randomization test.

Chapter 3

Neuroimaging Data Analysis

Summary. This chapter gives a brief introduction of neuroimaging data-analysis and the use of multiple-testing in this field. In particular, we provide some background information on functional Magnetic Resonance Imaging (fMRI), one of the major techniques in functional neuroimaging. We present the data processing pipeline for fMRI, from acquisition to preprocessing and performing statistical analysis. We revisit the two approaches for statistical analysis: univariate and multivariate inference, with the particular application to neuroimaging analysis in encoding and decoding tasks. For a more in-depth survey of this topic, readers can refer to [Lin08] or the book of [PMN12].

3.1 Functional Magnetic Resonance Imaging (fMRI)

Neuroimaging refers to the use of various techniques to either directly or indirectly image brain structure and function. The field has important applications in clinical neuroscience. MRI has been utilised to study brain structures in patients with neurological or phsychiatric diseases, which may be useful for diagnosis, or to understand the mechanisms that underlie the disease. In this thesis, we will focus on the application with functional Magnetic Resonance Imaging (MRI), one of the major neuroimaging techniques. As its name suggests, this technique uses magnetic resonance imaging to measure brain activity by measuring changes in the local oxygenation of blood, which can be understood as a proxy for local brain activity. Functional MRI is a specific MRI contrast, relying on the BOLD signal to image brain activations. It can be used to characterize the function of various brain regions or networks and the mental processes in which they are involved [VSP+18]. A typical experimental setup for fMRI can be such as showing the subjects a picture of an object, or asking them to perform a certain task, such as reading a sentence out loud. The ultimate goal is to investigate how certain regions of the brain are activated in response to those particular tasks. Figure 3.1 illustrates the common pipeline when doing fMRI data processing, from acquisition to statistical analysis.

Acquisition The most common method of fMRI, introduced by [OLKT90] in the early 90s. As stated above, this method relies on the fact that when neurons in the brain become active, the amount of blood flowing through that area is increased, which in turn increase the level of blood oxygenation. It is called the Blood Oxygenation Level Dependent (BOLD) signal. With an MRI machine, we can detect this change in oxygenation, and it would be the signal used in fMRI. FMRI studies therefore first involve recording the BOLD signal in the brains of subjects in a scanner. In this first step, called acquisition, the subject, lies in an MRI scanner and performs a predefined task. Typically, in each time interval of 2-seconds called repetition time (TR), a 3-dimensional image of the subject's brain will acquired, This

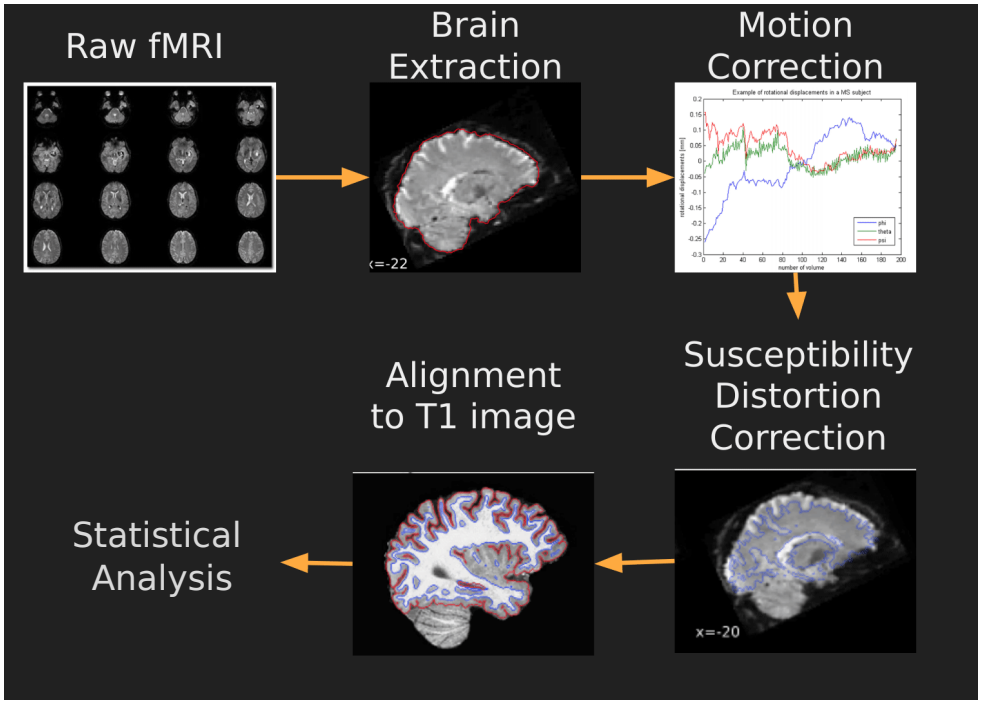


Figure 3.1: Common fMRI data processing pipeline. Picture taken from https://carpentries-incubator.github.io/SDC-BIDS-fMRI/aio/index.html.

image contains a BOLD signal measurement for each pixel, more commonly called brain voxel in neuroimaging literature. The whole acquisition therefore results in a 4-dimensional data array.

Remark 3.1. *The indirect measurement of MRI is in contrast with other neuroimaging techniques, like Electro-Encephalography (EEG) and Magneto-Encephalography (MEG), which directly measures electric of magnetic field created by neuronal activity.*

Preprocessing Raw fMRI data suffers from limitations and display some artifacts, that can partly be corrected by suitable preprocessing steps. In particular, the signal can be noisy, due to instrumental errors, head movement, or other physiological processes such as heart beating. Moreover, since fMRI experiments typically involve scanning multiple subjects, there is inherent variability in the data. The statistical analysis of fMRI data involves multiple assumptions, which without preprocessing step would not hold. We need to do *slice timing correction*, because it is typically assumed that the voxels in the brain volume are acquired at the same time, while in reality the BOLD signal is acquired in sequentially generated slices: a temporal interpolation is necessary to solve this timing issue. The artifacts created by head movement could be reduced by performing *realignment*, or finding the best possible alignment between input image and some target image (*e.g.* the mean image). Then, motion correction consists in estimating rigid body movement estimation of the brain volume. We can also perform co-registration, which is a rigid registration (6 degrees of freedom) performed to get all fMRI images in the same spatial position. For group analysis in functional MRI, we need to assume the same anatomical brain region for all subjects. However, this requires a correction for the different head position during acquisition, then a compensation for the variability across individuals in brain size and shape. Therefore, one performs normalization, ie register each subject's brain anatomy to a standard template, usually the Montreal Neurological Institute (MNI) template, by using some type of nonlinear mapping. This step is important for a

consistent analysis of the fMRI data. It is also a common practice to apply spatial smoothing with the data, usually with a Gaussian kernel, since it removes the inter-subject variability.

3.2 Statistical Analysis of fMRI Data

After the raw data is preprocessed, we can perform the statistical analysis. It is often composed of two successive steps: first-level analysis of data from each individual subject, followed by second-level analysis in which we combine multiple results obtained across subjects.

3.2.1 First-level Analysis

The goal of the first-level analysis is to estimate how each predictor (mental condition descriptors) contributes to the variability observed in the voxel's BOLD signal time-series. The most popular method to model BOLD signal is Generalized Linear Model (GLM), first used by [FHW+94]. Similar to the linear relationship introduced in Eq. 2.3, in this approach, the time-series associated with each voxel is modeled as a weighted sum of one or more known predictor variables plus an error term. More formally, we have

$$\mathbf{X} = \mathbf{Y}\mathbf{B} + \mathbf{E}, \tag{3.1}$$

where $\mathbf{X} \in \mathbb{R}^{T\times p}$ is the matrix of BOLD signal, $\mathbf{Y} \in \mathbb{R}^{T\times m}$ the design matrix contains descriptors of mental conditions, $\mathbf{B} \in \mathbb{R}^{m\times p}$ the parameters (regression coefficients) needed to estimate, and $\mathbf{E} \in \mathbb{R}^{n\times m}$ the error term (noise). Here p is the total number of brain voxels, m the total number of conditions, T the length of the experiment. To be more precise, design matrix $\mathbf{Y}$ is the result of convolving neuronal activities in the volume of interest with $h(\cdot)$, the *Haemodynamic response function* (HRF), acting as a convolution operator:

$$\mathbf{Y} \overset{\text{def.}}{=} h \star \mathbf{D}, \tag{3.2}$$

where $\mathbf{D}$ is the mental condition descriptor, or neuronal activities in voxels, before HRF convolution. With such linear equation, if $\mathbf{Y}$ is full rank, a straightforward solution for the estimator $\hat{\mathbf{B}}$ is OLS

$$\hat{\mathbf{B}} = (\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}\mathbf{X}. \tag{3.3}$$

After obtaining $\hat{\mathbf{B}}$, we calculate the z-maps of the brain, or the map contains normalized test statistics for each voxel. The reason for this is the same as for any type of hypothesis test: the effect size, or magnitude of $\hat{\mathbf{B}}$, rarely contains meaningful information without knowing its statistical significance. A straightforward way is to apply mass-univariate inference technique: we want to see if each voxel is truly activated given the mental condition, *without* taking into account correlation with other voxels [Lin08, PMN12]. First, for each voxel j and condition c, the (co)variance of $\hat{\mathbf{B}}_{c,j}$ is estimated by

$$v_{c,j} = \hat{\sigma}_j(\mathbf{Y}^T\mathbf{Y})^{-1}_{c,c} \tag{3.4}$$

with $\hat{\sigma}_j = \|\mathbf{X}_{:,j} - \mathbf{X}\hat{\mathbf{B}}_{:,j}\|_2^2/(p-m)$ the estimated residual variance. Then the t-statistics of voxel j and with each condition c is

$$T_{c,j} = v_{c,j}^{-1/2}\hat{\mathbf{B}}_{c,j}, \tag{3.5}$$

and we convert the t-statistics to z-scores to remove the dependence on degrees of freedom, as the t-statistics follow a Student-t distribution with $T-p-1$ degrees of freedom under the null hypothesis. The matrix $\mathbf{Z} \in \mathbb{R}^{p\times m}$ consists of z-scores are called z-maps.

Applications of Multiple Hypothesis Testing for fMRI data-analysis The z-maps output from first level analysis may be used to select voxels of interest for a given stimuli from the univariate point of view. For example, we may want to test whether the voxels of the map $Z_c \in \mathbb{R}^p$ that are activated given the mental simulation c at significant level $\alpha < 0.05$. Since the z-score follows standard normal distribution, the conversion to p-values is straightforward, and we can compare p-values with a significance level α. We thus obtain a selection of voxels that are triggered preferentially by condition c, and a localization of the brain regions recruited by this condition. However, as presented in Example 2.4 of previous chapter, such a procedure will yield many false positives, since the number of tests is too large (total of p number of tests) compared with the number of observations. This is where multiple hypothesis testing theory has to be used: we need to correct p-values for controlling multiple testing metric, such as FWER or FDR. Figure 3.2 is a good visualisation of this issue. The raw z-scores provide qualitative information, but do not give any statistical guarantee for the significance of the voxel. Although controlling at the same significance level $\alpha < 0.05$, the next three figures show different results. Figure 3.2b is thresholded z-maps with unadjusted p-values, and is quite spurious compare to the next two figures, where we adjust p-values to control FDR and FWER. The figure also highlights the problem of over-conservativeness of Bonferroni's procedure for controlling FWER: almost all the tests are declared as non-significant, *i.e.* most of the voxels are not declared as activated with the presence of stimuli.

Remark 3.1 (All-resolution inference for neuro-imaging)**.** *As briefly mentioned in section 2.2.2, there exists a recent line of works of [GS11, GMKS19] and [BNR20], which introduced an alternative multiple-testing procedure called* All-Resolutions Inference *(ARI). In short, an advantage of ARI compared to traditional procedures for controlling FWER/FDR is based on a "user-agnostic" principle: she chooses the subsets (collections) of rejected hypotheses, then ARI returns the associated quality criterion. The framework provides statistical guarantee* simultaneously *for a given multiple testing metric,* e.g. *FWER, for any such possible subset of the brain, no matter large or small, using closed testing procedure of [MEG76]. In other words, the user can apply any data-driven region selection and estimate the proportion of true discoveries of any subregion, or clusters, from the same dataset. This is especially relevant for neuro-imaging, where we want to quantify the proportion of truly active voxels within selected clusters, in contrast with the dimension-reduction technique presented in section 2.7.2, where we would only be able to select active clusters of brain voxels, or* low-resolution *inference. Indeed, there exists the work of [RFW+18] which discussed at length the application of ARI framework for neuro-imaging. We would not go into all the detail of ARI here, but the reader can refer to aforementioned works for a better understanding of this framework.*

3.2.2 Second-level Analysis

The main goal of second-level analysis is to study the link between brain activation and a condition (a behavior or a disease status) in a group of subjects. However, it depends whether we want to predict brain activity maps from a condition, *i.e.* encoding, or the opposite: predict conditions from brain activity maps, *i.e.* decoding.

Encoding The objective of encoding is to relate the presented stimuli, or mental conditions, to the subject's brain activity. To do so, we first need z-maps for several subjects, which can be obtained following the procedure in first-level analysis. We will assume there is a total of n subjects participate in the study. We first select two conditions c_1 and c_2 then stack the corresponding z-maps Z_1 and Z_2 of all subjects row-wise, forming the matrix $\bar{Z} \in \mathbb{R}^{2n \times p}$. With the same notation of $\mathbf{y} \in \mathbb{R}^{2n}$ the vector contains descriptors of mental conditions, the problem becomes

$$\bar{Z} = \mathbf{y}\boldsymbol{\beta}^T + \boldsymbol{\varepsilon}, \tag{3.6}$$

in which $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of coefficient contains information relates the z-maps with mental conditions, and $\boldsymbol{\varepsilon}$ the noise vector. Note that the design can be more complex, including covariates (age, sex, demographic information, genetics or behavioral data) to study interactions between these and the experimental manipulation. In that case, the design becomes a matrix $\boldsymbol{Y}$ Using the same technique as in first-level analysis, we can find the estimation of $\boldsymbol{\beta}$. Notice that because the z-maps and $\boldsymbol{\beta}$ is fitted per single brain voxel independently, we are still in the regime of univariate analysis. This setting, or marginal analysis of encoding, is problematic in the sense that for a specific task, conditional effects, or interactions between voxels, are not taken into account [HGF+01, CS03, WP20].

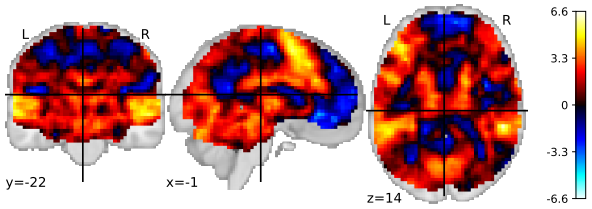
Decoding In contrast, by using the whole brain images as input, we can infer more information. More specifically, we want to do hypothesis test on each brain voxel with the mental stimulation, *conditioned on all other voxels*. This is the purpose of decoding, a type of multivariate inference. It has been presented in detail in Chapter 2, where the reader can refer to. Usually, this input for decoding task are the maps of estimated coefficients $\hat{\beta}$ or z-maps $\mathbf{Z}$ obtained in the first-level analysis, stacked together by rows. The targets $\mathbf{y}$ are mental conditions (stimuli/behaviour), similar to encoding task. They can be either a binary/discrete form of labels, which corresponds to category in classification problem, for example in [HGF+01] where we want to classify brain maps associated with watching face versus house, or continuous form of regression problem, as in [MWP+07] using brain maps to predict age of subjects. The linear relationship assumption is also widely used for decoding:

$$\mathbf{y} = \mathbf{Z}\mathbf{w}^0 + \sigma\boldsymbol{\varepsilon}, \tag{3.7}$$

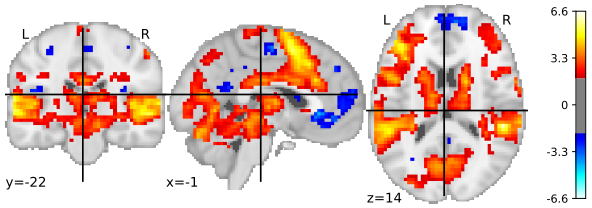
or in binary classification case,

$$\mathbf{y} = sign(\mathbf{Z}\mathbf{w}^0 + \sigma\boldsymbol{\varepsilon}), \tag{3.8}$$

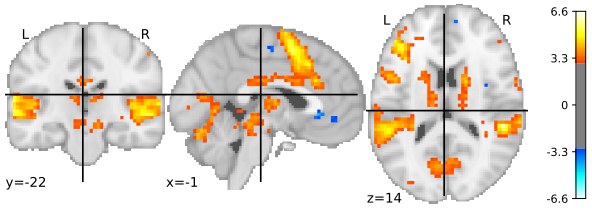
where $\mathbf{w}^0 \in \mathbb{R}^P$ is the true coefficients map, $\boldsymbol{\varepsilon}$ the vector of noise, often assumed to be Gaussian, σ the noise magnitude, and $sign(x)$ is the sign function, outputting -1 if x is negative and $+1$ if x is positive. In this thesis, we focus on this decoding setting. As stated before, fMRI decoding faces problems similar to multivariate inference: the data are very high-dimensional and can be noisy. The number of brain voxels p is typically much larger than the number of observations n: for typical fMRI data, p can be as large as over hundred thousands, while n is only of several hundreds. Moreover, it is often understood that neighboring voxels interacts together given a mental condition, *i.e.* there is strong correlation between variables. Therefore, finding an estimation of $\mathbf{w}^0$ with theoretical guarantees is highly non-trivial, and we also want to perform variable selection/multiple hypothesis. As we will see later in chapters 5 and 6, we mitigate these problems by dimension reduction with clustering, and the use of regularized estimation methods.



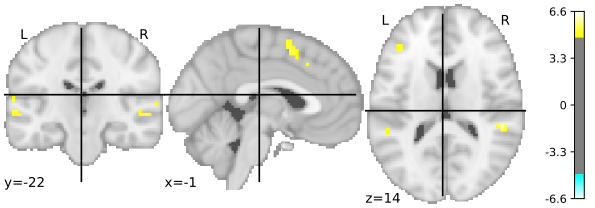
(a) Raw map of z-scores



(b) Unadjusted p-values, controlling False Positive Rate at level $\alpha < 0.05$



(c) Adjusted p-values, controlling FDR at level $\alpha < 0.05$



(d) Adjusted p-values, controlling FWER at level $\alpha < 0.05$

Figure 3.2: An example of hypothesis testing on second-level analysis z-maps of fMRI data. From uppermost to lowermost: raw unthresholded z-maps, thresholded with associated p-value $p < 0.05$, adjusted for multiple correction for controlling FDR at level $\alpha < 0.05$ using BH procedure, and for controlling FWER at level $\alpha < 0.05$ using Bonferroni procedure. Output from running Python example script on Nilearn website (`https://nilearn.github.io/`).

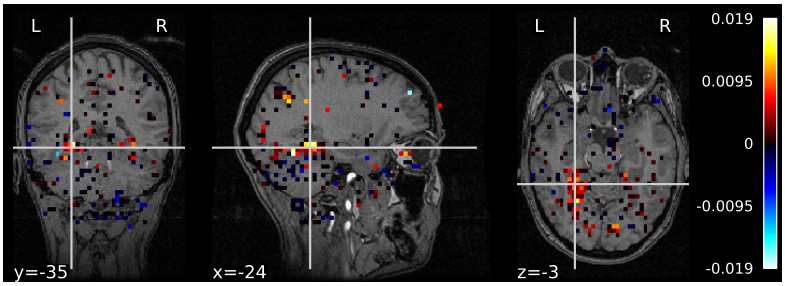


Figure 3.3: An example of decoding weight map for classification task, decoding fMRI activity associated with *face images* vs fMRI activity associated with *cat images* [HGF+01]. Image output from running Python example script on Nilearn website (`https://nilearn.github.io/`).

Part II

Contributions

Chapter 4

Aggregation of Multiple Knockoffs

This chapter has been published in the *Proceedings of 37th International Conference on Machine Learning* (ICML) [NCTA20] as a joint work with Sylvain Arlot, Bertrand Thirion and Jerome-Alexis Chevalier.

Summary. We develop an extension of the knockoff inference procedure, introduced by [BC15]. This new method, called aggregation of multiple knockoffs (AKO), addresses the instability inherent to the random nature of knockoff-based inference. Specifically, AKO improves both the stability and power compared with the original knockoff algorithm while still maintaining guarantees for false discovery rate control. We provide a new inference procedure, prove its core properties, and demonstrate its benefits in a set of experiments on synthetic and real datasets.

4.1 Background

In many fields, multivariate statistical models are used to *fit* some outcome of interest through a combination of measurements or features. For instance, one might predict the likelihood for individuals to declare a certain type of disease based on genotyping information. Besides prediction accuracy, the inference problem consists in defining which measurements carry useful features for prediction. More precisely, we aim at conditional inference (as opposed to marginal inference), that is, analyzing which features carry information *given* the other features. This inference is however very challenging in high-dimensional settings.

Among the few available solutions, knockoff-based (KO) inference [BC15, CFJL18] consists in introducing noisy copies of the original variables that are independent from the outcome conditional on the original variables, and comparing the coefficients of the original variables to those of the knockoff variables. This approach is particularly attractive for several reasons: *i)* it is not tied to a given statistical model, but can work instead for many different multivariate functions, whether linear or not; *ii)* it requires a good generative model for features, but poses few conditions for the validity of inference; and *iii)* it controls the false discovery rate (FDR, [BH95]), a more useful quantity than multiplicity-corrected error rates.

Unfortunately, KO has a major drawback, related to the random nature of the knockoff variables: two different draws yield two different solutions, leading to large, uncontrolled fluctuations in power and false discovery proportion across experiments (see Figure 4.1 below). This makes the ensuing inference irreproducible. An obvious way to fix the problem is to rely on some type of statistical aggregation, in order to consolidate the inference results. Such procedures have been introduced by [GZ19b]

and by [EK19], but they have several limitations: the computational complexity scales poorly with the number B of bootstraps, while the power of the method decreases with B. In high-dimensional settings that we target, these methods are thus only usable with a limited number of bootstraps.

In this work, we explore a different approach, that we call aggregation of multiple knockoffs (AKO): it rests on a reformulation of the original knockoff procedure that introduces intermediate p-values. As it is possible to aggregate such quantities even without assuming independence [MMB09], we propose to perform aggregation at this intermediate step. We first establish the equivalence of AKO with the original knockoff aggregation procedure in case of one bootstrap (Proposition 4.1). Then we show that the FDR is also controlled with AKO (Theorem 4.2). By construction, AKO is more stable than (vanilla) knockoff; we also demonstrate empirical benefits in several examples, using simulated data, but also genetic and brain imaging data. Note that the added knockoff generation and inference steps are embarrassingly parallel, making this procedure no more costly than the original KO inference.

Notation Let $[p]$ denote the set $\{1, 2, \dots, p\}$; for a given set given set $\mathcal{A}$, $|\mathcal{A}| \overset{\text{def.}}{=} \mathbf{card}(\mathcal{A})$; matrices are denoted in bold uppercase letter, while vectors in bold lowercase letter and scalars normal character. An exception for this is the vector of knockoff statistic $\mathbf{W}$, in which we follow the notation from the original paper of [BC15].

Problem Setting Let $\mathbf{X} \in \mathbb{R}^{n \times p}$ be a design matrix corresponding to n observations of p potential explanatory variables $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^p$, with its target vector $\mathbf{y} \in \mathbb{R}^n$. To simplify the exposition, we focus on sparse linear models, as [BC15] and [CFJL18]:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\varepsilon} \tag{4.1}$$

where $\boldsymbol{\beta}^* \in \mathbb{R}^p$ is the true parameter vector, $\sigma \in \mathbb{R}^+$ the unknown noise magnitude, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ some Gaussian noise vector. Yet, it should be noted that the algorithm does not require linearity or sparsity. Our main interest is in finding an estimate $\widehat{\mathcal{S}}$ of the true support set $\mathcal{S} = \{j \in [p] : \beta_j^* \neq 0\}$, or the set of important features that have an effect on the response. As a consequence, the complementary of the support $\mathcal{S}$, which is denoted $\mathcal{S}^c = \{j \in [p] : \beta_j^* = 0\}$, corresponds to null hypotheses. Identifying the relevant features amounts to simultaneously testing

$$\mathcal{H}_0^j : \beta_j^* = 0 \quad \text{versus} \quad \mathcal{H}_a^j : \beta_j^* \neq 0, \quad \forall j = 1, \dots, p.$$

Specifically, we want to bound the proportion of false positives among selected variables, that is, control the false discovery rate (FDR, [BH95]) under certain predefined level α:

$$\text{FDR} = \mathbb{E}\left[\frac{|\widehat{\mathcal{S}} \cap \mathcal{S}^c|}{|\widehat{\mathcal{S}}| \vee 1}\right] \leq \alpha \in (0, 1)\,.$$

Knockoff Inference Introduced originally by [BC15], the knockoff filter is a variable selection method for multivariate models with theoretical control of FDR. [CFJL18] expanded the method to work in the case of (mildly) high-dimensional data, with the assumption that $\mathbf{x} = (x_1, \dots, x_p) \sim P_X$ such that P_X is known. The first step of this procedure involves sampling extra null variables that have a correlation structure similar to that of the original variables, with the following formal definition.

Definition 4.1 (Model-X knockoffs, [CFJL18])**.** *The model-X knockoffs for the family of random variables* $\mathbf{x} = (x_1, \dots, x_p)$ *are a new family of random variables* $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_p)$ *constructed to satisfy the two properties*:

1. *For any subset* $\mathcal{K} \subset \{1, \dots, p\}$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(\mathcal{K})} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$, *where the vector* $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(\mathcal{K})}$ *denotes the swap of entries* x_j *and* $\tilde{x}_j$ *for all* $j \in \mathcal{K}$, *and* $\overset{d}{=}$ *denotes equality in distribution.*

2. $\tilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}$.

A test statistic is then calculated to measure the strength of the original variables versus their knockoff counterpart. We call this the knockoff statistic $\mathbf{W} = \{W_j\}_{j=1}^p$, that must fulfill two important properties.

Definition 4.2 (Knockoff statistic, [CFJL18])**.** *A knockoff statistic* $\mathbf{W} = \{W_j\}_{j \in [p]}$ *is a measure of feature importance that satisfies the two following properties*:

1. *It depends only on* $\mathbf{X}, \tilde{\mathbf{X}}$ *and* $\mathbf{y}$

$$\mathbf{W} = f(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}) .$$

2. *Swapping the original variable column* $\mathbf{x}_j$ *and its knockoff column* $\tilde{\mathbf{x}}_j$ *switches the sign of* W_j*:*

$$W_j([\mathbf{X}, \tilde{\mathbf{X}}]_{swap(S)}, y) = \begin{cases} W_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}) \text{ if } j \in \mathcal{K}^c \\ -W_j([\mathbf{X}, \tilde{\mathbf{X}}], y) \text{ if } j \in \mathcal{K} , \end{cases}$$

for all $\mathcal{K} \subset [p]$.

Following previous works on the analysis of the knockoff properties [ACC17, RRJW20], we make the following assumption about the knockoff statistic. This is necessary for our analysis of knockoff aggregation scheme later on.

Assumption 4.1 (Null distribution of knockoff statistic)**.** The knockoff statistic defined in Definition 4.2 are such that $\{W_j\}_{j \in \mathcal{S}^c}$, are independent and follow the same distribution $\mathbb{P}_0$.

Remark 4.1. *As a consequence of [CFJL18, Lemma 2] regarding the signs of the null* W_j *as i.i.d. coin flips, if Assumption 4.1 holds true, then* $\mathbb{P}_0$ *is symmetric around zero.*

One such example of knockoff statistic is the Lasso-coefficient difference (LCD). The LCD statistic is computed by first making the concatenation of original variable and knockoff variables $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, then solving the Lasso problem [Tib96]:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2p}}{\operatorname{argmin}} \left\{ \frac{1}{2} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{4.2}$$

with $\lambda \in \mathbb{R}$ the regularization parameter, and finally to take:

$$\forall j \in [p] , \qquad W_j = |\widehat{\beta}_j| - |\widehat{\beta}_{j+p}| . \tag{4.3}$$

This quantity measures how strong the coefficient magnitude of each original covariate is against its knockoff, hence the name Lasso-coefficient difference. Clearly, the LCD statistic satisfies the two properties stated in Definition 4.2.

Finally, a threshold for controlling the FDR under given level $\alpha \in (0, 1)$ is calculated:

$$\tau_+ = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \alpha \right\} , \tag{4.4}$$

and the set of selected variables is $\widehat{\mathcal{S}} = \{j \in [p] : W_j \geq \tau_+\}$.

Instability in Inference Results Knockoff inference is a flexible method for multivariate inference in the sense that it can use different loss functions (least squares, logistic, etc.), and use different variable importance statistics. However, a major drawback of the method comes from the random nature of the knockoff variables $\tilde{\mathbf{X}}$ obtained by sampling: different draws yield different solutions (see Figure 4.1 in Section 4.4.1). This is a major issue in practical settings, where knockoff-based inference is used to prove the conditional association between features and outcome.

4.2 Aggregation of Multiple Knockoffs

4.2.1 Algorithm Description

One of the key factors that lead to the extension of the original (vanilla) knockoff filter stems from the observation that knockoff inference can be formulated based on the following quantity.

Definition 4.1 (Intermediate p-value)**.** *Let* $\mathbf{W} = \{W_j\}_{j\in[p]}$ *be a knockoff statistic according to Definition 4.2. For* $j = 1, \ldots, p$*, the intermediate p-value* π_j *is defined as*:

$$\pi_j = \begin{cases} \dfrac{1 + \#\{k : W_k \leq -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \leq 0\,. \end{cases} \tag{4.5}$$

We first compute B draws of knockoff variables, and then knockoff statistics. Using Eq. (4.5), we derive the corresponding empirical p-values $\pi_j^{(b)}$, for all $j \in [p]$ and $b \in [B]$. Then, we aggregate them for each variable j in parallel, using the quantile aggregation procedure introduced by [MMB09]:

$$\bar{\pi}_j = \min\left\{1, \frac{q_\gamma\big(\{\pi_j^{(b)} : b \in [B]\}\big)}{\gamma}\right\} \tag{4.6}$$

where $q_\gamma(\cdot)$ is the γ-quantile function. In the experiments, we fix $\gamma = 0.3$ and $B = 25$. The selection of these default values is explained more thoroughly in Section 4.4.1.

Finally, with a sequence of aggregated p-values $\bar{\pi}_1, \ldots, \bar{\pi}_p$, we use Benjamini-Hochberg step-up procedure (BH, [BH95]) to control the FDR.

Definition 4.2 (BH step-up, [BH95])**.** *Given a list of p-values* $\bar{\pi}_1, \ldots, \bar{\pi}_p$ *and predefined FDR control level* $\alpha \in (0, 1)$*, the Benjamini-Hochberg step-up procedure comprises three steps*:

1. *Order p-values such that:* $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \cdots \leq \bar{\pi}_{(p)}$.
2. *Find:*
$$\widehat{k}_{BH} = \max\left\{k : \bar{\pi}_{(k)} \leq \frac{k\alpha}{p}\right\}. \tag{4.7}$$
3. *Select* $\widehat{\mathcal{S}} = \{j \in [p] : \bar{\pi}_{(j)} \leq \bar{\pi}_{(\widehat{k}_{BH})}\}$.

This procedure controls the FDR, but only under independence or positive-dependence between p-values [BY01]. As a matter of fact, for a strong guarantee of FDR control, one can consider instead a threshold yielding a theoretical control of FDR under arbitrary dependence, such as the one of [BY01]. We call BY step-up the resulting procedure. Yet, we use BH step-up procedure in the experiments of Section 4.4, as we observe empirically that the aggregated p-values $\bar{\pi}_j$ defined in Equation (4.5) does not deviate significantly from independence (details in supplementary material).

Definition 4.3 (BY step-up, [BY01]). *Given an ordered list of p-values as in step 1 of BH step-up* $\bar{\pi}_{(1)} \leq \bar{\pi}_{(2)} \leq \cdots \leq \bar{\pi}_{(p)}$ *and predefined level* $\alpha \in (0,1)$*, the Benjamini-Yekutieli step-up procedure first finds*:

$$\widehat{k}_{BY} = \max \left\{ k \in [p] : \bar{\pi}_{(k)} \leq \frac{k\beta(p)\alpha}{p} \right\}, \tag{4.8}$$

with $\beta(p) = (\sum_{i=1}^{p} 1/i)^{-1}$*, and then selects*

$$\widehat{\mathcal{S}} = \left\{ j \in [p] : \bar{\pi}_{(j)} \leq \bar{\pi}_{(\widehat{k}_{BY})} \right\}.$$

[BR09] later on introduced a general function form for $\beta(p)$ to make BY step-up more flexible. However, because we always have $\beta(p) \leq 1$, this procedure leads to a smaller threshold than BH step-up, thus being more conservative.

Algorithm 4.1: AKO – Aggregation of multiple knockoffs

Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, B – number of bootstraps ; $\alpha \in (0,1)$ – target FDR level

Output: $\widehat{S}_{AKO}$ – Set of selected variables index

for $b = 1$ **to** B **do**
  $\tilde{\mathbf{X}}^{(b)} \leftarrow$ SAMPLING_KNOCKOFF($\mathbf{X}$)
  $\mathbf{W}^{(b)} \leftarrow$ KNOCKOFF_STATISTIC($\mathbf{X}, \tilde{\mathbf{X}}^{(b)}, \mathbf{y}$)
  $\boldsymbol{\pi}^{(b)} \leftarrow$ CONVERT_STATISTIC($\mathbf{W}^{(b)}$) // Using Eq. (4.5)
end for

for $j = 1$ **to** p **do**
  $\bar{\pi}_j \leftarrow$ QUANTILE_AGGREGATION $\left(\{\pi_j^{(b)}\}_{b=1}^{B} \right)$ // Using Eq. (4.6)
end for

$\widehat{k} \leftarrow$ FDR_THRESHOLD($\alpha, (\bar{\pi}_1, \bar{\pi}_2, \ldots, \bar{\pi}_p)$) // Using either Eq. (4.7) or Eq. (4.8)

Return: $\widehat{S}_{AKO} \leftarrow \left\{ j \in [p] : \bar{\pi}_j \leq \bar{\pi}_{\widehat{k}} \right\}$

The AKO procedure is summarized in Algorithm 4.1. We show in the next section that with the introduction of the aggregation step, the procedure offers a guarantee on FDR control under mild hypotheses. Additionally, the numerical experiments of Section 4.4 illustrate that aggregation of multiple knockoffs indeed improves the stability of the knockoff filter, while bringing significant statistical power gains.

4.2.2 Related Work

To our knowledge, up until now, there have been few attempts to stabilize knockoff inference. Earlier work of [SQL15] rests on the same idea of generating multiple knockoff bootstrap as ours, but relies on the linear combination of the so-called *one-bit p-values* (introduced as a means to prove the FDR control in original knockoff work of [BC15]). As such, the method is less flexible since it requires a specific type of knockoff statistic to work. Furthermore, it is unclear how this method would perform in high-dimensional settings, as it was only demonstrated in the case of $n > p$. More recently, the work of [HH18] incorporates directly multiple bootstraps of knockoff statistics for FDR thresholding without the need of p-value conversion. Despite its simplicity and convenience as a way of aggregating knockoffs, our simulation study in Section 4.4.1 demonstrates that this method somehow fails to control FDR in several settings.

In a different direction, [GZ19b] and [EK19] have introduced *simultaneous knockoff* procedure, with the idea of sampling several knockoff copies at the same time

instead of doing the process in parallel as in our work. This, however, induces a prohibitive computational cost when the number of bootstraps increases, as opposed to the AKO algorithm that can use parallel computing to sample multiple bootstraps at the same time. In theory, on top of the fact that sampling knockoffs has cubic complexity on runtime with regards to number of variables p (requires covariance matrix inversion), simultaneous knockoff runtime is of $\mathcal{O}(B^3p^3)$, while for AKO, runtime is only of $\mathcal{O}(Bp^3)$ and $\mathcal{O}(p^3)$ with parallel computing. Moreover, the FDR threshold of simultaneous knockoff is calculated in such a way that it loses statistical power as the number of bootstraps increases, when the sampling scheme of vanilla knockoff by [BC15] is used. We have set up additional experiments in Appendix to illustrate this phenomenon. In addition, the threshold introduced by [EK19] is only proven to have a theoretical control of FDR in the case where $n > p$.

4.3 Theoretical Results

We now state our theoretical results about the AKO procedure.

4.3.1 Equivalence of Aggregated Knockoff with Single Sampling and Vanilla Knockoff

First, when $B = 1$ and $\gamma = 1$, we show that AKO+BH is equivalent to vanilla knockoff.

Proposition 4.1 (Proof in Appendix)**.** *Assume that for all $j, j' = 1, \dots, p$,*

$$\mathbb{P}(W_j = W_{j'}, \quad W_j \neq 0, \quad W_{j'} \neq 0) = 0$$

that is, non-zero LCD statistics are distinct with probability 1. Then, single bootstrap version of aggregation of multiple knockoffs ($B = 1$), using $\gamma = 1$ and BH step-up procedure in Definition 4.2 for calculating FDR threshold, is equivalent to the original knockoff inference by [BC15].

Remark 4.1. *Although Proposition 4.1 relies on the assumption of distinction between non-zero W_js for all $j = 1, \dots, p$, the following lemma establishes that this assumption holds true with probability one for the LCD statistic up to further assumptions.*

Lemma 4.1 (Proof in Appendix)**.** Define the equi-correlation set as:

$$\widehat{J}_\lambda = \left\{ j \in [p] : \mathbf{x}_j^\top (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) = \lambda/2 \right\}$$

with $\widehat{\boldsymbol{\beta}}, \lambda$ defined in Eq. (4.2). Then we have:

$$\mathbb{P}\left(W_j = W_{j'}, W_j \neq 0, W_{j'} \neq 0, \ \operatorname{rank}(X_{\widehat{J}_\lambda}) = |\widehat{J}_\lambda| \right) = 0 \tag{4.9}$$

for all $j, j' \in [p] : j \neq j'$. In other words, assuming $\mathbf{X}_{\widehat{J}_\lambda}$ is full rank, then the event that LCD statistic defined in Eq. (4.3) is distinct for all non-zero value happens almost surely.

4.3.2 Validity of Intermediate P-values

Second, the fact that the π_j are called "intermediate p-values" is justified by the following lemma.

Lemma 4.2. If Assumption 4.1 holds true, and if $|\mathcal{S}^c| \geq 2$, then, for all $j \in \mathcal{S}^c$, the intermediate p-value π_j defined by Eq. (4.5) satisfies:

$$\forall t \in [0, 1] \qquad \mathbb{P}(\pi_j \leq t) \leq \frac{\kappa p}{|\mathcal{S}^c|} t$$

where $\kappa = \dfrac{\sqrt{22} - 2}{7\sqrt{22} - 32} \leq 3.24$.

Proof. The result holds when $t \geq 1$ since $\kappa p \geq p \geq |\mathcal{S}^c|$ and a probability is always smaller than 1. Let us now focus on the case where $t \in [0, 1)$, and define $m = |\mathcal{S}^c| - 1 \geq 1$ by assumption. Let F_0 denote the c.d.f. of $\mathbb{P}_0$, the common distribution of the null statistics $\{W_k\}_{k \in \mathcal{S}^c}$, which exists by Assumption 4.1. Let $j \in \mathcal{S}^c$ be fixed. By definition of π_j, when $W_j > 0$ we have:

$$\begin{aligned}
\pi_j &= \frac{1 + \#\{k \in [p] : W_k \leq -W_j\}}{p} \\
&= \frac{1 + \#\{k \in \mathcal{S} : W_k \leq -W_j\}}{p} \\
&\quad + \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\
&\qquad\qquad \text{(since } W_j > 0 > -W_j\text{)} \\
&\geq \frac{m}{p} \widehat{F}_m(-W_j) + \frac{1}{p} \qquad (4.10)
\end{aligned}$$

where $\forall u \in \mathbb{R}$, $\widehat{F}_m(u) \overset{\text{def.}}{=} \dfrac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq u\}}{m}$ is the empirical cdf of $\{W_k\}_{k \in \setminus \{j\}}$. Therefore, for every $t \in [0, 1)$,

$$\begin{aligned}
&\mathbb{P}(\pi_j \leq t) \\
&= \mathbb{P}(\pi_j \leq t \text{ and } W_j > 0) + \underbrace{\mathbb{P}(\pi_j \leq t \text{ and } W_j \leq 0)}_{=0 \text{ since } \pi_j = 1 \text{ when } W_j \leq 0} \qquad (4.11) \\
&= \mathbb{E}\left[\mathbb{P}(\pi_j \leq t \mid W_j) \mathbb{1}_{W_j > 0}\right] \\
&\leq \mathbb{E}\left[\mathbb{P}\left(\frac{m}{p}\widehat{F}_m(-W_j) + \frac{1}{p} \leq t \mid W_j\right) \mathbb{1}_{W_j > 0}\right] \text{ by (4.10)} \\
&\leq \mathbb{P}\left(\frac{m}{p}\widehat{F}_m(-W_j) + \frac{1}{p} \leq t\right). \qquad (4.12)
\end{aligned}$$

Notice that W_j has a symmetric distribution around 0, as shown by Remark 4.1, that is, $-W_j$ and W_j have the same distribution. Since W_j and $\{W_k\}_{k \in \mathcal{S}^c \setminus \{j\}}$ are independent with the same distribution $\mathbb{P}_0$ by Assumption 4.1, they have the same joint distribution as $F_0^{-1}(U), F_0^{-1}(U_1), \ldots, F_0^{-1}(U_m)$ where $U, U_1, \ldots, U_m$ are independent random variables with uniform distribution over $[0, 1]$, and F_0^{-1} denotes the generalized inverse of F_0. Therefore, Eq. (4.12) can be rewritten as

$$\mathbb{P}(\pi_j \leq t) \leq \mathbb{P}\left(\frac{m}{p}\widetilde{F}_m\big(F_0^{-1}(U)\big) + \frac{1}{p} \leq t\right) \qquad (4.13)$$

$$\text{where} \qquad \forall v \in \mathbb{R}, \qquad \widetilde{F}_m(v) \overset{\text{def.}}{=} \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{F_0^{-1}(U_k) \leq v}.$$

Notice that for every $u \in \mathbb{R}$,

$$\begin{aligned}
\widehat{G}_m(u) &\overset{\text{def.}}{=} \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{U_k \leq u} \leq \frac{1}{m} \sum_{k=1}^{m} \mathbb{1}_{F_0^{-1}(U_k) \leq F_0^{-1}(u)} \\
&= \widetilde{F}_m\big(F_0^{-1}(u)\big)
\end{aligned}$$

since F_0^{-1} is non-decreasing. Therefore, Eq. (4.13) shows that

$$\begin{aligned}
\mathbb{P}(\pi_j \leq t) &\leq \mathbb{P}\left(m\widehat{G}_m(U) \leq tp - 1\right) \\
&= \int_0^1 \mathbb{P}\left(m\widehat{G}_m(u) \leq tp - 1\right) \mathrm{d}u. \qquad (4.14)
\end{aligned}$$

Now, we notice that for every $u \in (0,1)$, $m\widehat{G}_m(u)$ follows a binomial distribution with parameters (m,u). So, a standard application of Bernstein's inequality [BLM13, Eq. 2.10] shows that for every $0 \le x \le u \le 1$,

$$\begin{aligned}\mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) &\le \exp\left(\frac{-m^2(u-x)^2}{2mu + \frac{m(u-x)}{3}}\right) \\ &= \exp\left(\frac{-3mx\left(\frac{u}{x}-1\right)^2}{\frac{7u}{x}-1}\right).\end{aligned}$$

Note that for every $\lambda \in (0, 1/7)$, we have

$$\forall w \ge \frac{1-\lambda}{1-7\lambda} \ge 1\,, \qquad \frac{w-1}{7w-1} \ge \lambda$$

hence $\forall u \ge x\dfrac{1-\lambda}{1-7\lambda}$,

$$\mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) \le \exp\left[-3m\lambda x\left(\frac{u}{x}-1\right)\right].$$

As a consequence, $\forall \lambda \in (0, 1/7)$,

$$\begin{aligned}&\int_0^1 \mathbb{P}\left(m\widehat{G}_m(u) \le mx\right) \mathrm{d}u \\ &\le \frac{1-\lambda}{1-7\lambda}x + \int_{\frac{1-\lambda}{1-7\lambda}x}^1 \exp\big[-3m\lambda(u-x)\big]\mathrm{d}u \\ &\le \frac{1-\lambda}{1-7\lambda}x + \int_{\frac{6\lambda}{1-7\lambda}x}^{+\infty} \exp(-3m\lambda v)\mathrm{d}v \\ &\le \frac{1-\lambda}{1-7\lambda}x + \frac{1}{3m\lambda}\exp\left(-3m\lambda\frac{6\lambda}{1-7\lambda}x\right) \\ &\le \frac{1-\lambda}{1-7\lambda}x + \frac{1}{3m\lambda}\,.\end{aligned}$$

Taking $x = (tp-1)/m$, we obtain from Eq. (4.14) that $\forall \lambda \in (0, 1/7)$

$$\begin{aligned}\mathbb{P}(\pi_j \le t) &\le \frac{1-\lambda}{1-7\lambda}\frac{tp-1}{m} + \frac{1}{3m\lambda} \\ &= \frac{1-\lambda}{1-7\lambda}\frac{tp}{m} + \left(\frac{1}{3\lambda} - \frac{1-\lambda}{1-7\lambda}\right)\frac{1}{m}\,. \qquad (4.15)\end{aligned}$$

Choosing $\lambda = (5-\sqrt{22})/3 \in (0, 1/7)$, we have $\frac{1}{3\lambda} = \frac{1-\lambda}{1-7\lambda}$ hence the result with

$$\kappa = \frac{1-\lambda}{1-7\lambda} = \frac{\sqrt{22}-2}{7\sqrt{22}-32} \le 3.24.$$

□

Remark 4.2. *If the definition of π_j is replaced by*

$$\pi_{j,c} \overset{def.}{=} \begin{cases} \dfrac{c + \#\{k : W_k \le -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \le 0 \end{cases} \qquad (4.16)$$

for some $c > 0$, the above proof also applies and yields an upper bound of the form

$$\forall t \ge 0\,, \qquad \mathbb{P}(\pi_{j,c} \le t) \le \kappa(c)t$$

for some constant $\kappa(c) > 0$. It is then possible to make $\kappa(c)$ as close to 1 as desired, by choosing c large enough. Lemma 4.2 corresponds to the case $c = 1$.

Remark 4.3. *We also prove in Appendix that if $p \to +\infty$ with $|\mathcal{S}| \ll p$, then for every $j \geq 1$ such that $\beta_j^* = 0$, π_j is an asymptotically valid p-value, that is,*

$$\forall t \in [0,1]\,, \qquad \limsup_{p \to +\infty} \mathbb{P}(\pi_j \leq t) \leq t\,. \tag{4.17}$$

Yet, proving our main result (Theorem 4.2) requires a non-asymptotic bound such that the one of Lemma 4.2.

Remark 4.4. *In fact, the factor κ in Lemma 4.2 and later on in Theorem 4.2 can be reduced to 1, with the usage of [RW05, Lemma 1]. We thank Etienne Roquain, one of the reviewers for this thesis, for pointing us to this result. Readers can refer to Sec. 4.6.6 for detailed statements and proofs of the results.*

4.3.3 FDR control for AKO

Finally, the following theorem provides a non-asymptotic guarantee about the FDR of AKO with BY step-up.

Theorem 4.1. If Assumption 4.1 holds true and $|\mathcal{S}^c| \geq 2$, then for any $B \geq 1$ and $\alpha \in (0,1)$, the output $\widehat{\mathcal{S}}_{AKO+BY}$ of aggregation of multiple knockoff (Algorithm 4.1), with the BY step-up procedure, has a FDR controlled as follows:

$$\mathbb{E}\left[\frac{|\widehat{S}_{AKO+BY} \cap \ \mathcal{S}^c|}{|\widehat{S}_{AKO+BY}| \vee 1}\right] \leq \kappa\alpha$$

where $\kappa \leq 3.24$ is defined in Lemma 4.2.

Sketch of the proof. The proof of [MMB09, Theorem 3.3], which itself relies partly on [BY01], can directly be adapted to upper bound the FDR of $\widehat{S}_{AKO+BY}$ in terms of quantities of the form $\mathbb{P}(\pi_j^{(b)} \leq t)$ for $j \in \mathcal{S}^c$ and several $t \geq 0$. Combined with Lemma 4.2, this yields the result. A full proof is provided in Appendix. □

Note that Theorem 4.2 loses a factor κ compared to the nominal FDR level α. This can be solved by changing α into α/κ in the definition of $\widehat{S}_{AKO+BY}$. Nevertheless, in our experiments, we do not use this correction and find that the FDR is still controlled at level α.

4.4 Experiments

Compared Methods We make benchmarks of our proposed method aggregation of multiple knockoffs (AKO) with $B = 25, \gamma = 0.3$ and vanilla knockoff (KO), along with other recent methods for controlling FDR in high-dimensional settings, mentioned in Section 4.2.2: *simultaneous knockoff*, an alternative aggregation scheme for knockoff inference introduced by [GZ19b] (KO-GZ), along with its variant of [EK19] (KO-EK); the *knockoff statistics aggregation* by [HH18] (KO-HH); and *debiased Lasso* (DL-BH) [JJ19].

4.4.1 Synthetic Data

Simulation Setup Our first experiment is a simulation scenario where a design matrix $\mathbf{X}$ ($n = 500, p = 1000$) with its continuous response vector $\mathbf{y}$ are created following a linear model assumption. The matrix is sampled from a multivariate normal distribution of zero mean and covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p \times p}$. We generate $\mathbf{\Sigma}$ as a symmetric Toeplitz matrix that has the structure:

$$\mathbf{\Sigma} = \begin{bmatrix} \rho^0 & \rho^1 & \dots & \rho^{p-1} \\ \rho^1 & \ddots & \dots & \rho^{p-2} \\ \vdots & \dots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \dots & \rho^0 \end{bmatrix}$$

where the $\rho \in (0, 1)$ parameter controls the correlation structure of the design matrix. This means that neighboring variables are strongly correlated to each other, and the correlation decreases with the distance between indices. The true regression coefficient $\boldsymbol{\beta}^*$ vector is picked with a sparsity parameter that controls the proportion of non-zero elements with amplitude 1. The noise $\boldsymbol{\varepsilon}$ is generated to follow $\mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_n)$ with its magnitude $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2/(\text{SNR}\|\boldsymbol{\varepsilon}\|_2)$ controlled by the SNR parameter. The response vector $\mathbf{y}$ is then sampled according to Eq. (4.1). In short, the three main parameters controlling this simulation are correlation ρ, sparsity degree k and signal-to-noise ratio SNR.

Aggregation Helps Stabilizing Vanilla Knockoff To demonstrate the improvement in stability of the aggregated knockoffs, we first do multiple runs of AKO and KO with $\alpha = 0.05$ under *one simulation* of $\mathbf{X}$ and $\mathbf{y}$. In order to guarantee a fair comparison, we compare 100 runs of AKO, each with $B = 25$ bootstraps, with the corresponding 2500 runs of KO. We then plot the histogram of FDP and power in Figure 4.1. For the original knockoff, the false discovery proportion varies widely and has a small proportion of FDP above $0.2 = 4\alpha$. Besides, a fair amount of KO runs returns null power.

On the other hand, AKO not only improves the stability in the result for FDP —the FDR being controlled at the nominal level $\alpha = 0.05$— but it also improves statistical power: in particular, it avoids catastrophic behavior (zero power) encountered with KO.

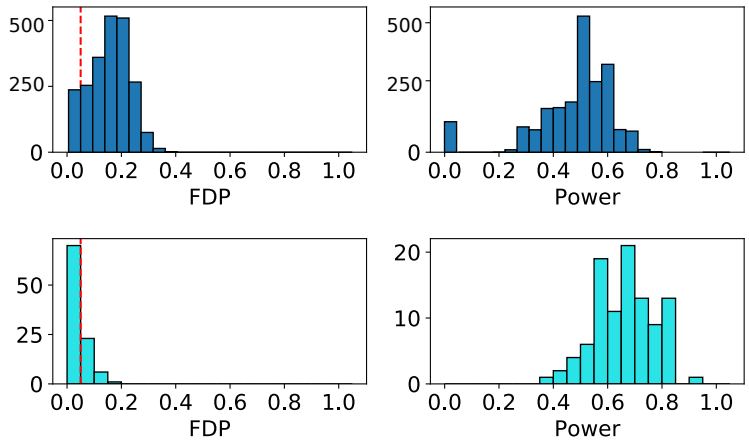


Figure 4.1: **Histogram of FDP and power for 2500 runs of KO (blue, top row) vs. 100 runs of AKO with B = 25 (teal, bottom row) under the same simulation**. Simulation parameter: SNR $= 3.0, \rho = 0.5,$ sparsity $= 0.06$. FDR is controlled at level $\alpha = 0.05$.

Inference Results on Different Simulation Settings To observe how each algorithm performs under various scenarii, we vary each of the three simulation parameters while keeping the others unchanged at default value. The result is shown

in Figure 4.2. Compared with KO, AKO improves statistical power while still controlling the FDR. Noticeably, in the case of very high correlation between nearby variables ($\rho > 0.7$), KO suffers from a drop in average power. The loss also occurs, but is less severe for AKO. Moreover, compared with simultaneous knockoff (KO-GZ), AKO gets better control for FDR and a higher average power in the extreme correlation (high ρ) case. Knockoff statistics aggregation (KO-HH), contrarily, is spurious: it detects numerous truly significant variables with high average statistical power, but at a cost of failure in FDR control, especially when the correlation parameter ρ gets bigger than 0.6. Debiased Lasso (DL-BH) and KO-EK control FDR well in all scenarii, but are the two most conservative procedures.

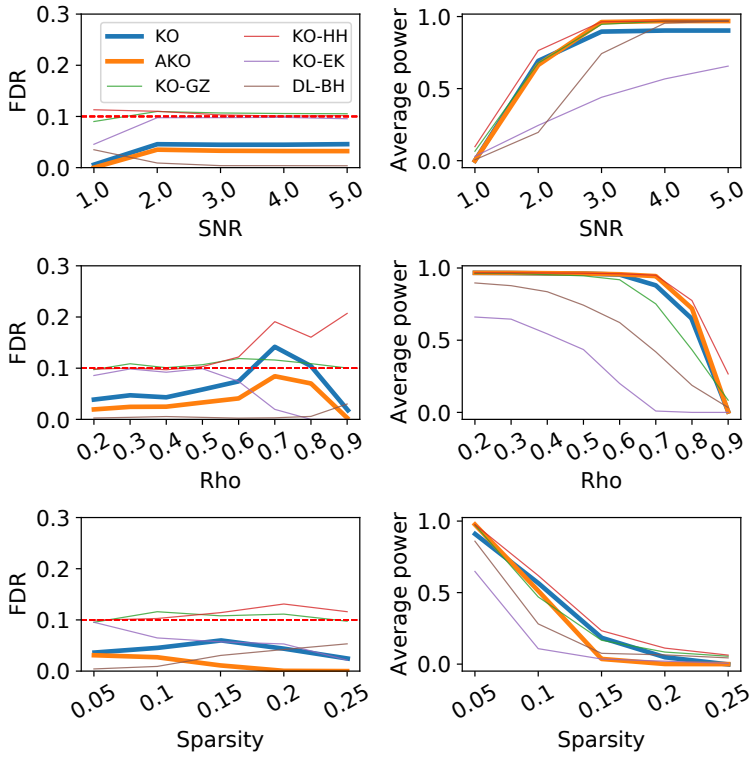


Figure 4.2: **FDR (left) and average power (right) of several methods for 100 runs with varying simulation parameters**. For each varying parameter, we keep the other ones at default value: SNR $= 3.0, \rho = 0.5$, sparsity $= 0.06$. FDR is controlled at level $\alpha = 0.1$. The benchmarked methods are: aggregation of multiple knockoffs (AKO – ours); vanilla knockoff (KO); simultaneous knockoff by [GZ19b] (KO-GZ) and by [EK19] (KO-EK); knockoff statistics aggregation (KO-HH); debiased-Lasso (DL-BH).

Choice of B and γ for AKO. Figure 4.3 shows an experiment when varying γ and B. FDR and power are averaged across 30 simulations of fixed parameters: SNR=3.0, $\rho = 0.7$, sparsity=0.06. Notably, it seems that there is no further gain in statistical power when $B > 25$. Similarly, the power is essentially equal for γ values greater than 0.1 when $B \geq 25$. Based on the results of this experiment we set the default value of $B = 25, \gamma = 0.3$.

4.4.2 GWAS on Flowering Phenotype of *Arabidopsis thaliana*

To test AKO on real datasets, we first perform a genome-wide association study (GWAS) on genomic data. The aim is to detect association of each of 174 candidate genes with a phenotype **FT_GH** that describes flowering time of *Arabidopsis*

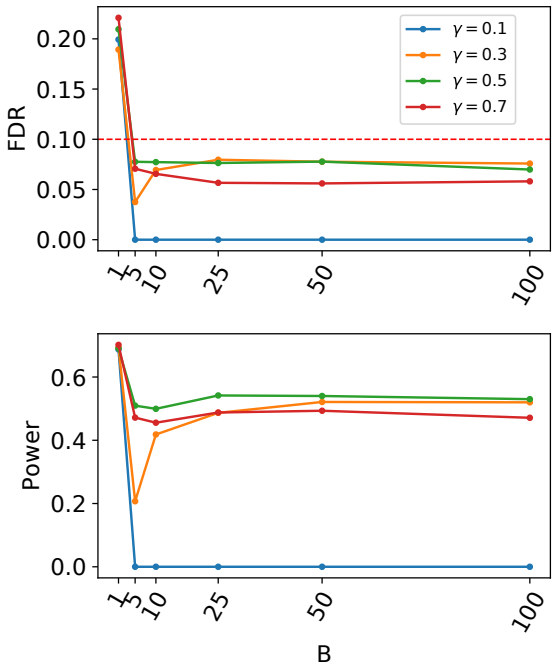


Figure 4.3: **FDR and average power for 30 simulations of fixed parameters: SNR=3.0, $\rho = 0.7$, sparsity=0.06.** There is virtually no gain in statistical power when $B > 25$ and when $\gamma \geq 0.1$.

thaliana, first done by [AHV+10]. Preprocessing is done similarly to [AGS+13]: 166 data samples of 9938 binary SNPs located within a $\pm 20-$kilobase window of 174 candidate genes that have been selected in previous publications as most likely to be involved in flowering time traits. Furthermore, we apply the same dimension reduction by hierarchical clustering as [SCAV19] to make the final design matrix of size $n = 166$ samples $\times$ $p = 1560$ features. We list the detected genes from each method in Table 6.2.

Table 4.1: **List of detected genes associated with phenotype FT_GH**. Empty line (—) signifies no detection. Detected genes are listed in well-known studies dated up to 20 years ago.

Method	**Detected Genes**
AKO	AT2G21070, AT4G02780, AT5G47640
KO	AT2G21070
KO-GZ	AT2G21070
DL-BH	—

The three methods that rely on sampling knockoff variables detect AT2G21070. This gene, which is responsible for the mutant FIONA1, is listed by [KKY+08] to be vital for regulating period length in the *Arabidopsis* circadian clock. FIONA1 also appears to be involved in photoperiod-dependent flowering and in daylength-dependent seedling growth. In particular, the time for opening of the first flower for FIONA1 mutants are shorter than the ones without under both long and short-day conditions. In addition to FIONA1 mutant, AKO also detects AT4G02780 and AT5G47640. It can be found in studies dating back to the 90s [SCS98] that AT4G02780 encodes a mutation for late flowering. Meanwhile, AT5G47640 mutant

delay flowering in long-day but not in short-day experiments [CBE+07].

4.4.3 Functional Magnetic Resonance Imaging (fMRI) analysis on Human Connectome Project Dataset

Human Connectome Project (HCP900) is a collection of neuroimaging and behavioral data on 900 healthy young adults, aged 22–35. Participants were asked to perform different tasks inside an MRI scanner while blood oxygenation level dependent (BOLD) signals of the brain were recorded. The analysis investigates what brain regions are predictive of the subtle variations of cognitive activity across participants, conditionally to other brain regions. Similar to genomics data, the setting is high-dimensional with $n = 1556$ samples acquired and 156437 brain voxels. A voxel clustering step that reduces data dimension to $p = 1000$ clusters is done to make the problem tractable.

When decoding brain signals on HCP subjects performing a foot motion experiment (Figure 4.4, left), AKO recovers an anatomically correct anti-symmetric solution, in the motor cortex and the cerebellum, together with a region in a secondary sensory cortex. KO only detects a subset of those. Moreover, across seven such tasks, the results obtained independently from DL-BH are much more similar to AKO than to KO, as measured with Jaccard index of the resulting maps (Figure 4.4, right). The maps for the seven tasks are represented in Appendix. Note that the sign of the effect for significant regions is readily obtained from the regression coefficients, with a voting step for bootstrap-based procedures.

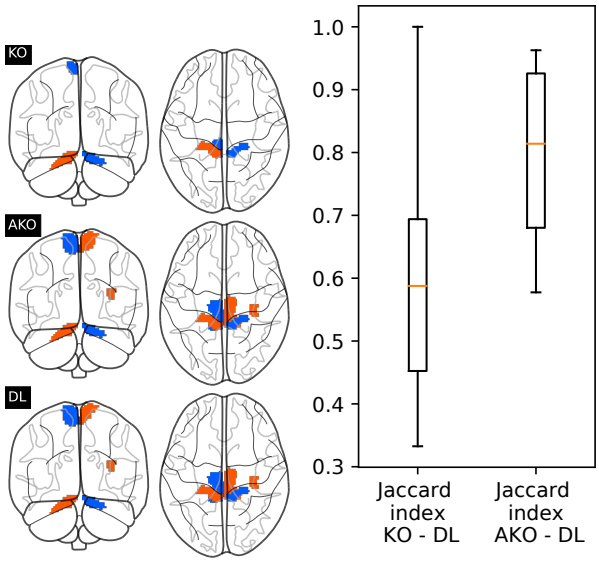


Figure 4.4: **Detection of significant brain regions for HCP data (900 subjects).** (left) Selected regions in a left or right foot movement task. **Orange**: brain areas with positive sign activation. **Blue**: brain areas with negative sign activation. Here the AKO solution recovers an anatomically correct pattern, part of which is missed by KO. (right) Jaccard index measuring the Jaccard similarity between the KO/AKO solutions on the one hand, and the DL solution on the other hand, over 7 tasks: AKO is significantly more consistent with the DL-BH solution than KO.

4.5 Discussion

In this work, we introduce a p-value to measure knockoff importance and design a knockoffs bootstrapping scheme that leverages this quantity. With this we are able to tame the instability inherent to the original knockoff procedure. Analysis shows that aggregation of multiple knockoffs retains theoretical guarantees for FDR control. However, *i)* the original argument of [BC15] no longer holds (see Appendix); *ii)* a factor κ on the FDR control is lost; this calls for tighter FDR bounds in the future, since we always observe empirically that the FDR is controlled without the factor κ. Moreover, both numerical and realistic experiments show that performing aggregation results in an increase in statistical power and also more consistent results with respect to alternative inference methods.

The quantile aggregation procedure from [MMB09] used here is actually conservative: as one can see in Figure 4.2, the control of FDR is actually stricter than without the aggregation step. Nevertheless, as often with aggregation-based approaches, the gain in accuracy brought by the reduction of estimator variance ultimately brings more power.

We would like to address here two potential concerns about FDR control for AKO+BH. The first one is when the $\{W_j\}_{j\in\mathcal{S}^c}$ are not independent, hence violating Assumption 4.1. In the absence of a proof of Theorem 4.2 that would hold under a general dependency, we first note that several schemes for knockoff construction (for instance, the one of [CFJL18]) imply the independence of $(\mathbf{x}_i - \tilde{\mathbf{x}}_i)_{i\in[p]}$, as well as their pseudo inverse. These observations do not establish the independence of W_j. Yet, intuitively, the Lasso coefficient of one variable should be much more associated with its knockoff version than with other variables, so it should not be much affected by these other variables, making the Lasso-coefficient differences weakly correlated if not independent. Moreover, in the proof of Lemma 4.2 and Theorem 4.2, Assumption 4.1 is only used for applying Bernstein's inequality, and several dependent versions of Bernstein's inequality have been proved [Sam00, MPR09, HS17, among others]. Similarly, the proof of Eq. (4.17) only uses Assumption 4.1 for applying the strong law of large numbers, a result which holds true for various kinds of dependent variables (for instance, [Abd18], and references therein). Therefore we conjecture that independence in Assumption 4.1 can be relaxed into some mixing condition. Overall, given that the unstability of KO with respect to the KO randomness is an important drawback (see Figure 4.1), we consider Assumption 4.1 as a reasonable price to pay for correcting it, given that we expect to relax it in future works.

The second potential concern is that Theorem 4.2 is for AKO with $\widehat{k}$ computed from the BY procedure, while BH step-up may not control the FDR when the aggregated p-values $(\bar{\pi}_j)_{j\in[p]}$ are not independent. We find empirically that the $(\bar{\pi}_j)_{j\in[p]}$ do not exhibit spurious Spearman correlation (Figure B.2 in Appendix) under a setting where the W_j satisfy a mixing condition. This is a mild assumption that should be satisfied, especially when each feature X_j only depends on its "neighbors" (as typically observed on neuroimaging and genomics data). It is actually likely that the aggregation step contributes to reducing the statistical dependencies between the $(\bar{\pi}_j)_{j\in[p]}$. Eventually, we note that BH can be replaced by BY procedure [BY01] in case of doubt.

To conclude on these two potential concerns, let us emphasize that the FDR of AKO+BH with $B > 1$ is always below α (up to error bars) in *all* the experiments we did, including preliminary experiments not shown in this article, which makes us confident when applying AKO+BH on real data such as the ones of Sections 4.4.2–4.4.3.

A practical question of interest is to handle the cases where $n \ll p$, that is, the number of features overwhelms the number of samples. Note that in our experiments, we had to resort to a clustering scheme of the brain data and to select some genes. A possible extension is to couple this step with the inference framework, in order to take into account that for instance the clustering used is not given but *estimated*

from the data, hence with some level of uncertainty.

The proposed approach introduces two parameters: the number B of bootstrap replications and the γ parameter for quantile aggregation. The choice of B is simply driven by a compromise between accuracy (the larger B, the better) and computation power, but we consider that much of the benefit of AKO is obtained for $B \approx 25$. Regarding γ, adaptive solutions have been proposed [MMB09], but we find that choosing a fixed quantile (0.3) yields a good behavior, with little variance and a good sensitivity.

Appendix

The Appendix is organized as follows. First, the main theoretical results of the article are proved:

- Proof of Proposition 4.1: AKO+BH with $B = 1$ and $\gamma = 1$ is equivalent to vanilla KO.
- Proof of Lemma 4.1: for Lasso-coefficient differences, the non-zero W_j are distinct.
- Proof that the π_j are *asymptotically* valid p-values (without any multiplicative correction): Lemma 4.3.
- Statement and proof of a new general result about FDR control with quantile-aggregated p-values: Lemma 4.4.
- Proof of Theorem 4.2.

Second, the results of some additional experiments are reported:

- Additional experiments to show that the KO-GZ alternative aggregation procedure by [GZ19b] has decreasing power when the number κ of knockoff vectors $\tilde{\mathbf{x}}$ considered simultaneously increases (we compare $\kappa = 2$ with $\kappa = 3$). We show empirically that this is not the case for AKO with respect to B.
- Empirical evidence for the near independence of p-values π_j.
- Additional figures for HCP 900 experiments.

4.6 Detailed Proofs

4.6.1 Proof of Proposition 4.1

We begin by noticing that the function $f : \mathbb{R}^+ \to \mathbb{Z}^+$, $f(x) = \dfrac{\#\{k : W_k \leq -x\}}{p}$ is decreasing in x. This means the first step of both FDR control step-up procedures, that involves ordering the intermediate p-values ascendingly, is the same as arranging the knockoff statistic in descending order: $W_{(1)} \geq W_{(2)} \geq \cdots \geq W_{(p)}$. Therefore from Eq. (4.7) and the definition of π_j we have:

$$\widehat{k} = \max\left\{k : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{p} \leq \frac{k\alpha}{p}\right\}$$

(note that we can exclude all the $\pi_{(k)} = 1$ due to the fact that $\forall\, k \in [p], \alpha \in (0,1) : k\alpha/p < 1$).

This can be written as:

$$\widehat{k} = \max\left\{k : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{\#\{i : W_{(i)} \geq W_{(k)}\}} \leq \alpha\right\},$$

since $\#\{i : W_{(i)} \geq W_{(k)}\} = k$ because $\{W_{(j)}\}_{j\in[p]}$ is ordered descendingly and because of the assumption that non-zero LCD statistics are distinct. Furthermore, finding the maximum index k of the descending ordered sequence is equivalent to finding the minimum value in that sequence, or

$$\widehat{k} = \min\left\{W_{(k)} > 0 : \frac{1 + \#\{i : W_{(i)} \leq -W_{(k)}\}}{\#\{i : W_{(i)} \geq W_{(k)}\}} \leq \alpha\right\},$$

since all $W_{(j)} \leq 0$ (corresponding with $\pi_{(k)} = 1$) have been excluded. Without loss of generality, we can write:

$$\widehat{t}_+ = \min\left\{t > 0 : \frac{1 + \#\{i : W_i \leq -t\}}{\#\{i : W_i \geq t\}} \leq \alpha\right\}.$$

This threshold $\widehat{t}_+$ is exactly the same as the definition of threshold τ_+ in Eq. (4.4) from the original KO procedure. □

4.6.2 Proof of Lemma 4.1

Setting Let $\mathbf{X} \in \mathbb{R}^{n\times q}$, $\boldsymbol{\beta}^* \in \mathbb{R}^q, \lambda > 0, \sigma > 0$ be fixed. Define

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$$
$$\forall \boldsymbol{\beta} \in \mathbb{R}^q, \qquad L(\boldsymbol{\beta}) := \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 .$$

with $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma\mathbf{I}_n)$ the Gaussian noise and $\|\cdot\|_p$ the L_p norm.

Classical Optimization Properties Since L is convex, non-negative, and tends to $+\infty$ at infinity, its minimum over $\mathbb{R}^q$ exists and is attained (although may not be unique). Since L is convex, its minima are characterized by a first-order condition:

$$\widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^q}\{L(\boldsymbol{\beta})\} \Leftrightarrow \qquad 0 \in \partial L(\boldsymbol{\beta})$$

which is equivalent to

$$\begin{cases} \exists\, \widehat{\mathbf{z}} \in [-1,1]^q : \mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}^\top\mathbf{y} - \dfrac{\lambda}{2}\widehat{\mathbf{z}} \\ \quad \forall j \text{ s.t. } (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0,\ \widehat{\mathbf{z}}_j = \operatorname{sign}((\widehat{\boldsymbol{\beta}}_\lambda)_j) \end{cases} \tag{4.18}$$

As shown by [Gir14, Section 4.5.1] for instance, the fitted value $\widehat{f}_\lambda \in \mathbb{R}^n$ is uniquely defined:

$$\exists!\ \widehat{f}_\lambda \in \mathbb{R}^n \quad \text{such that} \quad \forall \widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^q}\{L(\boldsymbol{\beta})\}, \qquad \widehat{f}_\lambda = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} .$$

As a consequence, the equicorrelation set

$$\widehat{J}_\lambda := \left\{j \in \{1,\dots,q\} : |\mathbf{x}_j^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}})| = \lambda/2\right\}$$

is uniquely defined. We also have,

$$\forall\ \widehat{\boldsymbol{\beta}}_\lambda \in \operatorname*{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^q}\{L(\boldsymbol{\beta})\}, \qquad \left\{j : (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0\right\} \subset \widehat{J}_\lambda \tag{4.19}$$

(but these two sets are not necessarily equal, and the former set may not be uniquely defined).

Note that for every set $J \subset \{1,\dots,q\}$ such that $\forall j \notin J$, $(\widehat{\boldsymbol{\beta}}_\lambda)_j = 0$, we have $\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda = \mathbf{X}_J(\widehat{\boldsymbol{\beta}}_\lambda)_J$ so that $(\mathbf{X}^\top\mathbf{X}\widehat{\boldsymbol{\beta}}_\lambda)_J = \mathbf{X}_J^\top\mathbf{X}_J(\widehat{\boldsymbol{\beta}}_\lambda)_J$. As a consequence, taking $J = \widehat{J}_\lambda$, by eq. (4.18) and (4.19), any minimizer $\widehat{\boldsymbol{\beta}}_\lambda$ of L over $\mathbb{R}^q$ satisfies

$$\mathbf{X}_{\widehat{J}_\lambda}^\top\mathbf{X}_{\widehat{J}_\lambda}(\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda} = \mathbf{X}_{\widehat{J}_\lambda}^\top\mathbf{y} - \frac{\lambda}{2}\widehat{\mathbf{z}}_{\widehat{J}_\lambda} \tag{4.20}$$

for some $\widehat{\mathbf{z}}_{\widehat{J}_\lambda} \in \{-1,1\}^{\widehat{J}_\lambda}$.

If the matrix $\mathbf{X}_{\widehat{J}_\lambda}^\top \mathbf{X}_{\widehat{J}_\lambda}$ is non-singular (that is, if $\mathbf{X}_{\widehat{J}_\lambda}$ is of rank $|\widehat{J}_\lambda|$), then the argmin of L is unique [Gir14, Section 4.5.1].

Result 4.6.1. *For every $\boldsymbol{\alpha} \in \mathbb{R}^q \backslash \{0\}$, the event*

$$\operatorname{rank}(\mathbf{X}_{\widehat{J}_\lambda}) = |\widehat{J}_\lambda|\,, \quad \boldsymbol{\alpha}^\top \widehat{\boldsymbol{\beta}}_\lambda = 0 \quad \textit{and} \quad \exists\, j \in \{1,\dots,q\}\,, \quad \alpha_j (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0\,, \tag{4.21}$$

where $\{\widehat{\boldsymbol{\beta}}_\lambda\} = \operatorname{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^q} \{L(\boldsymbol{\beta})\}$ is well-defined by the first property, has probability zero.

Proof. Let Ω be the event defined by Eq. (4.21). If Ω holds true, then there exists some $J \subset \{1,\dots,q\}$ and some $\widehat{\mathbf{z}} \in \{-1,1\}^q$ such that $\operatorname{rank}(X_J) = |J|$, $(\widehat{\boldsymbol{\beta}}_\lambda)_{J^c} = 0$, and

$$\mathbf{X}_J^\top \mathbf{X}_J (\widehat{\boldsymbol{\beta}}_\lambda)_J = \mathbf{X}_J^\top \mathbf{y} - \frac{\lambda}{2} \widehat{\mathbf{z}}_J\,.$$

Indeed, this is a consequence of Eq. (4.19) and (4.20), by taking $J = \widehat{J}_\lambda$ and $\mathbf{z}$ such that $\mathbf{z}_{\widehat{J}_\lambda} = \operatorname{sign}\left((\widehat{\boldsymbol{\beta}}_\lambda)_{\widehat{J}_\lambda}\right)$. Therefore, using that $\mathbf{X}_J^\top \mathbf{X}_J$ is non-singular, we get

$$\begin{aligned} & (\widehat{\boldsymbol{\beta}}_\lambda)_J = M(J)\boldsymbol{\varepsilon} + v(J,\mathbf{z}) \\ \text{where} \quad & M(J) := (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \mathbf{X}_J^\top \\ \text{and} \quad & v(J,\mathbf{z}) := (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \mathbf{X}_J^\top \mathbf{X}\boldsymbol{\beta}^* - (\mathbf{X}_J^\top \mathbf{X}_J)^{-1} \frac{\lambda}{2} \widehat{\mathbf{z}}_J\,, \end{aligned}$$

hence

$$\boldsymbol{\alpha}^\top \widehat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\alpha}_J^\top M(J)\boldsymbol{\varepsilon} + \boldsymbol{\alpha}^\top v(J,\mathbf{z})$$

follows a normal distribution with variance $\sigma^2 \boldsymbol{\alpha}_J^\top M(J) M(J)^\top \boldsymbol{\alpha}_J = \sigma^2 \|M(J)^\top \boldsymbol{\alpha}_J\|^2$. Now, on Ω, we also have the existence of some j such that $\alpha_j (\widehat{\boldsymbol{\beta}}_\lambda)_j \neq 0$. Since $(\widehat{\boldsymbol{\beta}}_\lambda)_{J^c} = 0$, we must have $j \in J$, which shows that $\boldsymbol{\alpha}_J \neq 0$.

Overall, we have proved that

$$\begin{aligned} & \Omega \subset \bigcup_{J\in\mathcal{J}, \mathbf{z}\in\{-1,1\}^q} \Omega_{J,z} \\ \text{where} \quad & \mathcal{J} := \{j \in \{1,\dots,q\} : \operatorname{rank}(X_J) = |J| \text{ and } \alpha_J \neq 0\} \\ \text{and} \quad & \Omega_{J,\mathbf{z}} := \left\{\boldsymbol{\alpha}_J^\top M(J)\boldsymbol{\varepsilon} + \alpha^\top v(J,\mathbf{z}) = 0\right\}\,. \end{aligned}$$

For every $J \in \mathcal{J}$, $M(J)^\top \boldsymbol{\alpha}_J \neq 0$ since $\alpha_J \neq 0$ and $M(J)$ is of rank $|J|$. As a consequence, for every $J \in \mathcal{J}$ and $\mathbf{z} \in \{-1,1\}^p$, $\mathbb{P}(\Omega_{J,\mathbf{z}})$ is the probability that a Gaussian variable with non-zero variance is equal to zero, so it is equal to zero. We deduce that

$$\mathbb{P}(\Omega) \leq \sum_{J\in\mathcal{J}, \mathbf{z}\in\{-1,1\}^q} \mathbb{P}(\Omega_{J,z}) = 0$$

since the sets $\mathcal{J}$ and $\{-1,1\}^q$ are finite. □

Applying Result 4.6.1 to the case where $\mathbf{X}$ concatenates the original p covariates and their knockoff counterparts (hence $q = 2p$), we get that, apart from the event where $\mathbf{X}_{\widehat{J}_\lambda}$ is not full rank, for every $j \in \{1,\dots,p\}$, W_j takes any fixed non-zero value with probability zero (with $\alpha_j = \pm 1$, $\alpha_{j+p} = \pm 1$, $\alpha_k = 0$ otherwise).

Similarly, the above lemma shows that for every $j \neq j' \in \{1,\dots,p\}$:

$$\mathbb{P}(\mathbf{X}_{\widehat{J}_\lambda} \text{ is full-rank and } \exists j \neq j', W_j = W_{j'}, W_j \neq 0, W_{j'} \neq 0) = 0.$$

As a consequence, with probability 1, all the non-zero W_j are distinct if $X_{\widehat{J}_\lambda}$ is full-rank. □

Remark 4.1. *The proof of Result 4.6.1 is also valid for other noise distributions: it only assumes that the support of the distribution of $\boldsymbol{\varepsilon}$ is not included into any hyperplane of $\mathbb{R}^n$.*

4.6.3 Asymptotic Validity of Intermediate P-values

We consider in this section an asymptotic regime where $p \to +\infty$.

Assumption 4.2 (Asymptotic regime $p \to +\infty$)**.** When p grows to infinity, n, $\mathbf{X}$, β^*, $\boldsymbol{\varepsilon}$ and $\mathbf{y}$ all depend on p implicitly. We assume that for every integer $j \geq 1$, $\mathbb{1}_{\beta_j^*=0}$ does not depend on p (as soon as $p \geq j$), and that

$$\frac{|\mathcal{S}|}{p} = \frac{\big|\{j \in [p] : \beta_j^* \neq 0\}\big|}{p} \xrightarrow[p\to+\infty]{} 0\,.$$

When making Assumption 4.1, we also assume that $\mathbb{P}_0$ does not depend on p.

Lemma 4.3. If Assumptions 4.1 and 4.2 hold true, then for all $j \geq 1$ such that $\beta_j^* = 0$, the empirical p-value π_j defined by Eq. (4.5) is a valid p-value asymptotically, that is,

$$\forall t \in [0,1]\,, \qquad \lim_{p\to+\infty} \mathbb{P}(\pi_j \leq t) \leq t\,.$$

Note that our proof of Theorem 4.2 in Section 4.6.5 relies on the use of Lemma 4.2 with t that can be of order $1/p$. Therefore, Lemma 4.3 above is not sufficient for our needs. Nevertheless, it still provides a interesting insight about the π_j, and justifies (asymptotically) their name, which is why we state and prove this result here.

Proof. By definition, $\pi_j \leq 1$ almost surely, so the result holds when $t = 1$. Let us now focus on the case where $t \in [0,1)$. Let F_0 denote the c.d.f. of $\mathbb{P}_0$, the common distribution of the null statistics $\{W_j\}_{1\leq j\leq p\,/\,\beta_j^*=0}$, which exists by Assumption 4.1. Let $j \geq 1$ such that $\beta_j^* = 0$ be fixed, and assume that $p \geq j$ is large enough so that $|\mathcal{S}^c| \geq 2$. Let $m = |\mathcal{S}^c| - 1 \geq 1$ as in the proof of Lemma 4.2. Note that m depends on p, and $m/p \to 1$ as $p \to +\infty$ by Assumption 4.2, hence $m \to +\infty$ as $p \to +\infty$.

By definition of π_j, when $W_j > 0$ we have:

$$\begin{aligned}
\pi_j &= \frac{1 + \#\{k \in [p] : W_k \leq -W_j\}}{p} \\
(\text{since } W_j > 0 > -W_j) &= \frac{1 + \#\{k \in \mathcal{S} : W_k \leq -W_j\} + \#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\
&\geq \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq -W_j\}}{p} \\
&= \frac{\widehat{F}_m(-W_j)}{\alpha_p}
\end{aligned} \tag{4.22}$$

where $\alpha_p \overset{\text{def.}}{=} \dfrac{p}{m}$ and for all $u \in \mathbb{R}$,

$$\widehat{F}_m(u) \overset{\text{def.}}{=} \frac{\#\{k \in \mathcal{S}^c \setminus \{j\} : W_k \leq u\}}{m}$$

is the empirical cdf of $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$.

Now, since $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$ are *i.i.d.* with distribution $\mathbb{P}_0$ by Assumption 4.1 , the law of large numbers implies that, for all $u \in \mathbb{R}$,

$$\widehat{F}_m(u) \xrightarrow[p\to+\infty]{\text{a.s.}} F_0(u)\,.$$

Since we assume $\lim_{p\to+\infty} |\mathcal{S}|/p = 0$, $\lim_{p\to+\infty} \alpha_p = 1$ and we get that for all $u \in \mathbb{R}$,

$$\frac{1}{\alpha_p}\widehat{F}_m(u) \xrightarrow[p\to+\infty]{\text{a.s.}} F_0(u)\,.$$

Since W_j is independent from $\{W_k\}_{k\in\mathcal{S}^c\setminus\{j\}}$, this result also holds true *conditionally to* W_j, with $u = -W_j$. Given that almost sure convergence implies convergence in distribution, we have: conditionally to W_j,

$$\frac{1}{\alpha_p}\widehat{F}_m(-W_j) \xrightarrow[p\to+\infty]{\text{(d)}} F_0(-W_j) \overset{(d)}{=} F_0(W_j) \tag{4.23}$$

where the latter equality comes from the fact that W_j has a symmetric distribution, as shown in Remark 4.1.

So, when $W_j > 0$, for every $t \in [0,1)$,

$$\begin{aligned}\limsup_{p\to+\infty} \mathbb{P}(\pi_j \le t \,|\, W_j) &\le \limsup_{p\to+\infty} \mathbb{P}\left(\frac{\widehat{F}_m(-W_j)}{\alpha_p} \le t \,\middle|\, W_j\right) \qquad \text{by Eq. (4.22)}\\ &\le \mathbb{1}_{F_0(W_j)\le t} \end{aligned} \tag{4.24}$$

by Eq. (4.23) combined with the Portmanteau theorem.

Therefore, for every $t \in [0,1)$,

$$\begin{aligned}\limsup_{p\to+\infty} \mathbb{P}(\alpha_p\pi_j \le t) &= \limsup_{p\to+\infty}\Big\{\mathbb{P}(\alpha_p\pi_j \le t \text{ and } W_j > 0) + \underbrace{\mathbb{P}(\alpha_p\pi_j \le t \text{ and } W_j \le 0)}_{\substack{=0 \text{ since } \alpha_p\ge 1>t\\ \text{and } \pi_j = 1 \text{ when } W_j\le 0}}\Big\}\\
&= \limsup_{p\to+\infty} \mathbb{E}[\mathbb{P}(\alpha_p\pi_j \le t \mid W_j)\mathbb{1}_{W_j>0}]\\
&\le \mathbb{E}\left[\limsup_{|\mathcal{S}^c|\to+\infty}\{\mathbb{P}(\alpha_p\pi_j \le t \mid W_j)\mathbb{1}_{W_j>0})\}\right]\\
&\le \mathbb{P}\big(F_0(W_j) \le t\big) \qquad \text{by Eq. (4.24)}\\
&\le t\,,\end{aligned}$$

which concludes the proof. □

4.6.4 A General FDR Control with Quantile-aggregated P-values

The proof of Theorem 4.2 relies on an adaptation of results proved by [MMB09, Theorems 3.1 and 3.3] about aggregation of p-values. The results of [MMB09], whose proof relies on the proofs of [BY01], are stated for randomized p-values obtained through sample splitting. The following lemma shows that they actually apply to any family of p-values.

Lemma 4.4. Let $(\pi_j^{(b)})_{1\le j\le p, 1\le b\le B}$ be a family of random variables with values in $[0,1]$. Let $\gamma \in (0,1]$, $\bar{\alpha} \in [0,1]$ and $\mathcal{N} \subset [p]$ be fixed. Let us define

$$\begin{aligned}\forall j \in [p]\,, \quad Q_j &\overset{\text{def.}}{=} \frac{p}{\gamma} q_\gamma\big(\{\pi_j^{(b)} : 1 \le b \le B\}\big) \quad \text{where} \quad q_\gamma(\cdot) \text{ is the } \gamma\text{-quantile function,}\\
\widehat{h} &\overset{\text{def.}}{=} \max\big\{i \in [p] : Q_{(i)} \le i\bar{\alpha}\big\} \quad \text{where} \quad Q_{(1)} \le \cdots \le Q_{(p)}\,,\\
\text{and} \quad \widehat{S} &\overset{\text{def.}}{=} \big\{j \in [p] : Q_j \le Q_{(\widehat{h})}\big\}\,.\end{aligned}$$

Then,

$$\mathbb{E}\left[\frac{|\widehat{S}\cap\mathcal{N}|}{|\widehat{S}|\vee 1}\right] \le \sum_{j=1}^{p-1}\left(\frac{1}{j} - \frac{1}{j+1}\right)F(j) + \frac{F(p)}{p} \tag{4.25}$$

$$\text{where} \qquad \forall j \in [p]\,, \qquad F(j) \overset{\text{def.}}{=} \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B}\sum_{i\in\mathcal{N}} \mathbb{P}\left(\pi_i^{(b)} \le \frac{j\bar{\alpha}\gamma}{p}\right)\,.$$

As a consequence, if some $C \geq 0$ exists such that

$$\forall t \geq 0, \forall b \in [B]\,, \forall i \in \mathcal{N}\,, \qquad \mathbb{P}\left(\pi_i^{(b)} \leq t\right) \leq Ct\,, \tag{4.26}$$

then we have

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] \leq \frac{|\mathcal{N}|C}{p}\left(\sum_{j=1}^{p} \frac{1}{j}\right)\bar{\alpha}\,. \tag{4.27}$$

Let us emphasize that Lemma 4.4 can be useful in general, well beyond knockoff aggregation. To the best of our knowledge, Lemma 4.4 is new. In particular, the recent preprint by [RD19b], that studies p-value aggregation procedures, focuses on FWER controlling procedures, whereas Lemma 4.4 provides an FDR control for a less conservative procedure.

Proof. For every $i, j, k \in [p]$, let us define

$$p_{i,j,k} = \begin{cases} \mathbb{P}\Big(Q_i \in \big((j-1)\bar{\alpha}, j\bar{\alpha}\big]\,,\, i \in \widehat{S} \text{ and } |\widehat{S}| = k\Big) & \text{if } j \geq 2 \\ \mathbb{P}\big(Q_i \in [0, \bar{\alpha}]\,,\, i \in \widehat{S} \text{ and } |\widehat{S}| = k\big) & \text{if } j = 1\,. \end{cases}$$

Then,

$$\begin{aligned} \frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1} &= \sum_{k=1}^{p} \mathbb{1}_{|\widehat{S}|=k} \frac{\sum_{i \in \mathcal{N}} \mathbb{1}_{i \in \widehat{S}}}{k} \\ &= \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S}} \\ &= \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \mathbb{1}_{|\widehat{S}|=k \text{ and } i \in \widehat{S} \text{ and } 0 \leq Q_i \leq k\bar{\alpha}} \end{aligned}$$

since $i \in \widehat{S}$ and $|\widehat{S}| = k$ implies that $Q_i \leq Q_{(\widehat{h})} \leq \widehat{h}\bar{\alpha} = k\bar{\alpha}$. Taking an expectation and writing that

$$\mathbb{1}_{0 \leq Q_i \leq k\bar{\alpha}} = \mathbb{1}_{Q_i \in [0, \bar{\alpha}]} + \sum_{j=2}^{k} \mathbb{1}_{Q_i \in ((j-1)\bar{\alpha}, j\bar{\alpha}]}\,,$$

we get —following the computations of [MMB09, proof of Theorems 3.3], which themselves rely on the ones of [BY01]—,

$$\begin{aligned} \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] &\leq \sum_{i \in \mathcal{N}} \sum_{k=1}^{p} \frac{1}{k} \sum_{j=1}^{k} p_{i,j,k} = \sum_{i \in \mathcal{N}} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{k} p_{i,j,k} \\ &\leq \sum_{i \in \mathcal{N}} \sum_{j=1}^{p} \sum_{k=j}^{p} \frac{1}{j} p_{i,j,k} = \sum_{j=1}^{p} \frac{1}{j} \underbrace{\sum_{i \in \mathcal{N}} \sum_{k=j}^{p} p_{i,j,k}}_{= \overline{F}(j) - \overline{F}(j-1)\mathbb{1}_{j \geq 2}} \end{aligned}$$

$$\text{where} \qquad \forall j \in \{1, \dots, p\}\,, \qquad \overline{F}(j) \overset{\text{def.}}{=} \sum_{i \in \mathcal{N}} \sum_{j'=1}^{j} \sum_{k=1}^{p} p_{i,j',k}\,.$$

Since the above upper bound is equal to

$$\overline{F}(1) + \sum_{j=2}^{p} \frac{1}{j}\big[\overline{F}(j) - \overline{F}(j-1)\big] = \sum_{j=1}^{p} \left(\frac{1}{j} - \frac{1}{j+1}\right)\overline{F}(j) + \frac{\overline{F}(p)}{p}\,,$$

$$\text{we get that} \qquad \mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] \leq \sum_{j=1}^{p} \left(\frac{1}{j} - \frac{1}{j+1}\right)\overline{F}(j) + \frac{\overline{F}(p)}{p}\,. \tag{4.28}$$

Notice also that

$$\overline{F}(j) = \sum_{i \in \mathcal{N}} \mathbb{P}(Q_i \leq j\bar{\alpha} \text{ and } i \in \widehat{S}) \leq \sum_{i \in \mathcal{N}} \mathbb{P}(Q_i \leq j\bar{\alpha}) \ .$$

Now, as done by [MMB09, proof of Theorems 3.1], we remark that $Q_i \leq j\bar{\alpha}$ is equivalent to

$$\frac{1}{B}\left|\left\{b \in [B] \,:\, \frac{p\pi_i^{(b)}}{\gamma} \leq j\bar{\alpha}\right\}\right| = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{p\pi_i^{(b)} \leq j\bar{\alpha}\gamma} \geq \gamma$$

so that

$$\begin{aligned}
\mathbb{P}(Q_i \leq j\bar{\alpha}) &= \mathbb{P}\left(\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{p\pi_i^{(b)} \leq j\bar{\alpha}\gamma} \geq \gamma\right) \\
&\leq \frac{1}{\gamma}\mathbb{E}\left[\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}_{p\pi_i^{(b)} \leq j\bar{\alpha}\gamma}\right] \qquad \text{by Markov inequality} \\
&= \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B} \mathbb{P}\left(p\pi_i^{(b)} \leq j\bar{\alpha}\gamma\right) .
\end{aligned}$$

Therefore,

$$\overline{F}(j) \leq \sum_{i \in \mathcal{N}} \frac{1}{\gamma}\frac{1}{B}\sum_{b=1}^{B} \mathbb{P}\left(p\pi_i^{(b)} \leq j\bar{\alpha}\gamma\right) = F(j)\,,$$

so that Eq. (4.28) implies Eq. (4.25).

If condition (4.26) holds true, then, for every $j \in [p]$,

$$F(j) \leq \frac{|\mathcal{N}|C}{\gamma}\frac{j\bar{\alpha}\gamma}{p} = \frac{|\mathcal{N}|C\bar{\alpha}}{p}j\,,$$

hence Eq. (4.25) shows that

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] \leq \sum_{j=1}^{p-1} \frac{F(j)}{j(j+1)} + \frac{F(p)}{p} \leq \frac{|\mathcal{N}|C\bar{\alpha}}{p}\sum_{j=1}^{p-1}\frac{1}{j+1} + \frac{|\mathcal{N}|C\bar{\alpha}}{p} = \frac{|\mathcal{N}|C\bar{\alpha}}{p}\left(\sum_{j=1}^{p}\frac{1}{j}\right),$$

which is the desired result. □

4.6.5 Proof of Theorem 4.2

We can now prove Theorem 4.2. We apply Lemma 4.4 with $\bar{\alpha} = \beta(p)\alpha$, $\mathcal{N} = \mathcal{S}^c$, so that $\widehat{S} = \widehat{\mathcal{S}}_{AKO+BY}$. Since the $\pi_j^{(b)}$, $b = 1, \ldots, B$, have the same distribution as π_j, by Lemma 4.2, condition (4.26) holds true with $C = p/|\mathcal{S}^c|$, and Eq. (4.27) yields the desired result. □

Note that an FDR control for AKO such as Theorem 4.2 cannot be obtained straightforwardly from the arguments of [BC15] and [CFJL18]. One key reason for this is that their proof relies on a reordering of the features according to the values of $(|W_j|)_{j\in[p]}$ [BC15, Section 5.2], such a reordering being permitted since the signs of the W_j are iid coin flips *conditionally to* the $(|W_j|)_{j\in[p]}$ [CFJL18, Lemma 2]. In the case of AKO, we must handle the $(W_j^{(b)})_{j\in[p]}$ *simultaneously for all* $b \in [B]$, and conditioning with respect to $(|W_j^{(b)}|)_{j\in[p],b\in[B]}$ may reveal some information about the signs of the $(W_j^{(b)})_{j\in[p]}$ as soon as $B > 1$. At least, it does not seem obvious to us that the key result of [CFJL18, Lemma 2] can be proved *conditionally to* the $(|W_j^{(b)}|)_{j\in[p],b\in[B]}$ when $B > 1$, so that the proof strategy of [CFJL18] breaks down in the case of AKO with $B > 1$.

4.6.6 Theoretical Results of AKO without Factor κ

As mentioned in Remark 4.4, the factor $\kappa \approx 3.24$ in Lemma 4.2 and Theorem 4.2 can be removed, thanks to the usage of the follow result from [RW05].

Lemma 4.5 (Lemma 1, [RW05])**.** Suppose that $Y_1, Y_2, \ldots, Y_B$ are exchangeable real-valued random variables; that is, their distribution is invariant under permutations. Let $\tilde{q}$ be defined by

$$\tilde{q} = \frac{1}{B}\left[1 + \sum_{i=1}^{B-1} \mathbb{1}_{Y_i \geq Y_B}\right].$$

Then $\mathbb{P}(\tilde{q} \leq u) \leq u$ for all $0 \leq u \leq 1$.

First, we state a version of Lemma 4.2 without factor κ.

Lemma 4.6. If Assumption 4.1 holds true, and if $|\mathcal{S}^c| \geq 2$, then, for all $j \in \mathcal{S}^c$, the intermediate p-value π_j defined by Eq. (4.5) satisfies:

$$\forall t \in [0,1] \qquad \mathbb{P}(\pi_j \leq t) \leq \frac{p}{|\mathcal{S}^c|} t$$

Proof. The result holds when $t \geq 1$ since $p \geq |\mathcal{S}^c|$ and a probability is always smaller than 1. Let us now focus on the case where $t \in [0,1)$, and let $j \in \mathcal{S}^c$ be fixed. From the formula of intermediate p-value defined by knockoff statistics conversion, defined in Eq. (4.5), when $W_j > 0$, we have

$$\begin{aligned}\pi_j \overset{\text{def.}}{=} \frac{1}{p}\left[1 + \sum_{k \neq j} \mathbb{1}_{-W_k \geq W_j}\right] &= \frac{|\mathcal{S}^c|}{p} \, \frac{1 + \sum_{k \neq j} \mathbb{1}_{-W_k \geq W_j}}{|\mathcal{S}^c|} \\ &\geq \frac{|\mathcal{S}^c|}{p} \underbrace{\frac{1 + \sum_{k \neq j, k \in \mathcal{S}^c} \mathbb{1}_{-W_k \geq W_j}}{|\mathcal{S}^c|}}_{\tilde{\mathbf{Q}}}.\end{aligned}$$

Observe that, under Assumption 4.1, and a result from [CFJL18, Lemma 3.3] that implies the common distribution of $\{W_k\}_{k \in \mathcal{S}^c}$ is symmetric around 0, we have that the vector $(W_j, -W_k, k \in \mathcal{S}^c \backslash \{j\}) \in \mathbb{R}^{|\mathcal{S}^c|}$ is an exchangeable random vector. Therefore, Lemma 4.5 applies to the quantity $\tilde{q} = \tilde{\mathbf{Q}}$, with $B = p, Y_i = -W_k, Y_B = W_j$. This leads to

$$\mathbb{P}(\pi_j \leq t) \leq \mathbb{P}\left(\frac{|\mathcal{S}^c|}{p} \, \tilde{\mathbf{Q}} \leq t\right) \leq \frac{p}{|\mathcal{S}^c|} \, t \quad \text{for all } t \in [0,1),$$

where the second inequality is obtained by applying Lemma 4.5. □

With this result, we have the following result for non-asymptotic guarantee of FDR control of Aggregation of Multiple Knockoffs with BY step-up, where the constant κ is removed. Proving this theorem is a straightforward application of Lemma 4.6 and Lemma 4.4.

Theorem 4.2. If Assumption 4.1 holds true and $|\mathcal{S}^c| \geq 2$, then for any $B \geq 1$ and $\alpha \in (0,1)$, the output $\widehat{\mathcal{S}}_{AKO+BY}$ of aggregation of multiple knockoff (Algorithm 4.1), with the BY step-up procedure, has a FDR controlled as follows:

$$\mathbb{E}\left[\frac{|\widehat{S}_{AKO+BY} \cap \mathcal{S}^c|}{|\widehat{S}_{AKO+BY}| \vee 1}\right] \leq \alpha.$$

4.7 Additional Experimental Results

4.7.1 Demonstration of Aggregated Multiple Knockoff vs. Simultaneous Knockoff

Using the same simulation settings as in Section 4.4.1 with $n = 500, p = 1000$ and varying simulation parameters to generate Figure 4.2 in the main text, we benchmark only aggregation of multiple knockoffs (AKO) with 5 and 10 bootstraps ($B = 5$ and $B = 10$) and compare with simultaneous knockoffs with 2 and 3 bootstraps ($\kappa = 2$ and $\kappa = 3$). Results in Figure 4.5 show that while increasing the number of knockoff bootstraps makes simultaneous knockoffs more conservative, doing so with AKO makes the algorithm more powerful (and in the worst case retains the same power with fewer bootstraps).

4.7.2 Empirical Evidence on the Independence of Aggregated P-values $\bar{\pi}$

Using the same simulation settings as in Section 4.4.1 with $n = 500, p = 1000, \rho = 0.6, \text{snr} = 3.0, \text{sparsity} = 0.06$ we generate 100 observations of *aggregated* p-values $\bar{\boldsymbol{\pi}}$. Then, we compute the Spearman rank-order correlation coefficient of the *Null* $\bar{\pi}_j$ for these 100 observations along with their two-sided p-value (for a hypothesis test whose null hypothesis is that two sets of data are uncorrelated).

The results are illustrated in Figure 4.6: the Spearman correlation values are concentrated around zero, while the distribution of associated p-values seems to be a mixture between a uniform distribution and a small mixture component consisting of mostly non-null p-values. This indicates near independence between the aggregated p-values using quantile-aggregation [MMB09], hence justifies our use of BH step-up procedure for selecting FDR controlling threshold in the AKO algorithm.

Remark 4.1. *Again, it is worth noticing that the empirical evidence we have shown is only done in a setting with a Toeplitz structure for the covariance matrix. However, as explained in the main text, this correlation setting is usually found in neuroimaging and genomics data. Hence, we believe that assuming short-distance correlations between the* $(X_j)_{j\in[p]}$ *is a mild assumption, which should be satisfied in the practical scenarios where we want to apply our algorithm.*

The decoding maps returned by the KO, AKO and DL inference procedures are presented in Figure 4.7. As quantified by the Jaccard index in the main text, we observe that the AKO solution is typically closer to an external method based on the desparsified lasso (DL). Moreover, AKO is also typically more sensitive than KO alone.

The seven classification problems are the following:

- Emotion: predict whether the participant watches an angry face or a geometric shape.
- Gambling: predict whether the participant gains or loses gambles.
- Motor foot: predict whether the participant moves the left or right foot.
- Motor hand: predict whether the participant moves the left or right hand.
- Relational: predict whether the participant matches figures or identified feature similarities.
- Social: predict whether the participant watches a movie with social behavior or not.
- Working Memory: predict whether the participant does a 0-back or a 2-back task.

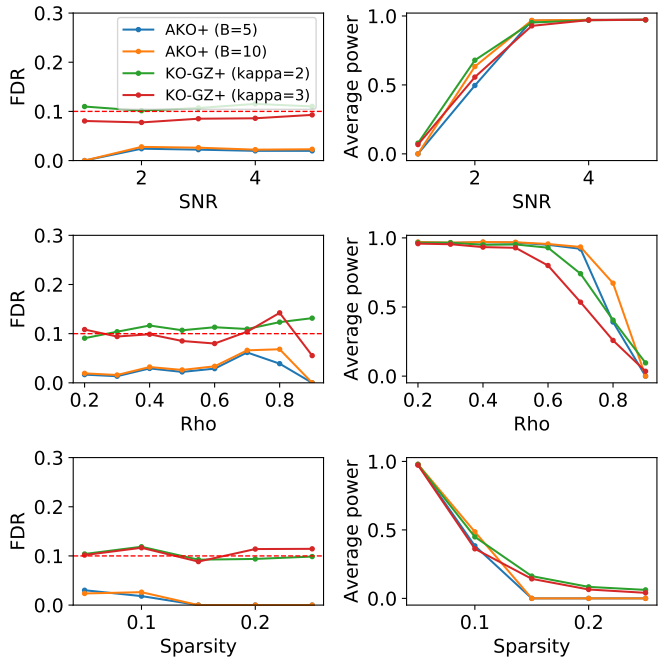


Figure 4.5: **Aggregation of multiple knockoffs ($B = 5$ and $B = 10$) vs. simultaneous knockoffs ($\kappa = 2$ and $\kappa = 3$).** A clear loss in statistical power is demonstrated in the latter method when increasing the number of bootstraps κ, while the former (AKO) shows the opposite: with $B = 10$ bootstraps there are small, yet consistent gains in the number of true detections compared to using only $B = 5$ bootstraps.

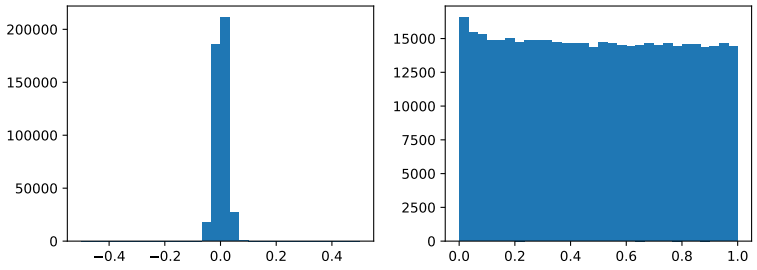


Figure 4.6: **Left: Histogram of Spearman correlation values for 100 samples of null aggregated p-values $\bar{\pi}_j$. Right: Histogram of corresponding p-values for the Spearman correlation**

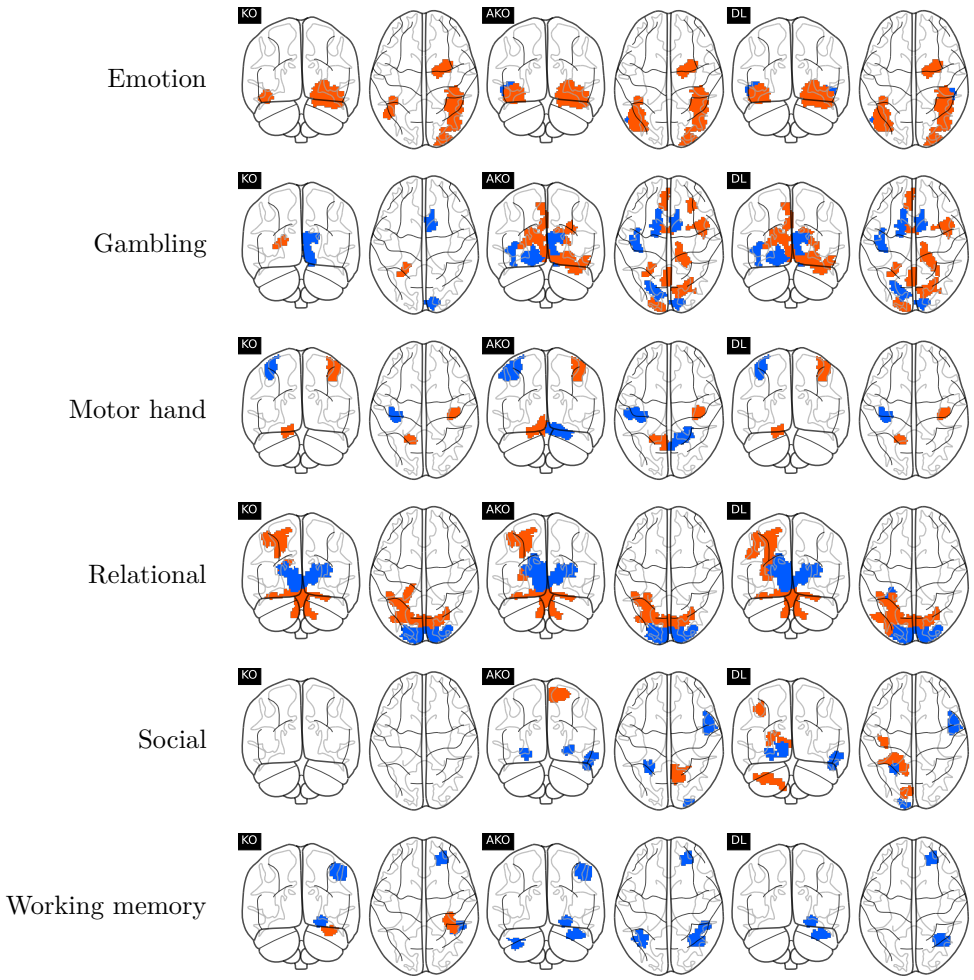


Figure 4.7: **Decoding maps obtained for seven classification tasks.** Emotion, gambling, motor foot, motor hand, relational, social and working memory refer to 7 binary tasks that were considered based on the HCP900 dataset. We observe that AKO is typically more sensitive than KO, and yields solution closer to that of an independent solution based on a desparsified-Lasso (DL) estimator.

Chapter 5

Ensemble of Clustered Knockoffs

This chapter is an extended version of the publication in the *Proceedings of 26th Conference on Information Processing in Medical Imaging* (IPMI 2019) [NCT19] as a joint work with Jerome-Alexis Chevalier and Bertrand Thirion.

Summary. Continuous improvement in medical imaging techniques allows the acquisition of higher-resolution images. When these are used in a predictive setting, a greater number of explanatory variables are potentially related to the dependent variable (the response). Meanwhile, the number of acquisitions per experiment remains limited. In such high dimension/small sample size setting, it is desirable to find the explanatory variables that are truly related to the response while controlling the rate of false discoveries. To achieve this goal, novel multivariate inference procedures, such as knockoff inference, have been proposed recently. However, they require the feature covariance to be well-defined, which is impossible in high-dimensional settings. In this paper, we propose a new algorithm, called Ensemble of Clustered Knockoffs, that allows to select explanatory variables while controlling the false discovery rate (FDR), up to a prescribed spatial tolerance. The core idea is that knockoff-based inference can be applied on groups (clusters) of voxels, which drastically reduces the problem's dimension; an ensembling step then removes the dependence on a fixed clustering and stabilizes the results. We benchmark this algorithm and other FDR-controlling methods on brain imaging datasets and observe empirical gains in sensitivity, while the false discovery rate is controlled at the nominal level.

5.1 Background

Medical images are increasingly used in predictive settings, in which one wants to classify patients into disease categories or predict some outcomes of interest. Besides predictive accuracy, a fundamental question is that of *opening the black box*, *i.e.* understanding the combinations of observations that explains the outcome. A particular relevant question is that of the importance of image features in the prediction of an outcome of interest, conditioned on other features. Such conditional analysis is a fundamental step to allow causal inference on the implications of the signals from image regions in this outcome; see e.g. [WMO+15] for the case of brain imaging. However, the typical setting in medical imaging is that of high-dimensional small-sample problems, in which the number of samples n is much smaller than the number of covariates p. This is further aggravated by the steady improvements in data resolution. In such cases, classical inference tools fail, both theoretically and

practically. One solution to this problem is to reduce the massive number of covariates by utilizing dimension reduction, such as clustering-based image compression, to reduce the number of features to a value close to n; see e.g. [BRvdGZ13]. This approach can be viewed as the bias/variance trade-off: some loss in the localization of the predictive features —bias— is tolerated as it comes with less variance —hence higher power— in the statistical model. This is particularly relevant in medical imaging, where localizing predictive features at the voxel level is rarely important: one is typically more interested in the enclosing region.

However, such a method suffers from the arbitrariness of the clustering step and the ensuing high-variance in inference results with different clustering runs, as shown empirically in [CST18]. The same work also introduced an algorithm called Ensemble of Clustered Desparsified Lasso (ECDL), based on the inference technique developed in [ZZ14], that provides p-values for each feature, and controls the Family Wise Error Rate (FWER), i.e. the probability of making one or more false discoveries. In applications, it is however more relevant to control the False Discovery Rate (FDR) [BH95], which indicates the expected fraction of false discoveries among all discoveries, since it allows to detect a greater number of variables. In univariate settings, the FDR is easily controlled by the Benjamini-Hochberg procedure [BH95], valid under independence or positive correlation between features. It is unclear whether this can be applied to multivariate statistical settings. A promising method which controls the FDR in multivariate settings is the so-called knockoff inference [BC15, CFJL18], which has been successfully applied in settings where $n \approx p$. However, the method relies on randomly constructed knockoff variables, therefore it also suffers from instability. Our contribution is a new algorithm, called Ensemble of Clustered Knockoffs (ECKO), that *i)* stabilizes knockoff inference through an aggregation approach; *ii)* adapts knockoffs to $n \ll p$ settings. This is achieved by running the knockoff inference on the reduced data and ensembling the ensuing results.

The remainder of our paper is organized as follows. Section 5.2 introduces necessary background of statistical inference with structured data, together with a formal introduction on knockoff inference and dimension reduction. Section 5.3 presents our main contribution, the Ensemble of Clustered Knockoffs algorithm. Section 5.4 introduces FDR^{δ}, a spatial-relaxation of the False Discovery Rate, and establishes a theoretical guarantee of controlling FDR^{δ} for ECKO. Section 5.5 describes the setup of our experiments with both synthetic and brain imaging data predictive problems, to illustrate the performance of ECKO, followed by details of the experimental results. Specifically, we benchmark this approach against the procedure proposed in [GSD+15], that does not require the clustering step, yet only provides asymptotic ($n \to \infty$) guarantees. We show the benefit of the ECKO approach in terms of both statistical control and statistical power. We conclude in Section 5.6 with a discussion of potential future research directions.

5.2 Preliminaries

5.2.1 High Dimensional Linear Models with Structured Data

Settings We consider the case where samples (rows) of the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ are *i.i.d.* and follow a centered Gaussian distribution, *i.e.* for all $i \in [n] \overset{\text{def.}}{=} \{1, \dots, n\}$, $\mathbf{X}_{i,*} \sim \mathcal{N}(0_p, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is the population covariance matrix of the variables. The columns of $\mathbf{X}$ refer to the explanatory variables, while the rows of $\mathbf{X}$ represent the different samples in the feature space. We focus on experimental settings in which the number of features p is much greater than the number of samples n. We assume $\mathbf{X}$ and $\mathbf{y}$ follows linear relationship

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \sigma\boldsymbol{\xi}, \tag{5.1}$$

where $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ is the vector or true parameter, $\boldsymbol{\xi} \sim \mathcal{N}(0, \mathbf{I}_n)$ is the noise vector that is independent from $\mathbf{X}$, and $\sigma > 0$ is the noise magnitude. Additionally, we denote the true support, the set of indices of non-zero true-parameters, by $\mathcal{S} \overset{\text{def.}}{=} \{k \in [p] \mid \beta_k^0 \neq 0\}$. Let $\hat{S}$ denotes an estimation of the support given a particular inference procedure. We also define the signal-to-noise ratio (SNR), which allows to assess the noise regime of a given experiment by

$$\text{SNR} = \frac{\|\mathbf{X}\boldsymbol{\beta}^0\|_2^2}{n^{-1/2}\sigma^2}. \tag{5.2}$$

A high SNR means that the signal magnitude is strong compared to the noise, hence it refers to an easier inference problem.

Structured Data In medical imaging and many other experimental settings, the data stored in the design matrix $\mathbf{X}$ relate to *structured signals*. More precisely, the features have a peculiar dependence structure that is related to an underlying spatial organization, for instance the spatial neighborhood in 3D images. Then, the features are generated from a random process acting on this underlying metric space. In our paper, the distance between the j-th and the k-th features is denoted by $d(j, k)$. Since the variables have a natural representation in a metric space (eg a lattice structure for an image), we first assume that the distance $d(\cdot, \cdot)$ between any two variables is known. We introduce the two following key assumptions.

Assumption 5.1 (Spatial homogeneity with distance δ)**.** *For all* $(j, k) \in [p] \times [p]$, $d(j, k) \leq \delta$ *implies that* $\boldsymbol{\Sigma}_{j,k} \geq 0$, *where* $\boldsymbol{\Sigma}_{j,k} \overset{def.}{=} Cov(\mathbf{x}_j, \mathbf{x}_k)$.

Assumption 5.2 (Sparse-smooth with distance δ)**.** *For all* $(j, k) \in [p] \times [p]$, $d(j, k) \leq \delta$ *implies that* $\text{sign}(\boldsymbol{\beta}_j^0) = \text{sign}(\boldsymbol{\beta}_k^0)$.

Note that 0 is implicitly considered as positive and negative in the above assumption. Assumption 5.1 states that two variables at a spatial distance smaller than δ are positively correlated, and Assumption 5.2 states that the weights of closely located variables share the same sign. Equivalently, the sparse-smooth assumption with parameter δ holds if the distance between the two closest weights of opposite sign is larger than δ.

5.2.2 Knockoff Inference

Initially introduced by [BC15] to identify variables in genomics, the knockoff filter is an FDP control approach for multivariate models. This method has been improved to work with mildly high-dimensional settings in [CFJL18], leading to the so-called model-X knockoffs, defined as follows.

Definition 5.1 (Model-X knockoffs [CFJL18])**.** *The model-X knockoffs for the family of random variables* $\mathbf{X} = (X_1, \dots X_p)$ *are a new family of random variables* $\tilde{\mathbf{X}} = (\tilde{X}_1, \dots, \tilde{X}_p)$ *constructed to satisfy the two properties:*

1. *For any subset* $\mathcal{K} \subset \{1, \dots, p\}$*:* $(\mathbf{X}, \tilde{\mathbf{X}})_{swap(\mathcal{K})} \overset{d}{=} (\mathbf{X}, \tilde{\mathbf{X}})$,
   where the vector $(\mathbf{X}, \tilde{\mathbf{X}})_{swap(\mathcal{K})}$ *denotes the swap of entries* X_j *and* $\tilde{X}_j$, $\forall j \in \mathcal{K}$
2. $\tilde{\mathbf{X}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{X}$ *where* $\mathbf{y}$ *is the response vector.*

In a nutshell, knockoff procedure first creates extra null variables that have a correlation structure similar to that of the original variables. A test statistic vector is then calculated to measure the strength of the original versus its knockoff counterpart. An example of such statistic is the lasso-coefficient difference (LCD) that we use in this paper:

Definition 5.2. *Knockoff procedure with intermediate p-values [BC15, CFJL18]*

1. *Construct knockoff variables, produce matrix concatenation:* $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$.

2. *Calculate Lasso-Coefficient Difference statistics by solving*

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{2p}} \frac{1}{2} \|\mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}]\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1,$$

   and then, for all $j \in [p]$*, take the difference* $W_j = |\hat{\beta}_j(\lambda)| - |\hat{\beta}_{j+p}(\lambda)|$.

3. *Compute the p-values* p_j*, for* $j \in [p]$*:*

$$p_j = \begin{cases} \dfrac{1 + \#\{k : W_k \leq -W_j\}}{p} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \leq 0\,. \end{cases} \tag{5.3}$$

4. *Given a desired FDR level* $\alpha \in (0, 1)$*, FDR control with p-values* $\{p_j\}_{j=1}^p$ *by Benjamini-Hochberg [BH95] or Benjamini-Yekutieli [BY01] procedure.*

Remark 5.1. *The above formulation is distinct from that of [BC15, CFJL18], but it is equivalent (as stated in Proposition 4.1). We use it to introduce the intermediate variables* p_j *for all* $j \in [p]$.

We restate as next an important assumption regarding null distribution of knockoff statistic, first presented in Chapter 4.

Assumption 5.3 (Assumption 4.1 – Null distribution of knockoff statistics)**.** *The knockoff statistic defined in Definition 5.2 are such that* $\{W_j\}_{j \in [p] \setminus \mathcal{S}}$ *are independent and follow the same distribution under the null hypothesis.*

5.2.3 Dimension reduction

Knockoff (KO) inference is intractable in high-dimensional settings, as knockoff generation requires the estimation and inversion of covariance matrices of size $(2p \times 2p)$. Hence, we leverage data structure by introducing a clustering step that reduces data dimension before applying KO inference. As in [CST18, CNTS21], assuming the features' signals are spatially smooth, it is relevant to consider a spatially-constrained clustering algorithm. By averaging the features with each clustering, we reduce the number of parameters from p to C, the number of clusters, where $C \ll p$. We introduce the formalization of dimension reduction by clustering the data into a number of groups (clusters) of variables that are spatially concentrated as follows.

Compressed representation Let $[C] \overset{\text{def.}}{=} \{1, \ldots, C\}$ be the set of group indices, with $C < p$. The set of all the groups is denoted by $\mathcal{G} = \{G_1, G_2, \ldots, G_C\}$. Every group representative variable is defined by the average of the covariates it contains.

Let $\mathbf{Z} \in \mathbb{R}^{n \times C}$ be the compressed random design matrix that contains the group representative variables in columns. Without loss of generality, assuming a suitable ordering of the columns of $\mathbf{X}$, dimension reduction can be formalized as:

$$\mathbf{Z} = \mathbf{X}\mathbf{A}, \tag{5.4}$$

where $\mathbf{A} \in \mathbb{R}^{p \times q}$ is the transformation matrix defined by:

$$\mathbf{A} = \begin{bmatrix} \alpha_1 \text{—} \alpha_1 & 0 \text{—} 0 & \ldots & 0 \text{—} 0 \\ 0 \text{—} 0 & \alpha_2 \text{—} \alpha_2 & \ldots & 0 \text{—} 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 \text{—} 0 & 0 \text{—} 0 & \ldots & \alpha_C \text{—} \alpha_C \end{bmatrix},$$

where $\alpha_c = 1/|G_c|$ for all $c \in [C]$.

This means the distribution of the i-th row of $\mathbf{Z}$ is given by $\mathbf{Z}_{i,*} \sim \mathcal{N}_q(0, \boldsymbol{\Upsilon})$, where $\boldsymbol{\Upsilon} = \mathbf{A}^\top \boldsymbol{\Sigma} \mathbf{A}$. The correlation between the groups r and l is given by

$$\mathrm{Cor}(\mathbf{Z}_{*,r}, \mathbf{Z}_{*,l}) = \frac{\boldsymbol{\Upsilon}_{r,l}}{\sqrt{\boldsymbol{\Upsilon}_{r,r} \boldsymbol{\Upsilon}_{l,l}}}.$$

As mentioned in [BRvdGZ13], under the Gaussian linear model assumption in Eq. (5.1), we have the following compressed representation:

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\theta}^0 + \sigma_\eta \boldsymbol{\eta}, \tag{5.5}$$

where $\boldsymbol{\theta}^0 \in \mathbb{R}^q$, $\boldsymbol{\eta} \sim \mathcal{N}(0, \mathbf{I}_n)$, $\sigma_\eta \geq \sigma > 0$ and $\boldsymbol{\eta}$ is independent from $\mathbf{Z}$. We now give a property of the weights of the compressed problem which is a consequence of [BRvdGZ13, Proposition 4.3].

Proposition 5.1 (Prop. 4.1, [CNTS21])**.** *Considering the Gaussian linear model in Eq. (5.1) and assuming:*

(i) for all $c \in [C]$, for all pairs (j,k) in $\in (G_c)^2$, $\boldsymbol{\Sigma}_{j,k} \geq 0$,

(ii) for all $c \in [C]$, for all $c' \in [C] \setminus \{c\}$, $\boldsymbol{\Upsilon}_{c,c'} = 0$,

(iii) for all $c \in [C]$, $\left(\boldsymbol{\beta}^0_j \geq 0 \text{ for all } j \in G_c\right)$ or $\left(\boldsymbol{\beta}^0_j \leq 0 \text{ for all } j \in G_c\right)$.

Then, in the compressed representation of Eq.(5.5), for $c \in [C]$, $\boldsymbol{\theta}^0_c \neq 0$ if and only if there exists $j \in G_c$ such that $\boldsymbol{\beta}^0_j \neq 0$. If such an index j exists then $\mathrm{sign}(\boldsymbol{\theta}^0_r) = \mathrm{sign}(\boldsymbol{\beta}^0_j)$.

Remark 5.2. *Assumption (i) holds when $\mathrm{Diam}(\mathcal{G}) \leq \delta$ under the spatial homogeneity assumption with parameter δ (Assumption 5.1). Assumption (ii) states the independence of the groups. A sufficient condition is when the covariance matrix $\boldsymbol{\Sigma}$ is block diagonal, which means that variables of different groups are independent. Besides [CNTS21], [BRvdGZ13] also made use of this assumption to establish the faithfulness of compressed model to original model. Assumption (iii) states that all the weights in a group share the same sign, which holds under Assumption 5.2 and when the clustering diameter is smaller than δ.*

Knockoff inference on the compressed model To perform Knockoff inference on the compressed problem Eq. (5.5), we run the procedure defined in Definition 5.2 with input data $\mathbf{Z}$ (instead of the design matrix $\mathbf{X}$) and $\mathbf{y}$. This outputs the cluster-wise intermediate p-values, denoted $\hat{\mathbf{p}}^{\mathcal{G}} = (\hat{p}^{\mathcal{G}}_c)_{c \in [C]}$, by the following formula

$$\hat{p}^{\mathcal{G}}_c = \begin{cases} \dfrac{1 + \#\{k \in [C] : W_k \leq -W_j\}}{C} & \text{if} \quad W_j > 0 \\ 1 \quad \text{if} \quad W_j \leq 0\,. \end{cases} \tag{5.6}$$

5.3 Ensemble of Clustered Knockoffs

Our main contribution is to combine dimension reduction by clustering with Knockoff inference. First, we compute B (randomized) clusterings $(C_b)_{b \in [B]}$ and B draws of knockoff variables (at the cluster level), and we derive the corresponding cluster-wise intermediate p-values $\hat{\mathbf{p}}^{\mathcal{G},(b)}$ for each $b \in [B]$, using Eq. (5.6). Our next aim is to derive p-values related to the variables of the original problem. To do so, for each $b \in [B]$, we first transfer the p-values from clusters (group of voxels) to features (voxels).

De-grouping Given a family of cluster-wise p-values $\hat{\mathbf{p}}^{\mathcal{G}}$, we set the p-value of the j-th variable to be equal to the p-value of its corresponding group, or, for all $j \in [p]$:

$$\hat{p}_j = \sum_{c \in [C]} \mathbb{1}_{\{j \in G_c\}}\, \hat{p}^{\mathcal{G}}_c. \tag{5.7}$$

Quantile-aggregation of p-values To create a stable inference results, we aggregate the above p-values $\hat{p}_j^{(b)}$, $b = 1, \ldots, B$, for each variable j in parallel, using the quantile aggregation procedure introduced in [MMB09]:

$$\bar{p}_j \overset{\text{def.}}{=} \frac{1}{\gamma} q_\gamma\big(\{\hat{p}_j^{(b)} \,:\, 1 \le b \le B\}\big), \tag{5.8}$$

for all $j \in [p]$; where $q_\gamma(\cdot)$ is the γ-quantile function.

The full Ensemble of Clustered Knockoffs algorithm is summarized in Algorithm 5.1, and graphically in Fig. 5.1.

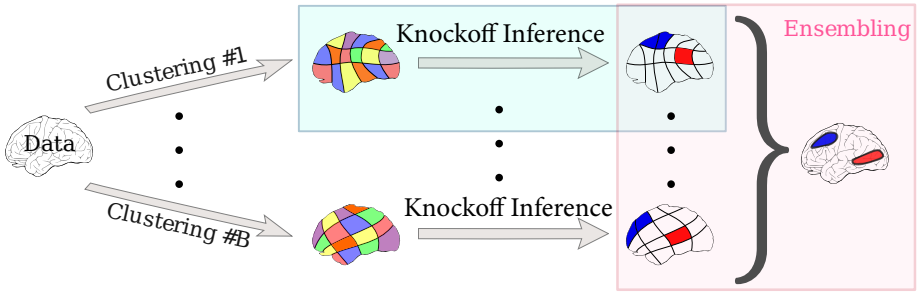


Figure 5.1: The ECKO algorithm. To create a stable inference result, we introduce ensembling steps both within each cluster level and at the voxel-level, across clusterings.

Algorithm 5.1: Ensemble of Clustered Knockoff

input : Data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, response vector $\mathbf{y} \in \mathbb{R}^n$;
param : B - number of sampling, C - number of clusters; α - Nominal FDR threshold;

for $b = 1, \ldots, B$ **do**
  $X^{(b)}_{\texttt{init}} = \texttt{resample}(\mathbf{X}, \mathbf{y})$
  $X^{(b)}_{\texttt{clustered}} = \texttt{clustering}(q, X^{(b)})$
  $\{W_c^{(b)}\}_{c=1}^{C} \leftarrow \texttt{Knockoffs}(X^{(b)}_{\texttt{clustered}}, \mathbf{y}, \alpha)$
  $p_c^{(b)} \leftarrow \dfrac{1 + \#\{k \in [C] : W_k^{(b)} \le -W_j^{(b)}\}}{C} \quad \forall c = 1, \ldots, C$
  $\tilde{p}_k^{(b)} \leftarrow \texttt{degroup}(p_c^{(b)})$ if $k \in G_j^{(b)} \quad \forall c = 1, \ldots, C$
end
for $j = 1, \ldots, p$ **do**
  $\hat{p}_j \leftarrow \texttt{quantile\_aggregated}(\tilde{p}_j^{(b)})$
end
$\widehat{k} \leftarrow \texttt{fdr\_threshold}((\hat{p}_1, \ldots, \hat{p}_p), \alpha,)$ `// Using either Benjamini-Hochberg or Benjamini-Yekutieli procedure`

return $\hat{S}_{\texttt{ECKO}} \leftarrow \{j \in [p] \mid \hat{p}_j \le \hat{p}_{\widehat{k}}\}$

Overall, ECKO procedure might result in a spatial inaccuracy of δ in the location of significant activity, δ being the diameter of the clusters. In the next section, we will incorporate the spatial relaxation δ into the concept of FDR, which we call FDR$^\delta$. We then provide a theoretical guarantee for ECKO controlling FDR$^\delta$ under a predefined level $\alpha \in (0, 1)$.

5.4 Theoretical Results

5.4.1 Spatial Relaxation of False Discovery Rate

With dimension reduction by clustering, and then de-group the variable from cluster-level to voxel-level, we introduce the false discovery rate (FDR) and a spatial generalization of the FDR, called FDR^δ. This quantity is important, since a desirable property of an inference procedure is to either control the FDR or the FDR^δ. A similar concept was introduced by [GZ19a], while [CST18, CNTS21] presented a spatial relaxation version of the Family-Wise Error Rate.

First, we present the classic definitions of False discovery proportion (FDP) and False Discovery Rate (FDR), introduced by [BH95].

Definition 5.1 (False discovery proportion (FDP) and False Discovery Rate (FDR))**.** *Given an estimate of the support $\hat{S}$ obtained from a particular inference procedure, the false discovery proportion is the ratio of the number selected features that do not belong to the support (false discoveries) divided by the number of selected features (discoveries):*

$$FDP = \frac{\#\{k \in \hat{S} : k \notin S\}}{\#\{k \in \hat{S}\}} \tag{5.9}$$

The FDR is the expectation of the FDP

$$FDR = \mathbb{E}[FDP]. \tag{5.10}$$

We now introduce some additional definitions, which is necessary for a formal definition of the spatially-relaxed False Discovery Rate.

Definition 5.2 (δ-null hypothesis)**.** *For each $j \in [p]$, the δ-null hypothesis $\mathcal{H}^\delta_{0,j}$, and the alternative hypothesis $\mathcal{H}^\delta_{\alpha,j}$ for linear relationship of Eq. (5.1) is defined as:*

- $\mathcal{H}^\delta_{0,j}$*: "for all $k \in [p]$ such that $d(j,k) \leq \delta$, $\boldsymbol{\beta}^0_k = 0$",*
- $\mathcal{H}^\delta_{\alpha,j}$*: "there exists $k \in [p]$ such that $d(j,k) \leq \delta$ and $\boldsymbol{\beta}^0_k \neq 0$"*

Definition 5.3 (δ-null region)**.** *The set of variable indices that verify the δ-null hypothesis is called the δ-null region and is denoted by $N^\delta(\boldsymbol{\beta}^0)$ (or simply N^δ):*

$$N^\delta(\boldsymbol{\beta}^0) = \left\{ j \in [p] \mid \textit{for all } k \in [p],\ d(j,k) \leq \delta \textit{ implies } \boldsymbol{\beta}^0_k = 0 \right\}.$$

Figure 5.2 illustrates the idea of the spatial tolerance δ for a true positive. Following Definition 5.3, all the variables located inside the region with radius δ around a significant variable (red dot) will be considered significant.

Definition 5.4 (FDP^δ and FDR^δ)**.** *Given a family of p-values $\{\hat{p}_j\}_{j=1}^p$ and an estimation of the support $\hat{\mathcal{S}}$, the FDP^δ is defined as:*

$$\text{FDP}^\delta = \frac{|\{N^\delta \cap \hat{\mathcal{S}}\}|}{|\hat{\mathcal{S}}| \vee 1}.$$

The FDR^δ is the expectation of the FDP^δ:

$$\text{FDR}^\delta = \mathbb{E}[\text{FDP}^\delta].$$

Remark 5.1. *One can notice that for $\delta = 0$, the FDR and the FDR^δ refer to same quantity* i.e. *$FDR^0 = FDR$.*

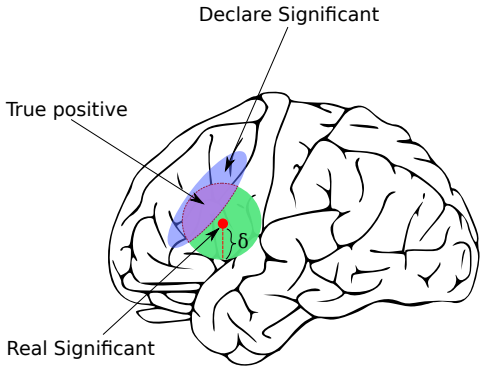


Figure 5.2: Illustration of spatial tolerance δ for a true positive.

5.4.2 Main Results

The following proposition states that within each clustered inference, the intermediate p-values output by Knockoff procedure are valid.

Proposition 5.1. *For any $c \in [C]$, under the compressed model of Eq. (5.5) and the null hypothesis $H_0(G_c) : \theta_c^0 = 0$, we assume that Assumption 5.3 holds true. Then, the p-value converted from knockoff statistics associated with the c-th cluster, defined by Eq. (5.6), denoted by $\hat{p}_c^{\mathcal{G}}$, satisfies*

$$\mathbb{P}(\hat{p}_c^{\mathcal{G}} \leq \alpha) \leq \frac{C}{|N_{\mathcal{G}}|}\alpha, \tag{5.11}$$

for all $\alpha \in (0, 1)$, where $N_{\mathcal{G}} \subset \mathcal{G}$ is the set of null clusters.

Before going to the proof of this proposition, we state as follows a result in [RW05], as it will be useful later on.

Lemma 5.1 (Lemma 1, [RW05])**.** Suppose that $Y_1, Y_2, \ldots, Y_B$ are exchangeable real-valued random variables; that is, their distribution is invariant under permutations. Let $\tilde{q}$ be defined by

$$\tilde{q} = \frac{1}{B}\left[1 + \sum_{i=1}^{B-1} \mathbb{1}_{Y_i \geq Y_B}\right].$$

Then $\mathbb{P}(\tilde{q} \leq u) \leq u$ for all $0 \leq u \leq 1$.

Proof of Prop 5.1. From the formula of intermediate p-value defined by knockoff statistics conversion, defined in Eq. (5.6), we have

$$\begin{aligned}\hat{p}_c^{\mathcal{G}} \overset{\text{def.}}{=} \frac{1}{C}\left[1 + \sum_{k \neq j} \mathbb{1}_{-W_k \geq W_j}\right] &= \frac{|N_{\mathcal{G}}|}{C}\,\frac{1 + \sum_{k \neq j} \mathbb{1}_{-W_k \geq W_j}}{|N_{\mathcal{G}}|} \\ &\geq \frac{|N_{\mathcal{G}}|}{C}\,\underbrace{\frac{1 + \sum_{k \neq j, k \in N_{\mathcal{G}}} \mathbb{1}_{-W_k \geq W_j}}{|N_{\mathcal{G}}|}}_{\tilde{\mathbf{Q}}}.\end{aligned}$$

Observe that, under Assumption 5.3, and a result from [CFJL18, Lemma 3.3] that implies the common distribution of $\{W_k\}_{k \in N_{\mathcal{G}}}$ is symmetric around 0, we have that the vector $(W_j, -W_k, k \in N_{\mathcal{G}} \backslash \{j\}) \in \mathbb{R}^C$ is an exchangeable random vector. Therefore, Lemma 5.1 applies to the quantity $\tilde{q} = \tilde{\mathbf{Q}}$, with $B = C, Y_i = -W_k$, and

$Y_B = W_j$. This leads to

$$\mathbb{P}(\hat{p}_c^{\mathcal{G}} \leq \alpha) \leq \mathbb{P}\left(\frac{|N_{\mathcal{G}}|}{C}\ \tilde{\mathbf{Q}} \leq \alpha\right) \leq \frac{C}{|N_{\mathcal{G}}|}\ \alpha \quad \text{for all } \alpha \in (0,1),$$

where the second inequality is obtained by applying Lemma 5.1. □

The following proposition is from [CNTS21], which states the validity of de-grouped intermediate p-values from Eq. (5.7).

Proposition 5.2 (Prop. 4.2, [CNTS21])**.** *Let $\hat{\mathbf{p}}^{\mathcal{G}}$ be the family of cluster-wise p-values given by Definition 5.2 with input data $\mathbf{Z}, \mathbf{y}$. Then, under the assumptions of Proposition 5.1 and assuming that the diameter of each cluster is smaller than δ, each element of the family $\hat{\mathbf{p}}$ defined by Eq. (5.7) satisfies*

$$\mathbb{P}(\hat{p}_j \leq \alpha) \leq \alpha, \quad \textit{for all } j \in N^{\delta}, \textit{ and } \ \alpha \in (0,1).$$

Intuitively, since all the clusters that intersect the δ-null region have low p-value with low probability, one can conclude that with low probability, all the variables of the δ-null region also have low p-value.

Corollary 5.1. *As a consequence of Proposition 5.1 and Proposition 5.2, for all $j \in [p]$, if we set the p-value of the j-th variable to be equal to the knockoff intermediate p-value of its corresponding group*

$$\hat{p}_j = \sum_{c \in [C]} \mathbb{1}_{\{j \in G_c\}}\ \hat{p}_c^{\mathcal{G}},$$

then we have, under the assumptions of Proposition 5.1 and assuming that the clustering diameter is smaller than δ, each element of the family $\hat{p}$ satisfies the following property:

$$\textit{for all } j \in N^{\delta}, \ \textit{for all } \alpha \in (0,1), \ \mathbb{P}(\hat{p}_j \leq \alpha) \leq \frac{p}{|N^{\delta}|}\ \alpha.$$

Remark 5.2. *As Corollary 5.1 is a consequence of Proposition 5.2, it makes use of assumptions of Proposition 5.1. This means that a necessary condition for Corollary 5.1 to work (and later on Theorem 5.1 that uses this Corollary in the proof) is that the compressed model is faithful to the original model,* i.e. *for $c \in [C]$, $\boldsymbol{\theta}_c^0 \neq 0$ if and only if there exists $j \in G_c$ such that $\boldsymbol{\beta}_j^0 \neq 0$.*

We can now state the main theorem, which establishes the control of FDR^{δ} for Ensemble of Clustered Knockoffs.

Theorem 5.1 (FDR^{δ} control by the Ensemble of Clustered Knockoffs procedure)**.** *Under the Gaussian linear model given in Eq. (5.1), and assume the following.*

(i) Assumption 5.1, 5.2, and 5.3 hold true for all clusterings.

(ii) All the clusters from all partitions considered have a diameter smaller than δ.

(iii) The uncorrelated cluster assumption, i.e. *assumption (ii) of Proposition 5.1, is verified for each clustering.*

Then, the estimation $\hat{\mathcal{S}}_{\texttt{ECKO}}$ of Ensemble of Clustered Knockoffs (Algorithm 5.1) using BY-procedure controls the FDR^{δ} *at the predefined level α,* i.e.

$$\text{FDR}^{\delta}(\hat{\mathcal{S}}_{\texttt{ECKO}}) \leq \alpha.$$

Before the proof, for the sake of completeness, we first rewrite the simplified version (stated as a consequence) of Lemma 4.4 with the system of notation in this chapter.

Lemma 5.2 (Simplified version of Lemma 4.4)**.** Let $(p_j^{(b)})_{1\leq j\leq p, 1\leq b\leq B}$ be a family of random variables with values in $[0,1]$ that satisfies the following property

$$\forall t \in (0,1), \forall b \in [B]\,, \forall i \in \mathcal{N}\,, \qquad \mathbb{P}\left(p_i^{(b)} \leq t\right) \leq Kt, \tag{5.12}$$

where K is some positive constant. Let $\gamma \in (0,1]$, $\alpha \in [0,1]$ and $\mathcal{N} \subset [p]$ be fixed. Let us define

$$\forall j \in [p]\,, \quad \bar{p}_j \overset{\text{def.}}{=} \frac{p}{\gamma} q_\gamma\big(\{p_j^{(b)} \,:\, 1 \leq b \leq B\}\big) \quad \text{where} \quad q_\gamma(\cdot) \text{ is the } \gamma\text{-quantile function,}$$

$$\widehat{h} \overset{\text{def.}}{=} \max\big\{i \in [p] \,:\, \bar{p}_{(i)} \leq i\bar{\alpha}\big\} \quad \text{where} \quad \bar{p}_{(1)} \leq \cdots \leq \bar{p}_{(p)}\,,$$

$$\text{and} \quad \widehat{S} \overset{\text{def.}}{=} \big\{j \in [p] \,:\, \bar{p}_j \leq \bar{p}_{(\widehat{h})}\big\}\,.$$

Then we have

$$\mathbb{E}\left[\frac{|\widehat{S} \cap \mathcal{N}|}{|\widehat{S}| \vee 1}\right] \leq \frac{|\mathcal{N}|K}{p}\left(\sum_{j=1}^{p} \frac{1}{j}\right)\bar{\alpha}. \tag{5.13}$$

Proof of Theorem 5.1. If Assumptions (i)-(iv) of Theorem 5.1 hold true, using Corollary 5.1, we have, for a family of p-values $(p_j^{(b)})_{1\leq j\leq p, 1\leq b\leq B}$ outputting from running line 1-7 of Algorithm 5.1

$$\forall t \in (0,1), \forall b \in [B]\,, \forall i \in N^\delta\,, \qquad \mathbb{P}\left(p_i^{(b)} \leq t\right) \leq \frac{pt}{|N^\delta|}.$$

We therefore have the sufficient condition in Eq. (5.12) to use Lemma 5.2 with $K = p/|N^\delta|$. From Eq. (5.13), take $\mathcal{N} = N^\delta$, we have

$$\mathrm{FDR}^\delta(\hat{\mathcal{S}}) \overset{\text{def.}}{=} \mathbb{E}\left[\frac{|\widehat{S} \cap N^\delta|}{|\widehat{S}| \vee 1}\right] \leq \left(\sum_{j=1}^{p} \frac{1}{j}\right)\alpha.$$

since $|N^\delta| \leq p$ by Definition 5.3 of the δ-null region. We conclude the proof by replacing $\hat{\mathcal{S}} = \hat{\mathcal{S}}_{\texttt{ECKO}}$, since ECKO estimates the support $\hat{\mathcal{S}}_{\texttt{ECKO}}$. Control under arbitrary dependence of the p-values is guaranteed by Benjamini-Yekutieli procedure. □

5.5 Empirical Results

5.5.1 Alternative approaches

In the present work, we benchmark CKO/ECKO with two alternative approaches: the ensemble of clustered desparsified lasso (ECDL) [CST18] and the APT framework from [GSD+15]. As we already noted, ECDL is structured as ECKO. The main differences are that it relies on desparsified lasso rather than knockoff inference and returns p-values instead of q-values. The APT approach was proposed to return feature-level p-values for binary classification problems (though the generalization to regression is straightforward). It directly works at the voxel level, yet with two caveats:

- Statistical control is granted only in the $n \to \infty$ limit
- Unlike ECDL and ECKO, it is unclear whether the returned score represents marginal or conditional association of the input features with the output.

For both ECDL and APT, the returned p-values are converted to q-values using the standard BHq procedure. The resulting q-values are questionable, given that BHq is not valid under negative dependence between the input q-values [BH95]; on the other hand, practitioners rarely check the hypothesis underlying statistical models. We thus use the procedure in a black-box mode and check its validity a posteriori.

Remark 5.1. *In this section, we use a slightly different version of CKO/ECKO, which was presented in [NCT19].*

5.5.2 Synthetic data

To demonstrate the improvement of the proposed algorithm, we first benchmark the method on 3D synthetic data set that resembles a medical image with compact regions of interest that display some predictive information. The size of the true weight vector $\boldsymbol{\beta}^0$ is $50 \times 50 \times 50$, with 5 regions of interest (ROIs) of size $6 \times 6 \times 6$. A design matrix $\mathbf{X}$ that represents random brain signal is then sampled according to a multivariate Gaussian distribution. Finally, the response vector $\mathbf{y}$ is calculated following linear model assumption with Gaussian noise, which is configured to have $SNR \approx 3.6$, similar to real data settings. An average precision-recall curve of 30 simulations is calculated to show the relative merits of single cluster Knockoffs inference versus ECKO and ECDL and APT. Furthermore, we also vary the Signal-to-Noise Ratio (SNR) of the simulation to investigate the accuracy of FDR control of ECKO with different levels of difficulty in detecting the signal.

5.5.3 Real MRI dataset

We compare single-clustered Knockoffs (CKO), ECKO and ECDL on different MRI datasets downloaded from the Nilearn library [APE+14]. In particular, the following datasets are used:

- **Haxby** [HGF+01]. In this functional-MRI (fMRI) dataset, subjects are presented with images of different objects. For the benchmark in our study, we only use brain signals and responses for images related to faces and houses of subject 2 ($n = 216, p = 24083$).
- **Oasis** [MWP+07]. The original collection include data of gray and white matter density probability maps for 416 subjects aged 18 to 96, 100 of which have been clinically diagnosed from mild to moderate Alzheimer's disease. The purpose for our inference task is to find regions that predict the age of a subject ($n = 400, p = 153809$).

We chose $q = 500$ in all experiments for the algorithms that require clustering step (KO, ECKO and ECDL). In the two cases, we start with a qualitative comparison of the returned results. The brain maps are ternary: all regions outside $\hat{S}$ are zeroed, while regions in $\hat{S}$ get a value of $+1$ or -1, depending on whether the contribution to the prediction is positive or negative. For ECKO, a vote is performed to decide whether a voxel is more frequently in a cluster with positive or negative weight.

5.5.4 Results

5.5.4.1 Synthetic data.

A strong demonstration of how ECKO makes an improvement in stabilizing the single-clustering Knockoffs (CKO) is shown in Fig. 5.3. There is a clear distinction between selection of the orange area at lower right and the blue area at upper right in the CKO result, compared to the ground truth. Moreover, CKO falsely discovers some regions in the middle of the cube. By Contrast, ECKO's selection is more similar to the true 3D weight cube. While it returns a wider selection than ECKO, ECDL also claims more false discoveries, most visibly in the blue area on upper-left corner. At the same time, APT returns adequate results, but is more conservative than ECKO. Fig. 5.6a is the result of averaging 30 simulations for the 3D brain synthetic data. ECKO and ECDL obtain almost identical precision-recall curve: for a precision of at least 90%, both methods have recall rate of around 50%. Meanwhile, CKO falls behind, and in fact it cannot reach a precision of over 40% across all recall rates. APT yields the best precision-recall compromise, slightly above ECKO and ECDL.

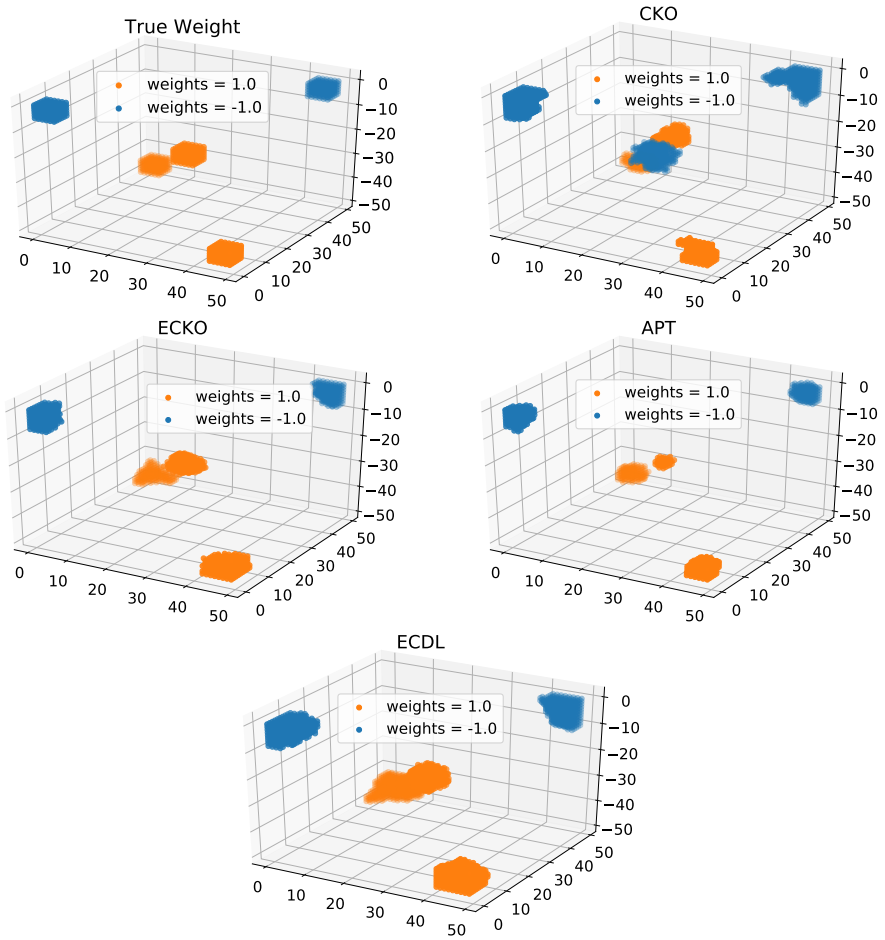


Figure 5.3: Experiments on simulated data: Original 3D weight vector (top left) and inference results from CKO vs. ECKO. The single CKO run has markedly different solutions to the ground truth. Meanwhile, ECKO's solution is closer to the ground truth in the sense that altogether, it is more powerful than APT and also more precise than ECDL.

When varying SNR (from 2^{-1} to 2^5) and investigating the average proportion of false discoveries (δ−FDR) made over the average of 30 simulations (Fig. 5.6b), we observe that CKO fails to control FDR$^\delta$ at nominal level 10% in general. Note that accurate FDR$^\delta$ control would be obtained with larger δ values, but this makes the whole procedure less useful. The ECDL controls FDR$^\delta$ at low SNR level. However, when the signal is strong, ECDL might select more false positives. ECKO, on the other hand, is always reliable —albeit conservative— keeping FDR below the nominal level even when SNR increases to larger magnitude.

Oasis & Haxby dataset When decoding the brain signal on subject 2 of the Haxby dataset using response vector label for watching 'Face vs. House', there is a clear resemblance of selection results between ECKO and ECDL. Using an FDR threshold of 10%, both algorithms select the same area (with a difference in size), namely a face responsive region in the ventral visual cortex, and agree on the sign of the effect. However, on Oasis dataset, thresholding to control the FDR at 0.1 yields empty selection with ECDL and APT, while ECKO still selects some voxels. This potentially means that ECKO is statistically more powerful than ECDL and APT.

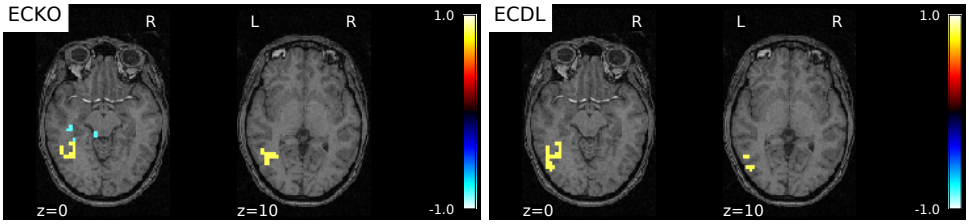


Figure 5.4: Comparison of results for 2 ensembling clustered inference methods on Haxby dataset, nominal FDR$^{\delta}$=0.1. The results are similar to a large extent. No voxel region is detected by APT, therefore we omit to show the selection outcome of the method.

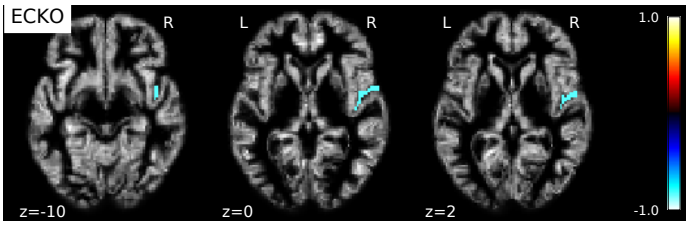


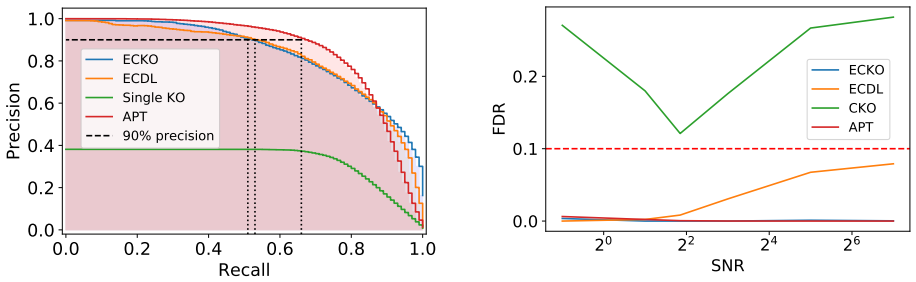
Figure 5.5: Results of ECKO inference on Oasis dataset, nominal FDR$^{\delta}$=0.1. ECKO is the only method to detect significant regions. The temporal region detected by ECKO would be detected by other approaches using a less conservative threshold.

5.6 Discussion

In this work, we proposed an algorithm that makes False Discovery Rate (FDR) control possible in high-dimensional statistical inference. The algorithm is an integration of clustering algorithm for dimension reduction and aggregation technique to tackle the instability of the original knockoff procedure. Evaluating the algorithm on both synthetic and brain imaging datasets shows a consistent gain of ECKO with respect to CKO in both FDR control and sensitivity. Furthermore, Under assumptions on the structure of data, we also prove the non-asymptotic statistical guarantee for ECKO, yet requires the δ-relaxation for FDR.

The number of clusters represents a bias-variance trade-off: increasing it can reduce the bias (in fact, the value of δ), while reducing it improves the conditioning for statistical inference, hence the sensitivity of the knockoff control. We set it to 500 in our experiments. Learning it from the data is an interesting research direction.

On a different note, we acknowledge the limitation of the assumptions in Proposition 5.1. In particular, assuming the covariance is null between clusters (Assumption ii) while being positive between delta-neighbors (Assumption i) hold true for *each clustering*, while necessary for the proof of ECKO controlling FDR$^{\delta}$, can be problematic for many realistic applications, where the correlation structure of data does not follow block-diagonal matrix. The work of [CNTS21] has provided a relaxation of this assumption, but only for the Debiased Lasso. Relaxing this assumption for Clustered/Ensemble of Clustered Knockoff Inference, therefore, is a natural extension for future work.



(a) (b)

Figure 5.6: (a) Average Precision-recall curve (for $SNR \approx 3.6$) and (b) SNR-FDR curve of 30 synthetic simulations. Nominal FDR control level is 10%. ECKO shows substantially better results than CKO and is close to ECDL. APT obtains a slightly better Precision-recall curve. ECKO, ECDL and APT successfully control FDR under nominal level 0.1 where as CKO fails to.

Chapter 6

CRT-Logit: a Conditional Randomization Test for High-Dimensional Logistic Regression

Summary. Inferring the relevant variables of a classification model with correct confidence levels is a central but difficult task in statistics. Despite the core role of logistic regression, currently, there is no good solution to allow accurate inference on the coefficients in the high-dimensional regime – where the number of features p is *much* larger than the number of samples n. In this work, we adapt the recent line of works on Conditional Randomization Test (CRT) to tackle this problem. The original CRT algorithm shows promise as a way to output p-values while making as few assumptions on the distribution of the test statistics as possible. However, it comes at prohibitive computational cost for a practical usage. To tackle this issue, we combine the distillation operation from [LKJR20] with a decorrelation step that adapts it to the geometry of the logistic regression. We call this method `CRT-logit`. Our empirical results suggest that this adaptation works well in mildly high dimension regime, where $n/p \in (0.5, 1.0)$. To address higher dimensional regimes, active dimension reduction procedures are needed. We present a solution by combining clustering and a p-value aggregation technique, that is well-suited for spatially smooth data. With the tradeoff of an additional indeterminacy on the location of predictive features, we show that inference can be performed with reasonable power, while keeping a relaxed version of the False Discovery Rate controlled. We demonstrate the effectiveness of `CRT-logit` on synthetic simulations, along with large-scale brain-imaging and genomics datasets.

6.1 Background

Logistic regression is one of the most popular tools in modern application of statistics and machine learning, partly due to its relative algorithmic simplicity. The method belongs to the class of *generalized linear models* that handle discrete outcomes, *i.e.* classification problems. In this work, we focus on the binary classification problem, where one observation of the responses $y \in \{0, 1\}$ and the data vectors $\mathbf{x} \in \mathbb{R}^p$ follows the relationship:

$$\mathbb{P}(y = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{x}^T \boldsymbol{\beta}^0)}, \tag{6.1}$$

where $\boldsymbol{\beta}^0$ is the vector of true regression coefficients.

In the classical setting, in which the number of samples n is greater than the number of features p, an estimation $\hat{\boldsymbol{\beta}}$ of the true signals $\boldsymbol{\beta}^0$ can be obtained using *maximum likelihood estimation* (MLE). The asymptotic behaviour and derivation of the test statistic, confidence intervals and p-values of the MLE has been well studied, *e.g.* in [CH79]. The availability of p-values for the test statistics makes it possible to rely on multiple hypothesis testing, where one wants to test which variables have a non-zero effect on the outcome, conditionally to the remaining variables:

$$\text{(null) } \mathcal{H}_0^j : x_j \perp y \mid \mathbf{x}_{-j} \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_\alpha^j : x_j \not\perp y \mid \mathbf{x}_{-j},$$

for each feature $j \in [p] \overset{\text{def.}}{=} \{1, \dots, p\}$ and $\mathbf{x}_{-j} \overset{\text{def.}}{=} \{x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p\}$. Equivalently, under the setting in Eq. (6.1), we have:

$$\text{(null) } \mathcal{H}_0^j : \beta_j^0 = 0 \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_\alpha^j : \beta_j^0 \neq 0.$$

However, this line of classical analysis cannot be applied to the high-dimensional regime, where p is larger than n, as argued in [SC19, ZSC20]. These two works showed that in the regime where $\lim_{n,p\to\infty} n/p = \kappa$, the MLE estimator exists only when $\kappa > 2$. We note that this type of analysis is done asymptotically, and to the best of our knowledge, there exists no finite-sample analysis of MLE in high-dimensional regime. The problem becomes even harder with the ℓ_1-regularization added to the likelihood function, i.e. using penalized estimators.

In this work, we focus our attention to the case where p is larger or much larger than n, or high-dimensional setting. This setting is typical in modern applications of pattern recognition, *e.g.* in brain-imaging or genomics, with the number of features as large as hundreds of thousands, but the number of samples stays at most few thousands. The family of methods we consider here is the Conditional Randomization Test (CRT), introduced in [CFJL18]. CRT relies on generating multiple noisy copies of original variables to output empirical p-values in high-dimensional inference problems. However, prohibitive computational cost makes CRT an impractical approach, as discussed at length in [CFJL18, TVZ+18, BWBS19, LKJR20]. There have been several lines of research that attempt to fix this problem, most notably the distilled Conditional Randomization Test (dCRT) [LKJR20]. This work introduced a *distillation operator* as a replacement of the randomized sampling step to compute the importance statistics (see Section 6.2 for more details). To reduce the computation cost of $\mathcal{O}(p^4)$ of vanilla CRT further, [LKJR20] combines distillation operator with an initial screening step, where potentially unimportant variables are removed before the calculation of test statistics. In other words, the number of variables from p is reduced to a much smaller number $\hat{k}$, the cardinality of the screening set. This makes dCRT having an iteration cost of $\mathcal{O}(kp^3)$.

The distillation operator provides a way to output p-values for multiple types of regression and classification, assuming convergence to Gaussian of the test statistic in large-sample setting. Yet, our demonstration in Section 6.2 shows that the originally proposed dCRT test-statistic for logistic regression does not behave as well as intended. In particular, its null distribution deviates markedly from standard normal, even in mildly high-dimension, where one has $n/p \in (0.5, 1)$. We therefore provide a correction for the dCRT, inspired by the decorrelation method presented in [NL17]. The decorrelation step makes the null-distribution of the test statistics much closer to standard normal, and thus increases the statistical power of the method. When the dimension of the data is mildly high, *i.e.* when $n/p \in (0.5, 1)$, our study on toy datasets suggests that type-I errors and statistical power of dCRT depend heavily the ℓ_1 regularization parameter. To the best of our knowledge, this phenomenon that has only been discussed in the case of linear regression [ZZ14]. Furthermore, in a regime where n/p grows much smaller than 0.5, our empirical observation suggests that it is impossible to do statistically-controlled feature selection while maintaining reasonable power. To tackle this issue, we incorporate a randomized clustering step

into our algorithm, coupled with a p-value aggregation technique. We demonstrate the effectiveness of our method with a reasonably high statistical power while controlling the False Discovery Rate (FDR) [BH95], a popular metric for measuring type-I error, in the case $n/p \geq 0.5$, or its spatially relaxation version δ-FDR when using clustering inference algorithms in the case $n/p < 0.5$.

Related works The closest cousin of the Conditional Randomization Test is *Model-X Knockoff Filter*. The method is a recent breakthrough in the False Discovery Rate (FDR) control literature. It relies on the creation of additional noisy features, called knockoffs, to calculate variable-importance statistics. Furthermore, the method requires only one fit of some high-dimensional estimator, subject to a hyperparameter selection that most often requires cross-validation. Thus, it has a iteration complexity of $\mathcal{O}(p^3)$ with the Lasso test statistic, hence much lower than the original CRT. Another work similar to distilled CRT is the *Holdout Randomization Test (HRT)* [TVZ+18], which is also an extension of vanilla CRT. While still requiring multiple sampling of noisy variables, HRT partially solves the computational issue of original CRT by doing heavy model fitting only once on one part of the dataset, and test statistics calculation on the other part that does not require refitting the model. However, this method relies on data-splitting, hence inherently suffers from a loss of statistical power. A parallel work with HRT is the *Conditional Permutation Tests (CPT)* [BWBS19], which is a non-parametric alternative to the CRT. As the name suggests, the method relies on a random shuffling mechanism applied on original variables, instead of multiple sampling of new variables. The authors argued that this makes CPT more robust to model mis-specification. A recent work of [YYMD21] proposes a method called *SLOE*, which adapts the analysis of [ZSC20], but only for the mildly high-dimensional case where $\lim_{n,p\to\infty} n/p \to \kappa \in (1, 2)$.

On a separate note, we notice the similarity of dCRT [LKJR20] with debiased Lasso [JM14, vdGBRD14, ZZ14]. These lines of work suggest that the ℓ_1-regularization optimization solution $\hat{\boldsymbol{\beta}}^{\text{LASSO}}$ [Tib96] is a biased estimator of the true signals $\boldsymbol{\beta}^0$ in the high-dimensional regime. Therefore, they proposed a debiasing formula for the estimator, which involves correcting $\hat{\boldsymbol{\beta}}^{\text{LASSO}}$ with a quantity proportional to regression residuals and diagonal elements of the estimated inverse matrix of population covariance $\boldsymbol{\Sigma}$. With this correction, the asymptotic distribution of $(\hat{\boldsymbol{\beta}}^{\text{LASSO}} - \boldsymbol{\beta}^0)$ is standard normal, and one can compute the test statistic and p-value associated with each variable. Debiased Lasso and dCRT have one operation that resembles each other: the distillation operator for each variable j in dCRT (Algorithm 6.2) and the nodewise Lasso for precision matrix estimation [vdGBRD14, Section 2.1.1]. This similarity is also conjectured in the recent work of [CMW20], demonstrated with a set of empirical benchmarks.

We present the time complexity of aforementioned methods in Table 6.1.

Table 6.1: **Time complexities of related methods with CRT-logit**, where p is the dimension size (number of variables), B is the number of sampling runs, and $\hat{k} \ll p$ the cardinality of the screening set (see Section 6.2.2 for more details).

Methods	**Time (Iteration) Complexity**	**References**
Debiased Lasso	$\mathcal{O}(p^3)$	[ZZ14, vdGBRD14, JM14]
Knockoff Filter	$\mathcal{O}(p^3)$	[BC15, CFJL18]
CRT	$\mathcal{O}(Bp^4)$ (with B the number of sampling)	[CFJL18]
Distilled-CRT	$\mathcal{O}(\hat{k}p^3)$	[LKJR20]
CRT-logit	$\mathcal{O}(\hat{k}p^3)$	**(this work)**

Summary of our contributions The two key contributions of our work are as follows.

1. We improve the Conditional Randomization Test to be more efficient in mildly high-dimensional logistic regression, *i.e.* when $n/p \in (0.5, 1.0)$, by decorrelating the randomized test statistics. We called the method CRT-logit.
2. We upscale the CRT-logit to work in extremely high-dimensional setting (defined by when $n/p < 0.5$), with dimension reduction by clustering and aggregation across randomized clusters. This is an extension of [CNTS21] to logistic regression.

Outline

- Section 6.2 presents the Conditional Randomization Test: the original formulation, the distillation version, and our extension of the method.
- Section 6.3 is an extensive empirical study of our proposed algorithm CRT-logit and some alternative methods for high-dimensional statistical inference. We provide empirical demonstrations on both simulated and realistic datasets, where we perform brain-imaging analysis and Genome Wide Association Study (GWAS).
- We conclude the paper with Section 6.4 by discussing limitations of CRT-logit and some open directions based on our observations.

6.2 Methodology

6.2.1 Preliminaries

Notation We will use the following system of notations. We denote matrices, scalar variables and sets by bold uppercase letters, script lowercase letters, and script letters, respectively (e.g. $\mathbf{X}$, x, and $\mathcal{X}$). The i-th row of matrix $\mathbf{X}$ will be denoted $\mathbf{X}_{i,*}$, the j-th column $\mathbf{X}_{*,j}$ and (i,j)-th element $\mathbf{X}_{i,j}$. For any integer p, we represent the set $\{1,\dots,p\}$ by $[p]$. For each $\mathbf{x} \in \mathbb{R}^p$ and $j \in [p]$, we denote $\mathbf{x}_{-j} \overset{\text{def.}}{=} \{x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_p\}$, or the vector of one observation after removing variable j. The cumulative distribution function (CDF) of the standard Gaussian distribution will be denoted $\Phi(\cdot)$. The indicator function of a random event $\mathcal{A}$ will be denoted $\mathbf{1}_{\mathcal{A}}$. We write $x \asymp y$ if $cy \leq x \leq Cy$ for some positive constants c and C.

Problem setting We will work exclusively with binary classification where the responses vector are denoted $\mathbf{y} \in \{0,1\}^n$ and data matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$ (consists of n observations of p features). Throughout the paper, we will assume that the data $\{\mathbf{X}_{i,*}\}_{i=1}^n$, are *i.i.d.* and follows a distribution with zero mean and population covariance matrix $\mathbf{\Sigma}$. Moreover, we assume that $\mathbf{X}_{i,*}$ and $\mathbf{y}_i$ follow the logistic relationship in Eq. (6.1). We denote the support set $\mathcal{S} \overset{\text{def.}}{=} \{j \in [p] : \beta_j^0 \neq 0\}$ and assume that it is sparse, *i.e.* $\text{card}(\mathcal{S}) = k \ll p$. Furthermore, $\hat{\mathcal{S}} \overset{\text{def.}}{=} \{j \in [p] : \hat{\beta}_j \neq 0\}$ indicates an estimation of $\mathcal{S}$, where $\hat{\beta}_j$ is an estimate of the true signal β_j^0. We try to obtain it through a regularized MLE estimator:

$$\hat{\boldsymbol{\beta}}^{\text{MLE}} = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{X}_{i,*}^T\boldsymbol{\beta}))\right] + \lambda\|\boldsymbol{\beta}\|_1. \tag{6.2}$$

One of the multiple testing metrics used in many applications is the *False Discovery Rate*, introduced by [BH95]. Given an estimate of the support $\hat{\mathcal{S}}$, the false discovery proportion (FDP) is the ratio of the number of selected features that do not

belong to the true support $\mathcal{S}$, divided by the cardinality of $\hat{\mathcal{S}}$, or the total number of selected features. The False Discovery Rate is the expectation of the FDP. In other words, we can write the formula for FDP and FDR as follows:

$$\text{FDP}(\hat{\mathcal{S}}) = \frac{\#\{j : j \in \hat{S}, j \notin S\}}{\text{card}(\hat{S}) \vee 1} \qquad \text{and} \qquad \text{FDR}(\hat{\mathcal{S}}) = \mathbb{E}[\text{FDP}].$$

6.2.2 Distilled Conditional Randomization Test and its Extension to High-Dimensional Logistic Regression

The concept of Conditional Randomization Test was originally introduced in the model-X knockoff paper [CFJL18] as one way to output valid empirical p-values using knockoff variables. The knockoff filter is another recent advance in high-dimensional statistical inference that aims at controlling the False Discovery Rate. The principle of the knockoff filter is first to sample noisy copies $\tilde{\mathbf{X}}_{*,j}$ of variable $\mathbf{X}_{*,j}$, given a sampling mechanism $P_{j|-j}$. The method is *distribution free*, in the sense that we place no specific assumption on the distribution of the test statistic inferred from dataset $\mathcal{D}$. However, this means in general, there is no mechanism to derive the p-value p_j from the knockoff statistic T_j for each $j = 1, \ldots, p$ (we do note that there exists a later work by [NCTA20] that proposed a conversion of knockoff statistics to p-values). This motivates the introduction of the CRT. We present the pseudo-code for CRT in Algorithm 6.1.

Algorithm 6.1: Conditional Randomization Test [CFJL18]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, with $\mathbf{X} \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, number of sampling runs B, test statistic T_j, conditional distribution $P_{j|-j}$ for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;
for $j = 1, 2, \ldots, p$ **do**
  for $b = 1, 2, \ldots, B$ **do**
    1. Generate $\tilde{\mathbf{X}}_{*,j}^{(b)}$, a knockoff sample from $P_{j|-j}$;
    2. Compute test statistics T_j for original variable and $\tilde{T}_j^{(b)}$ for knockoff variables;
  end
  Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

The original CRT requires running B times high-dimensional inference for each variable j. Assuming we use the Lasso program to compute T_j, CRT runtime would be $\mathcal{O}(Bp^4)$, since the Lasso has an optimization cost of $\mathcal{O}(p^3)$ with coordinate descent algorithm when $n < p$ [HTF09, pp. 93]. This runtime is prohibitive when p grows large. Moreover, a decently large B is required to make the empirical distribution of the p-values smooth enough.

Distillation Conditional Randomization Test Reducing the computational cost of CRT is the main motivation of several works [TVZ+18, BWBS19, LKJR20]. One of them is the introduction of the distilled-CRT (dCRT) in [LKJR20]. The main appeal of this method is that it can output p-values analytically. Lasso-dCRT improves upon vanilla CRT by bypassing the multiple sampling step used to infer each variable. In other words, the knockoffs sampling step is eliminated, which leads to a reasonable reduction of the computation cost. However, this hinges on the

assumption that the distribution of the test statistics is known. In particular, the authors assumed that the test statistic T_j follows a standard Gaussian under the null hypothesis. Moreover, while the loop of multiple knockoff sampling has been removed, run-time complexity of dCRT is still $\mathcal{O}(p^4)$. To deal with this problem, [LKJR20] introduced a screening step to eliminate potentially unimportant variables before distillation. The p-values of unselected variables are set to 1.0, with the assumption that the screening step estimated well the true support, or more specifically, $\mathcal{S} \subset \hat{\mathcal{S}}^{\text{SCREENING}}$. The runtime is reduced to $\mathcal{O}(\hat{k}p^3)$, where $\hat{k} \overset{\text{def.}}{=} \text{card}(\hat{\mathcal{S}}^{\text{SCREENING}})$ is the number of selected variables by. In high-dimension, it is usually expected that the ground truth signal $\boldsymbol{\beta}^0$ is sparse, *i.e.* $\hat{k} \ll p$. Therefore, using distillation only on the estimated support set $\mathcal{S}$ reduces computational cost drastically.

Distillation operator The key idea behind dCRT is the distillation operation: for each variable j, we want to distill all the conditional information of the remaining variables $\mathbf{X}_{-j}$ to $\mathbf{x}_j$ and to $\mathbf{y}$. Here, $\mathbf{X}_{-j}$ is the data matrix $\mathbf{X}$ with column $\mathbf{X}_{*,j}$ removed. This is done via either least-squares loss minimization (original dCRT) or logistic regression (ours) with ℓ_1-regularization to enforce sparsity. The authors of [LKJR20] suggest to perform distillation as follows. First, for each variable j, we solve the lasso problem by regressing $\mathbf{X}_{*,j}$ on $\mathbf{X}_{-j}$,

$$\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{2} \|\mathbf{X}_{*,j} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1. \tag{6.3}$$

For distillation of variable j and the binary response $\mathbf{y}$ with logistic relationship, [LKJR20] briefly suggested to solve a penalized estimation problem similar to Eq. (6.2):

$$\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \sum_{i=1}^{n} \log\left[1 + \exp(-y_i(\mathbf{X}_{i,-j}^T \boldsymbol{\beta}))\right] + \lambda \|\boldsymbol{\beta}\|_1. \tag{6.4}$$

Then, a test statistic is calculated for each $j = 1, \ldots, p$:

$$T_j = \frac{\sqrt{n}(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{X}_{*,j} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})}{\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|_2 \left\|\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}}\right\|_2}, \tag{6.5}$$

which they assume to follow a Gaussian distribution asymptotically, conditional to $\mathbf{y}$ and $\mathbf{X}_{-j}$:

$$T_j \xrightarrow[n \to +\infty]{(d)} \mathcal{N}(0, 1)\,. \tag{6.6}$$

It follows that we can output a p-value for each variable j by taking

$$p_j = 2\left[1 - \Phi(|T_j|)\right], \tag{6.7}$$

where $\Phi(\cdot)$ is the cumulative distribution function (CDF) of standard normal distribution. A pseudo-code for dCRT is presented in Algorithm 6.2.

Remark 6.1. *An important point to note is that CRT and dCRT are just methods to output p-values, and not tied to a specific error-rate metric such as the False Discovery Rate.*

Decorrelating test statistics for high-dimensional logistic regression While the formula of the test statistic in Eq. (6.5) is adequate for linear regression, it is not truly satisfactory in the setting we are considering. More specifically, both the calculation of regression residuals $\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}$ and test statistics T_j do not take into account the non-linear relationship between $\mathbf{X}$ and the binary response $\mathbf{y}$. The first row of Figure 6.1 plots the qq-plot of the test statistics T_j for logistic regression, which shows that even in the classical regime where $n > p$, its distribution is far from standard normal. In this work, we adapt the *decorrelated test score* of [NL17]

Algorithm 6.2: Lasso-Distillation Conditional Randomization Test [LKJR20]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, $X \in \mathbb{R}^{n \times p}$, $\mathbf{y} \in \mathbb{R}^n$, test statistic T_j for each $j = 1, \ldots, p$;

OUTPUT vector of p-values $\{p_j\}_{j=1}^p$;

$\hat{\mathcal{S}}^{\text{SCREENING}} = \{j : j \in [p], \hat{\beta}_j^{\text{MLE}} \neq 0\}$ // Using Eq. (6.2)

for $j \in \hat{\mathcal{S}}^{SCREENING}$ **do**

1. Distill information of $\mathbf{X}_{-j}$ to $\mathbf{X}_{*,j}$ and to $\mathbf{y}$ by finding:
   - $\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \sum_{i=1}^n \log\left[1 + \exp(-y_i(\mathbf{X}_{i,-j}^T \boldsymbol{\beta}))\right] + \lambda \|\boldsymbol{\beta}\|_1$
   - $\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{2} \|\mathbf{X}_{*,j} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$
2. Obtain test statistic:
$$T_j = \frac{\sqrt{n}(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})}{\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|_2 \left\|\mathbf{X}_{*,j} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}}\right\|_2}$$
3. Compute (two-sided) p-value
$$p_j = 2\left[1 - \Phi(|T_j|)\right]$$

end

to improve the accuracy of the estimator in the case of logistic regression. As the name suggests, a decorrelation step for calculating the test statistic T_j is performed. This is preceded by an operation similar to distillation of [LKJR20]. There is a difference in the definition of $\hat{\boldsymbol{\beta}}^{d_{x_j}}$. Instead of solving normal Lasso program in Eq. (6.3), [NL17] proposed using the following scaled Lasso formula:

$$\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda_{dx}) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^{p-1}} \frac{1}{n} \sum_{i=1}^n \frac{\exp(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})}{[1 + \exp(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})]^2} (\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}\boldsymbol{\beta})^2 + \lambda_{dx}\|\boldsymbol{\beta}\|_1, \tag{6.8}$$

where $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}^{\text{MLE}}$ is defined in Eq. (6.2). On the other hand, we obtain $\hat{\boldsymbol{\beta}}_j^{d_y}$ from $\hat{\boldsymbol{\beta}}$ by omitting the j-th coefficient from it, *i.e.*

$$\hat{\boldsymbol{\beta}}^{d_y} = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p). \tag{6.9}$$

The equation for decorrelated test score writes

$$T_j^{\text{decorr}} = -\frac{1}{\sqrt{n}} \hat{\mathbf{I}}_{j|-j}^{-1/2} \sum_{i=1}^n \left[y_i - \frac{1}{1 + \exp(-\mathbf{X}_{i,-j}\hat{\boldsymbol{\beta}}^{d_y})} \right] \left[\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}^T \hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda_{dx}) \right], \tag{6.10}$$

where $\hat{\mathbf{I}}_{j|-j}^{-1/2}$ is the estimated partial Fisher information matrix, that is estimated by:

$$\hat{\mathbf{I}}_{j|-j}^{-1/2} = \frac{1}{n} \sum_{i=1}^n \frac{\exp(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})}{[1 + \exp(\mathbf{X}_{i,*}\hat{\boldsymbol{\beta}})]^2} \left[\mathbf{X}_{i,j} - \mathbf{X}_{i,-j}\, \hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda_{dx}) \right]^2 \mathbf{X}_{i,j}. \tag{6.11}$$

Then we have, under additional assumptions, the following result [NL17]:

$$T_j^{\text{decorr}} \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0, 1). \tag{6.12}$$

A summary of CRT-logit can be found in Algorithm 6.3.

Algorithm 6.3: CRT-logit

INPUT dataset $(\mathbf{X}, \mathbf{Y})$, $\mathbf{X} \in \mathbb{R}^{n \times p}, \mathbf{y} \in \mathbb{R}^n$;
OUTPUT vector of p-values $\{p_j\}_{j=1}^p$;
$\hat{\boldsymbol{\beta}}^{\text{MLE}} \leftarrow$ `regularized_MLE`; `// Using Eq. (6.2)`
$\hat{\mathcal{S}}^{\text{SCREENING}} \leftarrow \{j : j \in [p], \hat{\beta}_j^{\text{MLE}} \neq 0\}$;
for $j \in \hat{\mathcal{S}}^{SCREENING}$ **do**
  1. $\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda_{dx}) \leftarrow$ `scaled_lasso`$(\mathbf{X}_{*,j}, \mathbf{X}_{*,-j}, \lambda_{dx})$ `// Using Eq. (6.8)`
  2. $\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{j-1}, \hat{\beta}_{j+1}, \ldots, \hat{\beta}_p)$ `// Using Eq. (6.9)`
  3. $T_j^{\text{decorr}} \leftarrow$ `decorrelated_test_score`$(\mathbf{X}, \mathbf{y}, j)$ `// Using Eq (6.10)`
  4. $p_j \leftarrow 2[1 - \Phi\left(|T_j^{\text{decorr}}|\right)]$
end
for $j \notin \hat{\mathcal{S}}^{SCREENING}$ **do**
  $p_j = 1$
end

Intuition on the decorrelated score Let $\ell(\boldsymbol{\beta})$ be the negative log-likelihood function based on parameter $\boldsymbol{\beta}$. In short, the decorrelated test score T_j^{decorr} is based on classical Rao's score function for testing $H_0^j : \beta_j^0 = 0$ versus $H_\alpha^j : \beta_j^0 \neq 0$. This score function relies on $\nabla_j \ell(\beta_j, \hat{\boldsymbol{\beta}}_{-j})$, or the partial derivative of the negative log-likelihood at β_j. Here, $\hat{\boldsymbol{\beta}}_{-j} = \text{argmin}_{\boldsymbol{\beta}_{-j} \in \mathbb{R}^{p-1}} \ell(\beta_j, \boldsymbol{\beta}_{-j})$ is a constrained maximum likelihood estimator of $\boldsymbol{\beta}_{-j}$ with fixed β_j. We have, under null hypothesis [CH79], that

$$n^{1/2} \nabla_j \ell(\beta_j, \hat{\boldsymbol{\beta}}_{-j}) \xrightarrow[n \to \infty]{(d)} \mathcal{N}(0, \mathbf{I}_{j|-j}). \tag{6.13}$$

Thus, the Rao's score function is given by:

$$T_j^{\text{Rao}} = n^{1/2} \nabla_j \ell(\beta_j, \hat{\boldsymbol{\beta}}_{-j}) \hat{\mathbf{I}}_{j|-j}^{-1/2},$$

where $\hat{\mathbf{I}}_{j|-j}^{-1/2}$ is a consistent estimator of $\mathbf{I}_{j|-j}^{-1/2}$. However, in high-dimensional regime, where $n < p$, this formulation will not work. To see this, consider the Taylor expansion of the score function $\nabla_j \ell(\beta_j, \cdot)$ for any estimator $\bar{\boldsymbol{\beta}}_{-j}$:

$$\nabla_j \ell(\beta_j, \bar{\boldsymbol{\beta}}_{-j}) = \nabla_j \ell(\beta_j, \boldsymbol{\beta}_{-j}^0) + \nabla_{j,-j}^2 \ell(\beta_j, \boldsymbol{\beta}_{-j}^0)(\bar{\boldsymbol{\beta}}_{-j} - \boldsymbol{\beta}_{-j}^0) + \texttt{Remaining}.$$

On the right hand side, the first term converges weakly to a normal distribution due to central limit theorem, but the second and `Remaining` term will not, due to estimation bias and sparsity effect of $\bar{\boldsymbol{\beta}}_{-j}$, as argued in [FK00].

Therefore, for each variable j, the authors of [NL17] proposed to "debias" the score function by correcting the impact of other terms, by

$$\nabla_j \ell(\beta_j, \boldsymbol{\beta}_{-j}) - \mathbf{I}_{j,-j} \mathbf{I}_{-j,-j}^{-1} \nabla_{-j} \ell(\beta_j, \boldsymbol{\beta}_{-j}), \tag{6.14}$$

where $\mathbf{I} = \mathbb{E}_{\boldsymbol{\beta}}[\nabla^2 \ell(\boldsymbol{\beta})]$ is the Fisher information matrix, $\mathbf{I}_{i,-j}$ is the i-th row of $\mathbf{I}$ after removing j-th element, and $\mathbf{I}_{-j,-j}$ the matrix after removing j-th row and column. Define $\mathbf{w}^T = \mathbf{I}_{j,-j} \mathbf{I}_{-j,-j}^{-1}$, we can estimate it by

$$\hat{\mathbf{w}} = \underset{\mathbf{w} \in \mathbb{R}^{p-1}}{\text{argmin}} \frac{1}{2n} \sum_{i=1}^n \left\{ \nabla_j \ell_i(\hat{\boldsymbol{\beta}}) - \mathbf{w}^T \nabla_{-j} \ell_i(\hat{\boldsymbol{\beta}}) \right\}^2 + \|\mathbf{w}\|_1, \tag{6.15}$$

where $\hat{\boldsymbol{\beta}}$ is obtained with Eq. (6.2). This formula coincides with the x_j-distillation formula in Eq. (6.8). A detailed explanation, including geometric insights for the derivation of Eq. (6.8), (6.9), (6.10) and (6.11) can be found in [NL17, Section 2].

Effectiveness of decorrelation step To show how decorrelating the test statistic actually helps, we set up a simulation with matrix $\mathbf{X}$ of $p = 400$ features and vary the number of samples $n \in \{200, 400, 800\}$. The binary response vector $\mathbf{y}$ is created following Eq. (6.21) instead of Eq. (6.1) (the reader can refer to our explanation in Remark 6.1). More specifically, the design matrix $\mathbf{X}$ is sampled from a multivariate normal distribution of zero mean and the covariance matrix $\mathbf{\Sigma} \in \mathbb{R}^{p\times p}$ as a symmetric Toeplitz matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} \rho^0 & \rho^1 & \dots & \rho^{p-1} \\ \rho^1 & \ddots & \dots & \rho^{p-2} \\ \vdots & \dots & \ddots & \vdots \\ \rho^{p-1} & \rho^{p-2} & \dots & \rho^0 \end{bmatrix}.$$

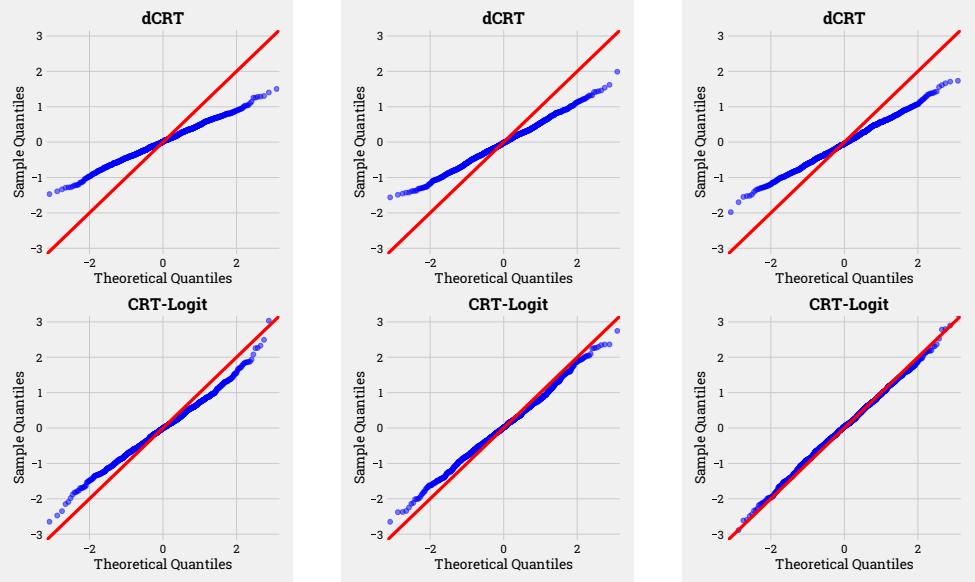
The parameter $\rho \in (0, 1)$ controls the correlation between neighboring features, and this correlation decreases quickly when the distance between feature indices increases. The true signal $\boldsymbol{\beta}^0$ is picked with a sparsity parameter $\kappa = \text{card}(\mathcal{S})/p = k/p$ that controls the proportion of non-zero elements with magnitude 2.0, *i.e.* $\beta_j = 2.0$ for all $j \in \mathcal{S}$. For the specific purpose of this experiment, non-zero indices of $\mathcal{S}$ will be kept fixed. The noise $\boldsymbol{\xi}$ (following Eq. 6.21) is *i.i.d.* normal $\mathcal{N}(0, \mathbf{I}_n)$ with magnitude $\sigma = \|\mathbf{X}\boldsymbol{\beta}^0\|_2/(\sqrt{n}\ \text{SNR})$, controlled by the SNR parameter. In short, the three main parameters controlling this simulation are correlation ρ, sparsity degree κ and signal-to-noise ratio SNR.

We then generate randomly 1000 datasets. For each, we run dCRT and CRT-logit algorithm and generate one sample of test statistics $\{T_j\}_{j=1}^p$ and $\{T_j^{\text{decorr}}\}_{j=1}^p$. We then pick 1000 samples of one null test statistic T_j and T_j^{decorr}, defined in Eq. (6.5) and (6.10), respectively, and plot the qq-plot of their empirical quantile versus the standard normal quantile.

From the results in Figure 6.1, we observe that the empirical null distribution of the test statistic is much similar to a standard normal when adding the decorrelation step. In particular, when the sample size n increases to 800, the decorrelated test statistic has empirical quantile almost inline with the theoretical quantile of the standard normal distribution, while the original dCRT test score still stays away from the 45-degree line. Again, we emphasize that the normality of T_j is crucial for the p-values calculation in Eq. (6.7). This outlines the importance of decorrelating step on T_j.

6.2.3 Setting the ℓ_1-Regularization Parameter of the $\mathbf{X}_{*,j}$-distillation

A core issue is the dependency of the statistical power and FDR of CRT-logit on the ℓ_1- regularization parameter λ_{dx} when doing Lasso distillation on x_j in Eq. (6.8). One might choose the heuristic value $\lambda_{\text{univ}} = \sqrt{n^{-1}\log p}$ with theoretical validity, as suggested in [NL17, vdGBRD14]. However, experimental results in Fig. 6.2 show that at $\lambda_{dx} = \lambda_{\text{univ}}$ (or $\log_{10}\lambda/\lambda_{\text{univ}} = 0.0$ with the labeling of the figure), we do not have the best possible FDR/Power with CRT-logit inference. For this experiment, we average the inference results of 100 simulations (with similar setting in Section 6.2.2) for different values of n and λ_{dx}, with p fixed. There is a clear phase transition in both FDR and average power when the regularization parameter λ_{dx} increases. In other words, we have found empirically that both FDR and power of the method are sensitive to the ℓ_1-regularization parameter. Preferably, one wants to return a high statistical power while controlling FDR under predefined level. Hence, it is necessary to choose λ_{dx} wisely. In a more practical scenario, we advise the usage of cross-validation for $\mathbf{X}_{*,j}$-distillation operator, as defined by Eq. (6.8). This means we would have to find p different values of λ_{dx} with cross-validation, and we reemphasize the importance of the screening step to reduce the number of computations. With



(a) $\mathbf{n = 200}$ (b) $\mathbf{n = 400}$ (c) $\mathbf{n = 800}$

Figure 6.1: **QQ-Plot for one null CRT statistic for logistic regression, with varying number of samples and a fixed number of variables** $p = 400$. This obtained over 1000 samplings. The theoretical quantile is obtained from a standard Gaussian distribution. The decorrelation step makes the empirical null distribution of the null statistics much closer to standard Gaussian. Default simulation parameters: SNR $= 3.0, \rho = 0.4,$ sparsity $= 0.06$. *Upper row: original dCRT statistic defined by Eq.* (6.5). *Bottom row: CRT-logit, with decorreleated test score defined by Eq.* (6.10) ***(our method)***.

the mildly high-dimensional setting where $n/p \in (0.5, 1)$,we recommend taking $\lambda_{dx} = \sqrt{10}\lambda_{\text{univ}}$, as observed from the results in Fig. 6.2.

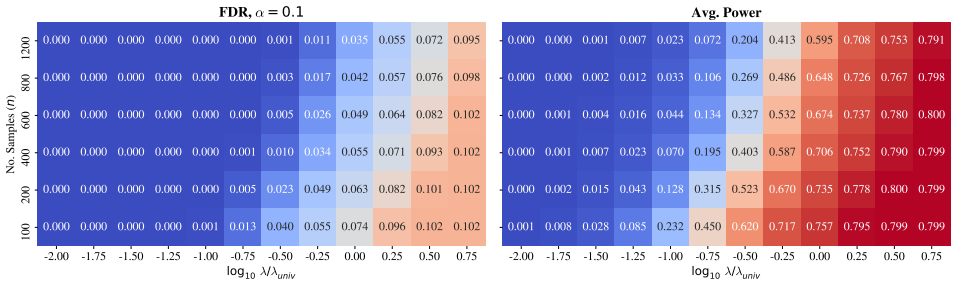


Figure 6.2: **FDR/Average Power of 100 runs of simulations while varying the number of samples and ℓ_1 regularization parameter and fixing the number of variables**. Note: λ_{dx} is scaled with the factor $\lambda_{\text{univ}} = \sqrt{\log(p)/n}$, *e.g.* the first value for regularization grid is $\lambda_{dx} = 10^{-2}\lambda_{\text{univ}}$. Default parameter (similar settings in Section 6.2.2): $p = 400$, SNR=3.0 (signal-to-noise ration), $\rho = 0.5$ (correlation), $\kappa = 0.05$ (sparsity). FDR is controlled at level $\alpha = 0.1$.

6.2.4 High-dimensional Statistical Inference with Spatial Relaxation

Ineffectiveness of CRT in extremely high-dimensional problems When the number of observations n is too small compared to the number of variables p,

e.g. when $n/p < 0.2$ as shown in Figure 6.3, the inference problem becomes too ill-posed. Indeed, the statistical power of both the original dCRT and our proposed solution CRT-logit decrease dramatically from a large value in the easy setting ($p < n$) to zero when $p > 1600$. The failure to detect any significant variable when the dimension of the problem becomes too high motivates us to propose a dimension reduction step using a clustering method. The works of [BRvdGZ13, CNTS21] have provided detailed discussions on this matter.

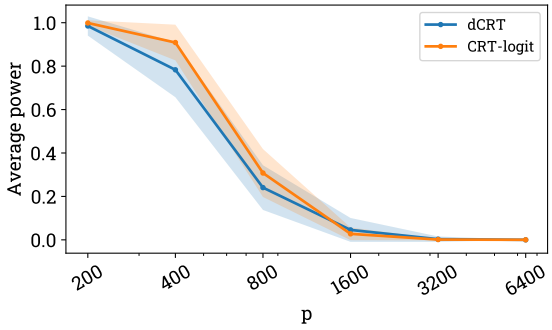


Figure 6.3: **FDR/Average Power of 100 runs of simulations while varying the number of variables p and fixing the number of observations** $n = 400$. Default parameter: SNR $= 2.0, \rho = 0.5, \kappa = 0.04$. FDR is controlled at level $\alpha = 0.1$. The experimental setup is similar to Section 6.2.2. Both methods (`dCRT`: original dCRT and `CRT-logit`: our version of CRT) perform well in easy settings where $n \geq p$, but cannot detect any variables when p becomes large compared to n.

Randomized Clustering To make the inference problem more tractable, we propose to work with clusters of correlated variables. By doing clustering and replacing groups of original variables with their mean value, the dimension of the problem goes from p to a much smaller value C. This means that we can potentially improve statistical power and reduce computational time simultaneously. Naturally, the choice of C comes with a trade-off: larger clusters (smaller C) are more easily selected by the inference algorithm, but will likely match poorly the geometry of the true signal $\boldsymbol{\beta}^0$, and conversely.

Remark 6.2. *In order for randomized clustering to work, we implicitly assume that true variables are mostly aggregated in a small number of clusters. Otherwise, there might exist many strong correlations between a null and a true variable, which leads to plenty of clusters being wrongly selected. A simpler (but stronger) version is to assume that true variables are located in a cluster-like pattern. Readers will see later in Section 6.3.2 that we create synthetic high-dimensional dataset following this assumption.*

Any clustering scheme can easily be randomized, simply by basing it on a random subsample of observations. For a more stable inference, we thus propose doing multiple runs of clustering on subsamples of observations, then aggregating the p-values output by CRT in each run with the aggregation technique of [MMB09]. In particular, assuming that for each variable j we have B p-values $p_j^{(b)}$ corresponding with runs $b = 1, \dots B$, the authors argued that a suitable function for aggregating these p-values relies on a quantile:

$$\bar{p}_j = \min \left\{ 1, \frac{q_\gamma\left(\{p_j^{(b)} : b \in [B]\}\right)}{\gamma} \right\}, \tag{6.16}$$

whereby $q_\gamma(\cdot)$ is the empirical γ-quantile function with fixed $\gamma \in (0,1)$. Since choosing a suitable γ might be difficult, [MMB09] also proposed an adaptive aggregation version:

$$\bar{p}_{j,\text{adapt}} = \min\left\{1, (1-\log\gamma_{min}) \inf_{\gamma\in(\gamma_{min},1)} \bar{p}_j\right\}, \tag{6.17}$$

with $\gamma_{min} \in (0,1)$ to be chosen. For a theoretical treatment of the ensemble of multiple clusters, we refer to the work of [CNTS21].

We summarize the procedure in Algorithm 6.4.

Algorithm 6.4: Ensemble of randomized clusters [CNTS21]

INPUT dataset $(\mathbf{X}, \mathbf{y})$, $\mathbf{X} \in \mathbb{R}^{n\times p}$, $\mathbf{y} \in \mathbb{R}^n$; parameters: B the number of subsampling runs, γ or γ_{min} the choice of quantile, C number of clusters, r subsample size;

OUTPUT vector of p-values $\{\bar{p}_j\}_{j=1}^p$;

for $b = 1, \ldots, B$ **do**

- $\tilde{\mathcal{D}}^{(b)} \leftarrow$ `uniformly_subsampling`$(\mathcal{D}, r)$ // $\tilde{\mathcal{D}}^{(b)} = \{\mathbf{x}_i, y_i\}_{i\in\mathcal{K}}$ where $\mathcal{K} \subset [n]$
- $\mathcal{Z}^{(b)} = \{\mathbf{z}_i, y_i\}_{i\in\mathcal{K}} \leftarrow$ `clustering`$(\tilde{\mathcal{D}}^{(b)}, C)$ // $\mathbf{z}_i \in \mathbb{R}^C$ where $C \ll p$
- $\{q_j^{(b)}\}_{j\in[C]} \leftarrow$ `inference`$(\mathcal{Z}^{(b)})$ // cluster-wise p-values
- $\{p_j^{(b)}\}_{j=1}^p \leftarrow$ `broadcast` $\left(\{q_j^{(b)}\}_{j\in[C]}, \mathcal{Z}^{(b)}, \tilde{\mathcal{D}}^{(b)}\right)$ // variables of same cluster get same p-value

end

$\bar{p}_j \leftarrow$ `aggregate`$\{p_j^{(b)}\}_{b\in[B]}$ for each $j = 1, \ldots, p$, with either Eq. (6.16) or Eq. (6.17)

False Discovery Rate with a spatial relaxation With the dimension reduction by grouping variables into clusters, it is necessary to introduce a spatial relaxation of the False Discovery Rate. We note that a similar extension has been introduced in [GZ19a]. We call this metric the FDR^δ, where δ is a positive integer. In short, the idea of adding δ is to add a spatial relaxation, where variables selected at a distance closer than δ from the true support $\boldsymbol{\beta}^0$ are not considered as false discoveries. This is due to the fact that when doing clustering of the design matrix, all variables located in the same cluster are selected. More precisely, δ should correspond to feature groups diameter. We note that a similar idea was presented in [CNTS21], but the latter considered Family-Wise Error Rate control. The formal definitions of FDP^δ and FDR^δ are as follows.

Definition 6.1. *Given an estimate of the support $\hat{\mathcal{S}}$, the spatially relaxed false discovery proportion with parameter $\delta > 0$, denoted FDP^δ, is the number selected features that are at a distance of more than δ from any feature of the true support, divided by the number of selected features:*

$$\text{FDP}^\delta = \frac{\#\{\ell \in \hat{\mathcal{S}} : \forall j \in \mathcal{S},\ d(j,\ell) > \delta\}}{\text{card}(\hat{\mathcal{S}}) \vee 1}, \tag{6.18}$$

where $d(j,\ell)$ is the distance between the j-th and the ℓ-th feature.

Definition 6.2. *The spatially relaxed false discovery rate with parameter $\delta > 0$, denoted FDR^δ, is the expectation of the FDP^δ:*

$$\text{FDR}^\delta = \mathbb{E}[\text{FDP}^\delta]. \tag{6.19}$$

Example 6.1. In brain imaging application (functional Magnetic Resonance Imaging —fMRI— data analysis), [CNS+21] proposed the following heuristic formula for choosing δ:

$$\delta_{\text{univ}} = \left(\frac{p}{2C}\right)^{1/3}, \tag{6.20}$$

where p is the number of variables (brain voxels) and C is the number of clusters. In other words, the value of δ varies approximately linearly with the cubic root of the average number of variables (or brain voxels) per cluster. Note that for genomics application, this heuristic formula does not apply, as variables are located on a one-dimensional lattice structure.

Remark 6.3. *The type of spatial relaxation to the error rate we consider in this work,* i.e. *the δ relaxation presented above, is possible in practical applications we focus on, since there always exists a natural distance measure on the set of variable $[p] = \{1, \dots, p\}$. More specifically, in brain imaging, the distance d corresponds to the Euclidean norm in some brain space (a subdomain of $\mathbb{R}^3$), and in genomics it corresponds to the ℓ_1 norm in $\mathbb{R}$. We also assume clusters of variables to be spatially consistent with d.*

6.3 Empirical Results

We provide benchmarks of the proposed CRT-logit algorithm along with most other methods mentioned in the introduction, in particular: model-X Knockoff (KO) [CFJL18], Debiased Lasso (dLasso) [ZZ14, JM14], Lasso-Distillation CRT (dCRT) [LKJR20]. We did not include SLOE, HRT and CPT as the provided open-source implementation are particularly slow, and that these methods do not perform well when $n < p$. Here is a summary of the set of experiments:

- Section 6.3.1 is based on a mildly high-dimensional ($n/p = 2/3$) synthetic dataset, where data display autocorrelation in 3D.
- Section 6.3.2 is also based on a synthetic dataset, but with much smaller ratio $n/p < 1/10$. The non-zero components of $\boldsymbol{\beta}^0$ are located in a cluster-like pattern, *i.e.* some non-zero groups followed by zero groups.
- Section 6.3.3 features inference on a semi-realistic dataset. The Human Connectome Project (HCP) is a functional magnetic resonance imaging (fMRI) dataset, which is an extremely high-dimensional setting with over 200,000 variables (brain voxels).
- Lastly, section 6.3.4 is a Genome-Wide Association Study (GWAS) of human brain cancer related gene, with a set of over 25000 genes as the number of variables and over 1000 of recorded patients.

For the latter three sets of experiments, we use clustering-based reduction of the inference procedure, additionally with the aggregation of multiple clustered CRT-logit. Following the explanation in Section 6.2.4, we aim to control the spatially relaxed δ-FDR in those set of benchmarks.

Remark 6.1. *As a slight caveat, in the synthetic and semi-synthetic experiment sections (6.3.1, 6.3.2, 6.3.3), we introduce an additional noise term to the logistic relationship equation (6.1):*

$$\mathbb{P}(y_i = 1 \mid \mathbf{x}_i) = \frac{1}{1 + \exp(-\mathbf{x}_i^T \boldsymbol{\beta}^0 + \sigma \xi_i)}. \tag{6.21}$$

where $\xi_i \sim \mathcal{N}(0, 1)$ is a standard Gaussian noise and $\sigma > 0$ the noise magnitude. There is a clear justification for the addition of noise: in most of the applications

we consider, data are often collected with measurement errors. In the case of brain-imaging, for example, recording the brain signal of the human subjects by magnetic resonance imaging scanners often includes noise caused either from the machine, or from the movement of the subjects during the scanning period [Lin08]. The formula in Eq. (6.21) has been used in previous works, e.g. *[BEG+15].*

6.3.1 Simulation: Mildly High-Dimensional Scenario

Our first experiment is a simulation scenario where a design matrix $\mathbf{X}$ ($n = 400, p = 600$) and a binary response vector $\mathbf{y}$ are created following the logistic relationship in Eq. (6.21). The data-generating process is the same as in Section 6.2.2, except indices of the non-zero elements of $\boldsymbol{\beta}^{)}$ are sampled uniformly at random, instead of keeping it fixed. To see how each algorithm performs under different settings, we vary each of the three simulation parameters while keeping the others unchanged at default value of SNR $= 4.0$, $\rho = 0.6$, $\kappa = 0.04$. We target a control of FDR at level 0.1, using Benjamini-Hochberg procedure.

Results in Figure 6.4 show that CRT-logit is the most powerful method while still controlling the FDR. Moreover, in the presence of higher correlations between nearby variables ($\rho > 0.6$), other methods suffer a drop in average power, but this is not as severe for CRT-logit. Debiased Lasso (`dlasso`) controls the FDR well in all settings, but is the most conservative procedure. We also do the same set of simulation but in

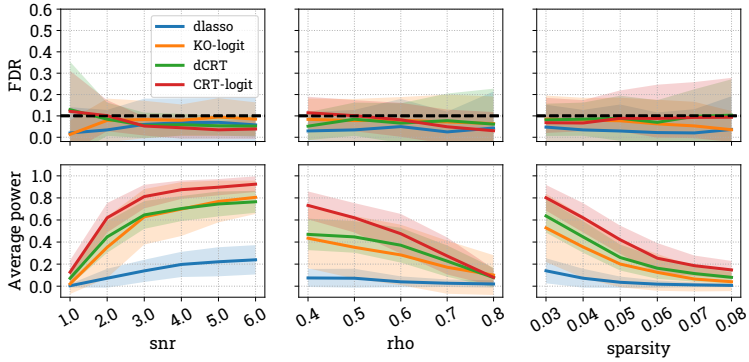


Figure 6.4: **FDR/Average Power of 100 runs of simulations across varying parameters in mildy high-dimensional settings**. Default parameter: $n = 400, p = 600, \text{SNR} = 2.0, \rho = 0.5, \kappa = 0.04$. FDR is controlled at level $\alpha = 0.1$. Methods: Debiased Lasso (`dlasso`), model-X Knockoff (`KO-logit`), original dCRT (`dCRT`), our version of CRT (`CRT-logit`).

a classical setting where $n = 600, p = 400$, whose results are shown in Figure 6.5. In this classical setting, we observe that CRT-logit remains the most powerful method, while Knockoffs with logistic loss (`KO-logit`) performs poorly. To investigate on this phenomenon, we plot the histogram of knockoff logistic in low-dimension for all results in Section 6.5.2. In short, in a large proportion of inference runs, in order to return zero false discovery proportion, `KO-logit` makes no discoveries. The same phenomenon has been observed in [NCTA20]. A reason is to keep the average FDP (or FDR) under pre-defined level at the cost of over-conservativeness.

6.3.2 Simulation: High-Dimensional Scenario With Clustered-Located Support

For the high-dimensional scenario, we decrease the ratio $n/p < 1/10$, with $n = 600$ and $p = 8000$. Instead of a uniform sampling like the previous section, non-zero components of the true signal $\boldsymbol{\beta}^0$ are localized in a cluster-like pattern, *i.e.* some non-zero

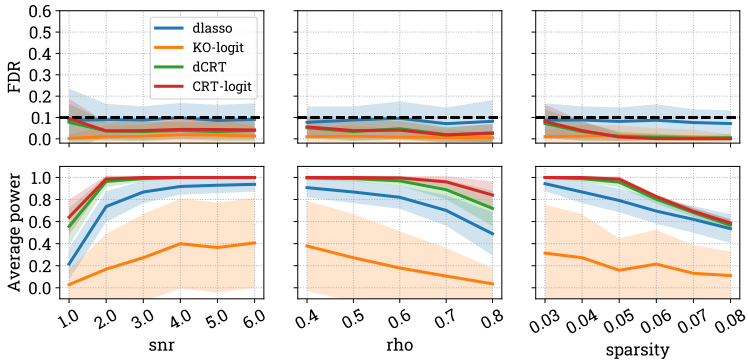


Figure 6.5: **FDR/Average Power of 100 runs of simulations in classical setting ($n > p$) across different parameters**. Default: $n = 600, p = 400$, snr $= 3.0, \rho = 0.5$, sparsity $= 0.04$. FDR is controlled at level $\alpha = 0.1$. Methods: Debiased Lasso (`dlasso`), model-X Knockoff (`KO-logit`), original dCRT (`dCRT`), our version of CRT (`CRT-logit`)

groups followed by zero groups. We still maintain the three main simulation parameters: SNR for noise magnitude, ρ for Toeplitz correlation structure and κ for sparsity degree of the true signal. We remark that this setting makes it almost impossible to detect significant variables without using dimension-reduction technique, as shown in the first two columns of Figure 6.6: both KO-logit ad CRT-logit are extremely conservative. Therefore, all the methods used involve a preliminary clustering step. The clustering version of Knockoff (`cKO-logit`) and CRT-logit (`cCRT-logit`) do reasonably well, although FDP$^{\delta}$ is not controlled at level 0.1 in the case of `cKO-logit`. With the randomization results of the clustering, we furthermore add the *aggregated version* for the CRT-logit. Notably, adding the aggregation step (`cCRT-logit-agg`) helps increasing the statistical power compared with the single-clustering version. The difference in inference results are demonstrated to be even more substantial in the next section.

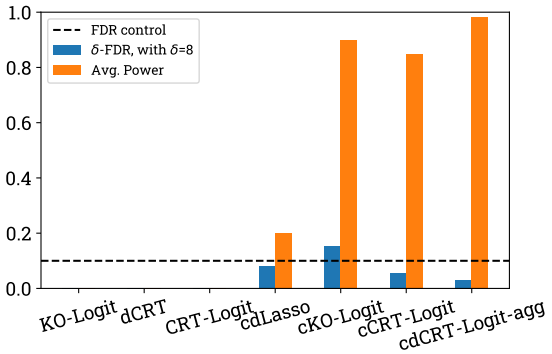


Figure 6.6: **FDR/Average Power of 100 runs of simulations for very high-dimensional scenario with cluster-located support.** Parameters: $n = 600$, $p = 8000$, $C = 500$, SNR $= 3.0$. FDR$^{\delta}$ is controlled at level $\alpha = 0.1$ and $\delta = 8$. The spatial tolerance δ is defined by taking $\delta = p/(2C)$, since the dimension of the support is 1D. Methods: KO-Logit, CRT-logit, (clustering version): Debiased Lasso (`cdlasso`), model-X Knockoff (`cKO-logit`), our version of CRT (`cCRT-logit`), aggregation of CRT (`cCRT-logit-agg.`).

6.3.3 Experiment with Brain-Imaging Datasets

Description The Human Connectome Project dataset (HCP) is a collection of brain imaging data on healthy young adult subjects with age ranging from 22 to 35. The participants performed different tasks while being scanned by a magnetic resonance imaging (MRI) device to record blood oxygenation level dependent (BOLD) signals of the brain. The aim of this analysis is to investigate which areas of the brain can predict cognitive activity across participants, while taking into account the information from other brain regions. The brain imaging modalities include, among others, resting-state fMRI (R-fMRI) and task-evoked fMRI (T-fMRI). In this work, we only deal with decoding the task-evoked fMRI dataset. More specifically, the input $\mathbf{X}$ is a set of 2mm statistical maps of 400 subjects across 8 cognitive tasks. These are called z-maps, as the data follow a standard normal distribution under the null hypothesis. The data have been *masked*, *i.e.* only voxels corresponding to brain regions are taken into acount. Each task in turn features 2 different contrasts, which effectively forms binary responses $\mathbf{y} \in \{0,1\}^n$: `relational` (relational vs match), `gambling` (reward vs. loss), `emotion` (emotional face vs shape outline), `social` (mental interaction vs. random interaction), `language` (story vs math), `motor hand` (left vs. right hand), `motor foot` (left vs. right foot), and `working memory` (remembering 2 block back vs. remembering 0 block back). The setting is high-dimensional with $n = 800$ samples, corresponding to 400 subjects with age ranging between 22 and 35 years old, while the total number of variables p is around $200,000$ brain voxels. As demonstrated in Section 6.3.2, it is impossible to achieve reasonable statistical power without doing dimension-reduction, hence we perform hierarchical clustering to reduce the data to $C = 500$ clusters. Following [NCT19, CNS+21], we use a hierarchical clustering scheme to group the variables into spatially connected and compact regions. We use Ward's minimum variance criterion that minimizes the total within-cluster variance, then average signals inside each cluster. This algorithm further incorporates connectivity constraints so that the resulting clusters are connected in the lattice geometry of the image. The main benefit of this approach is that it can capture well local correlations into spatial clusters, therefore reduce the correlations in the design matrix [VGT12].

Creating realistic ground-truth and response labels Since there is no ground truth for this dataset, we create synthetic true signals by fitting the data $\mathbf{X}$ and response $\mathbf{y}$ with a support vector classifier with linear kernel. The resulting $\hat{\boldsymbol{\beta}}^{\text{SVM}}$ will be considered as true regression coefficients for each task. Then, to avoid bias in simulating label $\hat{\mathbf{y}}$, the z-maps matrix $\mathbf{X}$ of one task are used in conjunction with the discriminative pattern map $\hat{\boldsymbol{\beta}}^{\text{SVM}}$ from the next task in the following order: `relational`, `gambling`, `emotion`, `social`, `language`, `motor_hand`, `motor_foot`, `working_memory (WM)`. For instance, we use $\hat{\boldsymbol{\beta}}^{\text{SVM}}$ of `gambling` with z-maps data matrix of `relational`, *i.e.* for all $i = 1, \dots, n$, given $\mathbf{x}_i^{\texttt{relational}}$,

$$\hat{y}_i \sim \text{Bern}\left(\frac{1}{1+\exp(-\mathbf{x}_i^{\texttt{relational}}\,\hat{\boldsymbol{\beta}}^{\text{SVM}}_{\texttt{gambling}} + \sigma\xi_i)}\right), \tag{6.22}$$

where $\text{Bern}(a)$ is a Bernoulli probability mass function that takes a value 1 with probability a, σ is a noise magnitude and ξ_i is a standard normal noise. Finally, we apply all inference algorithms on the semi-synthetic data $(\mathbf{X}, \hat{\mathbf{y}})$, and we evaluate their performance using the ground-truth $\hat{\boldsymbol{\beta}}$. This simulation setting is similar to [CNS+21], except that here we consider a classification and not a regression problem. It allows us to calculate the False Discovery Rate and average power with multiple runs of the inference (across tasks).

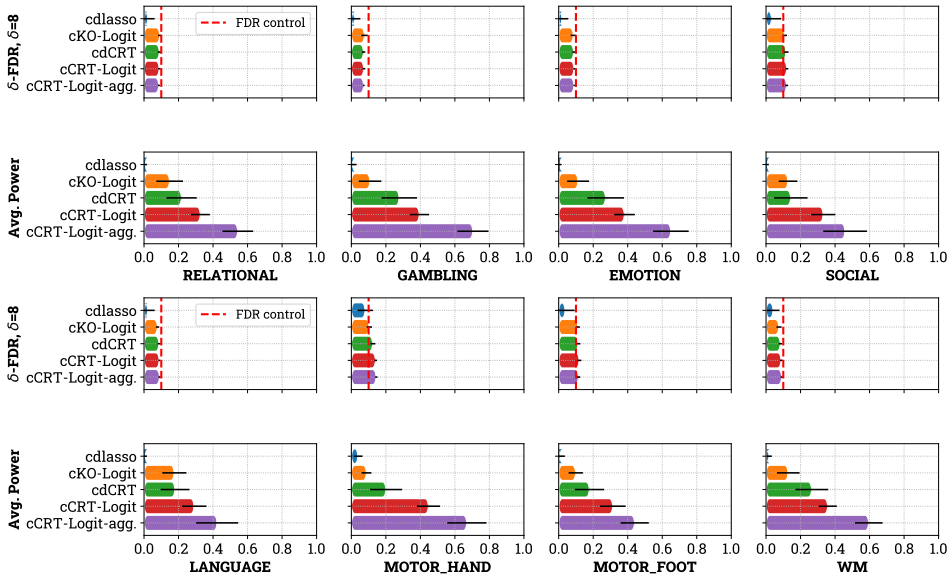


Figure 6.7: **FDR/Average Power of 50 runs of simulations on Human Connectome Project dataset.** Parameters: $n = 800$ (taken from 400 subjects), SNR $= 1.5$. FDR^{δ} is controlled at level $\alpha = 0.1$ and $\delta = 8$. Methods (clustering versions): Debiased Lasso (`cdlasso`), model-X Knockoff (`cKO-logit`), original dCRT (`cdCRT`), our version of CRT (`cCRT-logit`) and the aggregation of CRT-logit across clusterings (`cCRT-logit-agg.`)

Remark 6.2. *The* i.i.d. *assumption is technically broken in this experiment, where for each subject we analyze two images that represented data that are likely not independent. Yet, this remains a short-range correlation structure, and is not a strong challenge to the* i.i.d. *assumption.*

Results The overall trends in Figure 6.9 suggest that CRT-logit achieves a better recovery compared to KO or original CRT, which results in higher statistical power. Furthermore, a substantial improvement in the number of correct discoveries with CRT-logit are made when we add aggregation step across multiple clusterings. This gain comes with a good control of the FDR^{δ} under desired level $\alpha = 0.1$, with $\delta = 8$, following Eq. (6.20). We note that the clustered version of desparsified lasso (`cdlasso`) does not perform well in high-dimension. This might due to the fact that `cdlasso` also relies on the choice of the ℓ_1-regularization λ in the nodewise Lasso operation, similar to the $\mathbf{X}_{*,j}$-distillation of dCRT, as noted in Section 6.1. What makes the difference here is that instead of using cross-validation for setting λ for each variable j, as recommended in Section 6.2.3, a *fixed* value of $\lambda = 10^{-2}\lambda_{max}$ is used in the implementation of `cdlasso`. As the results in Fig. 6.2 have demonstrated, we strongly suspect this fixed value is not optimal, which makes the procedure powerless.

6.3.4 Genome-wide Association Study with Human Brain Cancer Dataset

Description The last in our benchmark is a Genome-wide Association Study (GWAS) for the human brain cancer dataset. The Cancer Genome Atlas (TCGA) [WCM+13, VSWZ18] is an open-access dataset that consists of several genomics studies on human cancer. We choose to analyze the Glioma cohort, which consists of $n = 1026$ patients across a wide age range, diagnosed with this type of brain tumor. There is a total of $p = 24776$ genes in the data matrix, recorded as copy number variations (CNVs) at the gene level in log ratio format. As with the brain-imaging

inference in Section 6.3.3, we use clustering to reduce the dimension to $C = 500$ clusters. For the response label $\mathbf{y}$, a long-term survivor (LTS) is defined as a patient who survived more than five years after diagnosis and would be labeled $y = 0$, and any patient who died within five years would be a short-term survivor (STS), labeled $y = 1$. The objective is to identify significant genes that contribute to classification of the LTS/STS status. Similar to the Human Connectome Project dataset, there is no real ground-truth for the TCGA Glioma. However, we have the list of mutations and the frequency of those detected in the diagnosed patients. We select the 1000 most frequent gene mutations that appeared in this list, *i.e.* the ground truth list consists of 1000 genes (variables).

Table 6.2: **List of detected genes associated with Glioma Cancer from the TCGA dataset** $n = 1026, p = 24776$. Empty line (—) signifies no detection. Methods listed in the table are the clustering version. Commonly detected genes between methods are put in bold text. Most detected genes are listed in the mutant list database that can be found in the recorded patients [VSWZ18].

Methods	**Detected Genes**
dLasso	—
KO	**ABCC10**, **ANK3**, **ANKRD30A**, CDH23, NOTCH2, PTEN, RET, SPAG17, **SPEN**, **SVIL**, ZMIZ1
Distilled-CRT	**ANK3**, **ANKRD30A**, CDH23, PTEN, RET, **SVIL**, ZMIZ1
CRT-logit	**ABCC10**, **ANKRD30A**, ANKRD30B, BCOR, EPHA3, F5, IL1RAPL2, LAMA3, NOTCH2, PCDH19, PHF3, PPL, SPAG17, **SPEN**, **SVIL**, USP9X
CRT-logit-aggregated	**ABCC10**, **ANKRD30A**, ANKRD30B, F5, IGF1R, IL1RAPL2, KIAA1217, LAMA3, LRRK1, MX2, MYO3A, NEBL, PCDH19, RIPK4, **SPEN**, **SVIL**, TP63

Results Table 6.2 lists the detected genes across methods. CRT-logit based methods find the largest number of genes. Moreover, most of selected genes in this table are found in the list of mutated genes found on recorded patients. Some genes are detected by all the benchmarked methods, most prominently `SPEN`, which is found on over 10 % of patients in the cohort. Furthermore, this gene is known to be associated not only with brain cancer, but also with other types of cancer in The Human Protein Atlas project [LCM+15]. Note that, in the absence of a ground-truth, this does not guarantee all genes found are associated with Glioma, but this experiment demonstrates the capability of CRT-logit in GWAS studies.

6.4 Discussion

In this work, we propose an adaptation of the Conditional Randomization Test (CRT) for logistic regression in the extremely high-dimensional regime. A major improvement of CRT-logit, our proposed algorithm, compared to original CRT, comes from

the decorrelation of test statistics to make their distribution closer to standard normal. Indeed, results from synthetic experiments in Figure 6.1 suggested that in mildly high-dimension (when $0.5 < n/p < 1.0$), the empirical null distribution of the CRT-logit's test statistic T^{decorr} is much similar to a standard normal compared to the original CRT test statistic. Moreover, empirical benchmarks in Section 6.3 demonstrated that CRT-logit performs better than related statistical inference methods, such as the Debiased Lasso or Model-X Knockoffs. In particular, CRT-logit is the most powerful method in our synthetic experiment with a mildly high-dimensional dataset in Section 6.3.1, while still keeping FDR controlled under predefined level $\alpha = 0.1$. For datasets with a large number of features but a small number of samples, *i.e.* $n \ll p$, we show empirically that a dimension-reduction technique via clustering increases much the statistical power, while still controlling the FDR^{δ}. Results in Section 6.3.2 showed a clear trend that without hierarchical clustering, CRT-logit and Knockoff Filter are statistically powerless. In both brain-imaging and genomics experiment with extremely high-dimensional data in Section 6.3.3 and Section 6.3.4, clustered CRT-logit and its ensemble of multiple clusters version demonstrate further a higher power than alternative methods.

On a related note, we observe that at least for some particular settings, *e.g.* in the experiment shown in Figure 6.2, choosing the ℓ_1-regularization parameter of the distillation step can strongly affect how variables are selected by CRT-logit. More specifically, Figure 6.2 showed that depending on the number of observations n and the number of variables p, the optimal choice for regularization parameter λ_{dx} can vary between $10^{-1/4}\lambda_{\text{univ}}$ and $10^{1/2}\lambda_{\text{univ}}$, where $\lambda_{\text{univ}} = \sqrt{(1/n)\log p}$. Thus, for practical applications, we recommend the usage of cross-validation for finding λ_{dx} in Eq. (6.8) and Eq. (6.2), *i.e.* in the process of calculating decorrelated test-statistics T^{decorr}.

We note that there exists some limitations to CRT-logit. The computational cost of CRT-logit, while faster than vanilla CRT proposed in [CFJL18], is still slower than alternative methods such as Knockoff Filter and Debiased Lasso, as shown in Table 6.1. Moreover, tuning the ℓ_1-regularization λ_{dx} parameter by cross-validation, as is often done, can further increase the computational cost of CRT-logit (and dCRT). Regarding the limitation of ensemble of clusters approach that adds a data reduction step to inference, as noticed in Remark 6.2, this method works best when true variables are located in cluster-like structure and correlations between neighboring null and true variables are not too strong.

On the other hand, our empirical results put forward several questions. First and foremost, there is a need for theoretical developments about the impact of the sparsity regularization parameter λ_{dx} on FDR and statistical power of CRT-logit, judging on the result shown in Figure 6.2. One can notice that there is no reason why the value selected by cross-validation, i.e. on goodness of fit grounds, is optimal or even suitable for error control, nor to optimize a detection/error control tradeoff. We however noticed that the value selected by cross-validation performed well in the situations we could probe. Yet the theory that would establish the suitability of this choice or ground improvements upon it is still missing. Second, there still lacks a theoretical analysis of the FDR-Power tradeoff in asymptotic setting for the CRT with generalized linear models. The closest work on this direction is [WJ20], which does the job for Lasso-based dCRT and Knockoffs in linear high-dimensional regime. However, the analysis of their work only applies in the complete *i.i.d.* setting, *i.e.* when the population covariance is $\mathbf{\Sigma} = \mathbf{I}_p$.

Despite these limitations and open questions, our empirical benchmarks on both simulated and real data show real promises of CRT-logit. Henceforth, we believe CRT-logit is competitive for practical settings that involve structured data, such as brain-imaging and genomics applications. We hope that the issues raised above would serve as key directions for future works.

6.5 Additional Experimental Details & Results

6.5.1 Role of Correlation in Inference Results

To investigate the role of autocorrelation parameter ρ in high-dimensional inference problem, we follow the experiment in Section 6.2.3, but varying ρ and fixing $n = 300, p = 400$ instead. Figure 6.8a shows that when there are strong correlations between variables, *e.g.* when $\rho > 0.7$, the method can either be conservative or fail to control FDR with Benjamini-Hochberg (BH) procedure [BH95]. On the other hands, results from Figure 6.8b show that using Benjamini-Yekutieli (BY) procedure [BY01], this issue is less severe, however the average powers are generally lower. It might be tempting to conclude that the FDR/Power tradeoff issue lies on the FDR controlling methods that we use. However, results from Figure 6.8c show a strong evidence against this idea. Even when there is no multiple testing correction for p-values, CRT-logit failed to control False Positive Rate at higher values of λ. This suggests the p-values outputting from CRT-logit are biased with poor choices of the ℓ_1-regularization parameter, therefore a good selection of λ is crucial.

6.5.2 Performance of Knockoff Logistic in Low-dimensional Setting

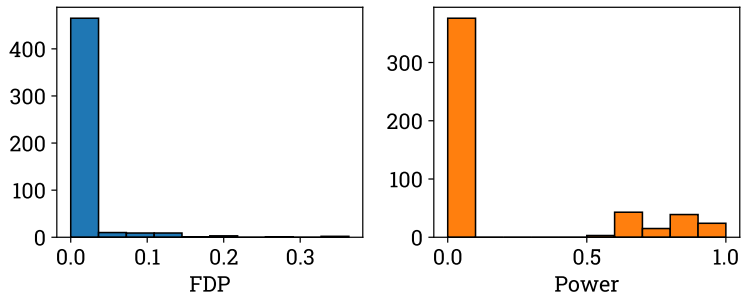
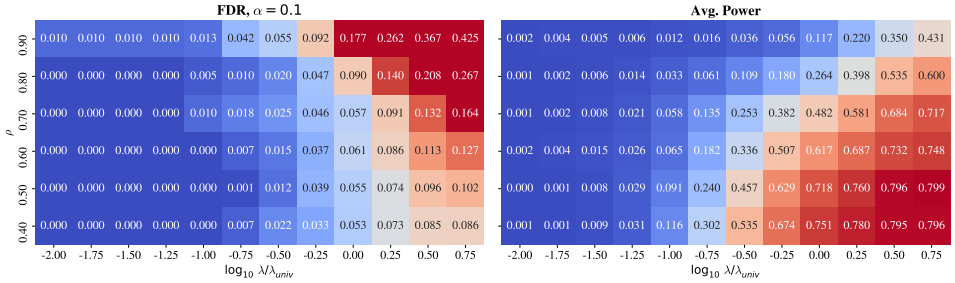
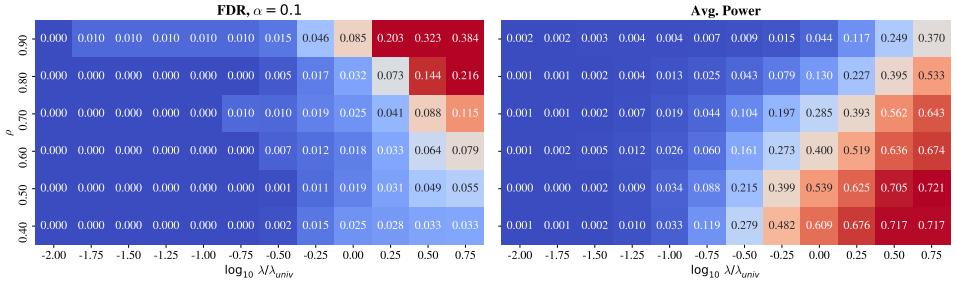


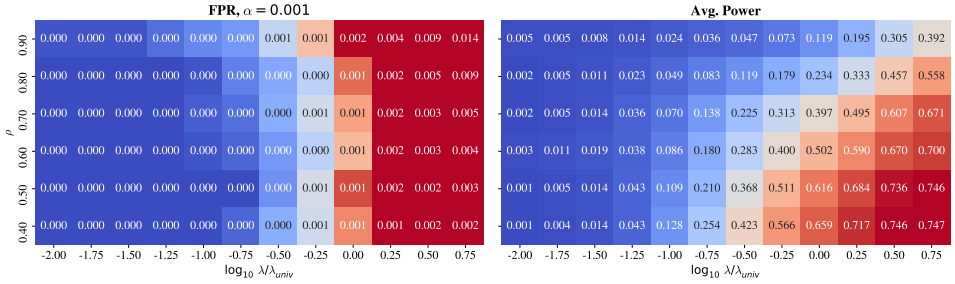
Figure 6.9: **FDP/Power of 100 runs of Knockoff Logistic across varying parameters in low-dimensional setting.** Default parameters: $n = 600, p = 400$, snr $= 3.0, \rho = 0.5$, sparsity $= 0.04$. FDR is controlled at level $\alpha = 0.1$.



(a) FDR control with Benjamini-Hochberg procedure, $\alpha = 0.1$



(b) FDR control with Benjamini-Yekutieli procedure, $\alpha = 0.1$



(c) False Positive Rate control, $\alpha = 0.001$

Figure 6.8: **FDR (FPR) /Average Power of 100 runs of simulations across varying number of correlation parameters ρ and ℓ_1-regularization parameter λ.** Note: λ is scaled with the factor $\lambda_{\text{univ}} = \sqrt{(1/n)\log(p)}$, *e.g.* the first value for regularization grid is $\lambda = \lambda_{\text{univ}} \times 10^{-2}$. Default parameter: $n = 300, p = 400, \text{SNR} = 3.0, \kappa = 0.05$.

Chapter 7

Conclusion

7.1 Summary

In this thesis, we have developed three new approaches for multiple testing/variable selection for high-dimensional datasets. The setting that we deal with is that of multivariate inference, which is challenging when the number of variables is (possibly orders of magnitude) larger than the number of observations. First, in Chapter 4, we presented Aggregation of Multiple Knockoffs (`AKO`), an extension of Knockoff Filter (KO). This algorithm fixes the inherent instability issue in knockoff inference due to the randomness of knockoff variables. `AKO` works by running multiple sampling of knockoff variables, converting knockoff statistics into p-values, then aggregating those p-values with quantile-aggregation technique. We learned through a set of simulated and brain-imaging experiments that `AKO` is more stable compared to vanilla KO, while gaining statistical power.

Next, in Chapter 5, we introduced an algorithm that combines randomized clustering and Knockoff Filter. We called this algorithm Ensemble of Clustered Knockoffs, or `ECKO`. The first step of `ECKO` is to reduce the dimension of datasets by randomized hierarchical clustering, which makes the statistical inference problem tractable. We then perform knockoff filter to select clusters (groups) of brain voxels which are significant, while aiming to control FDR. Similar to the idea of `AKO`, we perform multiple runs of clustering and knockoff inference, then combine it by p-value aggregation. To take into account the lower resolution when performing clustering, we introduce a new notion of FDR with a spatial tolerance δ. `ECKO` shows favorable performance in fMRI data-analysis applications, compared with alternative methods: high statistical power while still controlling $deltafdr$.

Finally, in Chapter 6, we presented a procedure for conditional independence testing that works with non-linear relationship between covariates and responses in high-dimension, in particular for logistic regression. It is an extension of the Conditional Randomization Test (CRT) [CFJL18], a procedure that can output p-values in high-dimension, but comes at prohibitive computational cost. Our algorithm, CRT-logit, combines the idea of distillation operation from dCRT of [LKJR20] and decorrelation operation from [NL17]. It fixes the computational issue of vanilla CRT, while returning higher statistical power in benchmarks with high-dimension synthetic dataset. Additionally, we present a solution that combines clustering and p-values aggregation, similar to the idea of `ECKO`. Results from neuro-imaging and genomics datasets suggest that `CRT-logit` performs well compared to competitive methods.

7.2 Perspectives

We present here some future extensions of the works in this thesis.

Relaxation of theoretical results for `AKO` As shown in Theorem 4.2 of Chapter 4, the non-asymptotic FDR control of Aggregation of Multiple Knockoffs has a factor $\kappa > 3$. There is almost certainly a possibility to tighten the non-asymptotic FDR bound, since we always observe empirically that the FDR is controlled without this factor. Moreover, the proofs in Chapter 4 rely on assumption 4.1, which involves the independence of knockoff statistics. However, in practical applications, this assumption might not hold, which calls for a theoretical relaxation. We remark that the original proof for FDR controlling of Knockoff Filter in [BC15] does not require this assumption. Indeed, their proof relied on martingale theory by assigning the FDR controlling threshold τ of Eq. (2.19) as an optimal stopping time. We find that adapting this technique to `AKO` is not straightforward, since `AKO` relies on aggregation of p-values converted from knockoff statistics. A potential adaptation could resuse the analysis of the proof of Prop. 4.1, where we stated the equivalence of Knockoff Filter's FDR threshold with that of BH step-up procedure.

FDR controlling for proposed algorithms without reliance on independence assumption The independence between p-values is required in this thesis' proposed algorithms, as we use Benjamini-Hochberg (BH) procedure to control FDR. However the independence of the decision statistics is hard to check and not guaranteed in realistic scenarios. Because this is a common issue for all FDR-controlling methods that rely on the BH procedure, `AKO` (Chapter 4), `ECKO` (Chapter 5) and `CRT-logit` (Chapter 6) all suffer from the same limitations. We could potentially use Benjamini-Yekutieli (BY) procedure, however the inference results will be more conservative. As a result, making the algorithms presented in this thesis work with relaxed independence assumption while still maintaining *reasonable* statistical power is a good direction for future studies. Another possibility is to adapt the p-values output by the proposed three algorithms to FDP control by post hoc bounds, following the works of [GS11, BNR20].

Theoretical analyses of FDR-Power tradeoff for `CRT-logit` and `AKO` under general dependence Lastly, we note that the presentation of `CRT-logit` (conditional randomization test for logistic regression) in Chapter 6 leans more towards empirical side. Therefore, we would like to develop a thorough theoretical analysis of the algorithm, along with the asymptotic FDR-Power tradeoff for CRT with generalized linear models. As noted in Section 6.4, there exists a recent work of [WJ20] for the analysis of dCRT, but only for independent variables case. One possibility is to adapt the Approximate Message Passing (AMP) technique from [WYBS20], in which they analyze the FDR-Power tradeoff for Lasso with a general covariance structure. However, their work only concerns the asymptotic regime, where the ratio n/p converges to a fixed quantity. Another possibility is to borrow the technique from [CMW20] with the usage of Gordon's min-max inequality. An advantage of this approach is that it works in the finite sample case, *i.e.* non-asymptotic regime. On a related note, we remark that there already exists an adaptation of AMP technique for the analysis of FDR-Power with knockoff Lasso-coefficient difference statistics in the work of [WSB+20]. We could adapt their analysis to the aggregation of multiple knockoffs (`AKO`), but this might require a modification on how p-values are aggregated.

Part III

Synthèse en Français

Chapter 8

Contributions

Cette thèse est divisée en deux parties : la première partie présente les informations de base nécessaires et la motivation de le deuxième partie, qui traite principalement des nouvelles procédures de test d'hypothèse/sélection de variables procédures de sélection de variables.

En particulier, pour la partie de fond, nous commençons par une introduction aux tests d'hypothèse au Chapitre 2. Le lecteur se familiarisera avec le système de notations, le problème de la test multiple et la difficulté de faire de l'inférence multivariée en haute dimension. haute dimension. Cela repose sur une distinction essentielle entre l'inférence conditionnelle et marginale. Plus précisément, nous nous concentrons sur les mesures populaires des tests multiples, telles que taux d'erreur en fonction de la famille ou les taux de fausse découverte, ainsi que les procédures de contrôle de ces mesures. métriques. Nous passons ensuite aux avancées récentes en matière de tests multiples à haute dimension, à savoir les filtres de contrefaçon et le test de randomisation conditionnelle. Nous concluons ce chapitre par des techniques d'agrégation qui permettent de stabiliser les résultats d'inférence des procédures de tests multiples. résultats d'inférence des procédures de tests multiples.

Chapitre 3 présente l'imagerie par résonance magnétique résonance magnétique fonctionnelle (IRMf), l'une des principales techniques de neuro-imagerie. L'analyse des données d'IRMf permet de comprendre le fonctionnement de notre cerveau, et l'analyse des données d'IRMF permet de comprendre le fonctionnement de notre cerveau et sera l'une des principales applications des procédures de sélection de variables que cette thèse a présentées. cette thèse a présenté. Nous passerons brièvement en revue le pipeline d'analyse des données d'IRMf : de l'acquisition au traitement, à la modélisation et enfin aux méthodes d'inférence statistique sur l'IRMf. traitement, la modélisation et enfin les méthodes d'inférence statistique sur les données IRMf. et enfin les méthodes d'inférence statistique sur les données IRMf.

La partie contribution de cette thèse est organisée autour de trois grandes majeures, qui sont présentées comme suit.

8.1 Résultats de l'inférence stabilisante du filtre Knockoff Filter

Knockoff filter [BC15, CFJL18] est l'une des les méthodes récemment introduites pour les tests multiples, dans le but de contrôler le le taux de fausses découvertes (False Discovery Rate – FDR). Bien qu'elle soit relativement nouvelle dans la littérature, cette méthode jouit d'une grande popularité grâce à sa flexibilité dans les paramètres du modèle et fonctionne bien dans des contextes légèrement haute dimension. Cependant, l'un des principaux inconvénients du filtre knockoff est son instabilité : l'approche de la méthode repose sur l'échantillonnage de nouvelles copies bruyantes des variables originales. méthode repose sur l'échantillonnage de nouvelles

copies bruyantes des variables originales, Les résultats de l'inférence sont donc intrinsèquement aléatoires. Dans Chapitre 4, nous introduisons une méthodologie pour résoudre ce problème d'aléatoire.

Notre algorithme, `AKO`, rend la sélection des variables plus stable en combinant plusieurs passages de l'inférence de knockoff en utilisant la technique d'agrégation des valeurs p technique de [MMB09]. Nous fournissons une analyse théorique de `AKO`, et surtout nous prouvons la garantie de contrôler le FDR sous un niveau prédéfini. Enfin, nous effectuons une validation empirique de `AKO` sur un ensemble de données numériques, de données numériques, de neuro-imagerie et de génétique.

Travaux publiés Nguyen, T., Chevalier, J., Thirion, B. & Arlot, S.. (2020). Aggregation of Multiple Knockoffs. *In Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*. Available from https://proceedings.mlr.press/v119/nguyen20a.html.

8.2 Tests Multiples En Très Haute Dimension : Une Combinaison De Knockoff Filter Et Clustering Aléatoire

Dans l'analyse des données d'IRMf, le nombre de variables (voxels cérébraux) peut être supérieur à des centaines de milliers, alors que seul un petit nombre d'observations d'un maximum de 10 000 pixels est nécessaire. centaines de milliers, alors que seul un petit nombre d'observations, de quelques milliers au maximum, est disponible. plusieurs milliers est disponible. Dans un tel contexte, il faut trouver des variables explicatives qui sont réellement liées à la réponse tout en contrôlant le taux de fausses découvertes est un problème difficile, tant sur le plan informatique que statistique [SP20]. Dans Chapitre 5, nous introduisons un algorithme qui combine le clustering aléatoire et le filtre knockoff. La première étape de cet algorithme, le regroupement hiérarchique aléatoire, est cruciale : elle agit comme une étape de réduction de la dimension, rendant la tâche d'inférence réalisable dans une résolution inférieure pour les données d'IRMf. Nous effectuons ensuite un filtre knockoff pour sélectionner les clusters (groupes) de voxels du cerveau qui sont significatifs, tout en visant à contrôler le FDR. Encore une fois, pour rendre la solution plus stable, nous effectuons plusieurs exécutions du clustering et de l'inférence knockoff, puis nous les combinons en utilisant l'agrégation des valeurs de p. Nous avons appelé cet algorithme Ensemble of Clustered Knockoffs, ou `ECKO`. Plus important encore, pour prendre en compte la résolution plus faible lors de l'exécution du clustering, nous introduisons une nouvelle notion de FDR avec une tolérance spatiale $\delta > 0$. En bref, les voxels non activés (variables nulles) sélectionnés à une distance inférieure à δ des voxels activés (variables significatives) pour un stimuli ne seraient pas considérés comme des faux positifs. Nous appelons cette métrique FDR^{δ}, et introduisons la définition nécessaire de la concept. À l'aide de divers points de repère sur des ensembles de données de neuro-imagerie simulés et réalistes, nous démontrons que le `ECKO` donne de bons résultats par rapport aux méthodes contemporaines.

Travaux publiés T.-B. Nguyen, J.-A Chevalier, & B. Thirion (2019), `ECKO`: Ensemble of Clustered Knockoffs for Robust Multivariate Inference on fMRI Data. *In International Conference on Information Processing in Medical Imaging (pp. 454-466). Springer, Cham.*

8.3 Un Test De Randomisation Conditionnelle Pour La Régression Logistique En Haute Dimension

Dans le chapitre 6, nous introduisons une procédure de test d'indépendance conditionnelle qui fonctionne avec des relations non linéaires entre les covariables et les réponses en haute dimension, en particulier pour la régression logistique. Dans un cadre à haute dimension, le test de l'indépendance conditionnelle est hautement non trivial, car le conditionnement sur un grand nombre de variables est intrinsèquement difficile [SP20]. Le test de randomisation conditionnelle (TRC) est une autre tentative récente de s'attaquer à ce problème, en plus des filtres knockoff. Le concept de CRT peut être compris comme une procédure de randomisation pour échantillonner la la distribution empirique des statistiques de test pour chaque variable j, afin de de tester l'indépendance de j avec la réponse *conditionnellement* à toutes les autres variables. Elle a été discutée à l'origine dans l'article sur les knockoffs [CFJL18] comme une procédure alternative, si le statisticien veut produire des valeurs p avec des variables knockoffs. Malgré cela, le CRT a un coût de calcul prohibitif, en raison de l'échantillonnage multiple de variables fictives bruyantes pour chaque statistique de test calculée. Les travaux récents de [LKJR20] ont introduit une opération de distillation de la CRT, appelée dCRT, qui semble prometteuse pour corriger cet inconvénient crucial. Cependant, nos observations empiriques montrent que la dCRT rencontre un problème statistique lorsqu'elle est exécutée en régime non linéaire, en particulier la régression logistique. Malgré sa popularité, il n'y a actuellement aucune bonne solution pour permettre une inférence précise sur les coefficients de la régression logistique dans le régime à haute dimension. Pour résoudre ce problème, nous combinons l'opération de distillation de [LKJR20] avec une étape de décorrélation qui adapte cette opération au régime non linéaire. régime non linéaire, et appelons cette méthode `CRT-Logit`. De plus, nous présentons un algorithme qui combine le regroupement et l'agrégation des valeurs p pour la régression logistique, similaire à l'idée de `ECKO`. Les résultats obtenus à partir d'ensembles de données synthétiques, de neuro-imagerie et de génomique suggèrent que `CRT-logit` donne de bons résultats par rapport à la méthode concurrente.

Préprint T.-B. Nguyen, S. Arlot, & B. Thirion (2021+), Upscaling the Conditional Randomization Test to Extremely High-Dimensional Logistic Regression. *A soumettre.*

8.4 Conclusion

Enfin, nous concluons la thèse dans le Chapitre 7 avec un résumé des contributions et une perspective plus large sur les questions ouvertes qui peuvent être davantage abordées.

8.5 Autres Travaux

Il existe d'autres préprints/publications que nous énumérons ci-dessous, mais que nous ne présenterons pas dans ce manuscrit de thèse. mais que nous ne présenterons pas dans ce manuscrit de thèse.

- J.-A. Chevalier, T.-B. Nguyen, B. Thirion J. Salmon, Spatially relaxed inference on high-dimensional linear models (2021+). *A soumettre.*
- J.-A. Chevalier, T.-B. Nguyen, J. Salmon, G. Varoquaux, B. Thirion, Decoding with confidence: Statistical control on decoder maps. *In NeuroImage, Volume 234, 2021, 117921, ISSN 1053-8119.*

8.6 Logiciel

Pour favoriser la reproductibilité scientifique, nous avons développé une implémentation open-source d'une partie des procédures présentées dans cette thèse. Le logiciel est écrit en Python et peut être trouvé sur : https://ja-che.github.io/hidimstat/

Chapter 9

Contexte: Test d'hypothèses en grande dimension

Résumé. (Ce chapitre est une traduction partielle du chapitre 2) Dans ce chapitre, nous introduisons la motivation du thème principal de cette thèse: le problème des tests d'hypothèses pour l'inférence multivariée en grande dimension. En particulier, nous partons du problème classique du test d'hypothèse unique, puis nous passons aux approches du test d'hypothèses multiples, qui a trouvé sa popularité dans l'analyse de données modernes. Nous présentons également certaines métriques populaires pour les tests d'hypothèses multiples, ainsi que les procédures correspondantes pour les contrôler. Nous présentons ensuite la difficulté de faire de l'inférence statistique en grande dimension, où le nombre d'observations est bien inférieur au nombre de variables. Ensuite, nous discutons des différences entre les approches univariées et multivariées dans ce régime, et soulignons l'importance de ces dernières pour les applications. Nous présentons ensuite le filtre knockoff et le test de randomisation conditionnelle, deux méthodes populaires récemment introduites pour l'inférence multivariée en grande dimension. Comme discuté dans le Chapitre 8, ces deux méthodes servent d'épine dorsale à l'orientation principale de cette thèse. Nous terminons le chapitre par la description de deux ingrédients importants de notre approche: une technique de clustering aléatoire et d'agrégation des p-valeurs.

9.1 Détermination de l'importance d'une variable par rapport à la réponse: un problème classique

Historiquement, le concept de test d'hypothèse a été introduit il y a presque un siècle par les travaux de Ronald Fisher [Fis25, Fis36]. Le test d'hypothèse est une procédure qui utilise les données observées pour prendre des décisions concernant les propriétés de la loi inconnue des données. Plus formellement, la définition d'un test d'hypothèse est la suivante.

Définition 9.1 (Test d'hypothèse, [CB02])**.** Un test d'hypothèse est une règle qui spécifie:

- Pour quels échantillons on décide d'accepter $\mathcal{H}_0$, *l'hypothèse nulle*.
- Pour quels échantillons on décide de rejeter $\mathcal{H}_0$ et de choisir $\mathcal{H}_a$, *l'hypothèse alternative*.

Exemple 9.1. L'un des premiers résultats sur les tests d'hypothèses est décrit dans le livre de Ronald Fisher *Design of Experiments* en 1936 [Fis36]. Le célèbre statisticien voulait tester l'affirmation d'une collègue féminine selon laquelle elle pouvait

distinguer si le lait ou le thé avait été placé en premier dans la tasse, simplement en goûtant. Fisher a donc conçu une expérience consistant à donner à sa collègue huit tasses de thé, quatre de chaque variété (en mettant le lait avant le thé et vice versa), dans un ordre aléatoire. On pouvait alors se demander quelle était la probabilité qu'elle obtienne le nombre correct, mais juste par hasard. L'hypothèse nulle $\mathcal{H}_0$ était que sa collègue n'avait pas cette capacité, et l'hypothèse alternative $\mathcal{H}_a$ était qu'elle pouvait effectivement déterminer l'ordre de la préparation d'une tasse de thé.

Supposons que nous ayons un échantillon de n observations *i.i.d.* $X = X_1, \ldots, X_n$ provenant d'une distribution de données P_X avec une moyenne $\mu \in \mathbb{R}$, *i.e.* $X_i \sim P_X$. Le problème pourrait être de tester si

$$\mathcal{H}_0 : \mu \leq 0 \quad \text{ou} \quad \mathcal{H}_a : \mu > 0.$$

Généralement, pour effectuer un test d'hypothèse, nous spécifions une statistique de test, désignée par $T(X_1, \ldots, X_n)$.

Définition 9.2 (Statistique de test)**.** Une statistique de test est une fonction qui capture un aspect de l'échantillon d'observations.

Exemple 9.2. Un exemple de statistique de test est la moyenne de l'échantillon: $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i/n$.

Après avoir calculé la statistique de test, nous devons prendre une décision pour déclarer si le résultat du test d'hypothèse est *statistiquement significatif* ou non. Une façon de procéder consiste à comparer la statistique de test à un certain quantile de sa distribution nulle ; ce quantile est appelé *niveau de significativité* – noté α – la probabilité que nous rejetions l'hypothèse nulle $\mathcal{H}_0$ étant donné que le $\mathcal{H}_0$ est supposé être vrai. Alternativement, on peut rapporter le résultat d'un test d'hypothèse basé sur une *p-valeur*, définie comme suit.

Définition 9.3 (p-valeur)**.** Une *p-valeur* $p(X)$ est une statistique de test satisfaisant les propriétés suivantes:

1. $p(X) \in [0, 1]$.
2. (Domination stochastique) Lorsque P_X satisfait à $\mathcal{H}_0$, pour tout $t \in [0, 1]$,

$$\mathbb{P}_{X \sim P_X}(p(X) \leq t) \leq t.$$

La propriété 2 est importante dans le sens où elle rend une p-valeur *valide*. De plus, une conséquence de cette propriété est que si $p(X)$ est une p-valeur valide, il est possible de construire une statistique de test au niveau de signification α basée sur $p(X)$, pour tout $\alpha \in (0, 1)$. Intuitivement, on peut voir que plus la p-valeur $p(X)$ est petite, plus le statisticien peut rejeter avec confiance les résultats du test.

Exemple 9.3 (P-valeur normale bilatère [CB02])**.** L'un des tests les plus courants effectués par les statisticiens est sans doute le test de la moyenne d'un échantillon aléatoire $X_1, \ldots, X_n$ provenant d'une distribution $\mathcal{N}(\mu, 1)$. Nous voulons vérifier $\mathcal{H}_0 : \mu = \mu_0$ contre $\mathcal{H}_a : \mu \neq \mu_0$. La statistique de test est $T(X) = (\bar{X} - \mu_0)/\sqrt{n}$ où $\bar{X} = \sum_{i=1}^{n} X_i/n$. Sous l'hypothèse nulle, cette statistique de test suit la loi de Student avec $n - 1$ degrés. de Student avec $n - 1$ degrés de liberté. Par conséquent, nous obtenons la p-valeur bilatère à l'aide de la formule:

$$p(X) = 2\mathbb{P}(t_{n-1} \geq T(X)),$$

où t_{n-1} suit la loi de Student avec $n - 1$ degrés de liberté.

Dans tout problème de test, deux types d'erreurs peuvent être commises. On dit que l'on fait un *faux positif*, ou *erreur de type I*, chaque fois que l'on rejette une hypothèse nulle vraie. En revanche, une *faux négatif*, ou *erreur de type II* se produit lorsque on ne parvient pas à rejeter une hypothèse nulle fausse. Idéalement, on souhaite minimiser simultanément la fréquence des erreurs de type I et II. Malheureusement, cela n'est pas réalisable et il faut trouver un compromis entre ces deux types d'erreurs. Ce compromis implique généralement la minimisation des erreurs de type II tout en respectant une contrainte d'erreur de type I. La minimisation du nombre de faux négatifs peut également être comprise comme la maximisation de la quantité de vraies découvertes, qui est la puissance statistique du test d'hypothèse.

Définition 9.4 (Puissance statistique)**.** La puissance statistique d'un test d'hypothèse simple est la probabilité que le test rejette correctement l'hypothèse nulle $\mathcal{H}_0$ lorsque l'hypothèse alternative $\mathcal{H}_a$ est vraie.

9.2 Test d'hypothèses multiples

Bien que les tests simples restent toujours populaires et ont été largement étudiés théoriquement, de nos jours, face à un phénomène donné, les statisticiens posent rarement une seule une seule question dans la tâche d'inférence. C'est ainsi que se pose le problème de l'évaluation de la signification statistique de plusieurs caractéristiques simultanément, ou *tests d'hypothèses multiples*. Plus concrètement, considérons le problème général du test simultané de m hypothèses $\{\mathcal{H}_0^i\}_{i=1}^m$. Si l'on suppose que des tests statistiques sont disponibles pour chaque hypothèse individuelle, le problème est de savoir comment les combiner en une procédure de test simultané. Une approche naïve consisterait à ignorer la multiplicité et à tester simplement chaque hypothèse au niveau α, comme dans le cas univarié. Cependant, avec une telle procédure, la probabilité d'un ou plusieurs faux positifs augmente rapidement lorsque m devient grand, car

$$\begin{aligned} &\mathbb{P}\left(\text{faire au moins un faux positif en testant m hypothèses}\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{m} \text{faire un faux positif sur la } i\text{-ème hypothèse nulle}\right). \end{aligned} \tag{9.1}$$

En d'autres termes, nous risquons d'effectuer davantage de rejets erronés lorsque le nombre de tests grandit, ce qui correspond au cadre des données à haute dimension. L'exemple suivant illustre ce problème dans l'analyse des images cérébrales, mais il est également courant dans d'autres applications, comme la génomique ou les données sur la santé, où de nombreuses affirmations dans les études scientifiques ne peuvent être reproduites en raison d'une utilisation potentiellement incorrecte des procédures de test d'hypothèse [Ioa05, BIM+13, ENK16].

Exemple 9.4 (Approche naïve du test univarié en imagerie cérébrale)**.** Un exemple typique en analyse d'imagerie cérébrale : supposons $p = 100000$ de voxels cérébraux, dont seulement 2000 sont importants. Si un statisticien effectue naïvement le test d'importance pour chaque variable avec le taux de faux positifs standard de 5%, alors même s'il découvre toutes les variables importantes, il reste en moyenne environ 5000 ($5\% \times 98,000$) variables non importantes qui sont faussement détectées. Le résultat est une liste de variables découvertes (supposées importantes) dont environ 96% sont en fait sans importance. Cela signifie que l'affirmation selon laquelle la procédure contrôle la probabilité de faux rejets au niveau 5% est trompeuse. La figure 9.1 illustre ce problème.

L'exemple 9.4 montre qu'afin d'effectuer des tests multiples correctement, nous devons réduire le niveau de signification pour chaque test individuel. Naturellement, plus il y a d'hypothèses à tester simultanément, plus le niveau de signification est

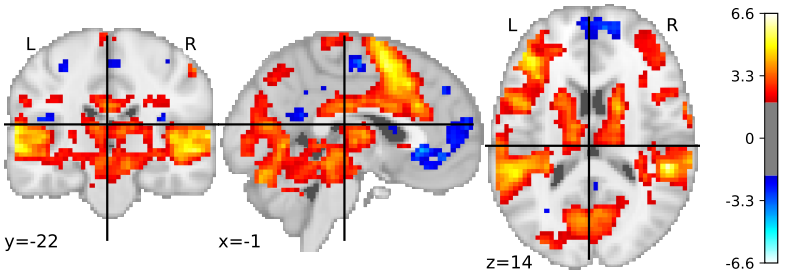


Figure 9.1: P-valeurs (non ajustées pour les tests multiples) associées à chaque voxel du cerveau au niveau de signification 5%. L'hypothèse à tester ici est de savoir quels voxels du cerveau sont activés lorsqu'une stimulation visuelle est présentée aux sujets humains. Les parties actives du cerveau sélectionnées sont bien plus nombreuses que le consensus général en neurosciences. Plus précisément, certaines zones activées du cerveau appartiennent au cortex auditif, qui n'est pas lié aux tâches de vision.

strict. Cette procédure s'appelle *correction des tests multiples*, et elle s'accompagne de plusieurs métriques plus adaptées. Nous pouvons utiliser les procédures de correction de la p-valeur pour ajuster le niveau de signification à ces métriques. Dans la section suivante, nous allons présenter deux des métriques les plus populaires: Taux d'erreur par famille (Family-Wise Error Rate – FWER) et Taux de fausses découvertes (False Discovery Rate – FDR).

	$\mathcal{H}_0$ vraie	$\mathcal{H}_a$ vraie	Total
$\mathcal{H}_0$ rejeté	V (faux positifs, erreur de type I)	S	R
$\mathcal{H}_0$ pas rejeté	U	T (faux négatifs, erreur de type II)	$m-R$
Total	m_0	$m-m_0$	m

Table 9.1: Notation des résultats possibles du test de m hypothèses

Remarque 9.1. Avec les notations du Tableau 9.1, la formule pour la puissance statistique dans les tests multiples est

$$\text{Power}(R) = \mathbb{E}\left[\frac{S}{m-m_0}\right]. \tag{9.2}$$

Remarque 9.2. Le lecteur peut se référer à [Roq15] comme une étude approfondie de la théorie des tests multiples, y compris certains résultats récents dans la littérature.

9.2.1 Taux d'erreur par famille (FWER)

Le taux d'erreur par famille est la probabilité de faire un ou plusieurs faux positifs lors de la réalisation de tests multiples, le terme "famille" désignant l'ensemble des hypothèses à tester simultanément. Les lecteurs peuvent peut-être reconnaître dans cette définition qu'il s'agit de la probabilité de part et d'autre de l'équation (9.1). Le FWER est une mesure classique de significativité dans les tests multiples. Au lieu d'essayer de contrôler la probabilité de faire une erreur de type I sur un test ne dépassant pas un niveau de signification statistique donné α, nous nous concentrons sur le contrôle du FWER en dessous de ce niveau :

$$\text{FWER} \leq \alpha.$$

Le contrôle du FWER est donc beaucoup plus fort que le contrôle du taux d'erreur de type I de chaque test d'hypothèse, puisque nous contrôlons la probabilité de

commettre une erreur de type I dans *n'importe quel* test, afin qu'elle soit inférieure à α. Les méthodes qui contrôlent cette métrique sont souvent obtenues par correction de la p-valeur de chaque test.

Définition 9.5 (Procédure de Bonferroni)**.** Soit $p_1, \dots, p_m$ les p-valeurs associées à la famille d'hypothèses $\mathcal{H}_0^1, \dots, \mathcal{H}_0^m$, alors la correction de Bonferroni rejette l'hypothèse nulle pour chaque $\mathcal{H}_0^i$ lorsque:

$$p_i \leq \frac{\alpha}{m}.$$

Théorème 9.1. La procédure de Bonferroni contrôle le FWER au niveau de signification α pour la famille d'hypothèses H_i, pour tous les $i = 1, \dots, m$.

Démonstration. Suivant les notations du Tableau 2.1, soit m_0 le nombre total d'hypothèses nulles qui sont vraies. À partir de la borne d'union et de la définition de la p-valeur, nous avons, pour la procédure de Bonferroni

$$\text{FWER} = \mathbb{P}\left[\bigcup_{i=1}^{m_0}\left(p_i \leq \frac{\alpha}{m}\right)\right] \leq \sum_{i=1}^{m_0}\mathbb{P}\left[p_i \leq \frac{\alpha}{m}\right] \leq m_0\frac{\alpha}{m} \leq \alpha.$$

□

La dénomination de cette méthode suit l'inégalité de Bonferroni [Bon36], une version généralisée de la borne d'union que nous utilisons pour prouver le contrôle de FWER. Cette méthode est un type de procédure *single-step*, c'est-à-dire que nous rejetons une hypothèse si sa p-valeur correspondante est inférieure à un seuil (qui, dans le cas de Bonferroni, est α/m). Une autre méthode de contrôle du FWER, la procédure de Holm, appartient à une autre classe de tests multiples appelée procédures *step-down*. Ces procédures reposent sur le seuillage des p-valeurs en fonction d'un ensemble de valeurs critiques $\alpha_1 \leq \cdots \leq \alpha_m$ avec $\alpha_i \in (0, 1)$, puis commencent par comparer la plus petite p-valeur avec le plus petit des α_i, α_1, et ainsi de suite. Un avantage de cette procédure par rapport à celle de Bonferroni est qu'en général, la procédure de Holm rend une puissance statistique plus élevée. Elle peut être décrite comme suit.

Définition 9.6 (Procédure de Holm)**.** Soit $p_1, \dots, p_m$ les p-valeurs (non corrigées) associées aux hypothèses $\mathcal{H}_0^1, \dots, \mathcal{H}_0^m$.

1. On ordonne de manière ascendante les m p-valeurs, notées $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, associées à $H_0^{(1)}, \dots, \mathcal{H}_0^{(m)}$.
2. Pour $i = 1, \dots, m$:
   - Si $p_{(i)} < \alpha/(m - i + 1)$, rejeter $\mathcal{H}_0^{(i)}$.
   - Sinon, si $p_{(i)} \geq \alpha/(m - i + 1)$, on arrête le processus et on accepte (on ne rejette pas) $\mathcal{H}_0^{(i)}, \dots, \mathcal{H}_0^{(m)}$.

Théorème 9.2. La procédure de Holm permet de contrôler le FWER sous le niveau de signification α.

Démonstration. Définissons $\mathcal{S}^c = \{i : \mathcal{H}_0^i \text{ vrai nul}\}$, donc $\text{card}\,(\mathcal{S}^c) = m_0$ avec notre système de notation. Soit j le plus petit indice satisfaisant $p_{(j)} = \min_{i \in \mathcal{S}^c} p_i$. Par définition, la procédure de Holm produit un faux positif si

$$p_{(j)} \leq \frac{\alpha}{m - j + 1},$$

ou

$$\min_{i \in \mathcal{S}^c} p_i \leq \frac{\alpha}{m - j + 1} \leq \frac{\alpha}{m_0},$$

puisque compte tenu de notre définition de p_j, on a $j \leq m - m_0 + 1$. En utilisant l'inégalité de Bonferroni:

$$\text{FWER} \leq \mathbb{P}\left[\min_{i \in \mathcal{S}^c} p_i \leq \frac{\alpha}{m_0}\right] \leq \sum_{i \in \mathcal{S}^c} \mathbb{P}\left[p_i \leq \frac{\alpha}{m_0}\right] \leq \alpha.$$

□ L'un des avantages de la correction de Bonferroni et de Holm est qu'aucune hypothèse sur la structure de corrélation des p-valeurs n'est nécessaire, puisque nous nous appuyons uniquement sur une borne d'union pour prouver la garantie théorique du contrôle du FWER. Cependant, le contrôle du FWER a un coût, celui d'être extrêmement conservatif lorsque l'ensemble d'hypothèses est important. Cela peut aller à l'encontre de l'objectif des praticiens, qui est de découvrir des caractéristiques qui présentent un effet intéressant. Dans la section suivante, nous examinerons un autre type de taux d'erreur qui tente de résoudre ce problème.

9.2.2 Taux de fausses découvertes (FDR)

L'article fondateur de Benjamini et Hochberg [BH95] a introduit un nouveau critère de tests multiples qui prend en compte la proportion de faux positifs, ou de fausses découvertes parmi toutes les hypothèses rejetées au lieu de la probabilité de faire des faux positifs comme avec le FWER. Depuis son introduction, le taux de fausses découvertes (FDR) est devenu une mesure populaire pour les tests multiples, grâce à un meilleur compromis entre le contrôle de l'erreur de type I et de type II que celui du FWER. Cela est en partie motivé par la disponibilité croissante de nombreux ensembles de données "larges", dans lesquels nous disposons de peu d'observations par rapport à un grand nombre de variables (ou d'hypothèses à tester), par exemple dans les données de génomique ou d'imagerie cérébrale. En d'autres termes, lorsque le nombre d'hypothèses m est important, par exemple lorsque l'on considère des ensembles de données à haute résolution, le contrôle du FWER permet de contrôler fortement les faux positifs, mais peut faire ressortir trop peu de variables statistiquement significatives (figure 3.2d). Les procédures de contrôle du FDR permettent de faire de nouvelles découvertes de manière moins restrictive, tout en maintenant un certain niveau de contrôle des erreurs, ce qui permet d'éviter une trop grande proportion de détections erronées. Tout d'abord, nous définissons la proportion de fausses découvertes (FDP) en nous fondant sur la notation introduite dans le tableau 2.1.

Definition 9.1 (Proportion de fausses découvertes (FDP) & taux de fausses découvertes (FDR) [BH95])**.**

(a) Proportion de fausses découvertes (False Discovery Proportion – FDP):

$$FDP = \frac{V}{R \vee 1} = \frac{\textit{nombre de fausses découvertes}}{\textit{nombre de découvertes}}$$

(b) Taux de fausses découvertes (False Discovery Rate – FDR):

$$FDR = \mathbb{E}\left[FDP\right] = \mathbb{E}\left[\frac{V}{R \vee 1}\right].$$

Nous devons nous attendre à ce que la proportion de fausses découvertes soit basée sur le fait que que la procédure de test est effectuée sur un échantillon d'observations, donc la FDP est aléatoire. En d'autres termes, le taux de fausses découvertes correspond à la proportion *moyenne* de découvertes qui sont fausses. Dans [BH95], une procédure pour contrôler le FDR a également été proposée. Cette procédure appartient à la famille des procédures *step-up*. Tout comme les procédures step-down,

comme la procédure de Holm mentionnée ci-dessus, les procédures step-up commencent également par ordonner les p-valeurs de manière ascendante, et les comparent avec un ensemble de valeurs critiques $\alpha_1 \leq \cdots \leq \alpha_m$, $\alpha_i \in (0,1)$. Cependant, les procédures step-up commencent la comparaison en utilisant la *plus petite* p-valeur significative avec le *plus grand* des α_i, ce qui est démontré avec la procédure de Benjamini-Hochberg (BH) comme suit.

Définition 9.7 (Procédure de Benjamini-Hochberg)**.** Soit $p_1, \ldots, p_m$ les p-valeurs associées aux hypothèses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$. Nous les réordonnons de manière ascendante, en notant $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, qui sont associées respectivement à $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}$.

1. Trouver $\widehat{k}_{BH}$ tel que

$$\widehat{k}_{BH} = \max\left\{k = [m] : p_{(k)} \leq \alpha_k\right\}, \quad \text{où } \alpha_k \overset{\text{def.}}{=} \frac{k\alpha}{m},$$

2. Si $\widehat{k}_{BH}$ existe, rejette $\mathcal{H}_0^{(1)}, \ldots \mathcal{H}_0^{(\widehat{k}_{BH})}$, sinon nous acceptons toutes les hypothèses.

Théorème 9.3. Lorsque les p-valeurs $p_1, \ldots, p_m$ sont indépendantes, la procédure BH contrôle la FDR au niveau α.

Comme le lecteur peut le remarquer, la procédure BH repose sur une hypothèse cruciale selon laquelle les p-valeurs sont *indépendantes*, ce qui est plutôt problématique dans des contextes pratiques. L'étude de [BY01] a fourni une discussion approfondie de ce problème, et a assoupli cette hypothèse. En particulier, ils ont prouvé que la FDR est toujours contrôlée lorsque la distribution conjointe des p-valeurs satisfait à une propriété appelée *dépendance positive en régression*, en plus du cas indépendant. Plus important encore, ce travail fournit également une procédure qui garantit le contrôle de la FDR dans le cas d'une *dépendance arbitraire*, c'est-à-dire dans toute structure de corrélation possible, qui est appelée procédure Benjamini-Yekutieli (BY). Nous décrivons cette procédure comme suit.

Définition 9.8 (Procédure de Benjamini-Yekutieli)**.** Soit $p_1, \ldots, p_m$ les p-valeurs associées aux hypothèses $\mathcal{H}_0^1, \ldots, \mathcal{H}_0^m$. Nous les réordonnons de manière ascendante, en notant $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$ et $\mathcal{H}_0^{(1)}, \ldots, \mathcal{H}_0^{(m)}$ les hypothèses correspondantes.

1. Trouver $\widehat{k}_{BY}$ tel que

$$\widehat{k}_{BY} = \max\left\{k \in [m] : p_{(k)} \leq \alpha_k, \right\} \quad \text{où } \alpha_k \overset{\text{def.}}{=} \frac{k\alpha}{m\sum_{i=1}^{m} 1/i}.$$

2. Si $\widehat{k}_{BY}$ existe, on rejette $\mathcal{H}_0^{(1)}, \ldots \mathcal{H}_0^{(\widehat{k}_{BH})}$. Sinon on accepte toutes les hypothèses.

Théorème 9.4. Sous toute structure de dépendance arbitraire des p-valeurs $p_1, \ldots, p_m$, la procédure BY contrôle le FDR au niveau α.

Remarque 9.3. Nous n'énoncerons pas ici la preuve complète que les procédures BH et BY contrôlent la FDR, sous les hypothèses indiquées ci-dessus, car elle est assez longue. La preuve du contrôle du FDR par la procédure BH peut être trouvée dans [BH95] en utilisant une preuve par récurrence. Une approche différente, qui utilise la théorie des martingales, est présentée dans [Sto02]. Pour prouver le contrôle FDR de la procédure BY, le lecteur peut se référer à [BY01].

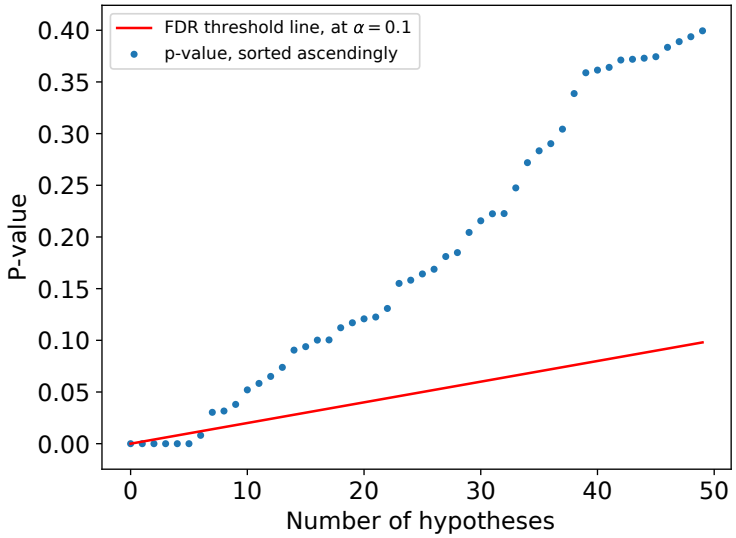


Figure 9.2: Procédure de seuillage FDR de Benjamini-Hochberg, la ligne rouge illustrant la fonction $(k\alpha)/m$, pour $k = 1, \ldots, m$. Nous rejetons toutes les hypothèses k qui ont des p-valeurs $p_{(k)} \leq \dfrac{k\alpha}{m}$, correspondant aux points situés sous la ligne rouge.

Problèmes potentiels de la FDR comme métrique de tests multiples Comme mentionné ci-dessus, bien que nous ayons le contrôle de la FDR, la FDP est toujours une variable aléatoire [Roq15]. Cela signifie qu'une procédure qui garantit un FDR $\leq \alpha$ peut toujours échouer à obtenir un FDP en dessous du niveau de signification α pour différents échantillons d'observations. Cette situation est problématique dans de nombreuses applications, par exemple en imagerie médicale, où il faut éviter de se tromper en déterminant si le patient est atteint ou non d'une maladie pulmonaire. Une solution naturelle à ce problème consiste à construire un intervalle de confiance pour le FDP. Plusieurs études vont dans ce sens, par exemple les premiers travaux de [GW06], ou les travaux plus récents de [GS11, GMKS19, BNR20] et [KR20] qui introduisent une procédure de tests multiples *post hoc*. Dans une telle procédure, le statisticien choisit librement la collection d'hypothèses rejetées, et la procédure de tests multiples renvoie le critère de qualité associé. Cela contraste avec les procédures traditionnelles de contrôle des FWER et FDR, qui renvoient la collection d'hypothèses rejetées en fonction d'un niveau *préspécifié*.

Limitations des procédures BH et BY Outre le problème de la FDR, il existe également des limites avec les deux procédures classiques procédures classiques de contrôle de cette métrique. La procédure de Benjamini-Hochberg repose sur l'indépendance ou la dépendance positive des statistiques de test, ce qui peut ne pas être le cas dans certains scénarios. D'autre part, bien que l'on puisse se fier à la procédure de Benjamini-Yekutieli pour de tels cas de dépendance arbitraire, le lecteur doit noter que le facteur $\sum_{i=1}^{m} 1/i$ dans la définition 2.5 de la procédure BY approche $\log m$ lorsque m devient grand, ce qui rend chaque seuil de p-valeur beaucoup plus petit. Cela signifie que, bien que nous ayons un contrôle fort de la FDR dans n'importe quelle structure de dépendance, BY est intrinsèquement beaucoup plus conservateur que la procédure BH, de manière très similaire à la conservativité de la procédure de Bonferroni qui ne nécessite aucune autre hypothèse sur la distribution de la statistique de test pour contrôler le FWER.

9.3 Inférence statistique en grande dimension

Définition du problème Supposons que nous ayons un échantillon d'observations avec n variables aléatoires i.i.d., chacune de la forme $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ avec la réponse $y \in \mathbb{R}$. De plus, ils suivent la relation linéaire

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^0 + \boldsymbol{\varepsilon} \tag{9.3}$$

avec $\mathbf{X} \in \mathbb{R}^{n \times p}$ est la matrice des covariables qui contient n lignes de $\mathbf{x}$, $\mathbf{y}$ le vecteur des réponses, $\boldsymbol{\beta}^0 \in \mathbb{R}^p$ le vecteur inconnu des coefficients de régression, $\boldsymbol{\varepsilon}$ le vecteur de bruit gaussien avec une amplitude de bruit de σ, c'est-à-dire que nous avons $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Nous supposerons en outre que la distribution de $\mathbf{X}$ est normale multivariée avec un vecteur de moyenne zéro et une covariance $\boldsymbol{\Sigma}$.

Test d'indépendance conditionnelle Outre les tests d'hypothèses multiples, un autre problème important auquel le statisticien doit faire face est le choix d'utiliser une approche *univariée* ou *multivariée* pour l'inférence statistique. Le problème de test énoncé dans la section 2.1 est un exemple d'approche univariée, ou inférence marginale. Comme son nom l'indique, dans cette approche, on ne s'intéresse qu'au test de la relation marginale d'une variable $\mathbf{x}_j$ par rapport à la réponse $\mathbf{y}$ sans prendre en compte l'interaction de j et des autres variables $\mathbf{X}_{-j} \overset{\text{def.}}{=} \{\mathbf{x}_1, \ldots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \ldots, \mathbf{x}_p\}$. Toutefois, cela n'est pas réaliste dans la plupart des applications pratiques, surtout de nos jours où les ensembles de données deviennent de plus en plus complexes grâce aux progrès des technologies d'acquisition. Par exemple, en neuro-imagerie, il est communément admis qu'il existe une corrélation positive entre des voxels cérébraux voisins. En particulier, lors d'une simulation mentale comme le visionnage d'une courte vidéo, nous observons généralement l'activation d'une région de plusieurs voxels cérébraux. Par conséquent, il est plus probable de formuler la question dans le sens d'une relation conditionnelle entre $\mathbf{y}$ et $\mathbf{x}_j$: le voxel cérébral $\mathbf{x}_j$ est-il activé lors de la présentation du stimulus mental $\mathbf{y}$, étant donné son interaction avec d'autres voxels cérébraux $\mathbf{X}_{-j}$? C'est ce qu'on appelle l'inférence multivariée, ou test d'indépendance conditionnelle. La figure 9.3 illustre cette idée dans le cadre des neurosciences.

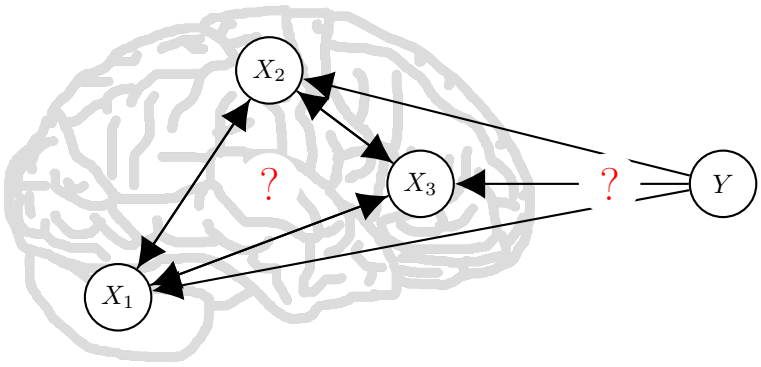


Figure 9.3: Illustration de l'interaction entre 3 voxels du cerveau en neurosciences cognitives et une variable comportementale Y (qui est typiquement déclenchée par un stimulus contrôlé). L'objectif est d'étudier si X_j est activé compte tenu de la condition mentale Y, conditionnellement à l'interaction avec les deux autres voxels cérébraux X_{-j}. La figure est une adaptation de [WMO+15, WP20].

Nous définissons formellement le problème du test d'hypothèse conditionnel comme suit. Soit $\mathcal{H}_0^1, \ldots \mathcal{H}_0^p$ la famille d'hypothèses à tester. Avec chaque variable j, nous voulons tester son indépendance avec la réponse $\mathbf{y}$, conditionnellement à d'autres

variables, ou en général:

$$\text{(null) } \mathcal{H}_0^j : \mathbf{x}_j \perp y_j \mid \mathbf{x}_{-j} \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_a^j : \mathbf{x}_j \not\perp y_j \mid \mathbf{x}_{-j}, \tag{9.4}$$

pour chaque caractéristique $j \in [p]$. De manière équivalente, en relation linéaire de l'Eq. 2.3, on a, pour chaque $j \in [p]$:

$$\text{(null) } \mathcal{H}_0^j : \beta_j^0 = 0 \quad \text{vs.} \quad \text{(alternative) } \mathcal{H}_a^j : \beta_j^0 \neq 0.$$

Une approche classique consiste à trouver d'abord l'estimateur $\hat{\boldsymbol{\beta}}$ en résolvant le problème d'estimation par les moindres carrés

$$\hat{\boldsymbol{\beta}} = \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2, \tag{9.5}$$

où $\|\cdot\|_2$ est la norme Euclidienne. Ce problème est facilement soluble lorsque $n > p$, ou le régime de "basse dimension". Plus précisément, nous disposons de la formule close pour $\hat{\boldsymbol{\beta}}$ lorsque en prenant la condition du premier ordre du problème d'optimisation:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{9.6}$$

où nous supposons que la matrice $\mathbf{X}^T\mathbf{X}$ est inversible (possible en faible dimension). Le calcul des statistiques de test $T(\mathbf{x}_j, y)$ pour chaque variable j, qui fait souvent intervenir $\hat{\beta}_j$, se fait donc naturellement. Avec cela, nous pouvons avoir des p-valeurs et faire la correction des tests multiples naturellement avec les procédures mentionnées dans la section 9.2. Cependant, lorsque $n < p$, ou en régime de haute dimension, cette matrice n'est pas inversible, et il y aura de multiples solutions au problème d'optimisation de l'éq. (9.5). De plus, le coût d'itération de l'approximation du pseudo-inverse de $\mathbf{X}^T\mathbf{X}$ est cubique en p, ce qui signifie que le calcul devient prohibitif lorsque p devient très grand. Nous devons donc penser à différentes approches pour nous adapter à ce régime, qui seront présentées dans la section suivante.

9.4 Filtre knockoff: une approche moderne pour contrôler le taux de fausses découvertes en grande dimension

Introduit dans le travail fondateur de [BC15] et développé par [CFJL18], le filtre knockoff est l'une des récentes percées dans la littérature sur les tests multiples. Le trait le plus distinctif de l'approche de cette méthode est la création de copies bruitées des variables originales, ce qui facilite le calcul des statistiques de test. Ces copies bruitées des variables sont appelées knockoffs. De manière plus formelle, supposons que nous continuions avec le cadre de la section précédente, c'est-à-dire la relation linéaire (9.3) avec n variables aléatoires *i.i.d.* , chacune de la forme $\mathbf{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p$ avec la réponse $y \in \mathbb{R}$. Nous cherchons à tester l'indépendance conditionnelle de la variable x_j avec y conditionnellement à toutes les autres variables $\mathbf{x}_{-j}$, de manière similaire au cadre formel de l'Eq. (9.4). La définition d'une variable knockoff est la suivante.

Définition 9.9 (Variable knockoff, [BC15, CFJL18])**.** $\tilde{\mathbf{x}}$ est une variable knockoff d'une famille de variables aléatoires $\mathbf{x} = (x_1, \ldots, x_p)$ si $\tilde{\mathbf{x}}$ satisfait aux deux propriétés suivantes:

1. Pour tout sous-ensemble $\mathcal{K} \subset \{1, \ldots, p\}$, $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{K})} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$, où le vecteur $(\mathbf{x}, \tilde{\mathbf{x}})_{\text{swap}(\mathcal{K})}$ représente la permutation des entrées x_j et $\tilde{x}_j$ pour tous les $j\mathcal{K}$, et $\overset{d}{=}$ représente l'égalité en loi.
2. $\tilde{\mathbf{x}} \perp\!\!\!\perp \mathbf{y} \mid \mathbf{x}$.

La première propriété de la Déf. 9.9 est appelée échangeabilité. La figure 9.4 illustre un exemple de variable knockoff avec $\mathcal{K} = j, k$ avec cette propriété d'échangeabilité. Pour échantillonner les knockoffs, [CFJL18] s'appuient sur l'hypothèse que la distribution des données P_X est connue. Ils ont appelé cela l'échantillonnage knockoff du second ordre, qui, comme son nom l'indique, utilise la moyenne (premier moment) et la variance (second moment) de $\mathbf{X}$ comme entrée:

$$\mathbb{E}[\tilde{\mathbf{X}}] = \mathbb{E}[\mathbf{X}], \quad \mathbb{E}[\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}] = \mathbf{\Sigma} \quad \text{and} \quad \mathbb{E}[\tilde{\mathbf{X}}^T\mathbf{X}] = \mathbf{\Sigma} - \text{diag}(\{)\mathbf{s}\}, \tag{9.7}$$

où $\mathbf{\Sigma}$ est la covariance de la population, diag$(\{)\mathbf{s}\}$ est une matrice diagonale agissant comme une perturbation qui fait de la matrice de corrélation entre $\mathbf{X}$ et $\tilde{\mathbf{X}}$ une matrice non triviale, *i.e.* différente de $\mathbf{\Sigma}$. Elle peut être calculée soit en résolvant un programme d'optimisation semi-définie, soit en utilisant une formule close. Si la distribution connue P_X est gaussienne, nous pouvons alors échantillonner $\tilde{\mathbf{X}}$ satisfaisant la condition ci-dessus par

$$\tilde{\mathbf{x}}_j \mid \mathbf{x}_j \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{V}), \tag{9.8}$$

où

$$\boldsymbol{\mu} = \mathbf{X} - \mathbf{X}\mathbf{\Sigma}^{-1}\text{diag}(\{)\mathbf{s}\} \tag{9.9}$$

$$\text{et } \mathbf{V} = 2\text{diag}(\{)\mathbf{s}\} - \text{diag}(\{)\mathbf{s}\}\Sigma^{-1}\text{diag}(\{)\mathbf{s}\}. \tag{9.10}$$

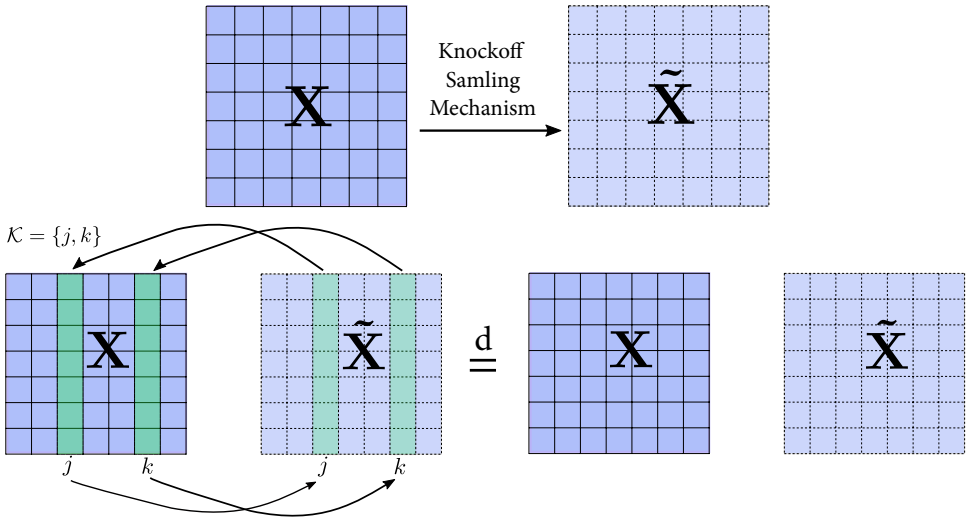


Figure 9.4: Une visualisation de l'échantillonnage des covariables knockoff (création de copies bruitées des variables originales) et de la propriété de permutation des variables. Voici quelques-unes des méthodes d'échantillonnage knockoff: knockoff de second ordre [CFJL18], Knockoff de Markov caché [SSC18], Knockoff de Metropolis-Hasting [BCJW20], Variantes de Deep Knockoff : Auto-encodeur et réseaux adversariaux génératifs [RSC18, JY19]

Après avoir obtenu les variables knockoff, il est crucial de calculer les statistiques de knockoff (ou score d'importance de knockoff), définies comme suit.

Définition 9.10 (Statistique de knockoff, [BC15, CFJL18])**.** Une statistique de knockoff $\mathbf{W} = \{W_j\}_{j=1}^p$ est une mesure de l'importance d'une caractéristique qui satisfait aux deux propriétés suivantes:

1. $\mathbf{W}$ est fonction de $\mathbf{X}, \tilde{\mathbf{X}}$ et $\mathbf{y}$ uniquement

$$\mathbf{W} = f(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y}).$$

2. L'échange de la colonne de variable originale $\mathbf{x}_j$ et de sa colonne fictive $\tilde{\mathbf{x}}_j$ change le signe de W_j, pour tout $\mathcal{K} \subseteq 1, \ldots, p$.

$$W_j\left([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(\mathcal{K})}, y\right) = \begin{cases} W_j\left([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}\right) & \text{if } j \in \mathcal{K}^c \\ -W_j\left([\mathbf{X}, \tilde{\mathbf{X}}], y\right) & \text{if } j \in \mathcal{K}, \end{cases}$$

L'idée est de prendre W_j telle qu'une grande valeur positive de W_j apporte une preuve contre l'hypothèse nulle $H_0^j : \boldsymbol{\beta}_j^0 = 0$. En particulier, si nous désignons par $\mathcal{S}$ l'ensemble de support vrai, défini comme $\mathcal{S} \stackrel{\text{def.}}{=} \{j \in [p] : \beta_j^0 \neq 0\}$ où β_j^0 le vrai coefficient de régression, la phrase précédente signifie que la probabilité $\mathbb{P}(W_j \geq k)$ avec $k > 0$ doit être grande lorsque $j \in \mathcal{S}$. À partir de la propriété d'échangeabilité, il est facile de vérifier que la distribution nulle de la statistique de knockoff W_j est symétrique autour de zéro. Cette propriété sera utile pour calculer le seuil de contrôle du FDR plus tard. Il existe plusieurs façons de définir une statistique knockoff qui satisfait à la définition, par exemple:

(1) $W_j = |\mathbf{x}_j^T y| - |\tilde{\mathbf{x}}_j^T y|$.

(2) $W_j = |\hat{\beta}^{\text{LS}}|_j - |\hat{\beta}^{\text{LS}}|_{j+p}$ où
$\hat{\boldsymbol{\beta}}^{\text{LS}} = \left(\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^T \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]\right)^{-1} \left[\mathbf{X}\ \tilde{\mathbf{X}}\right]^T \mathbf{y}$ est la solution classique des moindres carrés sur la matrice augmentée $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$.

Cependant, notez que ces suggestions ne fonctionnent que lorsque $n > 2p$, soit le cadre de la *fixed-X knockoff* originale, proposée dans [BC15]. Afin de calculer les statistiques de knockoff en grande dimension, [CFJL18] ont suggéré d'exécuter d'abord le programme Lasso sur $\left[\mathbf{X}\ \tilde{\mathbf{X}}\right]$:

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{2p}}{\text{argmin}} \left\{ \frac{1}{2} \|\mathbf{y} - \left[\mathbf{X}\ \tilde{\mathbf{X}}\right] \boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\}. \tag{9.11}$$

Ensuite, nous pouvons définir W_j en fonction des coefficients Lasso estimés:

(1) Prend $z_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$ pour $z_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$, puis on prend $W_j = (z_j \vee z_{j+p}) \times \text{sign}(z_j - z_{j+p})$. C'est ce qu'on appelle la *statistique de Lasso-max.*

(2) $W_j = |\widehat{\beta}_j(\lambda)| - |\widehat{\beta}_{j+p}(\lambda)|$, le *coefficient-différence de Lasso (LCD).*

Dans cette thèse, nous avons choisi de travailler exclusivement avec la statistique LCD et les knockoffs du modèle-X, car cette statistique offre une certaine flexibilité et a montré de bonnes performances par rapport à d'autres choix de statistiques de knockoffs [CFJL18, KLM20].

Après l'acquisition des statistiques de knockoff, l'étape finale consiste à calculer un seuil de test multiple dans le but de contrôler le FDR au niveau de signification $\alpha \in [0, 1]$.

$$\tau = \min\left\{ t > 0 : \widehat{\text{FDP}}(t) \leq \alpha \right\}, \quad \text{où } \widehat{\text{FDP}}(t) \stackrel{\text{def.}}{=} \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \tag{9.12}$$

et l'ensemble des variables sélectionnées est $\hat{\mathcal{S}} = \{j \in [p] : W_j \geq \tau\}$.

Remarque 9.4. Le seuil τ est en fait appelé `knockoff+` threshold dans la littérature originale [BC15, CFJL18]. Le nom est inspiré de la quantité $+1$ au numérateur de la quantité $\widehat{\text{FDP}}(t)$ dans l'Eq. (2.19) pour se distinguer de celle sans cette quantité. Nous n'avons pas besoin d'introduire deux notions différentes puisque seul le seuil `knockoff+` présente une garantie théorique pour le contrôle du FDR sous le niveau de signification α.

Voici une intuition à propos de la définition de $\widehat{\text{FDP}}(t)$ dans Eq. (9.12). La figure 9.5 montre l'histogramme d'un échantillon de statistiques de knockoff W_j pour chaque variable $j = 1, \ldots, p$. Pour tout $t > 0$, la zone rouge désigne le nombre de statistiques de test W_j qui sont inférieures ou égales à $-t$. Cette quantité peut être vue comme une estimation du nombre d'hypothèses nulles faussement rejetées. En effet, en raison de la symétrie de la distribution nulle de la statistique knockoff, cette quantité est égale au nombre de $W_j \geq t$, ou au nombre de H_j que nous avons faussement rejetés, c'est-à-dire les faux positifs. D'autre part, la zone bleue dénote l'estimation du support $\hat{\mathcal{S}}$, l'ensemble des hypothèses rejetées. C'est aussi le dénominateur de la formule de $\widehat{\text{FDP}}(t)$. Ainsi, par définition, si nous divisons la taille de la zone rouge par la taille de la zone bleue, nous aurons une estimation de la proportion de fausses découvertes, pour tout $t > 0$ donné. Nous notons également que si $j \in \mathcal{S}$, W_j sera positif avec une forte probabilité en raison de la propriété de la statistique du knockoff. Cela garantit que nous avons une bonne estimation du $\widehat{\text{FDP}}(t)$ avec la formule proposée.

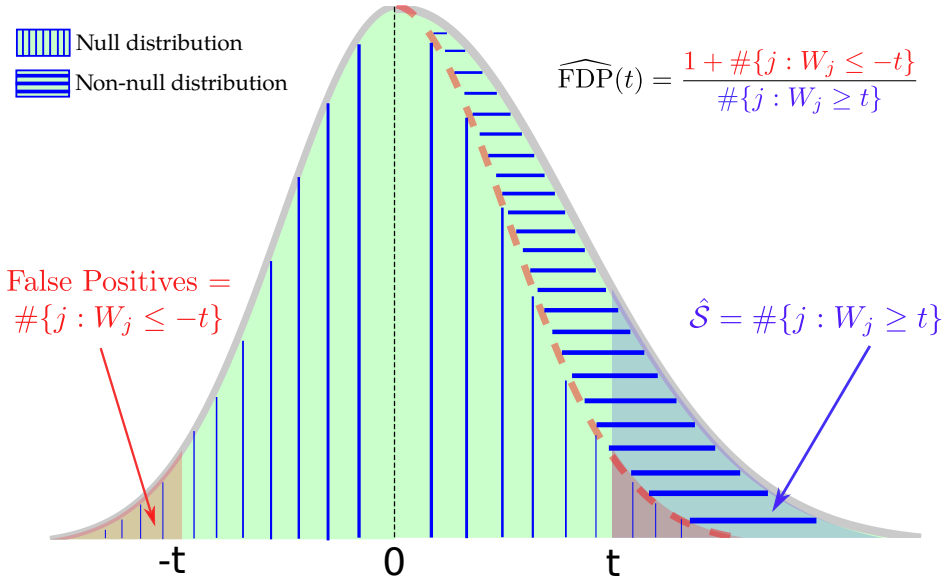


Figure 9.5: Illustration de la procédure d'estimation de la proportion de fausses découvertes avec les statistiques de knockoff : la figure montre un histogramme d'un échantillon de statistiques de knockoff. Le FDP est le rapport entre la taille de la zone rouge et la taille de la zone bleue de la queue droite de la distribution. Alors que la taille de la zone rouge à droite (qui indique le nombre de faux positifs) ne peut pas être calculée dans la réalité, elle peut être estimée par la zone rouge à gauche de la distribution en raison de la propriété de symétrie des statistiques knockoff.

Avantages du filtre knockoff Un avantage important du filtre knockoff est sa flexibilité: la méthodologie repose sur peu d'hypothèses. Par conséquent, nous pouvons utiliser knockoff sur une grande classe de problèmes et tester plusieurs statistiques de knockoffs. Par exemple, le statisticien peut sélectionner différents types de statistiques knockoff, dont certains sont énumérés ci-dessus, et passer de la résolution du programme Lasso standard au modèle linéaire généralisé avec régularisation ℓ_1. En outre, la flexibilité peut également provenir de la méthode de construction des knockoffs. On peut choisir un knockoff de second ordre comme suggéré dans l'article original de knockoff [CFJL18], ou à partir des extensions de [RSC18, JY19].

La construction knockoff proposée dans ces deux derniers travaux pourrait mieux évoluer en très haute dimension grâce aux architectures modernes de réseaux profonds. Pour prendre en compte l'interaction entre les variables latentes, ce qui est particulièrement vrai dans l'application de la génomique, [SSC18] a proposé le Hidden Markov Knockoff, et [BCJW20] a étendu l'idée avec le Fast Metropolis-Hasting Knockoff. Au-delà du régime du modèle-X, [BCS18] introduit une méthode plus robuste d'échantillonnage knockoff sans supposer que P_X est connue. Un réseau neuronal profond pourrait également être utilisé pour calculer la statistique du knockoff dans une relation non linéaire [LFLN18]. Nous renvoyons également les lecteurs aux travaux de [KLM20] pour une généralisation du filtre knockoff comme méthode de contrôle FDR. Un autre avantage du filtre knockoff est son coût de calcul relativement faible. Bien que certaines méthodes d'échantillonnage knockoff puissent être coûteuses, la proposition originale de knockoff de second ordre est peu coûteuse à calculer. La méthode ne nécessite également qu'une seule solution du problème d'optimisation pour le calcul de la statistique de test. Par rapport au test de permutation, par exemple, nous aurions besoin de plusieurs exécutions pour obtenir des p-valeurs empiriques significatives avec des coûts beaucoup plus élevés, bien que techniquement le filtre knockoff ne fournisse pas de p-valeurs pour chaque variable.

Quelques inconvénients du filtre knockoff Si l'inférence knockoff est en général une méthode souple et rapide, elle souffre d'un problème crucial. L'introduction de copies bruitées $\tilde{\mathbf{X}}$ signifie qu'il existe un caractère aléatoire inhérent à la méthode. Nous illustrons cette idée dans la figure 9.6 avec une expérience de données synthétiques. Tout d'abord, nous générons un ensemble de données synthétiques de 400 échantillons de $\mathbf{x}_j$ qui suit une distribution gaussienne multivariée avec une dimension de 600, et y généré selon la relation linéaire de l'équation (2.3) avec $\boldsymbol{\beta}^0$ épars. Nous effectuons ensuite 2500 exécutions du filtre knockoff sur ce même ensemble de données, et nous traçons l'histogramme du FDP résultant et de la puissance statistique. Bien que la FDR, moyenne de la FDP, soit contrôlée sous le niveau $\alpha = 0.1$, nous observons une grande variabilité dans les deux métriques. Cela suggère que le filtre knockoff est très instable dans les résultats d'inférence. Stabiliser la méthode est donc une question non triviale, et cette thèse fournit une des solutions à ce problème. Ce nouvel algorithme est basé sur une technique d'agrégation des p-valeurs, et sera présenté dans le chapitre 4.

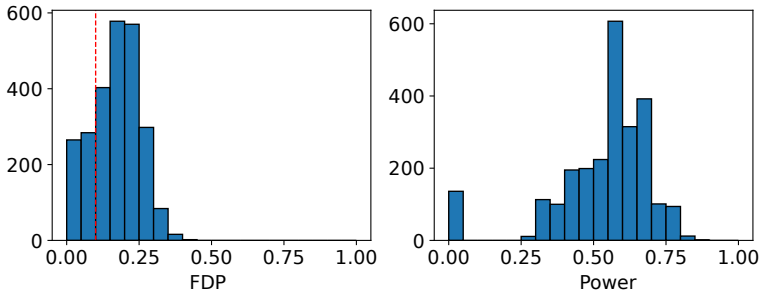


Figure 9.6: Histogramme du FDP et de la puissance statistique de 2500 exécutions du modèle X knockoff sur l'ensemble de *même* données synthétiques; $n = 400, p = 600$. Le FDR est contrôlé au niveau $\alpha = 0.1$. Comme indiqué ci-dessus, le FDP varie considérablement et peut être beaucoup plus élevé que α. Mais la puissance varie également de façon spectaculaire, et reste même à zéro dans une fraction des simulations.

9.5 Produire des p-valeurs valides en grande dimension via le test de randomisation conditionnelle

Introduit dans le même ouvrage qui présentait les knockoffs modèle-X [CFJL18], le *test de randomisation conditionnelle (Conditional Randomization Test – CRT)* est une approche alternative pour les tests multiples qui utilise les variables knockoff. L'idée est d'avoir un test de randomisation qui tire d'abord un nouvel échantillon $\tilde{\mathbf{x}}_j$ de chaque variable $\mathbf{x}_j$ selon la loi conditionnelle $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$. Ensuite, soit $T_j \overset{\text{def.}}{=} T(\mathbf{x}_j, \mathbf{y})$ une statistique de test calculée en utilisant $\mathbf{x}_j$ et $\mathbf{y}$. L'étape suivante consiste à trouver cette statistique, généralement en résolvant un problème d'optimisation avec régularisation. En fait, ces deux premières étapes sont similaires à celles de l'inférence knockoff: échantillonnage de copies bruitées de $\mathbf{x}_j$ et calcul des statistiques knockoff avec le programme Lasso. Cependant, contrairement au filtre knockoff, l'objectif du CRT en tant que test de randomisation est de produire une p-valeur empirique en exécutant ces deux étapes plusieurs fois. Nous comptons ensuite le nombre de résultats de test bruités $\tilde{T}_j \overset{\text{def.}}{=} T(\tilde{\mathbf{x}}_j, \mathbf{y})$ qui sont supérieurs au résultat de test original T_j. L'idée est que sous l'hypothèse nulle d'indépendance conditionnelle entre $\mathbf{y}$ et $\mathbf{x}_j$ étant donné $\mathbf{x}_{-j}$, T_j et $\tilde{T}_j$ sont identiquement distribués, donc, en prenant

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B} \tag{9.13}$$

où B est le nombre d'échantillonnages, nous pouvons avoir une bonne mesure de preuve contre la nullité si $\hat{p}_j$ est petit, ce qui se traduit par une valeur significativement plus grande de T_j contre $\tilde{T}_j$, conditionnellement à $\mathbf{x}_{-j}$. Notez que cette formule de p-valeurs empiriques est fondée sur le même principe que les p-valeurs des tests de permutation. Une description complète du CRT est présentée dans l'algorithme 9.1.

Algorithm 9.1: Conditional Randomization Test [CFJL18]

INPUT dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, with $\mathbf{x}_i \in \mathbb{R}^p$, number of sampling run B, test statistic T_j, conditional ditribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$ for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{\hat{p}_j\}_{j=1}^p$;
for $j = 1, 2, \ldots, p$ **do**
  for $b = 1, 2, \ldots, B$ **do**
    1. Generate $\tilde{\mathbf{x}}_j^{(b)}$, a knockoff sample from $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$;
    2. Compute test statistics for original variable T_j and for knockoff variables $\tilde{T}_j$;
  end
  Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j^{(b)} \geq T_j}}{1 + B}$$

end

Le CRT est une méthode puissante dans le sens où nous pouvons échantillonner à partir de n'importe quel type de distribution conditionnelle de n'importe quelle statistique de test, une flexibilité que l'on a également avec le filtre knockoff. Cependant, comme pour la plupart des types de tests de randomisation, le CRT est coûteux à réaliser. Il nécessite le calcul de p-valeurs aléatoires pour toutes les variables et pour

un nombre B d'échantillonnages suffisamment grand. De plus, pour chaque échantillonnage et calcul des statistiques de test T_j, nous devrions prendre en compte la dimensionnalité complète des données, ce qui signifie qu'une complexité algorithmique de $\mathcal{O}(p^3)$ est requise. Par conséquent, la complexité algorithmique de la CRT est de $\mathcal{O}(Bp^4)$, ce qui est prohibitif. La réduction du coût de calcul est donc l'un des principaux objectifs de certaines extensions du CRT. Nous allons brièvement passer en revue trois des variantes du CRT, appelées Test de randomisation de retenue (Holdout Randomization Test – HRT), Test de permutation conditionnelle (CPT) et Test de randomisation conditionnelle distillée (Distilled CRT – dCRT).

Test de permutation conditionnelle (Conditional Permutation Test) [BWBS19] Peut-être le cousin le plus proche, le test de permutation conditionnel (CPT) comporte presque le même nombre d'étapes que le CRT, à l'exception qu'au lieu d'utiliser la loi conditionnelle $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$, nous appliquons une fonction de fonction de permutation aux variables originales pour créer $\tilde{\mathbf{x}}_j$. Pour garantir la propriété d'inversion de signe de la statistique de test T_j et de $\tilde{T}_j$, cette fonction de permutation doit satisfaire certaines propriétés spécifiques, définies dans le même ouvrage. Les auteurs ont également fourni des arguments selon lesquels le CPT est plus robuste que le CRT en pratique, et le schéma de permutation peut dans certains cas être plus rapide que l'échantillonnage de la distribution conditionnelle des variables.

Test de randomisation de retenue (Holdout Randomization Test) [TVZ+18]. L'idée de base du test de randomisation de retenue (HRT) est de choisir la statistique de test $\tilde{T}_j$s de manière à minimiser le coût de calcul pour chacun des échantillonnages. Plus précisément, la différence est que nous divisons d'abord l'ensemble de données $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ en deux parties : les données de formation $\mathcal{D}_{train}$ et les données de test $\mathcal{D}_{test}$. L'ensemble de données d'entraînement sera utilisé pour trouver une estimation $\hat{\boldsymbol{\beta}}$ du vrai paramètre $\boldsymbol{\beta}^0$. Nous utilisons ensuite $\mathcal{D}_{test}$ pour évaluer la fonction de risque empirique, désignée par $L(\mathcal{D}_{test}, \hat{\boldsymbol{\beta}})$, qui sera également utilisée comme statistique de test T_j pour mesurer l'importance de la variable. Le calcul des p-valeurs empiriques est similaire à celui du CRT: pour chaque variable j, nous exécutons l'échantillonnage knockoff B fois, en utilisant la distribution conditionnelle $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$. Puis, à chaque exécution, nous remplaçons les échantillons de $\mathbf{x}_j$ dans $\mathcal{D}_{test}$ par ceux de $\tilde{\mathbf{x}}_j$, et recalculons la statistique de test, ou risque empirique $L(\tilde{\mathcal{D}}_{test}, \hat{\boldsymbol{\beta}})$. Une formule similaire aux p-valeurs empiriques du CRT est alors calculée. Le pseudo-code du HRT peut être trouvé dans l'algorithme 9.2.

Le HRT, tout en améliorant le temps de calcul en éliminant la partie d'ajustement de l'estimateur pour chaque échantillon, nécessite toujours l'échantillonnage de nouvelles copies bruitées pour chaque variable j dans chaque cycle d'échantillonnage b. Dans les ensembles de données à grande échelle, cela peut encore être très coûteux. De plus, la méthode HRT pose un problème différent de celui de la méthode CRT : la division de l'ensemble de données en deux parties peut potentiellement faire perdre à la méthode HRT une bonne partie de sa puissance statistique.

Distilled Conditional Randomization Test (dCRT) [LKJR20] Le Distilled CRT (dCRT) a pour but de corriger le coût de calcul prohibitif du CRT standard. Ceci est réalisé par un processus appelé *opérateur de distillation*. L'idée clé derrière la distillation est que les informations importantes d'une variable $\mathbf{x}_j$ sont "distillées" vers la réponse $\mathbf{y}$ et vers les $j-1$ variables restantes, c'est-à-dire la matrice de données $\mathbf{X}_{-j}$. Plus formellement, pour chaque variable j, nous effectuons la distillation en résolvant d'abord deux problèmes lasso :

$$\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{9.14}$$

Algorithm 9.2: Holdout Randomization Test [TVZ+18]

INPUT dataset $\mathcal{D} = \{\mathbf{x}_i, y_i\}_{i=1}^n$, a empirical risk function $L(\cdot)$, number of sampling run B, test statistic T_j, conditional ditribution $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$ for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\boldsymbol{\pi} = \{\hat{p}_j\}_{j=1}^p$;
Split $\mathcal{D}$ into two parts: $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$;
$\hat{\boldsymbol{\beta}} \leftarrow \texttt{fit}(\mathcal{D}_{train})$;
Compute empirical risk, used as test statistic: $T_j \overset{\text{def.}}{=} L(\mathcal{D}_{test}, \hat{\boldsymbol{\beta}})$;
for $j = 1, 2, \ldots, p$ **do**
    for $b = 1, 2, \ldots, B$ **do**
        1. Generate $\tilde{\mathbf{x}}_j^{(b)}$, a knockoff sample from $P_{\mathbf{x}_j|\mathbf{x}_{-j}}$;
        2. Replace samples of $\mathbf{x}_j$ in $\mathcal{D}_{test}$ with $\tilde{\mathbf{x}}_j^{(b)}$ to form $\tilde{\mathcal{D}}_{test}$;
        3. $\tilde{T}_j \overset{\text{def.}}{=} L(\tilde{\mathcal{D}}_{test}, \hat{\boldsymbol{\beta}})$;
    end
    Compute the empirical p-value

$$\hat{p}_j = \frac{1 + \sum_{b=1}^{B} \mathbf{1}_{\tilde{T}_j \geq T_j}}{1 + B}$$

end

et

$$\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 . \tag{9.15}$$

Ensuite, une statistique de test peut être calculée par la formule suivante

$$T_j \overset{\text{def.}}{=} T(x_j, y) = \left[(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})\right]^2 . \tag{9.16}$$

Une propriété supplémentaire intéressante de cette procédure est qu'elle peut directement produire une p-valeur analytique, puisque la statistique de l'importance de la caractéristique T_j est supposée suivre une distribution gaussienne, conditionnellement à $\mathbf{y}$ et $\mathbf{X}_{-j}$ [LKJR20]:

$$(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}}) \sim \mathcal{N}(0, \sigma^2 \left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|^2). \tag{9.17}$$

Il s'ensuit que nous pouvons obtenir la p-valeur exacte pour chaque variable j en prenant

$$p_j = 2\left[1 - \Phi\left(\frac{\sqrt{T_j}}{\hat{\sigma}\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|}\right)\right], \tag{9.18}$$

où $\Phi(\cdot)$ est la fonction de distribution cumulative (CDF) normale standard et $\hat{\sigma}$ est l'ampleur du bruit estimé.

Bien que [LKJR20] ait présenté le dCRT comme un cadre général permettant de produire des p-valeurs avec un choix flexible de l'opérateur de distillation, ce que nous avons présenté ci-dessus est une variante appelée Lasso Distillation CRT (Lasso-dCRT), sur laquelle les auteurs du travail se sont principalement concentrés dans leurs arguments théoriques et empiriques. Cela signifie que nous travaillons sous la relation linéaire de l'équation (9.3). Le Lasso-dCRT gagne sur le CRT standard en supprimant les étapes d'échantillonnage multiples pour chaque variable. Cela conduit à une réduction raisonnable du coût de calcul. Cependant, le lecteur peut toujours remarquer que la complexité d'exécution est toujours un polynôme d'ordre

4 en p, soit $\mathcal{O}(p^4)$. Pour résoudre ce problème, [LKJR20] a proposé une étape de sélection initiale en supposant que toutes les variables pertinentes sont sélectionnées au cours de cette étape, et a fixé toutes les p-valeurs des variables non sélectionnées à 1. Le temps d'exécution résultant est réduit à $\mathcal{O}(kp^3)$, où k est le nombre total de variables sélectionnées après le filtrage et est souvent beaucoup plus petit que p en régime de parcimonie. Plus formellement, l'étape de filtrage renvoie l'estimation du support $\hat{\mathcal{S}} := \{j \in [p] : \hat{\beta}_j \neq 0\}$, et k est la cardinalité de cet ensemble. Une description complète de l'algorithme est donnée dans l'algorithme 9.3.

Algorithm 9.3: Distilled Conditional Randomization Test (dCRT) [LKJR20]

INPUT dataset $\mathcal{D} = \{\mathbf{x_i}, y_i\}_{i=1}^n$, test statistic T_j for each $j = 1, \ldots, p$;
OUTPUT vector of p-values $\{p_j\}_{j=1}^p$;
$\hat{\mathcal{S}}^{\text{SCREENING}} = \{j : j \in [p], \hat{\beta}_j^{\text{LASSO}} \neq 0\}$;
for $j \in \hat{\mathcal{S}}^{SCREENING}$ **do**

1. Distill information of $\mathbf{X}_{-j}$ to $\mathbf{x}_j$ and to $\mathbf{y}$ by finding:
   - $\hat{\boldsymbol{\beta}}^{d_y}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{y} - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$
   - $\hat{\boldsymbol{\beta}}^{d_{x_j}}(\lambda) = \text{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{2} \|\mathbf{x}_j - \mathbf{X}_{-j}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$
2. Obtain test statistic:
$$T_j = \left[(\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y})^T(\mathbf{x}_j - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_{x_j}})\right]^2$$
3. Compute (two-sided) p-value
$$p_j = 2\left[1 - \Phi\left(\frac{\sqrt{T_j}}{\hat{\sigma}\left\|\mathbf{y} - \mathbf{X}_{-j}\hat{\boldsymbol{\beta}}^{d_y}\right\|}\right)\right]$$

end

Avantages et inconvénients du dCRT À notre connaissance, parmi toutes les variantes du test de randomisation conditionnelle, le dCRT est le seul à introduire une étape analytique pour calculer les p-valeurs au lieu de devoir procéder à un échantillonnage/permutation multiple. Par conséquent, le temps de calcul est réduit de manière significative. Cependant, comme nous l'avons déjà noté, la formule de distillation proposée dans l'Eq. (9.15) et l'Eq. (9.14) ne fonctionne que dans le cadre de la relation linéaire de l'Eq (9.3). De plus, la formule analytique des p-valeurs repose sur l'hypothèse que les statistiques de test T_j suivent une distribution normale standard. Cette hypothèse est assez forte et le fait qu'elle ne soit pas satisfaite, comme souvent dans les applications réelles, rendrait la procédure incorrecte. Dans le chapitre 6, nous étudions plus en détail ce phénomène pour la régression logistique en grande dimension. L'idée est de décorréler la formule de la statistique de test dCRT pour prendre en compte la relation non-linéaire entre $\mathbf{x}$ et y.

Bibliography

[Abd18] A. Abdesselam. The weakly dependent strong law of large numbers revisited. *arXiv e-prints*, page arXiv:1801.09265, January 2018.

[AC10] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Statistics Surveys*, 4(none), January 2010. arXiv: 0907.4728.

[ACC17] E. Arias-Castro and S. Chen. Distribution-free multiple testing. *Electron. J. Statist.*, 11(1):1983–2001, 2017.

[ADC21] J.-M. Azais and Y. De Castro. Multiple Testing and Variable Selection along Least Angle Regression's path. *arXiv:1906.12072 [cs, math, stat]*, July 2021. arXiv: 1906.12072.

[AGS+13] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013.

[AHV+10] S. Atwell, Y. S. Huang, B. J. Vilhjálmsson, G. Willems, M. Horton, Y. Li, D. Meng, A. Platt, A. M. Tarone, T. T. Hu, et al. Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. *Nature*, 465(7298):627, 2010.

[APE+14] A. Abraham, F. Pedregosa, M. Eickenberg, P. Gervais, A. Mueller, J. Kossaifi, A. Gramfort, B. Thirion, and G. Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[BBB+13] R. Berk, L. Brown, A. Buja, K. Zhang, and L. Zhao. Valid post-selection inference. *The Annals of Statistics*, pages 802–837, 2013.

[BC15] R. F. Barber and E. J. Candès. Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085, October 2015. arXiv: 1404.5609.

[BCJW20] S. Bates, E. Candès, L. Janson, and W. Wang. Metropolized knockoff sampling. *Journal of the American Statistical Association*, pages 1–15, 2020.

[BCS18] R. F. Barber, E. J. Candès, and R. J. Samworth. Robust inference with knockoffs. *arXiv:1801.03896 [stat]*, January 2018. arXiv: 1801.03896.

[BD08] T. Blumensath and M. E. Davies. Iterative thresholding for sparse approximations. *Journal of Fourier analysis and Applications*, 14(5-6):629–654, 2008.

[BEG+15] D. Bzdok, M. Eickenberg, O. Grisel, B. Thirion, and G. Varoquaux. Semi-Supervised Factored Logistic Regression for High-Dimensional Neuroimaging Data. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama,

and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[BH95] Y. Benjamini and Y. Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[BIM+13] K. S. Button, J. P. Ioannidis, C. Mokrysz, B. A. Nosek, J. Flint, E. S. Robinson, and M. R. Munafò. Power failure: why small sample size undermines the reliability of neuroscience. *Nature reviews neuroscience*, 14(5):365–376, 2013.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013.

[BNR20] G. Blanchard, P. Neuvial, and E. Roquain. Post hoc confidence bounds on false positives using reference families. *The Annals of Statistics*, 48(3):1281 – 1303, 2020.

[Bon36] C. Bonferroni. Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze*, 8:3–62, 1936.

[BR09] G. Blanchard and É. Roquain. Adaptive false discovery rate control under independence and dependence. *Journal of Machine Learning Research*, 10(Dec):2837–2871, 2009.

[BRvdGZ13] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, November 2013. arXiv: 1209.5908.

[BWBS19] T. B. Berrett, Y. Wang, R. F. Barber, and R. J. Samworth. The conditional permutation test for independence while controlling for confounders. *arXiv:1807.05405 [math, stat]*, May 2019. arXiv: 1807.05405.

[BY01] Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 08 2001.

[BZ21] P. C. Bellec and C.-H. Zhang. De-Biasing The Lasso With Degrees-of-Freedom Adjustment. *arXiv:1902.08885 [math, stat]*, July 2021. arXiv: 1902.08885.

[Bü13] P. Bühlmann. Statistical significance in high-dimensional linear models. *Bernoulli*, 19(4):1212–1242, September 2013. arXiv: 1202.1377.

[CB02] G. Casella and R. L. Berger. *Statistical inference*, volume 2. Duxbury Pacific Grove, CA, 2002.

[CBE+07] X. Cai, J. Ballif, S. Endo, E. Davis, M. Liang, D. Chen, D. DeWald, J. Kreps, T. Zhu, and Y. Wu. A Putative CCAAT-Binding Transcription Factor Is a Regulator of Flowering Timing in Arabidopsis. *Plant Physiology*, 145(1):98–105, 2007.

[CFJL18] E. Candès, Y. Fan, L. Janson, and J. Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[CH79] D. R. Cox and D. V. Hinkley. *Theoretical statistics*. CRC Press, 1979.

[CMW20] M. Celentano, A. Montanari, and Y. Wei. The Lasso with general Gaussian designs with applications to hypothesis testing. *arXiv:2007.13716 [cs, math, stat]*, July 2020. arXiv: 2007.13716.

[CNS+21] J.-A. Chevalier, T.-B. Nguyen, J. Salmon, G. Varoquaux, and B. Thirion. Decoding with confidence: Statistical control on decoder maps. *NeuroImage*, 234:117921, 2021.

[CNTS21] J.-A. Chevalier, T.-B. Nguyen, B. Thirion, and J. Salmon. Spatially relaxed inference on high-dimensional linear models. *arXiv:2106.02590 [math, stat]*, June 2021. arXiv: 2106.02590.

[CS03] D. D. Cox and R. L. Savoy. Functional magnetic resonance imaging (fMRI)"brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. *Neuroimage*, 19(2):261–270, 2003.

[CST18] J.-A. Chevalier, J. Salmon, and B. Thirion. Statistical Inference with Ensemble of Clustered Desparsified Lasso. *arXiv:1806.05829 [stat]*, June 2018. arXiv: 1806.05829.

[EK19] K. Emery and U. Keich. Controlling the FDR in variable selection via multiple knockoffs. *arXiv e-prints*, page arXiv:1911.09442, November 2019.

[ENK16] A. Eklund, T. E. Nichols, and H. Knutsson. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proceedings of the National Academy of Sciences*, 113(28):7900–7905, 2016.

[FHT10] J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.

[FHW+94] K. J. Friston, A. P. Holmes, K. J. Worsley, J.-P. Poline, C. D. Frith, and R. S. Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.

[Fis25] R. A. Fisher. *Statistical methods for research workers, Edinburg, Oliver and Boyd*. Auflage, 1925.

[Fis36] R. A. Fisher. *Design of experiments*, volume 1. British Medical Journal Publishing Group, 1936.

[FK00] W. Fu and K. Knight. Asymptotics for lasso-type estimators. *The Annals of statistics*, 28(5):1356–1378, 2000.

[Gir14] C. Giraud. *Introduction to high-dimensional statistics*. Chapman and Hall/CRC, 2014.

[GMKS19] J. J. Goeman, R. J. Meijer, T. J. Krebs, and A. Solari. Simultaneous control of all false discovery proportions in large-scale multiple hypothesis testing. *Biometrika*, 106(4):841–856, 2019.

[GS11] J. J. Goeman and A. Solari. Multiple testing for exploratory research. *Statistical Science*, 26(4):584–597, 2011.

[GSD+15] B. Gaonkar, R. T. Shinohara, C. Davatzikos, A. D. N. Initiative, et al. Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. *Medical image analysis*, 24(1):190–204, 2015.

[GW06] C. R. Genovese and L. Wasserman. Exceedance Control of the False Discovery Proportion. *Journal of the American Statistical Association*, 101(476):1408–1417, 2006. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

[GZ19a] J. R. Gimenez and J. Zou. Discovering Conditionally Salient Features with Statistical Guarantees. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2290–2298. PMLR, May 2019. ISSN: 2640-3498.

[GZ19b] J. R. Gimenez and J. Zou. Improving the Stability of the Knockoff Procedure: Multiple Simultaneous Knockoffs and Entropy Maximization. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2184–2192. PMLR, 16–18 Apr 2019.

[HGF+01] J. V. Haxby, M. I. Gobbini, M. L. Furey, A. Ishai, J. L. Schouten, and P. Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[HH18] L. Holden and K. H. Helton. Multiple Model-Free Knockoffs. *arXiv preprint arXiv:1812.04928*, 2018.

[HS17] H. Hang and I. Steinwart. A Bernstein-type inequality for some mixing processes and dynamical systems with an application to learning. *Ann. Statist.*, 45(2):708–743, 2017.

[HTF09] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. 2009. OCLC: 1058138445.

[Ioa05] J. P. Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[JJ19] A. Javanmard and H. Javadi. False discovery rate control via debiased lasso. *Electron. J. Statist.*, 13(1):1212–1253, 2019.

[JM14] A. Javanmard and A. Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[JY19] J. Jordon and J. Yoon. KnockoffGAN: Generating Knockoffs for Feature Selection using Generative Adversarial Networks. *International Conference on Learning Representations*, page 25, 2019.

[KKY+08] J. Kim, Y. Kim, M. Yeom, J.-H. Kim, and H. G. Nam. FIONA1 Is Essential for Regulating Period Length in the Arabidopsis Circadian Clock. *The Plant Cell*, 20(2):307–319, 2008.

[KLM20] Z. T. Ke, J. S. Liu, and Y. Ma. Power of FDR Control Methods: The Impact of Ranking Algorithm, Tampered Design, and Symmetric Statistic. *arXiv:2010.08132 [math, stat]*, October 2020. arXiv: 2010.08132.

[KR20] E. Katsevich and A. Ramdas. Simultaneous high-probability bounds on the false discovery proportion in structured, regression and online settings. *The Annals of Statistics*, 48(6):3465 – 3487, 2020.

[LCM+15] S. Légaré, L. Cavallone, A. Mamo, C. Chabot, I. Sirois, A. Magliocco, A. Klimowicz, P. N. Tonin, M. Buchanan, D. Keilty, et al. The estrogen receptor cofactor SPEN functions as a tumor suppressor and candidate biomarker of drug responsiveness in hormone-dependent breast cancers. *Cancer research*, 75(20):4351–4363, 2015.

[LFLN18] Y. Y. Lu, Y. Fan, J. Lv, and W. S. Noble. DeepPINK: reproducible feature selection in deep neural networks. *arXiv:1809.01185 [cs, stat]*, September 2018. arXiv: 1809.01185.

[Lin08] M. A. Lindquist. The Statistical Analysis of fMRI Data. *Statistical Science*, 23(4), November 2008.

[LKJR20] M. Liu, E. Katsevich, L. Janson, and A. Ramdas. Fast and Powerful Conditional Randomization Testing via Distillation. *arXiv:2006.03980 [stat]*, July 2020. arXiv: 2006.03980.

[LT15] J. R. Loftus and J. E. Taylor. Selective inference in regression models with groups of variables. *arXiv:1511.01478 [math, stat]*, November 2015. arXiv: 1511.01478.

[LTTT14] R. Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[MEG76] R. Marcus, P. Eric, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

[MGV+12] V. Michel, A. Gramfort, G. Varoquaux, E. Eger, C. Keribin, and B. Thirion. A supervised clustering approach for fMRI-based inference of brain states. *Pattern Recognition*, 45(6):2041–2049, June 2012. arXiv: 1104.5304.

[MM18] L. Miolane and A. Montanari. The distribution of the Lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv:1811.01212 [math, stat]*, November 2018. arXiv: 1811.01212.

[MMB09] N. Meinshausen, L. Meier, and P. Bühlmann. p-Values for High-Dimensional Regression. *Journal of the American Statistical Association*, 104(488):1671–1681, 2009.

[MPR09] F. Merlevède, M. Peligrad, and E. Rio. Bernstein inequality and moderate deviations under strong mixing conditions. In *High dimensional probability V: the Luminy volume*, volume 5 of *Inst. Math. Stat. (IMS) Collect.*, pages 273–292. Inst. Math. Statist., Beachwood, OH, 2009.

[MWP+07] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of cognitive neuroscience*, 19(9):1498–1507, 2007.

[NCT19] T.-B. Nguyen, J.-A. Chevalier, and B. Thirion. ECKO: Ensemble of Clustered Knockoffs for Robust Multivariate Inference on fMRI Data. In *International Conference on Information Processing in Medical Imaging*, pages 454–466. Springer, 2019.

[NCTA20] T.-B. Nguyen, J.-A. Chevalier, B. Thirion, and S. Arlot. Aggregation of multiple knockoffs. In *International Conference on Machine Learning*, pages 7283–7293. PMLR, 2020.

[NL17] Y. Ning and H. Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*, 45(1):158–195, February 2017. Publisher: Institute of Mathematical Statistics.

[OLKT90] S. Ogawa, T.-M. Lee, A. R. Kay, and D. W. Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24):9868–9872, 1990.

[PMN12] R. A. Poldrack, J. A. Mumford, and T. E. Nichols. *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge, 2012. OCLC: 861319956.

[RD19a] J. Romano and C. DiCiccio. Multiple Data Splitting for Testing. 2019.

[RD19b] J. P. Romano and C. DiCiccio. Multiple Data Splitting for Testing. Technical Report Technical Report 2019-03, Stanford University, Department of Statistics, April 2019. available at https://statistics.stanford.edu/research/multiple-data-splitting-testing.

[RFW+18] J. D. Rosenblatt, L. Finos, W. D. Weeda, A. Solari, and J. J. Goeman. All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796, 2018.

[Roq15] E. Roquain. Contributions to multiple testing theory for high-dimensional data. September 2015.

[RRJW20] M. Rabinovich, A. Ramdas, M. I. Jordan, and M. J. Wainwright. Optimal Rates and Tradeoffs in Multiple Testing. *Statistica Sinica*, 2020.

[RSC18] Y. Romano, M. Sesia, and E. J. Candès. Deep Knockoffs. *arXiv:1811.06687 [math, stat]*, November 2018. arXiv: 1811.06687.

[RTT15] S. Reid, J. Taylor, and R. Tibshirani. A general framework for estimation and inference from clusters of features. *arXiv:1511.07839 [stat]*, November 2015. arXiv: 1511.07839.

[RVN21] N. Randriamihamison, N. Vialaneix, and P. Neuvial. Applicability and Interpretability of Ward's Hierarchical Agglomerative Clustering With or Without Contiguity Constraints. *Journal of Classification*, 38(2):363–389, July 2021.

[RW05] J. P. Romano and M. Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.

[Sam00] P.-M. Samson. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Ann. Probab.*, 28(1):416–461, 2000.

[SC19] P. Sur and E. J. Candès. A modern maximum-likelihood theory for high-dimensional logistic regression. *Proceedings of the National Academy of Sciences*, 116(29):14516–14525, July 2019.

[SCAV19] L. Slim, C. Chatelain, C.-A. Azencott, and J.-P. Vert. kernelPSI: a Post-Selection Inference Framework for Nonlinear Variable Selection. In *International Conference on Machine Learning*, pages 5857–5865, 2019.

[SCS98] A. L. Silverstone, C. N. Ciampaglio, and T.-p. Sun. The Arabidopsis RGA Gene Encodes a Transcriptional Regulator Repressing the Gibberellin Signal Transduction Pathway. *The Plant Cell*, 10(2):155–169, 1998.

[SP20] R. D. Shah and J. Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020.

[SQL15] W. Su, J. Qian, and L. Liu. Communication-efficient false discovery rate control via knockoff aggregation. *arXiv preprint arXiv:1506.05446*, 2015.

[SSC18] M. Sesia, C. Sabatti, and E. J. Candès. Gene hunting with hidden Markov model knockoffs. *Biometrika*, 106(1):1–18, 08 2018.

[Sto02] J. D. Storey. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498, August 2002.

[Tib96] R. Tibshirani. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58, 1996.

[TT15] J. Taylor and R. J. Tibshirani. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25):7629–7634, June 2015.

[TVZ+18] W. Tansey, V. Veitch, H. Zhang, R. Rabadan, and D. M. Blei. The Holdout Randomization Test: Principled and Easy Black Box Feature Selection. *arXiv:1811.00645 [stat]*, November 2018. arXiv: 1811.00645.

[vdGBRD14] S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, June 2014. arXiv: 1303.0518.

[VGT12] G. Varoquaux, A. Gramfort, and B. Thirion. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. *Proceedings of the 29th International Conference on Machine Learning, Edinburgh, Scotland, UK, 2012*, page 8, 2012.

[VSP+18] G. Varoquaux, Y. Schwartz, R. A. Poldrack, B. Gauthier, D. Bzdok, J.-B. Poline, and B. Thirion. Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565, 2018.

[VSWZ18] S. V. Vasaikar, P. Straub, J. Wang, and B. Zhang. LinkedOmics: analyzing multi-omics data within and across 32 cancer types. *Nucleic Acids Research*, 46(D1):D956–D963, January 2018.

[WCM+13] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature genetics*, 45(10):1113–1120, 2013.

[WJ20] W. Wang and L. Janson. A Power Analysis of the Conditional Randomization Test and Knockoffs. *arXiv:2010.02304 [math, stat]*, October 2020. arXiv: 2010.02304.

[WMO+15] S. Weichwald, T. Meyer, O. Ozdenizci, B. Scholkopf, T. Ball, and M. Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.

[WP20] S. Weichwald and J. Peters. Causality in cognitive neuroscience: concepts, challenges, and distributional robustness. *arXiv:2002.06060 [q-bio, stat]*, July 2020. arXiv: 2002.06060.

[WR09] L. Wasserman and K. Roeder. High-dimensional variable selection. *The Annals of Statistics*, 37(5A):2178–2201, October 2009.

[WSB+20] A. Weinstein, W. J. Su, M. Bogdan, R. F. Barber, and E. J. Candès. A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic. *arXiv:2007.15346 [math, stat]*, July 2020. arXiv: 2007.15346.

[WYBS20] H. Wang, Y. Yang, Z. Bu, and W. Su. The Complete Lasso Tradeoff Diagram. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20051–20060. Curran Associates, Inc., 2020.

[YUFT18] M. Yamada, Y. Umezu, K. Fukumizu, and I. Takeuchi. Post selection inference with kernels. In *International Conference on Artificial Intelligence and Statistics*, pages 152–160. PMLR, 2018.

[YYMD21] S. Yadlowsky, T. Yun, C. McLean, and A. D'Amour. SLOE: A Faster Method for Statistical Inference in High-Dimensional Logistic Regression. *arXiv:2103.12725 [cs, math, stat]*, May 2021. arXiv: 2103.12725.

[ZSC20] Q. Zhao, P. Sur, and E. J. Candès. The Asymptotic Distribution of the MLE in High-dimensional Logistic Models: Arbitrary Covariance. *arXiv:2001.09351 [math, stat]*, January 2020. arXiv: 2001.09351.

[ZZ14] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssb.12026.