Reliable statistical inference : controlling the false discovery proportion in high-dimensional multivariate estimators

Alexandre Blain

université PARIS-SACLAY

Reliable statistical inference : controlling the false discovery proportion of high-dimensional estimators

L'inférence statistique fiable : contrôle de la proportion de fausses découvertes pour des estimateurs en grande dimension

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°580 Sciences et Technologies de l'Information et de la Communication (STIC)
Spécialité de doctorat : Informatique Mathématique
Graduate School : Informatique et sciences du numérique
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **Inria Saclay-Île-de-France** (Université Paris-Saclay, Inria), sous la direction de **Bertrand Thirion**, Directeur de Recherche, et la co-direction de **Pierre Neuvial**, Directeur de Recherche.

Thèse soutenue à Paris-Saclay, le 9 Décembre 2024, par

Alexandre Blain

Composition du jury

Membres du jury avec voix délibérative

Gilles Blanchard Professeur, Université Paris-Saclay	Président
Livio Finos Professor, University of Padova	Rapporteur & Examinateur
Matthieu Lerasle Professeur, École Polytechnique	Rapporteur & Examinateur
Thomas Nichols Professor, University of Oxford	Examinateur
Chloé-Agathe Azencott Professeure, École des Mines de Paris	Examinatrice

THESE DE DOCTORAT
NNT : 2024UPASG083

université PARIS-SACLAY

ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Titre : L'inférence statistique fiable : contrôle de la proportion de fausses découvertes pour des estimateurs en grande dimension

Mots clés : Inférence statistique, tests multiples, contrôle du taux de Fausses Découvertes, Knockoffs, IRMf, imagerie cérébrale

Résumé : La sélection de variables sous contrôle statistique est un problème fondamental rencontré dans divers domaines où les praticiens doivent évaluer l'importance des variables d'entrée par rapport à un résultat d'intérêt. Dans ce contexte, le contrôle statistique vise à limiter la proportion de fausses découvertes, c'est-à-dire la proportion de variables sélectionnées qui sont indépendantes du résultat d'intérêt. Dans cette thèse, nous développons des méthodes visant à assurer un contrôle statistique dans des contextes de grande dimension tout en conservant la puissance statistique. Nous présentons quatre contributions clés dans ce domaine de recherche. Premièrement, nous introduisons Notip, une méthode non paramétrique qui permet aux utilisateurs d'obtenir des garanties sur la proportion de vraies découvertes dans n'importe quelle région cérébrale. Cette procédure améliore la sensibilité de détection par rapport aux méthodes existantes tout en conservant le contrôle des fausses découvertes. Deuxièmement, nous étendons le cadre Knockoff en proposant KOPI, une méthode qui fournit un contrôle de la proportion de fausses découvertes (FDP) en probabilité plutôt qu'en espérance. KOPI est naturellement compatible avec l'agrégation de plusieurs tirages Knockoff, ce qui permet de prendre en compte la variabilité de l'inférence Knockoff traditionnelle. Troisièmement, nous développons un outil de diagnostic pour identifier les violations de l'hypothèse d'échangeabilité dans Knockoffs, accompagné d'une nouvelle méthode non paramétrique de génération de Knockoffs qui restaure le contrôle des fausses découvertes. Enfin, nous introduisons CoJER pour améliorer la prédiction conforme en fournissant un contrôle précis de la proportion de couverture fausse (FCP) lorsque plusieurs points de test sont pris en compte, garantissant des estimations d'incertitude plus fiables. CoJER peut également être utilisé pour agréger les intervalles de confiance fournis par différents modèles prédictifs, atténuant ainsi l'impact des choix de modélisation. Ensemble, ces contributions renforcent la fiabilité de l'inférence statistique dans des contextes de grande dimension tels que les données de neuroimagerie et de génomique.

Title : Reliable statistical inference : controlling the false discovery proportion of high-dimensional estimators

Keywords : Statistical inference, multiple testing, False Discovery Proportion control, Knockoffs, fMRI, brain imaging

Abstract : Statistically controlled variable selection is a fundamental problem encountered in diverse fields where practitioners have to assess the importance of input variables with regards to an outcome of interest. In this context, statistical control aims at limiting the proportion of false discoveries, meaning the proportion of selected variables that are independent of the outcome of interest. In this thesis, we develop methods that aim at statistical control in high-dimensional settings while retaining statistical power. We present four key contributions in this avenue of work. First, we introduce Notip, a non-parametric method that allows users to obtain guarantees on the proportion of true discoveries in any brain region. This procedure improves detection sensitivity over existing methods while retaining false discoveries control. Second, we extend the Knockoff framework by proposing KOPI, a method that provides False Discovery Proportion (FDP) control in probability rather than in expectancy. KOPI is naturally compatible with aggregation of multiple Knockoffs draws, addressing the randomness of traditional Knockoff inference. Third, we develop a diagnostic tool to identify violations of the exchangeability assumption in Knockoffs, accompanied by a novel non-parametric Knockoff generation method that restores false discoveries control. Finally, we introduce CoJER to enhance conformal prediction by providing sharp control of the False Coverage Proportion (FCP) when multiple test points are considered, ensuring more reliable uncertainty estimates. CoJER can also be used to aggregate the confidence intervals provided by different predictive models, thus mitigating the impact of modeling choices. Together, these contributions advance the reliability of statistical inference in high-dimensional settings such as neuroimaging and genomic data.

Acknowledgments

To the members of the jury.

I am extremely grateful to have you take part in this important milestone. Mathieu, Livio, I hope you enjoyed reading and reviewing my manuscript – I sincerely thank both of you for your kind words and comments, and also for submitting them on time! This made the manuscript submission process really smooth and I thank you dearly for that. Tom, Chloé, I admire both of you as scientists, especially for your contributions in statistics applied to neuroscience and medicine, it is truly an honor to have you in this jury. Gilles, I wouldn't have wished to have any other researcher as president of this jury. Much of this work is based on your ideas – your kindness and ability to propose ideas have always impressed me in our interactions.

A mes encadrants.
Pierre, j'ai une immense dette envers toi qui m'as donné ma première chance dans la recherche il y a maintenant plus de 5 ans – je n'ose pas imaginer à quel point ma vie serait différente si je ne t'avais pas rencontré! Dès nos premières interactions, ton humanité et ta bienveillance m'ont frappé. J'ai tout de suite eu le sentiment qu'on pouvait aborder tous les sujets, y compris les difficultés personnelles, et cela a été un grand soulagement. Au cours de ces discussions, tu m'as aussi appris à relativiser les diverses échéances qui marquent une thèse. Sur le plan scientifique, c'est ta rigueur en toutes circonstances qui m'a le plus impressionné. Même la veille d'une deadline ou dans des conditions compliquées, tu arrives toujours à réfléchir et à écrire proprement les choses. J'admire ton honneteté intellectuelle et ton engagement pour la recherche. Je te remercie pour tous les moments passés ensemble, ça a été un véritable honneur de travailler avec quelqu'un comme toi.

Bertrand, merci pour ton encadrement exceptionnel et pour tout le temps que tu as consacré à ma thèse. Plus j'ai avancé dans ma thèse et plus j'ai pris conscience de la chance que j'avais de t'avoir comme directeur. Je pense sincèrement que tu es l'une des personnes qui m'a appris le plus de choses dans ma vie: bien écrire, coder proprement, faire de belles figures, communiquer efficacement... la liste est longue. Scientifiquement, ta capacité à suivre les idées instantanément et à proposer des directions de recherche m'a vraiment impressionné. Je n'ai jamais eu le sentiment d'être perdu dans ma thèse et je sais maintenant que c'est une chance énorme. Humainement, si je devais ne retenir qu'une chose, c'est ton exemplarité. Le niveau d'exigence et de droiture que tu incarnes m'a énormément inspiré tout au long de ma thèse. On peut compter sur toi à toute heure du jour et de la nuit et dans n'importe quelle situation. Tu es un véritable modèle pour moi, et j'essaierai de me souvenir de tes enseignements durant toute ma vie.

A l'équipe Parietal et à mes amis. Merci pour ces trois années géniales et pour

votre bienveillance. Je crois sincérement que la vie de notre équipe est unique – j'estime avoir eu une chance immense de tomber sur cette équipe pour y faire ma thèse. Alex, merci d'avoir été comme un frère pour moi pendant ces trois ans, j'ai apprécié nos discussions scientifiques mais aussi toutes les autres. Julia et Théo, merci pour votre gentillesse et votre énergie de tous les jours. Florent et Thomas C., merci pour tous les moments passés à l'escalade et ailleurs. Léo, merci pour ton humour et pour toutes tes idées. Alexis, merci d'avoir été l'un de mes mentors pendant cette thèse: nos discussions sur Camus resteront un magnifique souvenir. Merci à tous les membres de notre beau groupe pour toutes nos discussions et activités: Apolline, Jovan, Ambroise, Félix, Samuel, Julie, Marie, Pierre-Louis, Sébastien, Jad, Célestin, Virginie, Jade, Jun, Lihu, Antoine, Guillaume, Mathieu D., Matthieu T., Maria, Gabriela, Fernanda, Himanshu, Pierre-Antoine, Louis, Cédric A., Cédric R., Melvine, Malo, Benoit, Raphael, Binh, Hugo et Ahmad. Je pense aussi aux PIs de l'équipe: Thomas, Marine, Gael, Alex G., Judith et Philippe. J'ai aussi une pensée particulière pour Olivier Grisel : ça a été un honneur d'apprendre auprès de toi.

Alex M., merci pour ta présence depuis la sixième jusqu'au jour de ma soutenance. Pierre et Ferdinand, merci pour tout, vous êtes comme des frères pour moi depuis le premier jour de sup. Fred, merci pour nos aventures ensemble en école, j'espère que tu t'éclates toujours autant à Wall Street! Julien, Roméo, Alex C. : merci pour notre conversation et pour nos débats sur le foot. Oscar, merci pour ton aide sur tous les projets en dernière année d'école et nos innombrables parties de baby.

A mes amis du Nord, Théo, Jérôme, Alexis, Benoit et Anthony: merci de m'avoir changé les idées pendant toute cette thèse. J'ai toujours espoir que le club pro FIFA reprenne avant 2050.

A mes professeurs et médecins. Merci pour votre soutien tout au long de ma vie. Je pense à Mme Azema, Mme Courouble et Mme Boutaleb au collège et au lycée. En classe préparatoire: M. Walbron, M. Mathurin et Mme Richard. M. Desmeules – vous êtes l'une des plus belles rencontres de ma vie, votre droiture et votre compétence m'inspire jusqu'à aujourd'hui. Je n'oublierai jamais ce que vous avez fait pour moi en 5/2.

Dr. Ploussard : je pense encore à votre exemple très souvent. Ce que vous représentez pour ma famille et moi est immense. Dr Dumont, votre intelligence et votre empathie m'impressionne. Avoir choisi de les mettre au service de ceux dont les autres détournent le regard vous honore. Dr. Puybaret, je ne vous oublierai jamais, reposez en paix.

A ma famille.
A ma soeur Inès, merci pour toutes nos aventures depuis le premier jour. Ma scolarité a commencé par la déception de te savoir déjà en CP et donc hors de la maternelle! Quel destin et quelle fierté que de finalement terminer nos thèses à deux ou trois mois d'intervalle. Ton travail aux urgences m'impressionne de jour en jour. J'espère que tu continuera toujours à admirer la beauté des fleurs en sortant de garde.

A mon père, qui m'a donné le gout des sciences et de l'informatique. J'admire ton honnêteté et ta curiosité intellectuelle et je crois sincérement que tu aurais fait un excellent chercheur. Au fil des années, je prends de plus en plus conscience de tout ce que tu as assumé sur tes épaules pour nous protéger et nous offrir une enfance paisible. Merci d'avoir toujours été là pour nous, même quand tu menais dans l'ombre des combats dont nous

n'avions pas idée.

A ma mère, la personne la plus aimante, droite et résiliente que je connaisse. Ton histoire et ton exemple forcent le respect. Tu as laissé ta vie de côté et t'es battue pour notre famille depuis nos premiers pas. Toi à qui on a si souvent voulu enlever le mérite de nos parcours, je l'écris ici pour l'éternité : de la première ligne de cahier d'écolier à la dernière ligne de cette thèse, j'ai écrit chaque mot de ma scolarité en ton honneur.

Abstract

Statistically controlled variable selection is a fundamental problem encountered in diverse fields where practitioners have to assess the importance of input variables with regards to an outcome of interest. In this context, statistical control aims at limiting the proportion of false discoveries, meaning the proportion of selected variables that are independent of the outcome of interest. In this thesis, we develop methods that aim at statistical control in high-dimensional settings while retaining statistical power. We present four key contributions in this avenue of work. First, we introduce Notip, a non-parametric method that allows users to obtain guarantees on the proportion of true discoveries in any brain region. This procedure improves detection sensitivity over existing methods while retaining false discoveries control. Second, we extend the Knockoff framework by proposing KOPI, a method that provides False Discovery Proportion (FDP) control in probability rather than in expectancy. KOPI is naturally compatible with aggregation of multiple Knockoffs draws, addressing the randomness of traditional Knockoff inference. Third, we develop a diagnostic tool to identify violations of the exchangeability assumption in Knockoffs, accompanied by a novel non-parametric Knockoff generation method that restores false discoveries control. Finally, we introduce CoJER to enhance conformal prediction by providing sharp control of the False Coverage Proportion (FCP) when multiple test points are considered, ensuring more reliable uncertainty estimates. CoJER can also be used to aggregate the confidence intervals provided by different predictive models, thus mitigating the impact of modeling choices. Together, these contributions advance the reliability of statistical inference in high-dimensional settings such as neuroimaging and genomic data.

Abstract en Français

La sélection de variables sous contrôle statistique est un problème fondamental rencontré dans divers domaines où les praticiens doivent évaluer l'importance des variables d'entrée par rapport à un résultat d'intérêt. Dans ce contexte, le contrôle statistique vise à limiter la proportion de fausses découvertes, c'est-à-dire la proportion de variables sélectionnées qui sont indépendantes du résultat d'intérêt. Dans cette thèse, nous développons des méthodes visant à assurer un contrôle statistique dans des contextes de grande dimension tout en conservant la puissance statistique. Nous présentons quatre contributions clés dans ce domaine de recherche. Premièrement, nous introduisons Notip, une méthode non paramétrique qui permet aux utilisateurs d'obtenir des garanties sur la proportion de vraies découvertes dans n'importe quelle région cérébrale. Cette procédure améliore la sensibilité de détection par rapport aux méthodes existantes tout en conservant le contrôle des fausses découvertes. Deuxièmement, nous étendons le cadre Knockoff en proposant KOPI, une méthode qui fournit un contrôle de la proportion de fausses découvertes (FDP) en probabilité plutôt qu'en espérance. KOPI est naturellement compatible avec l'agrégation de plusieurs tirages Knockoff, ce qui permet de prendre en compte la variabilité de l'inférence Knockoff traditionnelle. Troisièmement, nous développons un outil de diagnostic pour identifier les violations de l'hypothèse d'échangeabilité dans Knockoffs, accompagné d'une nouvelle méthode non paramétrique de génération de Knockoffs qui restaure le contrôle des fausses découvertes. Enfin, nous introduisons CoJER pour améliorer la prédiction conforme en fournissant un contrôle précis de la proportion de couverture fausse (FCP) lorsque plusieurs points de test sont pris en compte, garantissant des estimations d'incertitude plus fiables. CoJER peut également être utilisé pour agréger les intervalles de confiance fournis par différents modèles prédictifs, atténuant ainsi l'impact des choix de modélisation. Ensemble, ces contributions renforcent la fiabilité de l'inférence statistique dans des contextes de grande dimension tels que les données de neuroimagerie et de génomique.

Contents

Chapter 1

Overview

This PhD thesis focuses on false discoveries control in high-dimensional inference problems. This thesis is divided into two parts: in the first part, we introduce core concepts of statistical testing, conformal prediction and neuroimaging data analysis. In the second part, we present novel methods developed during this thesis. In the background part, we first introduce the notion of statistical testing in Chapter 2 and discuss the difficult problem of performing multiple tests at the same time. Classical multiple testing error rates are introduced, such as the Family-wise Error Rate (FWER) and the False Discovery Rate along with procedures that control these rates. We then move to Chapter 3; we dive into recently developed tools of the multiple testing literature that we use throughout this thesis. Namely, the Joint Error Rate, a general framework to control the False Discovery Proportion and the Knockoffs procedure, a novel idea to control the False Discovery Rate. In Chapter 4 we introduce conformal prediction, a popular framework for uncertainty quantification in prediction problems. Using conformal p-values, we relate this framework to the statistical testing literature. We then discuss the basics of neuroimaging data analysis from acquisition and preprocessing to statistical analysis in Chapter 5. FMRI data analysis is a central motivation in this thesis; most of the experimental work of this thesis' contributions is done on fMRI datasets.

Contributions of this thesis are organized around four papers, detailed in the following sections:

- Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492
- Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*
- Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024a). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*
- Blain, A., Thirion, B., and Neuvial, P. (2024b). Tight and reliable conformal prediction. *Under review*

1.1 Notip: Non-parametric True Discovery Proportion control for brain imaging

Cluster-level inference procedures are widely used for brain mapping. These methods compare the size of clusters obtained by thresholding brain maps to an upper bound under the global null hypothesis, computed using Random Field Theory or permutations. However, the guarantees obtained by this type of inference - i.e. at least one voxel is truly activated in the cluster - are not informative with regards to the extent of the signal therein. There is thus a need for methods to assess the amount of signal within clusters; yet such methods have to take into account that clusters are defined based on the data, which creates circularity in the inference scheme. This has motivated the use of *post hoc* estimates that allow statistically valid estimation of the proportion of activated voxels in clusters. In the context of fMRI data, the All-Resolutions Inference framework introduced in Rosenblatt et al., 2018 provides post hoc estimates of the proportion of activated voxels. However, this method relies on parametric threshold families, which results in conservative inference. In Chapter 6, we propose to adapt to data characteristics and obtain tighter false discovery control. For this we leverage randomization methods. We obtain *Notip*, for Non-parametric True Discovery Proportion control: a powerful, non-parametric method that yields statistical guarantees on the proportion of activated voxels in data-derived clusters. Numerical experiments demonstrate substantial gains in number of detections compared with state-of-the-art methods on 36 fMRI datasets. The conditions under which the proposed method brings benefits are also discussed.

Published work. Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492

1.2 False Discovery Proportion control for aggregated Knockoffs

Controlled variable selection is an important analytical step in various scientific fields, such as brain imaging or genomics. In these high-dimensional data settings, considering too many variables leads to poor models and high costs, hence the need for statistical guarantees on false positives. Knockoffs are a popular statistical tool for conditional variable selection in high dimension. However, they control for the expected proportion of false discoveries (FDR) and not their actual proportion (FDP). In Chapter 7 we present a new method, KOPI, that controls the proportion of false discoveries for Knockoff-based inference. The proposed method also relies on a new type of aggregation to address the undesirable randomness associated with classical Knockoff inference. We demonstrate FDP control and substantial power gains over existing Knockoff-based methods in various simulation settings and achieve good sensitivity/specificity tradeoffs on brain imaging and genomic data.

Published work. Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*

1.3 When Knockoffs fail: diagnosing and fixing non exchangeability of Knockoffs

Knockoffs are a popular statistical framework that addresses the challenging problem of conditional variable selection in high-dimensional settings with statistical control. Such statistical control is essential for the reliability of inference. However, knockoff guarantees rely on an exchangeability assumption that is difficult to test in practice, and there is little discussion in the literature on how to deal with unfulfilled hypotheses. This assumption is related to the ability to generate data similar to the observed data. To maintain reliable inference, we introduce a diagnostic tool based on Classifier Two-Sample Tests in Chapter 8. Using simulations and real data, we show that violations of this assumption occur in common settings for classical Knockoffs generators, especially when the data have a strong dependence structure. We show that the diagnostic tool correctly detects such behavior. To fix knockoff generation, we propose a nonparametric, computationally-efficient alternative knockoff construction, which is based on constructing a predictor of each variable based on all others. We show empirically that the proposed approach restores error control on simulated data.

Preprint. Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024a). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*

1.4 Tight and reliable conformal prediction

Split conformal prediction (SCP) offers a computationally efficient way to build confidence intervals in regression problems. Notably, most of the theory built around SCP is focused on the single test point problem. In real-life settings, inference sets consist of multiple points, which poses the question of coverage guarantees for many points simultaneously. While *on average*, the False Coverage Proportion (FCP) remains controlled, but it can fluctuate strongly around its mean. We show that when splitting a dataset multiple times, classical SCP may not control the FCP for up to 65% of splits. In Chapter 9 we propose CoJER, a novel method that achieves sharp FCP control in probability for conformal prediction, relying on knowledge of the distribution of conformal p-values under exchangeability. We show on extensive real data experiments that CoJER provides the announced coverage while standard SCP does not. Additionally, CoJER yields shorter interval than the *state-of-the-art* method and only slightly larger intervals than standard SCP.

Under review. Blain, A., Thirion, B., and Neuvial, P. (2024b). Tight and reliable conformal prediction. *Under review*

Part I

Background

Chapter 2

Statistical hypothesis testing

Summary. In this chapter, we review some fundamental concepts and methods of statistical hypothesis testing, which is the focus of this thesis. The aim of statistical hypothesis testing is to determine whether a certain hypothesis is sufficiently supported by the data at hand. Performing reliable statistical hypothesis testing is essential for drawing meaningful inferences from data, guiding decision-making, and building predictive models. For clarity, we start by presenting the elementary case where a single hypothesis is tested, before moving on to multiple testing procedures. In the context of our work, many hypotheses are tested simultaneously and performing reliable inference requires taking multiplicity into account.

Contents

2.1 Testing association between variables and outcome

The concept of statistical hypothesis testing was introduced in the early 20^{th} century in the seminal work of Pearson, 1900; Fisher, 1922. Statistical hypothesis testing aims at constructing a reliable inference procedure to decide whether the data at hand supports a given hypothesis. Denote H_0 the null hypothesis and H_1 the alternative hypothesis and the data at hand $X = (X_1, \ldots, X_n)$. Intuitively, the null hypothesis serves as the default position (e.g. no effect in a study) while the alternative hypothesis proposes that there is an effect or a difference, contradicting the null hypothesis. Then, a statistical hypothesis test can be defined as follows:

Definition 1 (Statistical hypothesis test)**.** A statistical hypothesis test is a decision rule that specifies whether or not to reject the null hypothesis H_0 in favor of the alternative hypothesis H_1.

This definition entails four possible scenarios, in which the null hypothesis is rejected (or not) rightfully (or not). Possible scenarios are summarized in the table below:

	H_0 true	H_0 false
H_0 not rejected	☺	False Negative
H_0 rejected	False Positive	☺

The probability of issuing a False Positive (type-I error) is called the **significance level** of the test and generally denoted α. In words, this is the probability of falsely rejecting the null hypothesis. On the other hand, the probability of obtaining a true positive is called the **power** of the test and is generally denoted $1 - \beta$. This is the probability of rightfully rejecting the null hypothesis. There exists a trade-off between these two quantities: in general, tests are calibrated such that they maximize power for a certain significance level (Neyman and Pearson, 1933).

The decision rule that constitutes a hypothesis test is generally based on a quantity derived of the data $T(X_1, \ldots, X_n) \in \mathbb{R}$. This quantity is called the **test statistic**. A notable type of test statistics are p-values, defined as follows:

Definition 2 (p-values)**.** A p-value is a test statistic $p(X)$ that satisfies:

1. $p(X) \in [0, 1]$.
2. (Sub-uniformity) If H_0 is true, for all $t \in [0, 1]$:

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

Statistical hypothesis tests are oftentimes formulated as p-values thresholding: the lower the p-value, the stronger the evidence against the null hypothesis H_0. The second property of this definition ensures that the p-value is *valid*: for any threshold t, sub-uniformity guarantees that the probability of making a false positive is at most t.

While p-values are ubiquitous in many scientific fields, they suffer from numerous misuses (Halsey et al., 2015; Sullivan and Feinn, 2012; Wasserstein and Lazar, 2016; Greenland et al., 2016). Misuses include circularity biases – running statistical testing procedures on data subsets which have been selected *after having seen the data* – and data snooping – running many statistical tests and reporting only those who exhibit a statistically significant result while concealing the others. A central motivation of this thesis is to provide valid inference procedures based on p-values.

2.2 Multiple hypothesis testing

In the era of modern machine learning and data science, it is increasingly common to test not just a single hypothesis, but rather many hypotheses simultaneously. This shift is driven by the vast amounts of data now available, allowing researchers to explore a multitude of relationships, patterns, and effects within a single study. However, this abundance of data also introduces substantial challenges. When many hypotheses are tested at once, the probability of obtaining statistically significant results purely by chance increases, potentially leading to false discoveries. This inflation can result in misleading conclusions, where the observed effects are not truly significant but rather artifacts of the testing process (Bender and Lange, 2001; Noble, 2009).

Example 1. An fMRI image consists of $100,000$ (or more) voxels, each of which represents a small volume of brain tissue. Researchers typically test the level of activity at each voxel to determine if it is significantly different from a baseline, to detect brain activity related to a specific task or stimulus. If each of these $100,000$ voxels is tested individually with a significance threshold of 5% (a common choice in hypothesis testing), then even if there is

no true signal—meaning no actual brain activity associated with the task or stimulus— 5% of these tests will be deemed significant purely by chance. In this scenario, that amounts to 5,000 voxels falsely identified as showing significant activity, despite nothing actually happening. Accounting for multiplicity is essential in all fMRI analyses (Bennett et al., 2009; Nichols, 2012).

The most straightforward approach to recover false discoveries control is to correct the significance level used for each test. Denoting p the number of hypotheses tested simultaneously, we use a significance level of α/p for each test. This amounts to the Bonferroni correction (Bonferroni, 1936):

Definition 3 (Bonferroni correction, Bonferroni, 1936)**.** Denote $(p_j)_{j\in[\![p]\!]}$ the p-values associated to the p tests. The Bonferroni procedure rejects the j^{th} null hypothesis $H_{0,j}$ when:

$$p_j \leq \frac{\alpha}{p}$$

The Bonferroni correction ensures that the probability of making at least **one false positive** is smaller than the original significance level α. This error rate is called the **Family-wise Error Rate (FWER)**. Note that controlling the FWER is much more stringent than controlling the type-I error for each hypothesis – here, the probability of making a false discovery is controlled *for all tests*. FWER control is satisfactory in applications where the cost of making *any* false discovery is intolerable. However, this strict control comes at the cost of statistical power since the significance threshold of the procedure shrinks as p grows large. Depending on the type of p-value used, making a discovery may become analytically impossible for small values of α.

To achieve better trade-offs between statistical power and type-I error control in multiple testing, Benjamini and Hochberg, 1995 proposed an alternative procedure which controls the *expected proportion* of false discoveries. First, let us define the False Discovery Proportion (FDP) and its expected value, the False Discovery Rate (FDR):

Definition 4 (False Discovery Proportion and False Discovery Rate)**.** Denote the set of true null hypotheses $\mathcal{H}_0$. For any rejection set $\hat{S} \subset [\![p]\!]$:

$$\mathrm{FDP}(\hat{S}) = \frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}, \quad \mathrm{FDR}(\hat{S}) = \mathbb{E}[\mathrm{FDP}(\hat{S})] = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\right].$$

Definition 5 (Benjamini-Hochberg procedure)**.** Denote $p_{(1)}, \ldots, p_{(p)}$ the sorted p-values. The Benjamini-Hochberg procedure consists in:

1. Finding $j_{BH} \in \max\left\{j \in [\![p]\!] \mid p_{(j)} \leq \frac{\alpha j}{p}\right\}$
2. Rejecting $H_{0,(1)}, \ldots, H_{0,(j_{BH})}$.

Property 1 (Positive Regression Dependency on Subset; PRDS, Benjamini and Yekutieli, 2001)**.** Denote an increasing set D, i.e. a set such that if $x \in D$ and $y \geq x$ then $y \in D$. Then, the data $\mathbf{X}$ is said to be PRDS – implicitly, on the set of true nulls – if $\forall j$ s.t. $H_{0,j}$ is true:

$$\mathbb{P}(\mathbf{X} \in D | X_j = x) \text{ is nondecreasing in } x.$$

The PRDS property notably holds in the case of positively correlated variables as shown in Benjamini and Yekutieli, 2001. Provided that the PRDS property is verified, the Benjamini-Hochberg (BH) procedure controls the FDR. This guarantee is much less stringent than FWER control, as $\alpha\%$ of false discoveries are tolerated *on average*. The BH procedure is extremely popular in many scientific fields. However, it suffers from important conceptual limitations, in addition to necessitating strong hypotheses on the data correlation structure:

FDR control is not FDP control (Korn et al., 2004; Efron, 2012; Roquain, 2015; Neuvial, 2020). The FDR is the *expected* proportion of false discoveries, rendering its control difficult to interpret. In practice, users generally have access to a single dataset on which they employ the BH procedure. For a single run of this procedure, the actual FDP may not be close to its mean (FDR). This is especially true in high correlation settings, where the FDR can become a poor representation of the underlying FDP distribution. As shown in Figure 2.1, for large values of equicorrelation ρ, either no hypotheses are rejected, or the FDP is much higher than the FDR level, with 10% of runs returning an FDP of 80% or more.

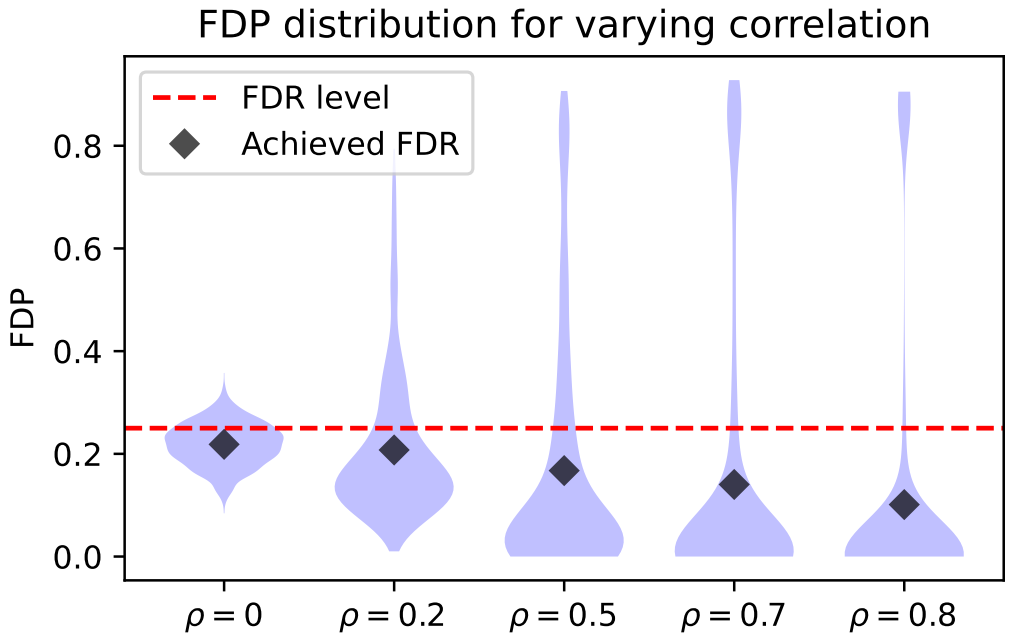


Figure 2.1: **FDP distribution for 1000 simulation runs and five correlation values.** We simulate $n = 500$ equicorrelated Gaussian samples of dimension $p = 1000$. We use 100 active variables for which some signal is added for half of the samples. Then, we compute p-values with a Welch test and perform inference using the BH procedure. Each violin plot represents the distribution of the FDP for a given correlation value ρ. Notice that the FDR is always controlled, but is an increasingly poor representation of the FDP distribution as ρ grows. For high values of ρ, either no hypotheses are rejected and FDP $= 0$, or else the FDP is much higher than the FDR level, with 10% of runs returning an FDP of 80% or more.

There has been much effort in the statistical community to achieve FDP control in probability (Genovese and Wasserman, 2002, 2004; Meinshausen, 2006; Fan et al., 2012; Goeman and Solari, 2011; Blanchard et al., 2020). While technically more challenging to

obtain, such control gives a guarantee that is directly interpretable.

Post hoc FDP control (Goeman and Solari, 2011; Blanchard et al., 2020; Katsevich and Ramdas, 2020). Another notable limitation of FDR controlling procedures is that they do not allow for *post hoc* inference: the procedure returns a set of hypotheses $\hat{S}$ for which the FDR is controlled, but cannot give any information on an arbitrary set S, potentially chosen by the user. FDR-controlling procedure can be misused in this sense, as users may apply it to subset of hypotheses chosen **after having seen the data**, rendering inference invalid (Kriegeskorte et al., 2009; Benjamini, 2020). As shown in Figure 2.2, performing inference on simulated null p-values using the BH procedure on a *data-dependent* subset leads to massive false positive inflation.

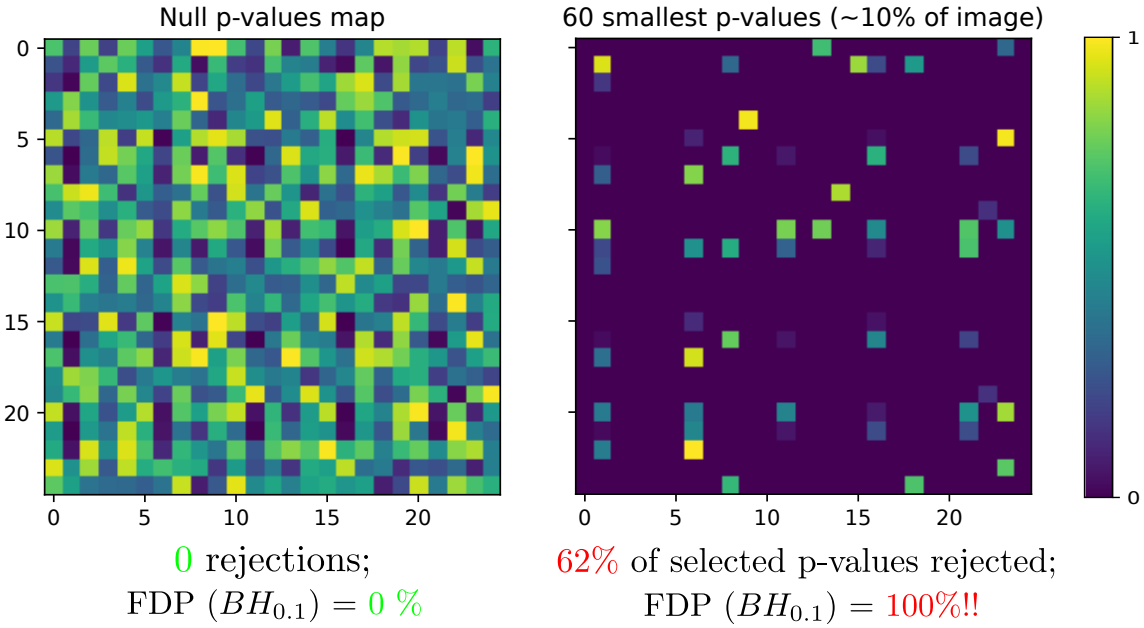


Figure 2.2: **Invalid selective inference using the BH procedure on simulated null p-values.** We simulate a 25×25 $2D$ map of null p-values drawn independently in $\mathcal{U}[0, 1]$. Performing inference using the BH procedure ($q = 0.1$) on the complete map yields no rejections and therefore a null FDP, which is the expected behavior. However, if inference is performed after having selected the 10% of smallest p-values, the BH procedure rejects 62% of the selected hypotheses with a 100% FDP. This illustrates that BH procedure does not support selective inference.

Chapter 3

Statistical control of False Discoveries

Summary. In this chapter, we present two core techniques used for the statistical control of False Discoveries and will be used in this thesis. First, the Joint Error Rate (JER) framework introduced in Blanchard et al., 2020 which offers a general approach to controlling the False Discovery Proportion (FDP) and supports post hoc inference. The JER framework relies on knowledge – or estimation – of the joint distribution of p-values under the null to build valid FDP upper bounds. Second, the Knockoffs method, introduced by Barber and Candès, 2015, constructs artificial variables that mirror the dependence structure of the original covariates, allowing for the control of the false discovery rate (FDR) in conditional independence testing. This approach is particularly effective in settings with highly correlated predictors.

Contents

3.1 The JER framework: a general approach to controlling the FDP

3.1.1 Post hoc FDP control

A post hoc upper bound V on the number of false positives is an integer-valued function of subsets S of hypotheses that satisfies:

$$\mathbb{P}\left(\forall S,\ |S \cap H_0| \leq V(S)\right) \geq 1 - \alpha\,. \tag{3.1}$$

Since $\mathrm{FDP}(S) = |S \cap H_0| \,/\, |S|$, obtaining a bound V satisfying (3.1) is strictly equivalent to obtaining a post hoc upper bound on the FDP. This equivalence will be used implicitly throughout this thesis.

As described in the seminal work of Goeman and Solari, 2011, the comparison between ordered p-values and $\left(t_k^{Simes}\right)_{k=1..p} = (\alpha k/p)_{k=1..p}$ can provide post hoc FDP control using *closed testing* (Marcus et al., 1976). Closed testing relies on having a valid α-level test for any intersection of hypotheses $\bigcap_{k\in S} H_k$ – these are called *local tests.* Once local tests have been performed, the procedure leverages set combinatorics to obtain FDP upper bounds. In practice, computing the required quantities can be impossible due to combinatorial complexity. To solve this issue Goeman and Solari, 2011 propose to use the Simes local test (Simes, 1986) which allows computational shortcuts, making the bound computable in linear time (Goeman et al., 2019).

The validity of the Simes test for independent data or positively correlated data relies on the Simes inequality (Simes, 1986):

$$\mathbb{P}\left(\exists k \in \{1,\dots,m_0\} : p_{(k:m_0)} < t_k^{Simes}\right) \leq \alpha\,. \tag{3.2}$$

The All-resolutions inference (ARI) method (Rosenblatt et al., 2018) provides a tighter post hoc bound that uses the thresholds $\alpha k/h(\alpha)$ instead of $t_k^{Simes} = \alpha k/m$ in (3.2), where $h(\alpha) \leq m$ is the so-called Hommel value (Hommel, 1986). $h(\alpha)$ represents an $1-\alpha$-level upper confidence bound on the number m_0 of true null hypotheses.

3.1.2 Joint Error Rate

While Simes based closed testing provides valid post hoc FDP upper bounds, it relies on an entirely parametric construction. Given that the Simes inequality is conservative for positively dependent p-values (Blanchard et al., 2020), and are thus suboptimal. An alternative construction of post hoc bounds has been introduced by Blanchard et al., 2020. Letting $R_k^{Simes} = \left\{i : p_i \leq t_k^{Simes}\right\}$, Equation (3.2) can be written as:

$$\mathbb{P}\left(\forall k, \left|R_k^{Simes} \cap H_0\right| \leq k-1\right) \geq 1-\alpha\,. \tag{3.3}$$

Equation (3.3) can be interpreted as the simultaneous control of all $k-$Family-Wise Error Rate (FWER), where the $k-$FWER is the probability of obtaining at least k false positives. Note that Equation (3.3) is exactly of the same form as Equation (3.1) but only valid for R_k and not all S. The bound for all S is obtained by interpolation, as each set R_k^{Simes} yields a valid FDP upper bound over any subset S as shown in Blanchard et al., 2020:

$$\begin{aligned}
|S \cap H_0| &= \left|S \cap \overline{R_k^{Simes}} \cap H_0\right| + \left|S \cap R_k^{Simes} \cap H_0\right| \\
&\leq \left|S \cap \overline{R_k^{Simes}}\right| + \left|R_k^{Simes} \cap H_0\right| \\
&= \sum_{i\in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + \left|R_k^{Simes} \cap H_0\right| \\
&\leq \sum_{i\in S} 1\left\{p_i(X) \geq t_k^{Simes}\right\} + k - 1 \\
&=: V_k^{Simes}(S)\,,
\end{aligned}$$

where the last inequality holds with probability at least $1-\alpha$ by (3.3).

The computation of $V_k^{Simes}(S)$ is illustrated in the top panels of Figure 3.1 for $k \in \{1, 3, 6\}$ on a toy example with 10 p-values. Since (3.3) holds simultaneously for all k, the minimum over k of all $V_k^{Simes}(S)$ is a valid upper bound on the false positives in S (Blanchard et al., 2020). Therefore, as illustrated in the bottom panel of Figure3.1, the final post hoc FDP upper bound is $V^{Simes}(S)/|S|$, where

$$V^{Simes}(S) = \min_{1 \le k \le |S|} \left\{ \sum_{i \in S} 1\left\{p_i(X) \ge t_k^{Simes}\right\} + k - 1 \right\}. \tag{3.4}$$

As noted by (Blanchard et al., 2020), the bound (3.4) coincides with the bound originally proposed by (Goeman and Solari, 2011). This can be generalized as follows by replacing $t^{Simes} := (t_k^{Simes})_{1 \le k \le m}$ with any threshold family $t := (t_k)_{1 \le k \le k_{max}}$ corresponding to $R_k = \{i : p_i \le t_k\}$.

The Joint Error Rate (JER) of the threshold family t is defined by (Blanchard et al., 2020) as:

$$JER(t) = \mathbb{P}\left(\exists k \in \{1, \dots, k_{max} \wedge m_0\} : p_{(k:m_0)} < t_k\right). \tag{3.5}$$

With this notation, both Equations 3.2 and 3.3 are equivalent to $\mathrm{JER}(t^{Simes}) \le \alpha$. By the interpolation argument outlined above, the bound

$$V^t(S) = \min_{1 \le k \le |S| \wedge k_{max}} \left\{ \sum_{i \in S} 1\left\{p_i(X) \ge t_k\right\} + k - 1 \right\} \tag{3.6}$$

provides a valid FDP upper bound for any threshold family t such that $\mathrm{JER}(t) \le \alpha$ (Blanchard et al., 2020). This bound can be calculated in $O(|S|)$ for a given set S using Algorithm 1 in Enjalbert-Courrech and Neuvial, 2022.

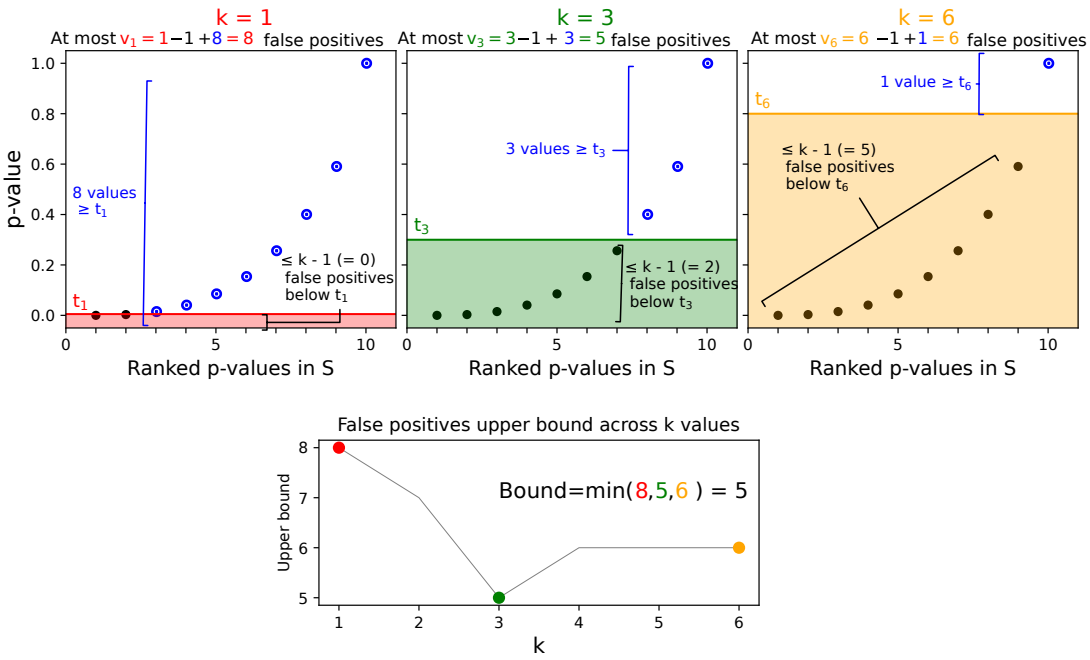


Figure 3.1: **Computation of the post hoc bound** (3.6) **on the number of false positives**, given a set S of 10 p-values and using a JER controlling threshold family. Top panels: computation of k-th bound $V_k(S) = \sum_{i \in S} 1\{p_i(X) \geq t_k\} + k - 1$ for 3 values of k, with horizontal colored lines representing the associated thresholds t_k. Bottom panel: The post hoc upper bound(3.6) corresponds to the minimum of all $V_k(S)$. In this example, the bound guarantees that the number of false positives in S is at most 5 with probability $> 90\%$.

3.1.3 Tighter FDP upper bounds via randomization

The Simes inequality (3.2) ensures JER control at level at most α for the threshold family $(\alpha k/m)_k$. While this control is sharp for independent p-values, it is overly conservative for positively dependent p-values (Blanchard et al., 2020), leading to conservative FDP bounds. The first degree of freedom that can be leveraged to obtain tighter bounds for a given α is to choose the least conservative threshold family among a pre-defined set of families. In the case of the Simes family, this is done by choosing the threshold family $(\lambda k/m)_k$ associated to the largest λ such that the following inequality (that is, JER control) holds:

$$\mathbb{P}\left(\exists k \in \{1, \dots, m_0\} : p_{(k:m_0)} < \frac{\lambda k}{m}\right) \leq \alpha\,. \tag{3.7}$$

In order to reach this goal more generally, we consider collections of threshold families called **templates** since their introduction in Blanchard et al., 2020. Formally, a template is set of functions $\lambda \mapsto (t_k(\lambda))_k$ such that any fixed value of λ corresponds to a threshold family. For example, the Simes template corresponds to the choice: $t_k(\lambda) = \lambda k/m$ for all $k = 1 \dots m$ and $\lambda > 0$.

The **calibration** procedure introduced in Blanchard et al., 2021, 2020 uses randomization (see Arlot et al., 2007) to obtain samples from the joint distribution of p-values under the null hypothesis. As the JER (3.5) is a function of this distribution, these so-called randomized p-values allow us to select the largest possible λ such that the JER is controlled.

Algorithm 1 describes how to compute such randomized p-values in the case of one-sample tests, using sign-flipping (Roche et al., 2007; Arlot et al., 2007). Randomized p-values can be obtained similarly for two-sample tests, using class label permutations instead of sign-flipping.

Algorithm 1: Computing randomized p-values using sign-flipping. For a number B of sign-flips, compute p-values using a one-sample t-test on the flipped data $X_{flipped}$.

Function `get_randomized_p_values`(X, B)**:**
  $n, p \leftarrow$ shape(X) `// n subjects, p features`
  $pval0 \leftarrow$ zeros(B, p) **for** $b \leftarrow 1$ **to** B **do**
    flip $\leftarrow$ diag(draw_random_vector$(\{-1, 1\}^n)$) `// matrix of shape (n, n)`
    $X_{flipped} \leftarrow$ flip $\cdot X$
    $pval0[b] \leftarrow$ one_sample_t_test$(X_{flipped}, 0)$ `// 0 = null hypothesis`
  end
  $pval0 \leftarrow$ sort_lines$(pval0)$ `// Sort each vector of randomized p-values`
  return $pval0$

Figure 3.2 illustrates the conservativeness of the parametric Simes template on real data and the benefit yielded by calibration using randomized p-values curves. Choosing $\lambda > \alpha$ in (3.4) leads to a less conservative bound. Note that, the more dependent the data, the more the parametric Simes bound is expected to be conservative, see e.g. Blanchard et al., 2020. Thus, calibration should be particularly useful for smooth data.

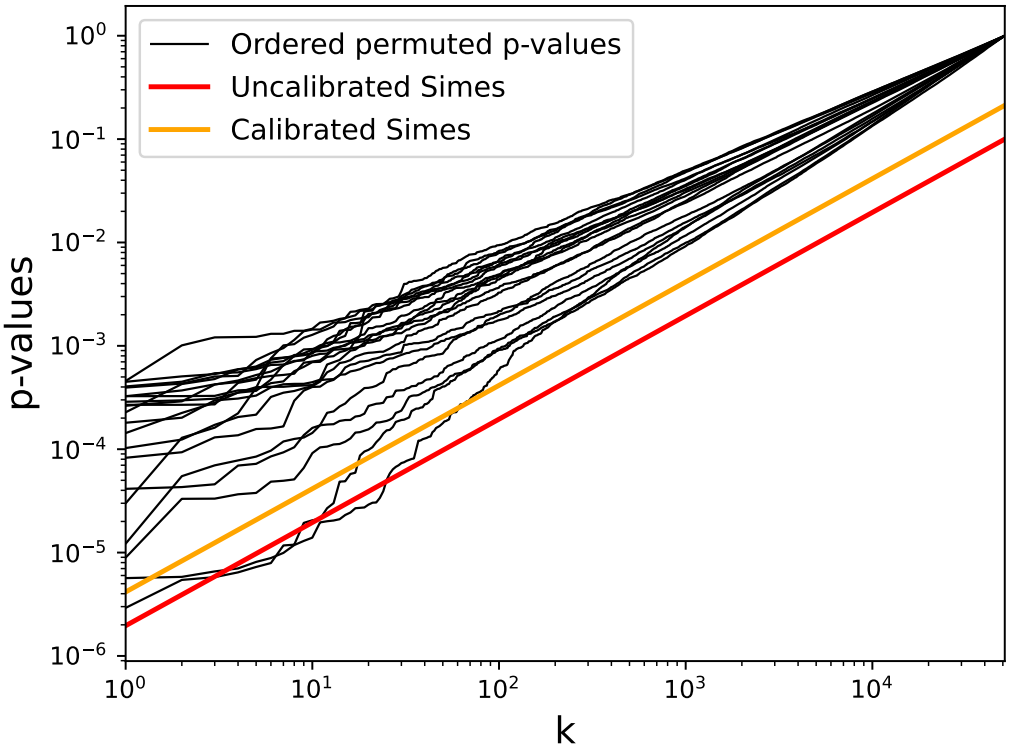


Figure 3.2: **Addressing the conservativeness of the Simes inequality by calibration**. A set of 20 randomized p-value curves are computed on real data (black curves). Two JER controlling families at level 10% are shown as colored lines. Both of them cross 2 curves (= 10% of all curves) which indeed corresponds to controlling the JER at level 10%. The uncalibrated Simes family (in red) is conservative since it is possible to choose higher threshold families that cross the same number of black curves. The calibrated Simes family (in orange) is the least possible conservative threshold family that crosses at most 2 curves.

While the ARI procedure corresponds to using Simes inequality without calibration[1] for JER control, calibration using the Simes template can be considered the state-of-the-art method for this problem (Blanchard et al., 2021, 2020). The bound obtained from this calibration procedure is equivalent to the bound considered in Andreella et al., 2020.

3.2 Knockoffs: FDR control in conditional variable selection

Traditional FDR control methods, like the Benjamini-Hochberg procedure, are effective for marginal testing using standard p-values, but do not support *conditional independence testing*. The intuition of testing for conditional independence – rather than for *marginal* independence – is to check if the relationship between a covariate and the outcome remains when controlling for the influence of all other covariates.

Example 2. Is X_1 (Cholesterol Level) independent of Y (Incidence of Cardiovascular Disease) given X_2 (Age Group)? Suppose that marginal testing results suggest that

[1] Rigorously, the ARI bound corresponds to using Simes inequality with the Hommel value h instead of m.

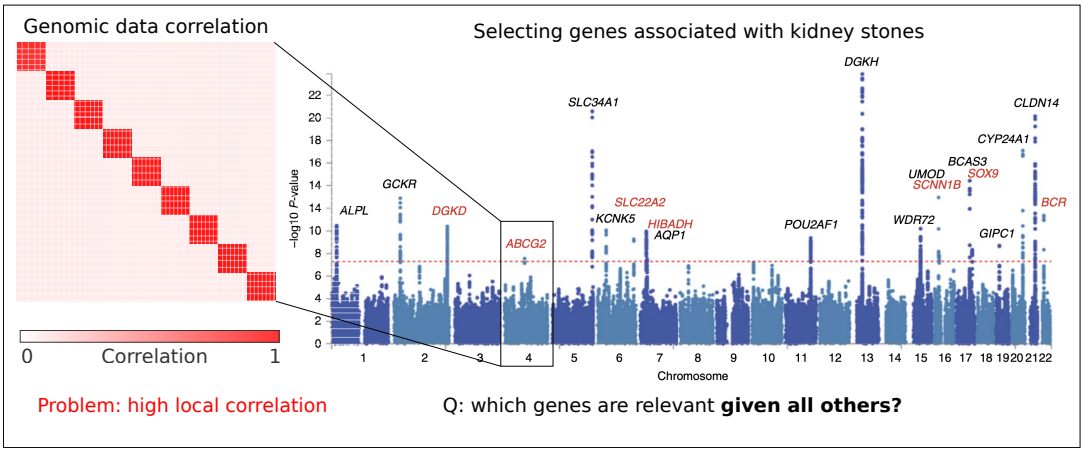


Figure 3.3: **Illustration of conditional variable selection.** Given a genomic dataset and kidney stones diagnoses (outcome), we aim at finding variables that add information w.r.t the outcome **given** all other variables. This problem is challenging since genes are highly correlated locally. Assessing the added information of specific genes given all others is crucial to improve the understanding of disease outcomes. Details about this Manhattan plot are available in Howles et al., 2019.

$X_1 \not\perp Y$, i.e. that cholesterol level and cardiovascular disease are linked. When performing conditional independence testing this relationship might change. For instance, among individuals under 50, the difference in cardiovascular disease rates between those with high and normal cholesterol might be negligible. However, among individuals aged 50 and over, the incidence of cardiovascular disease might be significantly higher for those with high cholesterol. This would suggest that the observed link between cholesterol levels and cardiovascular disease is strongly influenced by age, meaning that $X_1 \perp Y|X_2$. In this scenario, marginal independence testing suggests a direct link between high cholesterol levels and the risk of cardiovascular disease. However, conditional independence testing reveals that age plays a crucial role in this relationship, with older age groups being more susceptible.

Formally, we test simultaneously for all $j \in [\![p]\!]$:

$$H_{0,j} : y \perp x_j|\mathbf{x}_{-j} \quad \text{versus} \quad H_{1,j} : y \not\perp x_j|\mathbf{x}_{-j}.$$

In **conditional** variable selection, the objective is to assess the value of a particular variable w.r.t. an outcome **given a certain set of other variables** (König et al., 2021). This approach differs markedly from classical marginal inference, where variables are considered individually in relation to the outcome. With conditional variable selection, we examine whether a variable remains informative and relevant when considered alongside other variables. This is particularly crucial when dealing with datasets characterized by high levels of correlation among variables. Inference in the presence of correlations is illustrated with a genomics example of Genome-Wide Association Study (GWAS) in Figure 3.3, where correlation between variables is due to a biological phenomenon called linkage disequilibrium (Uffelmann et al., 2021).

This problem becomes difficult when there is a high level of correlation between variables on which one aims to perform inference. In machine learning contexts, the difficulty of identifying features that add unique predictive value when combined with others in large-correlation contexts has been well identified (Bickel et al., 2009; Mandozzi and Bühlmann,

2016; Goeman and Bühlmann, 2007; Javanmard and Montanari, 2014). High correlations are ubiquitous in many application areas, for example when the features of interest are derived from some biological measurements that inherently have complex dependencies. For instance, in brain mapping, practitioners may want to detect brain regions that are relevant for a given cognitive task given the rest of the brain (Weichwald et al., 2015), but have to deal with the strong dependencies observed in these data (Chevalier et al., 2021). In genomics, understanding which genes conditionally affect disease outcomes can lead to more effective disease prevention and intervention strategies (Sesia et al., 2019).

In this section, we describe the core elements of Knockoff inference as defined in Candès et al., 2018: valid Knockoff variables, statistics and inference. Then, we describe the classical Gaussian algorithm used to build Knockoffs and possible alternatives.

3.2.1 Knockoff inference

Knockoff inference, introduced in Barber and Candès, 2015; Candès et al., 2018, leverages noisy duplicates known as knockoff variables which serve as controls in the variable selection process. A key challenge in this method is to ensure that these knockoff variables maintain the same correlation structure as the original variables, while being conditionally independent of the outcome. This is essential to enable meaningful comparisons between the original variables and their knockoff counterparts, thereby identifying variables that provide relevant information regarding the outcome. Knockoff variables are defined as follows:

Definition 6 (Model-X Knockoffs, Candès et al., 2018)**.** For the family of random variables $\mathbf{x} = (x_1, \dots, x_p)$, Knockoffs are a new family of random variables $\tilde{\mathbf{x}} = (\tilde{x}_1, \dots, \tilde{x}_p)$ satisfying:

1. for any $S \subset [\![p]\!]$, $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)} \overset{d}{=} (\mathbf{x}, \tilde{\mathbf{x}})$

2. $\tilde{\mathbf{x}} \perp \mathbf{y} | \mathbf{x}$

where $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)}$ is obtained by swapping the entries x_j and $\tilde{x}_j$ for all $j \in S$.

Exchangeability is the first property of this definition. For any set of swapped variables, the joint distribution of real and knockoff variables must remain identical to the original one. In practice, this assumption is hard to enforce and to check especially for high-dimensional data distributions. A reasonable intuition is that knockoffs are meant to be used in situations where variables can easily be generated. We discuss existing work on Knockoff generation in Section 3.2.2. Violations of this assumption can result in invalid inference and massive false positives inflation – this is studied in depth in Chapter 8.

Statistical inference on Knockoff variables. Once valid knockoffs have been created, conditional variable selection can be performed. To distinguish variables that are substantially more important than their corresponding Knockoffs, a machine learning model is employed to generate importance scores for each variable and its respective knockoff. This step enables the identification of variables that offer valuable insights into the outcome, as they exhibit significant disparities in importance compared to their knockoffs. Quantitatively, this is done by computing Knockoff statistics $\mathbf{W} = (W_1, \dots, W_p)$ that are defined as follows.

Definition 7 (Knockoff Statistic, Candès et al., 2018)**.** A knockoff statistic $\mathbf{W} = (W_1, \dots, W_p)$ is a measure of feature importance that satisfies:

1. $\mathbf{W}$ depends only on $\mathbf{X}, \tilde{\mathbf{X}}$ and $\mathbf{y}$: $\mathbf{W} = g(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})$.

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

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

The most commonly used Knockoff statistic is the Lasso-coefficient difference (LCD) (Weinstein et al., 2020). This statistic is obtained by fitting a Lasso estimator (Tibshirani, 1996) on $[\mathbf{X}, \tilde{\mathbf{X}}] \in \mathbb{R}^{n \times 2p}$, which yields $\widehat{\boldsymbol{\beta}} \in \mathbb{R}^{2p}$. Then, the Knockoff statistic can be computed using $\widehat{\boldsymbol{\beta}}$:

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

This coefficient summarizes the importance of the original j^{th} variable relative to its own Knockoff: $W_j > 0$ indicates that the original variable is more important for fitting y than the Knockoff variable, meaning that the j^{th} variable is likely relevant. Conversely, $W_j < 0$ indicates that the j^{th} variable is probably irrelevant. We thus wish to select variables corresponding to large and positive W_j. Formally, the rejection set $\hat{S}$ of the Knockoffs method can be written $\hat{S} = \{j : W_j > T_q\}$, where:

$$T_q = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j \leqslant -t\}}{\#\{j : W_j \geqslant t\}} \leqslant q \right\}.$$

This definition of T_q ensures that the FDR is controlled at level q (Candès et al., 2018). Alternatively, inference can be performed using π-statistics, which quantify the evidence against each variable:

$$\pi_j = \begin{cases} \frac{1+|\{k:W_k \leq -W_j\}|}{p} \text{ if } & W_j > 0 \\ 1 \quad \text{if } & W_j \leq 0. \end{cases} \tag{3.8}$$

As noted by Nguyen et al., 2020, the vanilla Knockoffs procedure of Candès et al., 2018 is equivalent to using the Benjamini and Hochberg, 1995 procedure at level q on the vector of π-statistics $(\pi_j)_{j \in [\![p]\!]}$. The complete procedure is summarized in Figure 3.4.

3.2.2 Knockoff generation methods

Gaussian Knockoffs

Candès et al., 2018 have built Knockoffs that are provably exchangeable for Gaussian data. Provided an observed X with $X \sim \mathcal{N}(0, \Sigma)$, we sample from

$$\tilde{X} \mid X \overset{\mathrm{d}}{=} \mathcal{N}(\mu, \mathbf{V}),$$

where

$$\mu = X - X\boldsymbol{\Sigma}^{-1} \operatorname{diag}(s)$$
$$\mathbf{V} = 2 \operatorname{diag}(s) - \operatorname{diag}(s)\boldsymbol{\Sigma}^{-1} \operatorname{diag}(s)$$

With $\operatorname{diag}(s)$ any diagonal matrix. Then, if $\operatorname{diag}(s)$ is chosen such that $\mathbf{G}$ is positive semidefinite, we obtain valid Knockoffs since:

$$[X, \tilde{X}] \sim \mathcal{N}(0, \mathbf{G}), \quad \text{where } \mathbf{G} = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \operatorname{diag}(s) \\ \boldsymbol{\Sigma} - \operatorname{diag}(s) & \boldsymbol{\Sigma} \end{pmatrix}$$

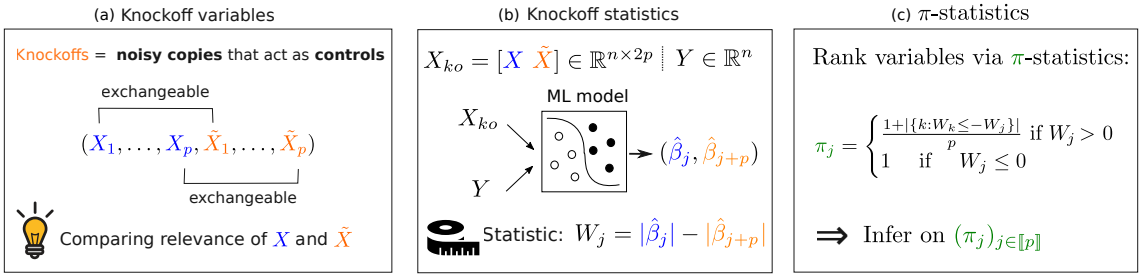


Figure 3.4: **Conditional variable selection via Knockoff variables.** Knockoff variables are noisy copies of the original variables that are used as controls for variable selection (Panel a). Conditionally on the original variables, knockoffs are independent of the outcome y. Importance scores for each variable are computed using the sensitivity estimate from machine learning model – typically, the corresponding coefficient of a Lasso fit. The Knockoff statistic is defined as the difference of modulus between the score associated with the real variable and its Knockoff counterpart (Panel b). Then, π-statistics are computed to rank variables and to perform inference (Panel c).

A major challenge is that the covariance matrix $\mathbf{\Sigma}$ is unknown in general and has to be estimated via shrinkage procedures such as Ledoit-Wolf shrinkage (Ledoit and Wolf, 2003) or Graphical Lasso (Friedman et al., 2008). Given a correctly estimated $\mathbf{\Sigma}$, examples of valid s include the equicorrelated construction $s_j^{\text{EQ}} = 2\lambda_{\min}(\Sigma) \wedge 1$ for all j.

In practice, testing exchangeability for non-Gaussian data is challenging. In the case of Gaussian data, since the actual covariance is not known but has to be estimated, exchangeability depends directly on the accuracy of the covariance estimate. Even if the data is Gaussian, its covariance matrix may be hard to estimate properly in high-dimensional regimes (Stein et al., 1972; Fukunaga, 2013). Indeed, such covariance estimation has to draw a difficult compromise between data fit and positive definiteness, along sparsity of the inverse covariance estimation from the Graphical Lasso. L2 shrinkage covariance estimation (Ledoit and Wolf, 2003) on the other hand, is known to lead to excessive bias (Belilovsky et al., 2017). In addition to the sheer computation cost, hyperparameter setting for the Graphical Lasso is challenging, and leads to difficult and costly parameter selection.

Existing alternatives to Gaussian Knockoffs

Other available methods to build Knockoffs include Monte-Carlo based methods (Sesia et al., 2019; Bates et al., 2021). These procedures focus on discrete distributions as in the case of Genome-Wide Association Studies (GWAS; Consortium et al., 2010). In the continuous setting, the method proposed in Bates et al., 2021 is equivalent to Gaussian Knockoffs.

Recent machine learning techniques such as deep learning have also been used to apprehend complex data dependence structures in the context of Knockoffs. Romano et al., 2020 propose using Deep Neural networks to build Knockoffs using a target covariance matrix estimated from the data as in the Gaussian algorithm. Additionally, the proposed algorithm aims at minimizing the correlation between original variables and Knockoffs to maximize power. Using similar ideas, Zhu et al., 2021; Liu and Zheng, 2018 propose using a Variational Auto-Encoders (VAE) to build Knockoffs as VAEs allow learning and sampling from complex data distributions. However, in the context of high-dimensional variable selection, the number of samples available is in most cases insufficient to properly train a

Deep Learning model. A limitation of these approaches is that they offer no theoretical guarantee on Knockoffs exchangeability.

In this chapter, we have explored two methods for controlling false discoveries in high-dimensional statistical settings: the Knockoffs framework and the Joint Error Rate (JER) framework. Through the examination of the Knockoffs method, we have learned that while it provides FDR control by generating artificial variables that mimic the correlation structure of the original covariates, this approach inherently involves a degree of randomness due to the random generation process of knockoffs. This randomness can lead to variability in results, which is an important concern for practitioners. Additionally, while Knockoffs are effective at controlling the FDR, they do not directly control the False Discovery Proportion (FDP). In certain runs of the procedure, the achieved FDP can be much higher than the announced control (FDR). We also discussed the common use of Gaussian Knockoffs. While efficient in many settings, this algorithm can fail in common cases: for instance, when the covariance matrix is poorly estimated or when the data are non-Gaussian. This is studied thoroughly in Chapter 8.

Moving forward, several questions need to be addressed. How can the randomness inherent in the Knockoffs generation process be mitigated to ensure more stable results? Is it possible to extend the Knockoffs framework to control the FDP directly, providing a more precise error control? Furthermore, how can we improve the robustness of knockoffs, particularly in non-Gaussian settings? These questions will be addressed in Chapters 7 and 8.

Chapter 4

Conformal prediction

Summary. In this chapter, we turn to conformal prediction, a method for quantifying predictive uncertainty in supervised learning. We begin by outlining the classical framework of prediction, highlighting the challenges associated with uncertainty quantification in traditional predictive models. Building on this foundation, we introduce conformal prediction, a versatile technique that provides finite-sample guarantees for predictive intervals. We focus on Split Conformal prediction, an extension of the classical approach that leverages data splitting to ensure computational efficiency while maintaining rigorous coverage properties. In this thesis, we will use Split Conformal prediction and improve upon its standard guarantees when considering the case of multiple test points.

Contents

4.1 Supervised learning and predictive uncertainty

The goal of supervised learning is to predict an outcome of interest $Y \in \mathcal{Y}$ given some features $X \in \mathcal{X}$. In this work, we focus on *regression problems*, meaning that Y is continuous. Predicting Y from X amounts to building a function f called a *predictor* such that $f(X)$ "is close to" Y. The quality of a predictor is measured using a loss function, defined as follows:

Definition 8 (Loss function)**.** A loss function $\ell : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is a function that quantifies the distance between two points (y, y') of the target space $\mathcal{Y}$. A typical loss function used in regression is the squared Euclidean distance $\ell(y, y') = (y - y')^2$.

Definition 9 (ℓ-risk)**.** The ℓ-risk of a predictor f is the expected value of the loss between $f(X)$ and Y:

$$\mathcal{R}_\ell(f) = \mathbb{E}[\ell(Y, f(X))].$$

Intuitively, we seek to build a predictor f that achieves optimality in the sense of minimizing the risk $\mathcal{R}_\ell$. In practice, computing $\mathbb{E}[\ell(Y, f(X))]$ is impossible – we only have access to n samples $(X_i, Y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$. We can leverage this dataset to learn a predictor which minimizes the *empirical* risk, defined as follows:

Definition 10 (Empirical ℓ-risk)**.** The ℓ-risk of a predictor f is the empirical mean of the loss between predictions $f(X_i)$ and true outcomes Y_i:

$$\widehat{\mathcal{R}}_{n,\ell}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(Y_i, f\left(X_i\right)\right).$$

Modern machine learning models achieve highly accurate predictions by minimizing the empirical ℓ-risk using vast amounts of data. These advancements have enabled machine learning systems to achieve unprecedented levels of performance across various applications, from image recognition to natural language processing and beyond. However, as these models are increasingly deployed in critical decision-making processes—such as healthcare, finance, and autonomous systems—the stakes have risen dramatically. In these high-stakes environments, the cost of incorrect predictions can be substantial, leading to potentially severe consequences.

Thus, it is not sufficient to rely solely on the accuracy of predictions; we must also understand the level of confidence we can have in these predictions. This is where predictive uncertainty (Smith, 2013; Abdar et al., 2021) becomes crucial. By quantifying the uncertainty associated with a prediction, we can make more informed decisions, better manage risks, and avoid overconfidence in the model's outputs.

Conformal prediction (Saunders et al., 1999; Vovk et al., 1999, 2005; Gammerman and Vovk, 2007; Shafer and Vovk, 2008; Lei and Wasserman, 2014; Sadinle et al., 2019; Foygel Barber et al., 2021) provides a mathematically grounded approach to addressing this need for predictive uncertainty. Unlike traditional methods that offer only point estimates, conformal prediction constructs prediction intervals or sets that are guaranteed to contain the true outcome with a specified probability. We now turn to describing precisely Split Conformal Prediction (SCP) which we will use in this thesis.

4.2 Split Conformal prediction

Split conformal prediction offers a simple and flexible approach to constructing reliable prediction intervals. This method offers the desired guarantee in a model-agnostic way at the cost of splitting the training data. The data is split into $\mathcal{D}_{train}, \mathcal{D}_{cal}$ such that $\mathcal{D}_{train} \cap \mathcal{D}_{cal} = \varnothing$. For the sake of simplicity, we slightly abuse notation and assume that n calibration samples are available and that we work in a regression framework. Define conformity scores $S_i = |Y_i - \hat{\mu}\left(X_i\right)|$ where $\hat{\mu}(x)$ is a point prediction of Y_i given $X_i = x$ learnt on $\mathcal{D}_{\text{train}}$.

The hypothesis underlying split conformal prediction is that the calibration set provides a realistic measure of the trained model's performance. Furthermore, evaluating conformity scores allows making no assumptions about data distributions or model characteristics, aside from scores' exchangeability. To obtain valid intervals, conformity scores have to be exchangeable across the calibration set and test point. Formally, exchangeability is defined as follows:

Assumption 1 (Conformity scores exchangeability, Vovk et al., 2005)**.** Denote $\mathcal{S}_{n+1}$ the symmetric group of $[\![1, n+1]\!]$. We assume that the conformity scores $(S_1, \ldots, S_{n+1})$ are exchangeable, i.e. that:

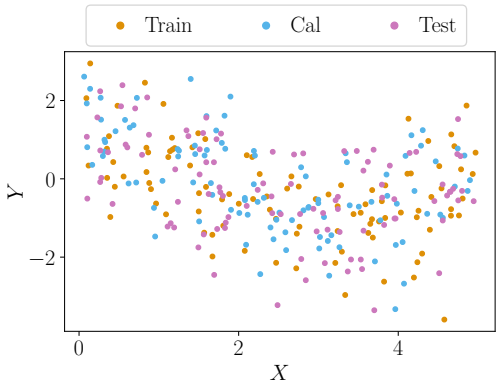
$$\forall \tau \in \mathcal{S}_{n+1}, \quad (S_1, \ldots, S_{n+1}) \overset{d}{=} (S_{\tau(1)}, \ldots, S_{\tau(n+1)})$$

This assumption is less stringent than imposing that test and calibration samples are i.i.d. – however, it is hard to check in practice and many real world applications involve non-stationary behavior of conformity scores (Gibbs and Candes, 2021). Achieving coverage beyond exchangeability has been the focus of much effort in the conformal prediction community (Tibshirani et al., 2019; Gibbs and Candes, 2021; Barber et al., 2023; Cauchois et al., 2024).

In this thesis, we focus on standard SCP. Provided that exchangeability holds, split conformal yields the following valid interval:

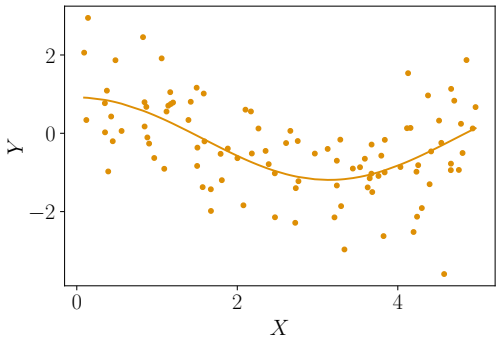
$$\widehat{C}_\alpha = \left[\hat{\mu}\left(X_{n+1}\right) \pm S_{(\lceil (n+1)(1-\alpha) \rceil)}\right].$$

Step 1



- Create a proper training set, a calibration set, and keep the test set, by randomly splitting the data set

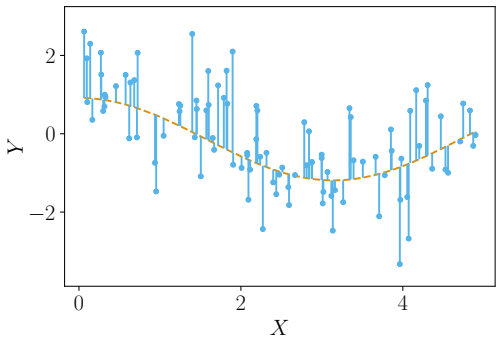
Step 2



On the proper training set:

- Learn (or get) $\hat{\mu}$

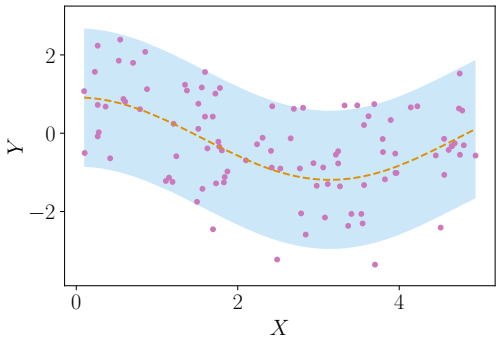
Step 3



On the calibration set:

- Predict with $\hat{\mu}$
- Get the |residuals|, a.k.a. conformity scores
- Compute the $(1-\alpha)$ empirical quantile of $\mathcal{S} = \{|\text{residuals}|\}_{Cal} \cup \{+\infty\}$, noted $q_{1-\alpha}(\mathcal{S})$

Step 4



On the test set:

- Predict with $\hat{\mu}$
- Build $\widehat{C}_{n,\alpha}(x)$: $[\hat{\mu}(x) \pm q_{1-\alpha}(\mathcal{S})]$

Figure 4.1: **Illustration of the Split Conformal Prediction procedure.** First, split the dataset into three separate datasets: training, calibration and test. Fit a predictor $\hat{\mu}$ on the training set – then, use it to formulate predictions on the calibration set. From these predictions, get the residuals and compute their $1-\alpha$ quantile. On the test set, use the predictor $\hat{\mu}$ and build the confidence interval $\widehat{C}_\alpha = \left[\hat{\mu}(X_{n+1}) \pm S_{(\lceil (n+1)(1-\alpha) \rceil)}\right]$. Figure from Zaffran, 2024.

This construction is detailed and thoroughly studied in Papadopoulos et al., 2002; Lei

et al., 2018. Alternatively, this interval can also be obtained by thresholding a so-called conformal p-value (Vovk et al., 2005; Lei et al., 2018). This p-value is denoted p_{n+1} and defined as follows:

Definition 11 (Conformal p-value,Vovk et al., 2005)**.** Given conformity scores $(S_i)_{i\in[\![n+1]\!]}$ the conformal p-value of test point (X_{n+1}, Y_{n+1}) is defined:

$$p_{n+1} := \frac{1}{n+1}\left(1 + \sum_{i=1}^{n} \mathbf{1}\left\{S_{n+1} \leq S_i\right\}\right)$$

Intuitively, this p-value quantifies how unlikely the observed conformity score is – under exchangeability. The scores computed on the calibration set serve as a reference to compute this likelihood. When building a confidence interval, we wish to include all possible values of y of high likelihood and exclude the values of y of low likelihood. This idea can be used to obtain the split conformal interval $\widehat{C}_\alpha$:

Remark 2. Split Conformal prediction intervals can equivalently be obtained by thresholding the conformal p-value, i.e.:

$$\widehat{C}_\alpha = \left[\hat{\mu}\left(X_{n+1}\right) \pm S_{(\lceil (n+1)(1-\alpha)\rceil)}\right] = \{p_{n+1} > \alpha\}$$

In this chapter, we have reviewed the framework of conformal prediction for assessing predictive uncertainty in supervised learning. We began by examining the classical approach to prediction via empirical risk minimization. We then explored conformal prediction, focusing on its ability to provide valid prediction intervals with finite-sample guarantees. Special attention was given to Split Conformal prediction, which offers a computationally efficient method while preserving the coverage properties essential for practical applications.

A question that remains open in the Split Conformal framework is the ability to control error rates when dealing with multiple points at the same time in the test set. We tackle this problem in Chapter 9.

Chapter 5

Neuroimaging data analysis

Summary. In this chapter, we give a brief overview of modern neuroimaging data analysis. Neuroimaging data analysis plays a major role in our understanding of the brain, and is also used for clinical diagnosis and neurosurgery. We focus on functional Magnetic Resonance Imaging (fMRI), a non-invasive modality which is widely used in cognitive neuroscience to understand how brain activity relates to functions such as memory, attention, emotion and many other domains of cognition. We first review the basics of acquisition and preprocessing of fMRI data before moving to statistical analysis. This is first done by obtaining statistical maps, that display the level of activity observed for different stimuli presented to subjects. Then, these statistical maps are combined in order to draw meaningful conclusions at the population scale. Central to this analysis is statistical hypothesis testing, which ensures that the interpretations drawn from the data are reliable and valid. Given the complexity and high dimensionality of neuroimaging data, addressing the risk of false positives is essential. This is where methods for multiple comparisons come into play, to provide error control when numerous statistical tests are conducted simultaneously; many existing frameworks aim at tackling this issue. Designing reliable statistical methods for fMRI inference is one of the main motivations of this thesis, and most of the numerical experiments of the contributions have been performed on fMRI data.

Contents

5.1 Functional MRI data

Functional Magnetic Resonance Imaging (fMRI) is a key tool in neuroscience due to its ability to non-invasively monitor brain activity by detecting blood oxygen level-dependent (BOLD) signals, which reflect changes in blood flow and oxygenation linked to neural activity (Logothetis, 2008). This capability makes fMRI a valuable resource for exploring the human brain's functional organization and understanding the neural mechanisms underlying various cognitive and behavioral processes.

In cognitive neuroscience, fMRI is widely used to investigate the neural basis of functions such as memory, language, perception, and decision-making. It enables researchers to identify the brain regions involved in specific tasks, providing insights into the functional architecture and connectivity of the brain (Cabeza and Nyberg, 2000; Smith et al., 2009). This information is crucial for advancing our understanding of how different brain areas work together to support complex cognitive functions and behaviors.

In clinical applications, particularly neurosurgery, fMRI plays a critical role in preoperative planning. Surgeons use fMRI data to localize essential brain areas involved in functions like speech, movement, and sensation, which helps to minimize the risk of damage to these regions during surgery (Hirsch et al., 2000). This approach enhances surgical outcomes by preserving critical cognitive and motor functions, making fMRI a standard tool in modern neurosurgical practice.

Additionally, fMRI is a valuable tool in population studies, providing insights into how brain function varies across different demographic groups, such as in aging populations or individuals with genetic predispositions. It is also used to investigate the neural correlates of psychiatric and neurological disorders, including depression, schizophrenia, and Alzheimer's disease, aiding in the identification of biomarkers and contributing to the development of targeted treatments (Paus, 2010; Matthews et al., 2006).

Acquisition. FMRI relies on the well-established observation that increased neural activity causes changes in metabolism and blood flow, leading to variations in the concentrations of oxyhaemoglobin (oxygen-carrying red blood cells) and deoxyhaemoglobin (oxygen-depleted red blood cells). These two forms of haemoglobin have distinct magnetic properties (diamagnetic and paramagnetic, respectively), that affect the local magnetic field differently. The MRI scanner detects these changes, known as the "Blood Oxygen Level Dependent" (BOLD) signal, to record brain activity.

During fMRI sessions, brain activity is measured over several minutes while participants perform cognitive tasks, with images typically acquired every 1-3 seconds (Repetition time, TR). A 3D brain image is composed of voxels (3D pixels), and the sequence of images during a session provides a time series of MRI signals for each voxel, sampled at every time repetition.

Preprocessing. Before analyzing fMRI images, preprocessing steps are necessary to ensure the accuracy and reliability of subsequent analyses. These steps include for instance motion correction, since small head movements by the subject during scanning can cause significant distortions in the data. Another common preprocessing step is artifact removal, since scanner issues, hardware malfunctions may generate artifacts that alter data. A necessary step to enable group comparisons is spatial normalization. Spatial normalization in fMRI is the process of aligning individual brain images to a common anatomical space or template to compensate for anatomical variability. While not a focus of this thesis, preprocessing is of crucial importance to obtain reliable downstream inference.

5.2 Statistical analysis of fMRI data

5.2.1 Obtaining statistical maps

Once fMRI data have been acquired and preprocessed, statistical analysis can be performed. One of the most commonly used statistical approaches in single subject fMRI analysis is the General Linear Model (GLM). A simple GLM can be written:

$$X_{fMRI} = X_d\beta + \epsilon,$$

where $X_{fMRI} \in \mathbb{R}^{t\times p}$ is the observed fMRI data reorganized as a matrix with t time steps in rows and p voxels in columns, $X_d \in \mathbb{R}^{t\times r}$ the design matrix containing r regressors across t time steps (e.g., task conditions or experimental variables) and $\beta \in \mathbb{R}^{r\times p}$ the matrix of regression coefficients that relates the observed signal to the r regressors and the activity of the p voxels of the brain. $\epsilon \in \mathbb{R}^{t\times p}$ models noise.

The β matrix can be estimated via ordinary least-squares or more complex techniques accounting for non-i.i.d. noise. Once this matrix is computed, *contrast maps* are derived using statistical testing. A contrast represents the difference between two conditions – contrast maps aim at determining whether the beta coefficient of certain voxels are significantly different for one type of stimulus compared to another. T-tests are commonly used, and the resulting t-statistics can be converted to p-values or z-values. Obtaining contrast maps completes the **first-level** analysis – an example of such analysis is displayed in Figure 5.1.

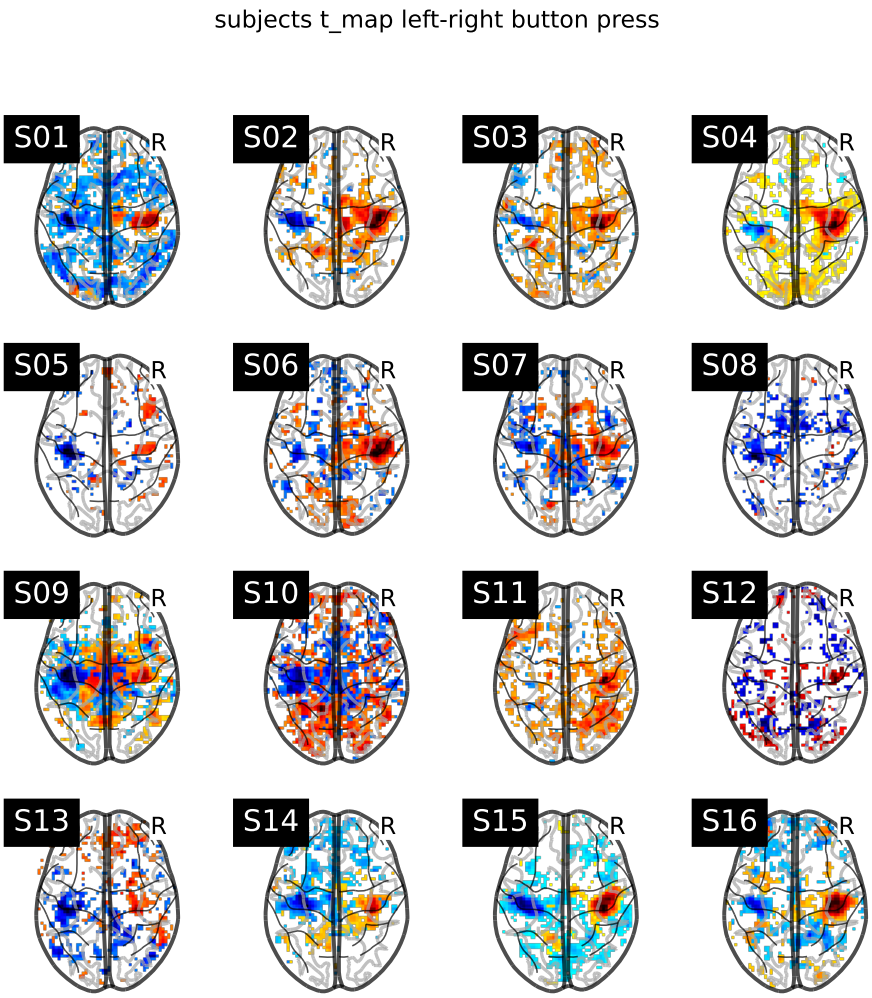


Figure 5.1: **First-level analysis on 16 subjects for a motor contrast.** Each of these 16 contrast maps was computed using a GLM fitted on the fMRI images of each subject. The maps displayed correspond to the "left-right button press" contrast. Figure from https://nilearn.github.io/stable/auto_examples/05_glm_second_level/plot_second_level_one_sample_test.html.

At this level of analysis, fMRI datasets are reduced to a set of volumes that represent BOLD signal differences between user-chosen contrasts. These images are then typically used in a *second-level analysis*. In this thesis, we use this type of input to draw conclusions at the population level.

5.2.2 Population-level analysis

Population-level analysis in fMRI is used to make group-level inferences by combining data from individual subjects. First-level contrast maps from all subjects are combined to perform statistical tests, allowing researchers to assess whether observed effects are consistent across the group. This approach enables the detection of group-level effects, such as the average response to a task or differences between groups, and accounts for

variability between subjects, making the results generalizable to the broader population (Worsley and Friston, 1995; Friston et al., 1996; Penny et al., 2003; Mumford and Nichols, 2009; Poldrack et al., 2011).

In practice, say that we are interested in comparing two contrasts using the fMRI images of n participants. Denoting $X \in \mathbb{R}^{n \times p}$ the aggregated contrast maps, $Y \in \{-1, 1\}^n$ the two contrasts, we seek to infer which voxels X_j are associated to the task at hand. In words, we wish to recover the voxels that exhibit a significantly different activation when comparing the two contrasts of interest. Let us formalise this problem using statistical hypothesis testing – we test simultaneously for all voxels X_j whether or not they are independent of the outcome y:

$$H_{0,j} : Y \perp X_j \quad \text{versus} \quad H_{1,j} : Y \not\perp X_j.$$

This is a severe multiple testing problem, as activation tests are performed on up to hundreds of thousands of voxels simultaneously. Many methods have been developed to address this inference problem using different statistical frameworks – we briefly review them next.

Mass univariate inference. The most straightforward approach to solving this inference problem is to compute parametric p-values for all voxels *marginally*. This results in a vector of p-values which can be used for inference. Obviously, since all hypotheses are tested simultaneously, multiplicity must be accounted for as explained in Chapter 2 using e.g. FDR control (Genovese et al., 2002). Alternatively, using the Bonferroni correction is also possible to obtain FWER control (Nichols and Hayasaka, 2003; Worsley, 2003).

While mass univariate approaches are theoretically grounded, they can be misused to produce invalid inference. A first common pitfall is that *selective inference* is not supported: if one were to apply e.g. the BH procedure to regions selected *after having seen the data*, massive false positive inflation may ensue (Kriegeskorte et al., 2009; Nieuwenhuis et al., 2011). This is a sizable issue since selecting brain regions (*post hoc* or not ? it is in general impossible to know) is common practice in fMRI.

Under the assumption that fMRI images are smooth 3D Gaussians, **Random Field Theory** (RFT; Adler, 2010) provides an elegant framework to study the distribution of extreme values in fMRI images. This can be done at the voxel-level or at the cluster level. RFT has been applied to fMRI images, yielding voxel-wise FWER controlling procedures (Friston et al., 1991; Worsley et al., 1992; Worsley, 1994).

Nonparametric approaches. Nonparametric approaches estimate statistical properties using the data itself rather than relying on parametric assumptions such as Gaussianity. Notably, permutation tests involve shuffling data labels or residuals under the null hypothesis to generate the null distribution of the largest test statistic – or equivalently smallest p-value, see e.g. Westfall and Young, 1993. This provides a robust way to control the family-wise error rate across thousands of brain voxels without assuming normality (Winkler et al., 2014). Permutations can also be used to obtain valid cluster-level inference (Smith and Nichols, 2009).

Cluster-level inference. An alternative type of inference to increase statistical power is to perform inference at cluster-level, rather than voxel-level (Poline and Mazoyer, 1993),

because brain activation is organised in compact regions *(clusters)* in the brain volume. Examples of voxel-level and cluster-level inference are shown in Figure 5.2. This type of inference tests whether regions above a given threshold are larger than expected under the null hypothesis, or whether the total amount of signal in these regions (Smith and Nichols, 2009) exceeds its expected value under a null distribution. However, this approach suffers from several problems (Eklund et al., 2016) such as the arbitrary choice of cluster-forming threshold (Woo et al., 2014), or the difficulty to establish a null distribution for cluster size and aggregated signal. Moreover, the null hypothesis corresponding to this procedure is **global**, meaning that there no signal across the whole brain in this regime. This entails a **spatial specificity paradox**: rejecting the global null hypothesis amounts to declaring that **at least one voxel** is activated. Therefore, when detecting larger clusters – which should indicate stronger signal – the spatial information about the signal therein worsens.

Conditional approaches. Conditional approaches aim at establishing that certain brain regions provide unique information on behavior. This is a hard problem as activation in certain regions may be better explained by correlation to a nearby region recruited in the cognitive process at hand (Weichwald et al., 2015). This is related to the conditional independence testing framework explained in Section 3.2. Existing methods rely on Knockoffs (Nguyen et al., 2019) or the Desparsified Lasso (Chevalier et al., 2021), a statistical method introduced in Zhang and Zhang, 2014; Van de Geer et al., 2014 which generalizes least-squares-based inference to high-dimensional settings. Chevalier et al., 2021, also relies on dimension reduction through clustering and randomization to stabilize the outcome.

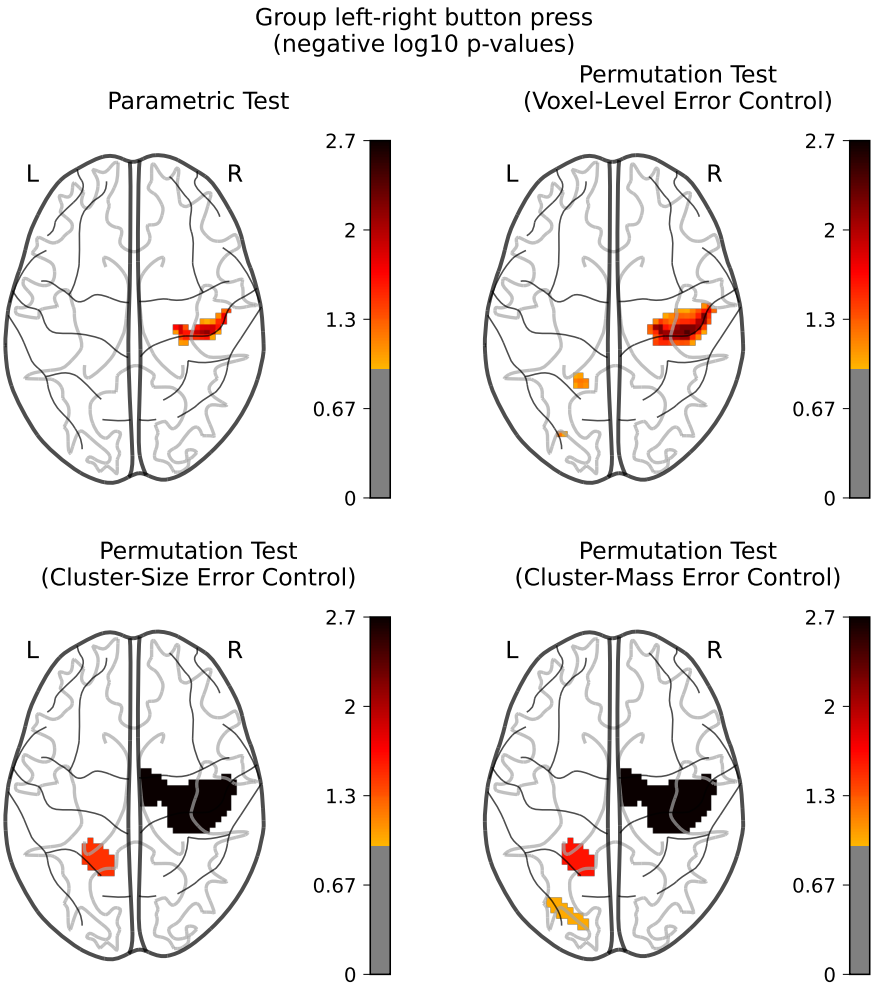


Figure 5.2: **Second-level analysis combining 16 subjects for a motor contrast.** The 16 contrast maps of Figure 5.1 are aggregated and used for statistical testing across all subjects using contrast "left-right button press". Both cluster-level and voxel-level methods are showcased, resulting in different selected voxels. Figure from https://nilearn.github.io/stable/auto_examples/05_glm_second_level/plot_second_level_one_sample_test.html.

Part II

Contributions

Chapter 6

Notip: Non-parametric True Discovery Proportion control for brain imaging

Summary. Cluster-level inference procedures are widely used for brain mapping. These methods compare the size of clusters obtained by thresholding brain maps to an upper bound under the global null hypothesis, computed using Random Field Theory or permutations. However, the guarantees obtained by this type of inference - i.e. at least one voxel is truly activated in the cluster - are not informative with regards to the strength of the signal therein. There is thus a need for methods to assess the amount of signal within clusters; yet such methods have to take into account that clusters are defined based on the data, which creates circularity in the inference scheme. This has motivated the use of *post hoc* estimates that allow statistically valid estimation of the proportion of activated voxels in clusters. In the context of fMRI data, the All-Resolutions Inference framework introduced in Rosenblatt et al., 2018 provides post hoc estimates of the proportion of activated voxels. However, this method relies on parametric threshold families, which results in conservative inference. In this chapter, we leverage randomization methods to adapt to data characteristics and obtain tighter false discovery control. We obtain *Notip*, for Non-parametric True Discovery Proportion control: a powerful, non-parametric method that yields statistically valid guarantees on the proportion of activated voxels in data-derived clusters. Numerical experiments demonstrate substantial gains in number of detections compared with state-of-the-art methods on 36 fMRI datasets. The conditions under which the proposed method brings benefits are also discussed.

Contents

In this chapter, we introduce the Notip procedure, a post hoc inference method with FDP control that adapts non-parametrically to data correlation . The use of non-parametric procedures also renders the inference robust to mis-specification of the statistics distribution. We study whether such a procedure can yield less conservative inference while offering the same statistical guarantees. We perform extensive experiments on dozens of fMRI datasets to compare the number of detections obtained by this approach with that of existing methods.

The chapter is organized as follows. The main contribution is the Notip method presented in Section 6.1: a nonparametric data-driven approach that relies on the JER framework to obtain sharper post hoc FDP control. Numerical experiments and results on fMRI data reported in Sections 6.2 and 6.3 show that substantial gains in the number of detections are obtained from the proposed method, while controlling the FDP of the detected regions at a fixed level. Finally, we discuss the benefits of our proposed methodology, and outline some possible limitations.

6.1 Data-driven templates and Notip procedure

The calibrated Simes family can lead to tighter post hoc bounds, yet it still relies on the Simes template, which is linear in k, as illustrated in Figure 3.2. Instead of only optimising λ for a given template shape (e.g. a linear shape for the Simes template), the second degree of freedom that can be exploited to achieve better statistical power while still controlling the JER is **to learn the template function, or, equivalently, its shape when displayed as a graph** Figure 3.2 illustrates that for small k, permuted p-value curves are not exactly linear. This suggests that using a non-linear template shape could be relevant for fMRI data. Several other parametric templates are considered in Andreella et al., 2020, but the authors report that none of these attempts outperformed the Simes template. An ideal template should approximately reproduce the shape of randomized p-values curves computed from real data. Therefore, we propose to **learn** a template directly from the data.

A related idea has been explored in Meinshausen, 2006. However, since the method proposed in that paper does not distinguish between the learning and calibration steps, it

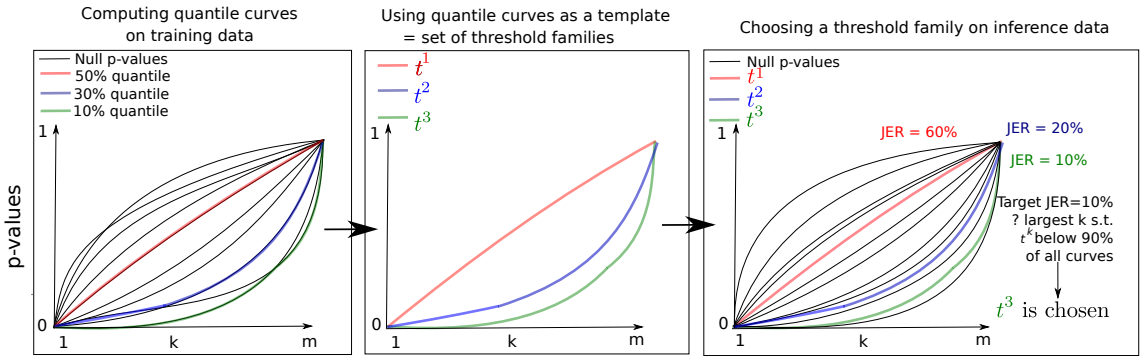


Figure 6.1: **Learning a template from training data and using this template for calibration on inference data.** Left panel: quantiles of randomized p-value curves are computed on training data. Middle panel: the resulting quantile curves are used as a template (the so-called learned template). Right panel: calibration is performed on inference data using the learned template. Notice that learned templates do not have a parametric shape (contrary e.g. to the Simes template), but follow the shape of sorted null p-values.

suffers from circularity biases, as noted by (Blanchard et al., 2020, Remark 5.3). Indeed, in the JER framework, the template has to be fixed a priori.
In order to address this issue, we propose to *learn* a template from an fMRI dataset that is independent from the datasets on which inference is performed. First, we compute B randomized p-value curves on training data using Algorithm 1 and extract quantile curves $t^b = (t^b_k)_k$ for $b = 1 \dots B$, as shown in the left panel of Fig. 6.1. These quantile curves are then viewed as a set of B sorted threshold families (middle panel), which is called a **learned template**. Note that it is indeed a template in the sense of Blanchard et al., 2020, that has been discretized over a set of B values.

After obtaining a learned template, calibration is performed on the inference data (i.e. any inference contrast) as would be done with a parametric template. This is shown in the right panel of Figure 6.1 and in Section 3.1. To perform calibration, we evaluate the empirical JER of all threshold families of the learned template. Then, we select the largest $b \in \{1, \dots B\}$ such that JER control holds on inference data for the threshold family t^b. To avoid evaluating the JER of all threshold families, this search is done by dichotomy in practice. The resulting method is called "Notip" for Non-parametric true discovery proportion. As described above, Notip requires a training dataset in order to learn the template. Note that learned templates do not have a parametric shape, but follow the shape of randomized p-value curves.

The calibration process depends on the parameter k_{max}, whose choice induces the following trade-off. On the one hand, since $JER((t^b_k)_{1 \le k \le K}) \le JER((t^b_k)_{1 \le k \le K'})$ for all $K \le K'$, choosing a smaller k_{max} allows calibration to choose a largest value of b in the dichotomy, leading to a less conservative family. On the other hand, a larger value for k_{max} leads to more thresholds considered in the min in the bound written in Equation 3.6, and hence to a possibly tighter bound. Guidelines to choose k_{max} as well as an informed default choice for fMRI data are given in Section 6.2.1.

The complete procedure is summarized in Algorithm 2, with lines 1-7 corresponding

to the training step and lines 8-20 corresponding to the inference step. The latter step requires the computation of the empirical JER for a given family, which is described in Algorithm 3.

Algorithm 2: Learning template on training data and calibrating on inference data. A template is learnt by computing permuted p-values and extracting quantile curves. Then, this template is used to perform calibration on inference data by choosing the least conservative family of the learned template that empirically controls the JER.

1 **Function** `learn_template_and_calibrate`$(X_{train}, X_{infer}, B_{train}, B_{infer}, \alpha, k_{max})$`:`
2 pvals$_{train}$ ← get_randomized_p_values(X_{train}, B_{train})
3 `// array of shape` (B_{train}, n_{voxels}) `// lines of pvals`$_{train}$ `are sorted`
4 **for** $b \leftarrow 1$ **to** B_{train} **do**
5 learned_templates[b] ← quantiles(pvals$_{train}$, b/B_{train})
6 **end**
7 pvals$_{infer}$ ← get_randomized_p_values(X_{infer}, B_{infer})
8 `// vector of shape` (B_{infer}, n_{voxels}) **for** $b \leftarrow 1$ **to** B_{train} **do**
9 $\widehat{JER}_b \leftarrow$ estimate_jer(pvals$_{infer}$, learned_templates[b], k_{max})
10 **end**
11 $b_{calibrated} \leftarrow \max\{b \in [1, B_{train}] \text{ s.t. } \widehat{JER}_b \leq \alpha\}$
12 `// Choose largest` b `such that JER control holds`
13 **if** $b_{calibrated} = 0$ **then**
14 **return** *Calibrated_Simes*
15 `// No suitable learned template found`
16 **end**
17 chosen_template ← learned_templates[$b_{calibrated}$]
18 **return** *chosen_template*

Algorithm 3: JER estimation on randomized p-values. The empirical JER is computed for a given template and a matrix of permuted p-values. This computation is directly based on Equation 3.5.

1 **Function** `estimate_jer`$(pvals, thr, k_{max})$`:`
2 (B_{infer}, p) ← shape(pvals)
3 $\widehat{JER} \leftarrow 0$
4 **for** $b' \leftarrow 1$ **to** B_{infer} **do**
5 **for** $i \leftarrow 1$ **to** k_{max} **do**
6 diff[i] ← pvals[b'][i] - thr[i] `// Check JER control at rank` i
7 **end**
8 **if** $min(diff) < 0$ **then**
9 $\widehat{JER} \leftarrow \widehat{JER} + 1/B_{infer}$ `// Increment risk if JER control event is violated`
10 **end**
11 **end**
12 **return** $\widehat{JER}$

Once Algorithm 2 has been run, according to (Blanchard et al., 2020, 2021), the bound defined in Equation 3.6 is a valid FDP upper bound. This bound can be computed on any subset of interest S in linear time in $|S|$ using Algorithm 1 in Enjalbert-Courrech and

Neuvial, 2022.

6.2 Experiments

6.2.1 Choice of k_{max}

The post hoc bound (3.6) is valid for any value of the parameter k_{max}, provided that this parameter is chosen *a priori* and not after data analysis (Blanchard et al., 2020). While some guidelines are given in the Discussion of Blanchard et al., 2020, the choice of k_{max} remains an open question. Equation 3.6 may be written as follows:

$$V(S) = \min_{1 \leq k \leq |S| \wedge k_{max}} V_k(S)\,, \tag{6.1}$$

where $V_k(S) = \sum_{i \in S} 1\{p_i(X) \geq t_k\} + k - 1$. Each $V_k(S)$ is itself an upper bound on the number of false positives in S. The choice of k_{max} implies a tradeoff. On the one hand, large values of k_{max} can seem advantageous because the minimum in (6.1) is taken on a larger set of values of k. On the other hand, when the thresholds t_k are obtained by calibration — as in Blanchard et al., 2020 or in the present chapter, a smaller k_{max} leads to larger values of (t_k) for a given k, and thus to a tighter bound V_k. Noting that $V_k(S) \geq k - 1$, the values of k such that $k > q|S|$ will yield $V_k(S)/|S| \geq q$ for any S. Therefore, these values of k are useless for obtaining a FDP bound less than q. This motivates a choice of k_{max} of the form

$$k_{max} = q_{max}|S_{max}|\,, \tag{6.2}$$

where q_{max} is the maximum proportion of false positives that can be tolerated by users and $|S_{max}|$ is the size of the largest set of voxels of interest.

In practice, the regions of interest are those in which a **high proportion of activated voxels** can be guaranteed. To be conservative, we set $q_{max} = 0.5$, which simply means that we are not interested in guaranteeing that the FDP is less than q for $q \geq 0.5$. In the case of fMRI, one is generally interested in sparse activation extent, as widespread effects are by definition not informative on the specific involvement of brain regions in the contrast of interest. As a default choice, we observe that most fMRI contrasts studied in the literature lead to less of 5% of the image domain to be declared activate, which amounts to setting $|S_{max}| = 0.05m$.

Finally, a reasonable choice seems to be $k_{max} = 0.5 * 0.05m = 0.025m$. In the context of the experiments we described where $m \simeq 50,000$, we settle for simplicity on using $k_{max} = 0.02m = 1,000$. This is the default value of k_{max} in the implementation we propose. To illustrate the effect of the choice of k_{max} we display the variation of the number of detections of all three methods on 36 fMRI datasets across 9 different inference settings for varying k_{max} in Figure 6.2. Except for extremely small or large values of k_{max} Notip is at worst slightly sub-optimal and $k_{max} = 0.02m$ is a safe default.

As noted in Blanchard et al., 2020, no choice of k_{max} uniformly outperforms others. For example, the above choice, which is motivated by the *prior*: "$|S_{max}| = 0.05m$", may be poorly adapted in situations where very large regions are considered.

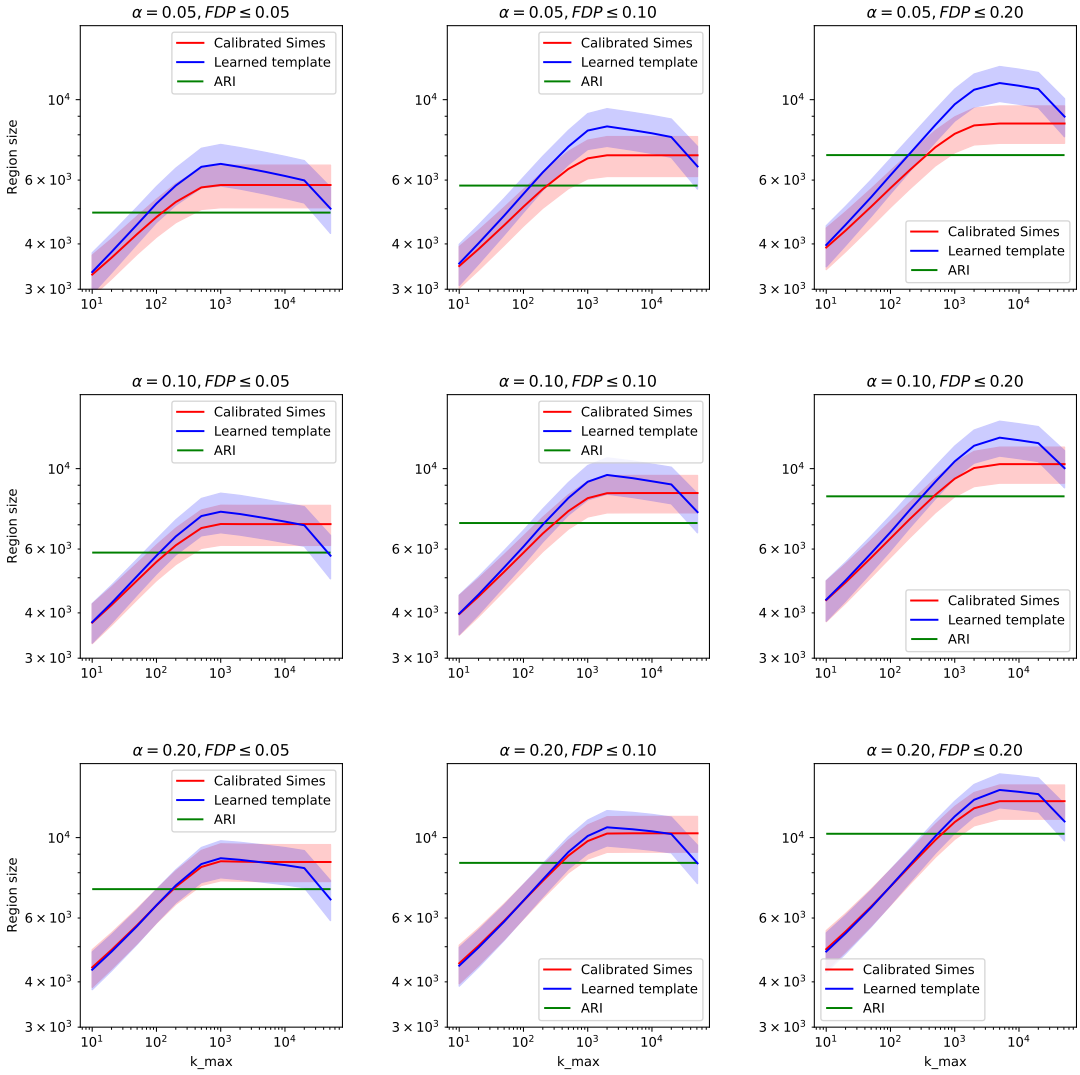


Figure 6.2: **Comparison of the number of detections between learned template and calibrated Simes for various k_{max} values with 5% error bands in log-log scale.** Notice that the chosen k_{max} largely influences the maximum size of the FDP controlling region for the learned template.

6.2.2 Data

FMRI data

To investigate the potential gain in number of detections yielded by using data-driven templates, we performed experiments on an fMRI dataset, collection 1952 (Varoquaux et al., 2018) of the Neurovault database (http://neurovault.org/collections/1952). This dataset is an aggregation of 20 different fMRI studies, consisting of statistical maps obtained at the individual level for a large set of contrasts. These images have been preprocessed using the procedure described in Varoquaux et al., 2018. In particular, they have been spatially normalized to MNI space using SPM12 software, and resampled to 3mm isotropic resolution. In the present case, the inference question concerns one-sample tests in group analyses, i.e. identifying what brain regions show a significant increase of activity for the contrast of interest, as opposed to the baseline, across participants. The group-level statistic and associated p-value are obtained through a one-sample t-test on the individual z-maps.

Collection 1952 only contains elementary *'versus baseline'* contrasts, with a massive amount of non-specific signal. In order to obtain meaningful inference examples, we paired them with control contrasts. A typical interesting contrast pair is "words vs baseline" vs "face vs baseline"; by subtracting these two contrasts, we obtain the more relevant "words vs face" contrast, which aims at uncovering brain regions with significantly higher or lower signal for word images than for face images stimuli.

To obtain consistent results, we excluded contrasts with too few subjects and/or trivial signal. The resulting list of 36 contrast pairs is given in Table 6.6.

In order to use data-driven templates on fMRI data, we have to choose a training set beforehand, on which we learn a template once and for all. The variability of the Notip method with regards to the choice of the training set is studied in Section 4.3. For the rest of the experiments, we use a single training set. Although learning a different template for each contrast pair would produce statistically valid inference, the computational cost would be high and this would lead to a loss in generality (i.e. the user would have to learn a template per inference contrast pair, instead of doing it once). For these experiments, we choose for training data a pair of contrasts with 113 subjects and 51199 voxels smoothed using $FWHM$ (full width at half-maximum) $= 4mm$ and at least 2% of active voxels (with probability $\geq 95\%$ according to ARI). This is the pair of contrasts with the lowest proportion of active voxels that we could find among contrast pairs with at least 100 subjects. This choice is referred to as the optimal template in the rest of the chapter. This template is learnt using $B_{train} = 10,000$ permutations and we choose $k_{max} = 1,000 \simeq \lfloor m/50 \rfloor$ for reasons detailed in Section 6.2.1. Note that we also apply the same choice of k_{max} when using the Simes template, so that both templates are compared on a fair basis.

Synthetic data

For some of the experiments described below, we have generated simulated data using the pyrft package: https://github.com/sjdavenport/pyrft. This package allows generates smooth noisy random fields that resemble fMRI data. In this controlled setup, the ground truth is known. An example of such simulated data can be found in Section 6.5.4. The simulation setting is the following, with π_0 the proportion of null voxels: $\alpha = 0.05,\ \pi_0 = 0.9,\ FWHM = 4mm,\ n_{train} = 100,\ n_{infer} = 50,\ q = 0.1,\ B_{train} = B_{infer} = 1000$.

Code to reproduce the experiments

Data manipulation is mostly performed through Nilearn v0.9.0, nibabel v.3.1.1. The proposed statistical methods are implemented in the sanssouci package: https://github.com/pneuvial/sanssouci.python. The experiments presented in this section can be reproduced using the code at: https://github.com/alexblnn/Notip. This repository contains a script per experiment.

The analysis we performed on this data can be divided into 6 main experiments that are detailed in the rest of this section.

6.2.3 Variation of the number of detections for different template types

To compare different choices of templates and investigate whether data-driven templates yield a gain in number of detections over existing methods, we compute the size of the largest possible region that satisfies a target error control for each choice of template on the 36 chosen contrast pairs. This is typically the type of inference that users aim for when applying FDR controlling procedures such as the Benjamini-Hochberg procedure. We denote by S_t the largest region (i.e. subset of voxels) such that its FDP upper bound is smaller than some user-defined value $q \in [0, 1]$, called the FDP budget. It corresponds to the maximum FDP that one is willing to tolerate in a given region. Formally, we solve the following optimisation problem for any template t:

$$|S_t| = \max_S |S| \quad \text{s.t.} \quad \frac{V_\alpha^t(S)}{|S|} \leq q\,, \tag{6.3}$$

where $V_\alpha^t(S)/|S|$ is the upper bound on the FDP at risk level α computed on S using the template t. By construction of the bound (3.6), the solution of (6.3) is a p-value level set, of the form $\{i/p_i \leq \tau\}$ for some τ (Blanchard et al., 2020, Section 7.4). As such, $|S_t|$ can be obtained in linear time in m using Algorithm 1 in Enjalbert-Courrech and Neuvial, 2022.

Then, we compute the relative size difference of S_t for all possible pairs of methods. Formally, **the variation of the number of detections** between the learned template (i.e., the Notip procedure) and the calibrated Simes template is defined as:

$$\frac{|S_{Learned}| - |S_{Simes}|}{|S_{Simes}|}$$

The calibration procedure on any a priori fixed template controls the JER (Blanchard et al., 2020, 2021). Therefore, it makes sense to compare the number of detections obtained by different template choices (i.e. ARI, calibrated Simes and learned template) for a given error control $1 - \alpha$. We compare the number of detections for several values of q, the FDP budget, for a given risk $\alpha = 5\%$.

We also perform the same experiment on the simulated data described in Section 6.2.2. In this case, since the ground truth is known, we can compare the empirical True Positive Rate (TPR) of all three methods. This quantity represents the proportion of true signal recovered by the template t for the region S_t defined in (6.3). Formally, we defined the TPR in S_t as the ratio of the lower bound on the true positives in S_t to the number of

truly activated voxels in S_t:

$$\mathrm{TPR}(S_t) = \frac{|S_t| - V_\alpha^t(S_t)}{|H_1|}.$$

Where $|H_1|$ corresponds to the number of truly activated voxels. As such, $\mathrm{TPR}(S_t)$ is an empirical measure of power for the template t.

6.2.4 Comparison with FDR control

The above experiment on the number of detections leads to a natural comparison based on the "BH region", that is the region obtained using the BH procedure that controls the FDR (= expected FDP). More precisely, we compare the size of the BH region to the size of FDP controlling regions. Conversely, we also compute FDP upper bounds on the BH region. This illustrates the difference between FDR control and FDP control with a concrete example.

6.2.5 Variation of the number of detections for low sample sizes

Because of the high cost of acquisition, many fMRI datasets comprise few subjects. This may lead to unstable behavior and limited statistical power. To study the impact of sample size on the inference procedure both at training and inference step, we perform two dual experiments. First, we compute the number of detections for the three possible methods as in Section 6.2.3 , with the difference that the template is learned using $n_{train} = 10$ subjects instead of $n_{train} = 113$. Second, we use the standard template with 113 subjects but this time infer on 25 pairs of fMRI contrasts with any number of subjects n_{infer}, varying from $n_{infer} = 8$ to $n_{infer} = 200$.

6.2.6 Sensitivity to the choice of training data

Since Notip requires learning a template on training data before performing inference, the choice of such data and its impact on the performance of the method is an important question. To assess this sensitivity quantitatively, we fix an fMRI contrast pair for inference. Then, we compare the number of detections for each template choice -as described in Section 6.2.3- using the 36 different fMRI contrast pairs as 36 different training sets for the Notip method. It should be noted that ARI and calibrated Simes do not depend on the chosen training set; their number of detections is computed once and for all. These 36 fMRI contrast pairs differ in several ways such as the number of subjects, the nature of the contrasts, the fMRI study or quantity of signal. This allows us to evaluate the robustness of Notip to poorly matched training and inference data. In this experiment, we also include the optimal template choice we used for all other experiments (i.e. least amount of signal and maximum number of subjects).

6.2.7 Influence of data smoothness

Another potential source of mismatch is the smoothing done in preprocessing of fMRI data. To assess the consequences on performance of a potential smoothing mismatch between training and inference data, we consider the case where the smoothing parameter FWHM is different in the training and inference data, using FHWM = 4mm for the training data and FWHM = 8mm for the inference data.

6.2.8 Using Notip on a single dataset

When learning a template on separate data is inconvenient, or to avoid the computational cost of learning the template, a natural idea is to use Notip on a single dataset. In such a setting, circularity biases may appear as in Meinshausen, 2006. The workaround that we propose to retain valid FDP control is to perform two independent rounds of randomization - one for training and one for inference. While this approach is formally not covered by the theoretical framework of Blanchard et al., 2020, we have performed experiments to assess its FDP control and power on the simulated data described in Section 6.2.2.

6.3 Results

6.3.1 Variation of the number of detections for different template types

A comparison of the number of detections obtained for the three possible methods at hand, i.e. ARI, calibrated Simes and Notip is displayed in Figure 6.3. To obtain this figure, we used 36 pairs of fMRI contrasts. The number n_{infer} of subjects ranged from 25 to 120 across inference contrast pairs.

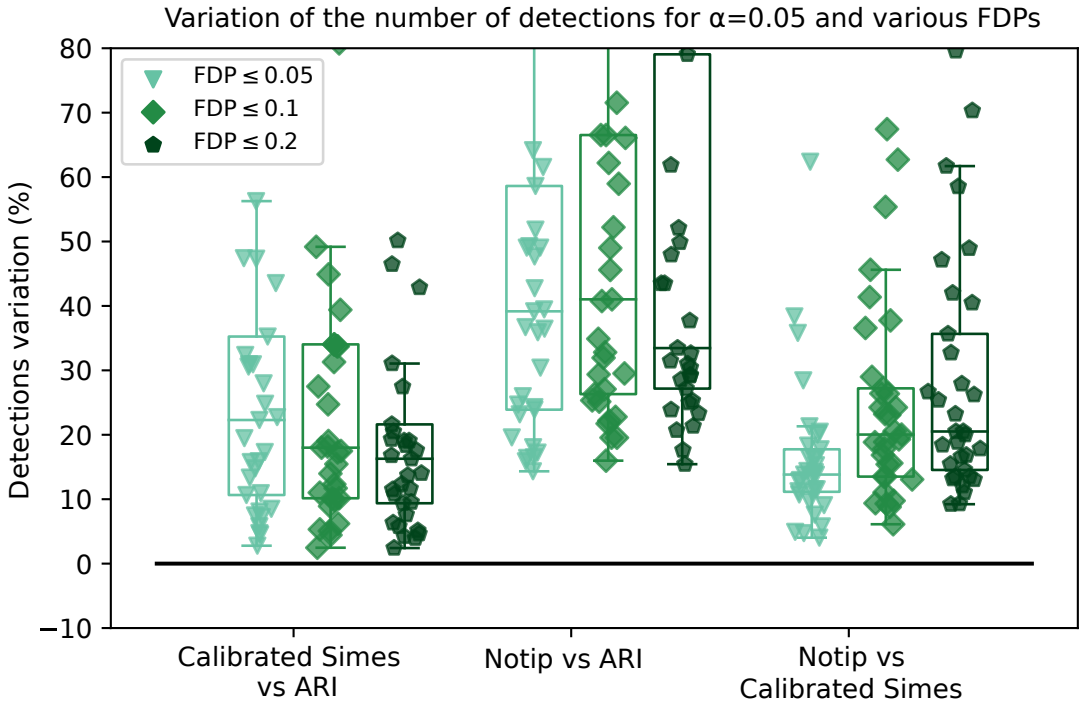


Figure 6.3: **Comparison of the number of detections between ARI, calibrated Simes and learned templates across 36 pairs of fMRI contrasts from Neurovault collection 1952.** After learning the template on a single contrast pair (see section 6.2), we perform inference on all 36 pairs. For each contrast pair, we compute the largest possible region that satisfies FDP$\leq q$ for $q \in \{0.05, 0.1, 0.2\}$ with risk level $\alpha = 0.05$.

In Figure 6.3, we notice that learned templates yield a substantial gain in detections compared to both other template choices for all target FDPs. On average, learned templates offer a $\sim$ **40% increase** in detections compared to the ARI method and a

$\sim$ **20% increase** compared to calibrated Simes. Gains in number of detections can vary largely across contrast pairs. This is essentially due to variance contained in the data, as all three methods exhibit similar TPR variability on simulated data (see Section 6.5.2). A concrete example of inference on fMRI data is shown in Figure 6.4.

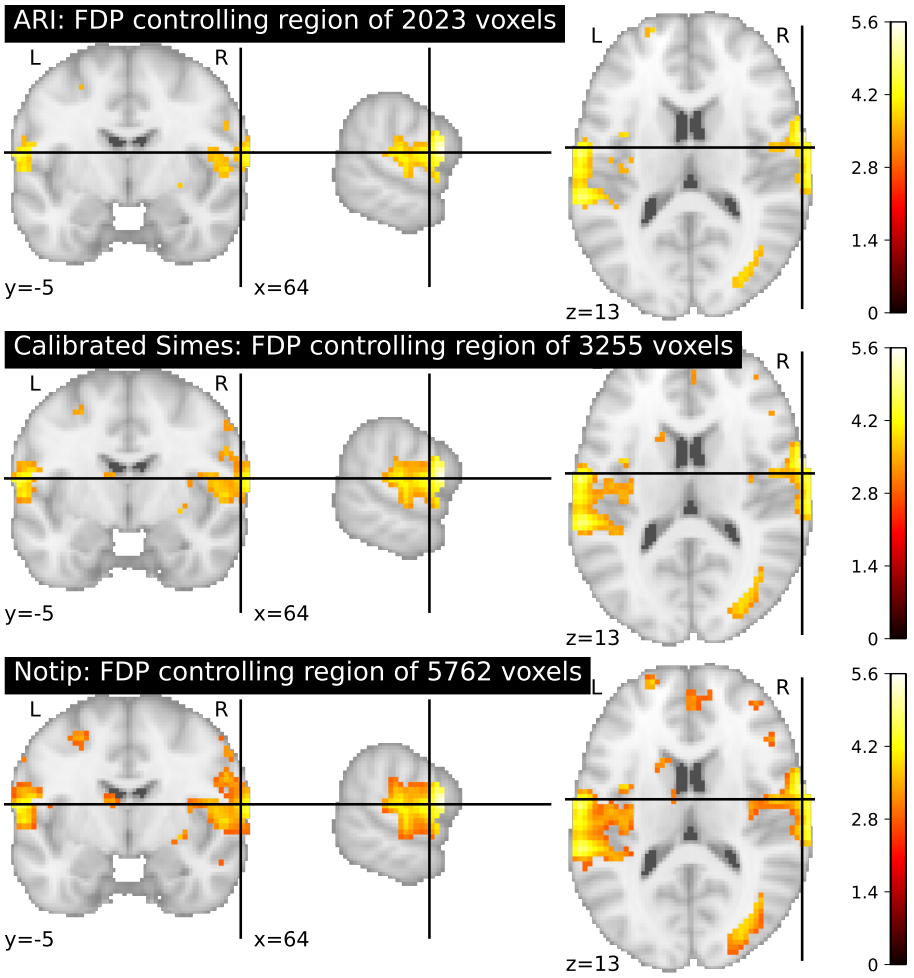


Figure 6.4: **Comparison of the number of detections between ARI, calibrated Simes and learned template on fMRI data.** For a pair of fMRI contrasts "look negative cue" vs "look negative rating" we compute the largest possible region such that FDP ≤ 0.1 with risk level $\alpha = 0.05$ for the three possible templates: ARI, calibrated Simes template and learned template. Notice that the number of detections is markedly higher (+ 77 %) using the learned template compared to the calibrated Simes template.

We have also performed the same experiment on simulated data. In this setting, we can report the actual TPR of the methods instead of region sizes. The empirical FDP for these simulations are reported in Figure 6.13.

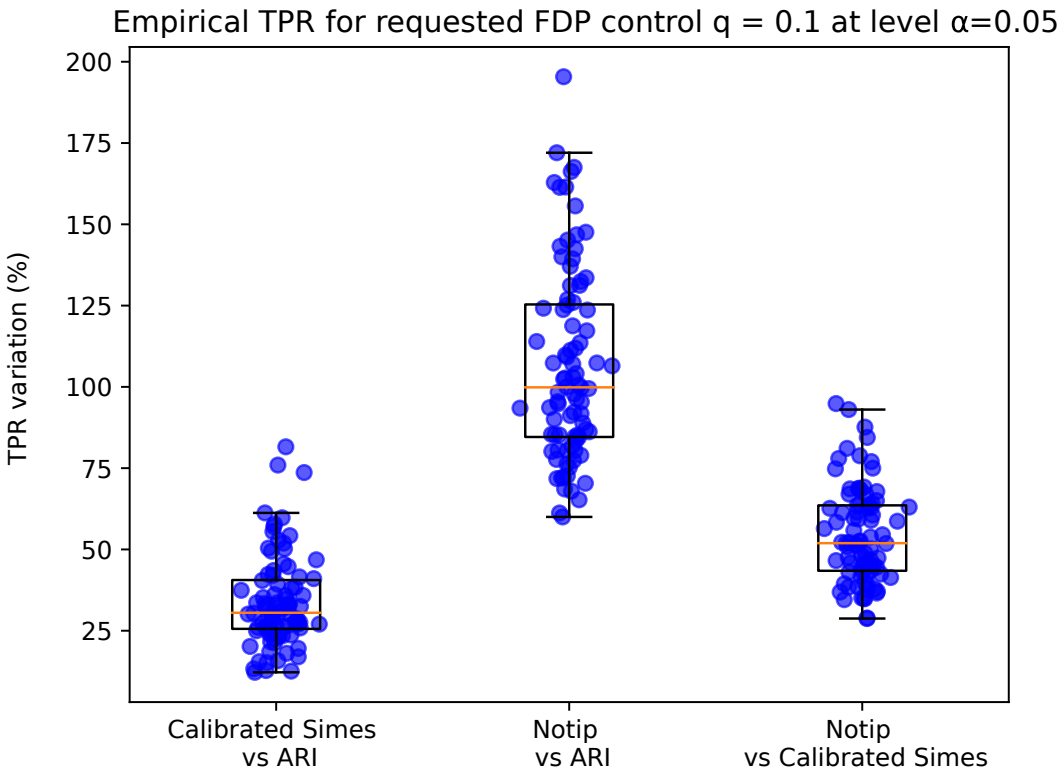


Figure 6.5: **TPR comparison for a FDP budget** $q = 0.1$ **at risk level** $\alpha = 0.05$. We run 100 simulations and report the TPR. Notice that Notip offers substantial gains in TPR compared to both ARI (100 % on average) and calibrated Simes (50 % on average).

Figure 6.5 illustrates the TPR gains achieved using Notip on simulated data compared to both ARI to both ARI (100 % on average) and calibrated Simes (50 % on average). Overall, simulations support the fact that Notip offers substantial performance gains compared to both ARI and calibrated Simes.

6.3.2 Comparison with FDR control

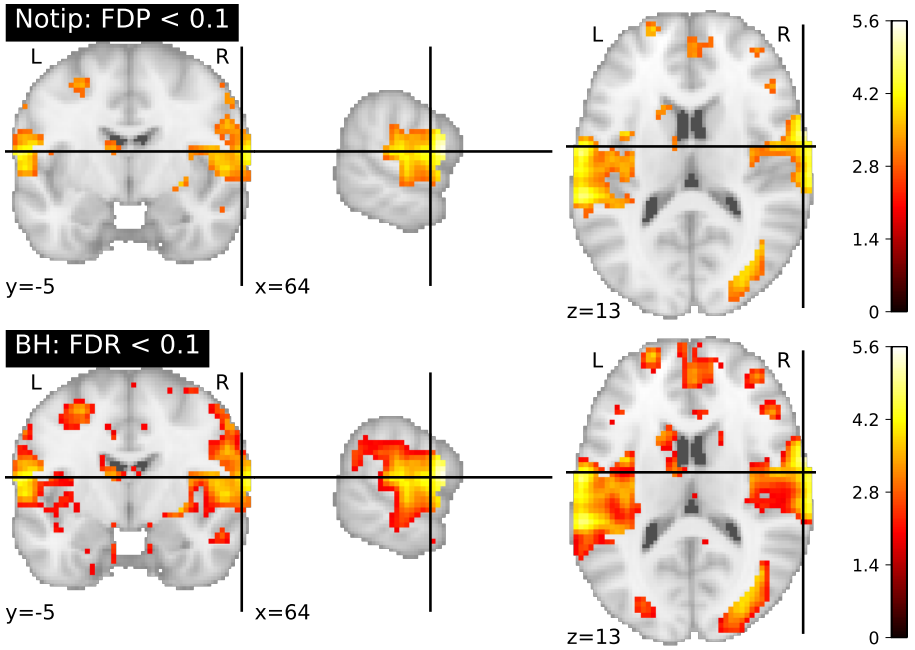


Figure 6.6: **Comparison of the number of detections between learned template and the BH procedure on fMRI data.** For a pair of fMRI contrasts "look negative cue" vs "look negative rating" we compute the largest possible region such that FDP ≤ 0.1 at risk level $\alpha = 0.05$ for the learned template and the largest possible region such that FDR ≤ 0.1 using the BH procedure. BH region size: 13814 voxels. Learned template region size: 5762 voxels.

Since FDR control is a much weaker guarantee than FDP control, it is expected that the BH procedure yields substantially more detections compared to FDP controlling procedures, as seen in Figure 6.6. However, FDP being the targeted guarantee, it is interesting to compute FDP upper bounds on the FDR controlling region yielded by BH. Concretely, we are trying to obtain a bound on the FDP of a region that only has a guarantee on its FDR. Table 6.1 shows the FDP upper bounds computed on the FDR controlling region using all three possible methods.

	ARI	Calibrated Simes	Notip
FDP Upper bound	61%	45%	25%

Table 6.1: **FDP upper bounds on the FDR controlling region obtained using the BH procedure (at level $q = 10\%$).** Notice that Notip yields smaller FDP bounds than ARI and calibrated Simes . This upper bound remains higher than the FDR guarantee (10%), which is more permissive by design.

Notip leads to a less conservative FDP upper bound than ARI and calibrated Simes. However, at risk level $\alpha = 5\%$, Notip is only able to guarantee that the FDP is less than 25% while the FDR is controlled at level 10%. This illustrates the difference between FDR control and FDP control, the latter being less permissive by design. Additionally, the guarantee offered by Notip is post hoc – such analysis would not be valid if we inverted

the roles of BH and Notip. While the BH procedure guarantees that the **expected** FDP is below 10%, Notip guarantees explicitly that the **actual** FDP is below 25% with high probability ($\geq 95\%$). It should be noted that on a single inference run, a guarantee on the **expected** FDP has no clear interpretation, whereas the guarantee on the **actual** FDP is directly interpretable.

6.3.3 Variation of the number of detections for low sample sizes

The above results demonstrate that data-driven templates yield consistent gains in number of detections over existing methods that offer the same guarantees. In this section we investigate whether these gains subsist in sub-optimal conditions. Namely, when the template is learned on very few subjects or if inference is done on experiments with few subjects. The first point is illustrated in Figure 6.7.

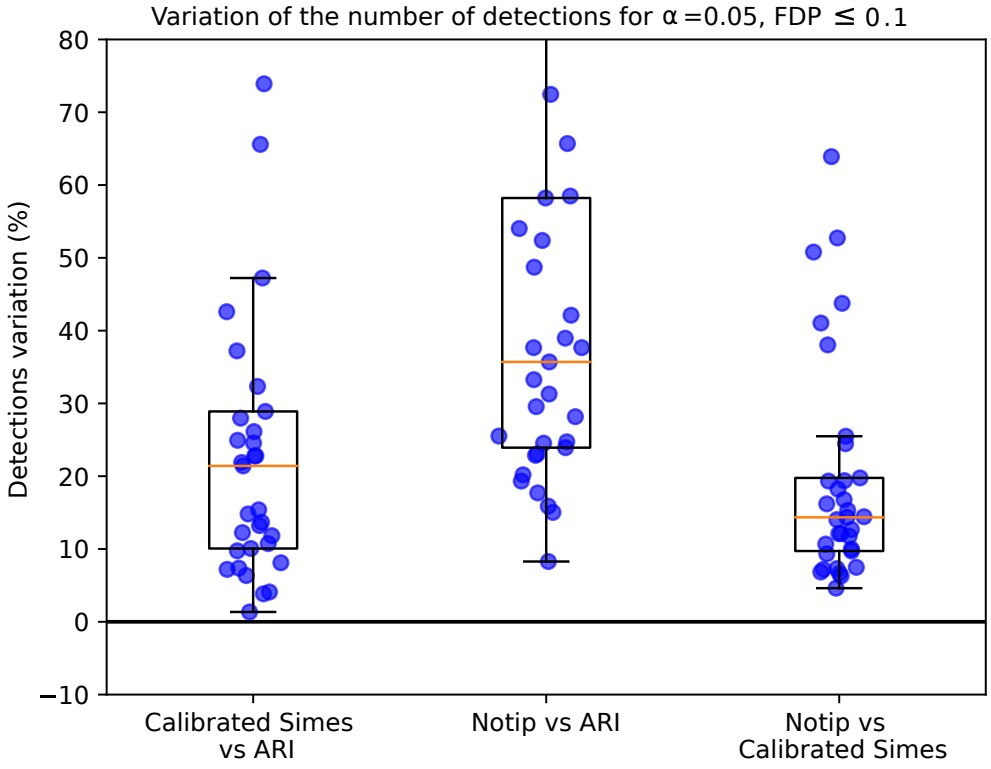


Figure 6.7: **Comparison of the number of detections between ARI, calibrated Simes and a learned template using a subsampled training set.** Here, the template is learned using $n_{train} = 10$ subjects instead of $n_{train} = 113$ subjects. Learned templates still perform better than the calibrated Simes template on average, but subsampling the training set leads to a sub-optimal number of detections , compared with Figure 6.3.

Unstable performance may occur when inferring on data with few subjects, even if the template is learned on a large number of subjects ($n_{train} = 113$ here). This is illustrated in Figure 6.8: gains in number of detections remain consistent- yet more variable for smaller sample sizes - across datasets with different number of subjects. As noted in Button et al., 2013, high variance is unavoidable when inferring on small datasets (e.g. $n_{infer} \leq 25$). For

a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.

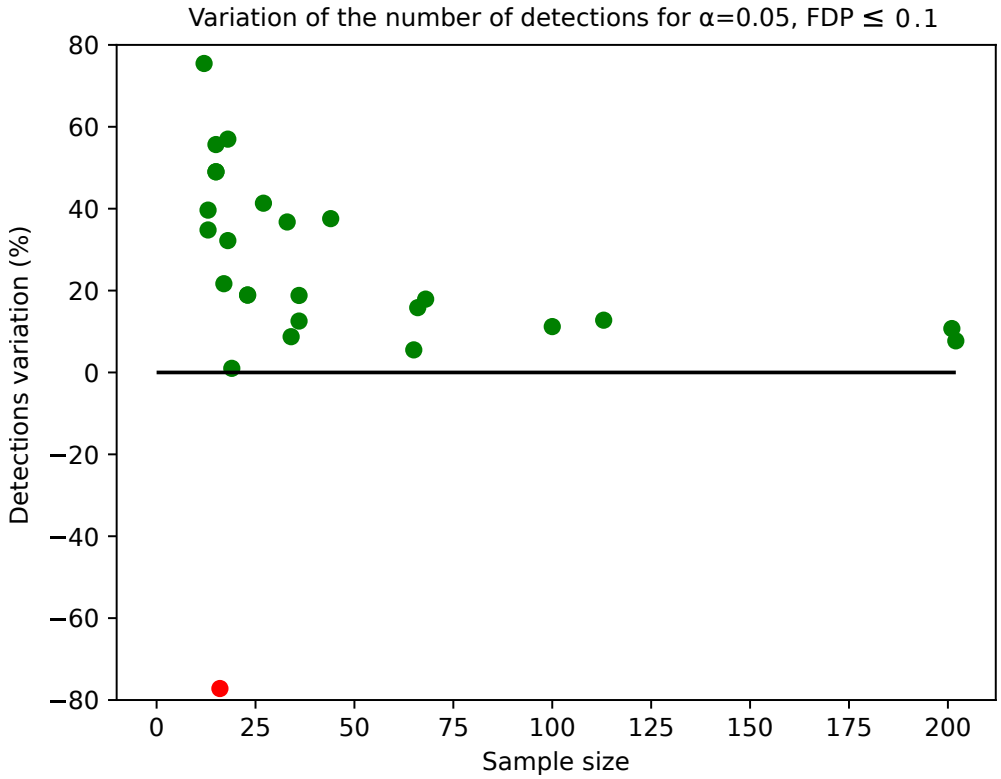


Figure 6.8: **Comparison of the number of detections between learned template and calibrated Simes for many contrast pairs with a different numbers of subjects.** The gains in number of detections remain consistent across datasets with different number of subjects. However, for a single dataset comprising 17 subjects, the learned template performs substantially worse than calibrated Simes.

6.3.4 Sensitivity to the choice of training data

Figure 6.9 displays the variation of the number of detections made by Notip compared to ARI and calibrated Simes using 36 different training sets. All training contrast pairs except one yield more detections than calibrated Simes, with gains ranging from 10% to 80%. This shows that the Notip procedure is robust to poorly matched training and inference data, since contrast pairs considered for training vary along many dimensions: number of subjects, nature of contrasts, fMRI study, quantity of signal... In the worst possible case, Notip performs marginally worse than calibrated Simes. Also note that the optimal template used in all other experiments (corresponding to the template learned from the training data with minimal signal and maximum number of subjects as described in Section 6.2.2) outperforms all other choices.

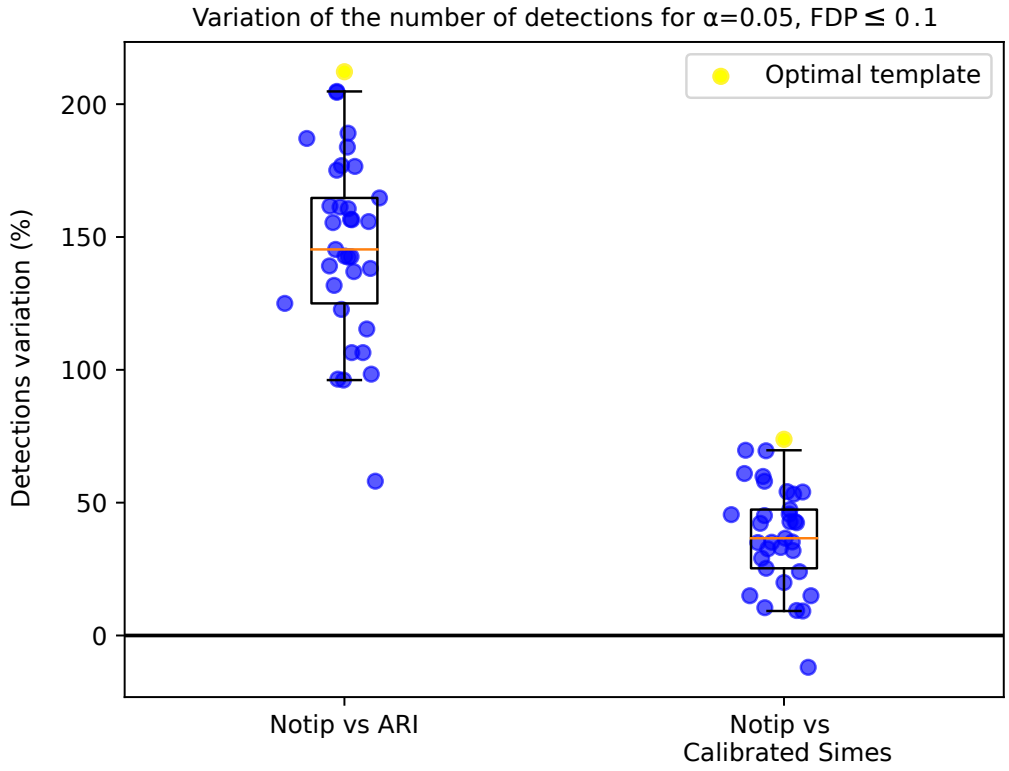


Figure 6.9: **Variation of the number of detections using many training sets**. For a fixed contrast pair "look negative cue" vs "look negative rating" and 36 different training contrast pairs, we compute the largest possible regions that ensure FDP ≤ 0.1 at risk level $\alpha = 0.05$. Note that for all training contrast pairs except one, Notip performs better than calibrated Simes, with gains ranging from 10 % to 80 % for the optimal template choice described in Section 4.1. In the worst case, Notip performs slightly worse than calibrated Simes.

6.3.5 Influence of data smoothness

We have seen in Figure 6.9 that Notip is robust to mismatches of training and inference data across different dimensions (number of subjects, quantity of signal...). We now examine the robustness of Notip with regards to a mismatch of the smoothing parameter between training and inference data.

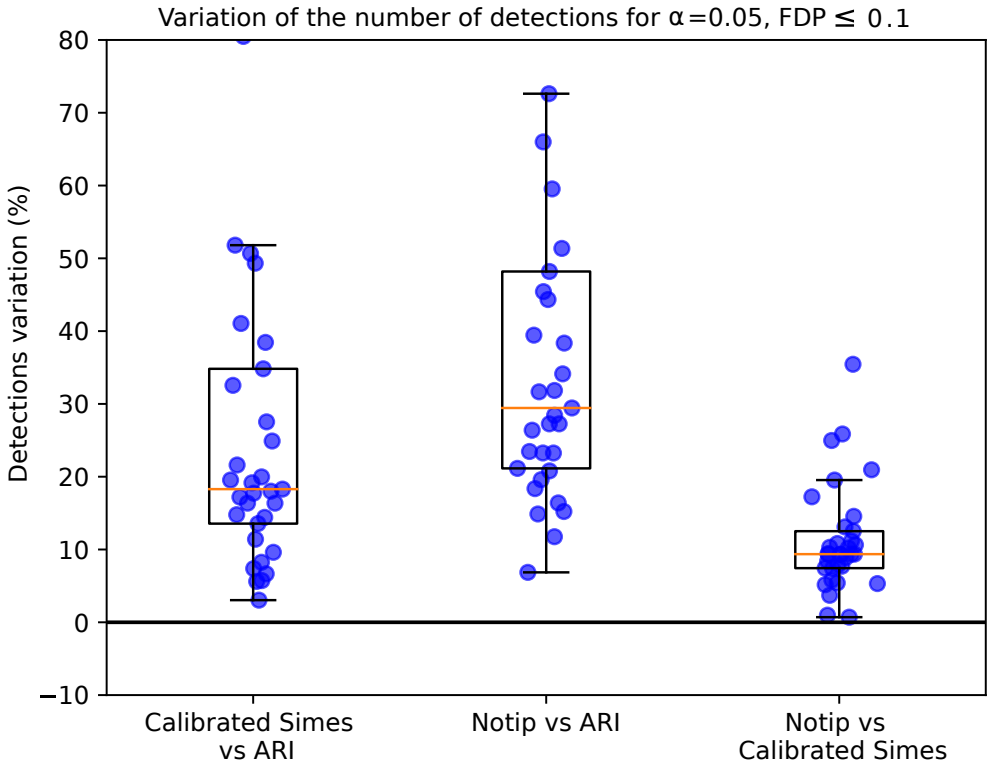


Figure 6.10: **An example of mismatch between the smoothing factors of training and inference data.** After learning the template on a single contrast pair (see Section 6.2) with smoothing full width at half maximum (FWHM) 4mm, we perform inference on all 36 pairs smoothed with FWHM 8mm. For each contrast pair, we compute the largest possible region that satisfies FDP control at level 0.1 with risk level $\alpha = 0.05$. The learned template still performs marginally better than calibrated Simes in this case, but gains are substantially lower in this regime.

Figure 6.10 shows that the smoothing parameter of the training data and the inference data should be matched for optimal performance. Otherwise performance gains relative to the calibrated Simes method are reduced, albeit still positive.

6.3.6 Using Notip on a single dataset

To assess whether using Notip with the same dataset for training and inference controls the FDP, and whether it yields performance gains compared to ARI and calibrated Simes, we performed 1000 simulations. For each of these runs, we report the empirical FDP and TPR of all three methods.

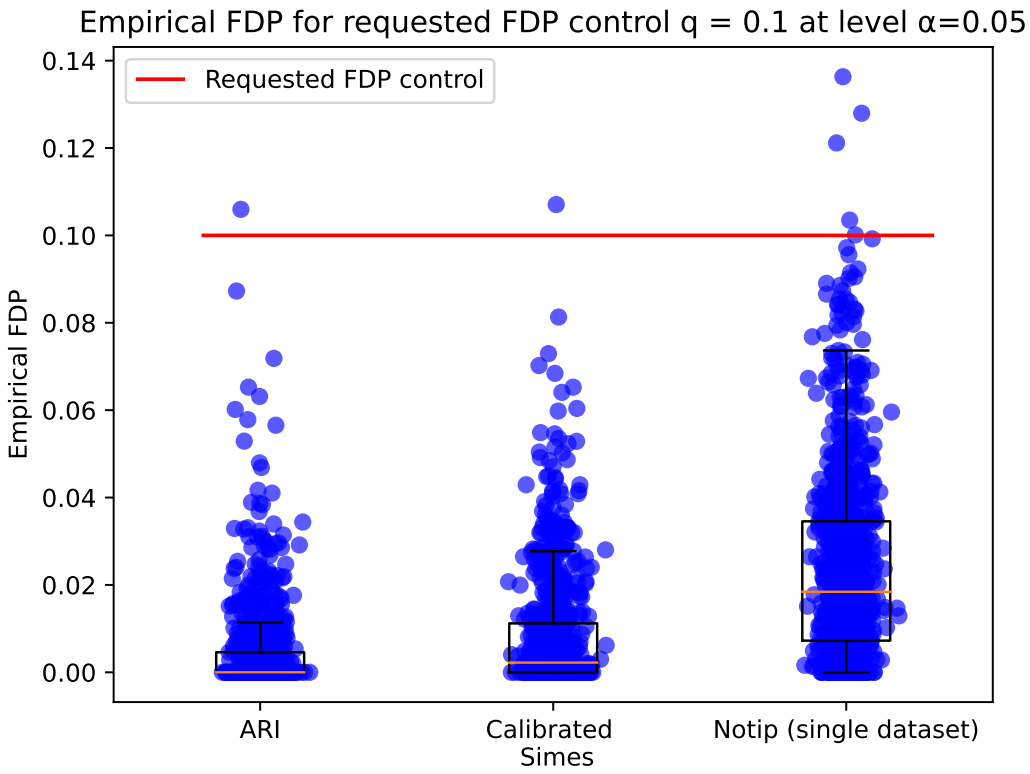


Figure 6.11: **False Discovery Proportion achieved for a FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$ **using Notip on a single dataset**. We run 1000 simulations and report the empirical FDP for each one. Notice that Notip (single dataset) controls the FDP at level $\alpha = 0.05$ since FDP control is violated for 5 runs, i.e. $0.5\% < 5\%$ of all simuations. As expected, ARI and calibrated Simes also control the FDP.

Notice that as seen in Figure 6.11, Notip (single dataset) indeed controls the FDP, as only 5 points are above the red line - i.e. the FDP was above the budget $q = 0.1$ in 0.5% of experiments ($< \alpha = 5\%$). As expected, ARI and calibrated Simes control the FDP more conservatively.

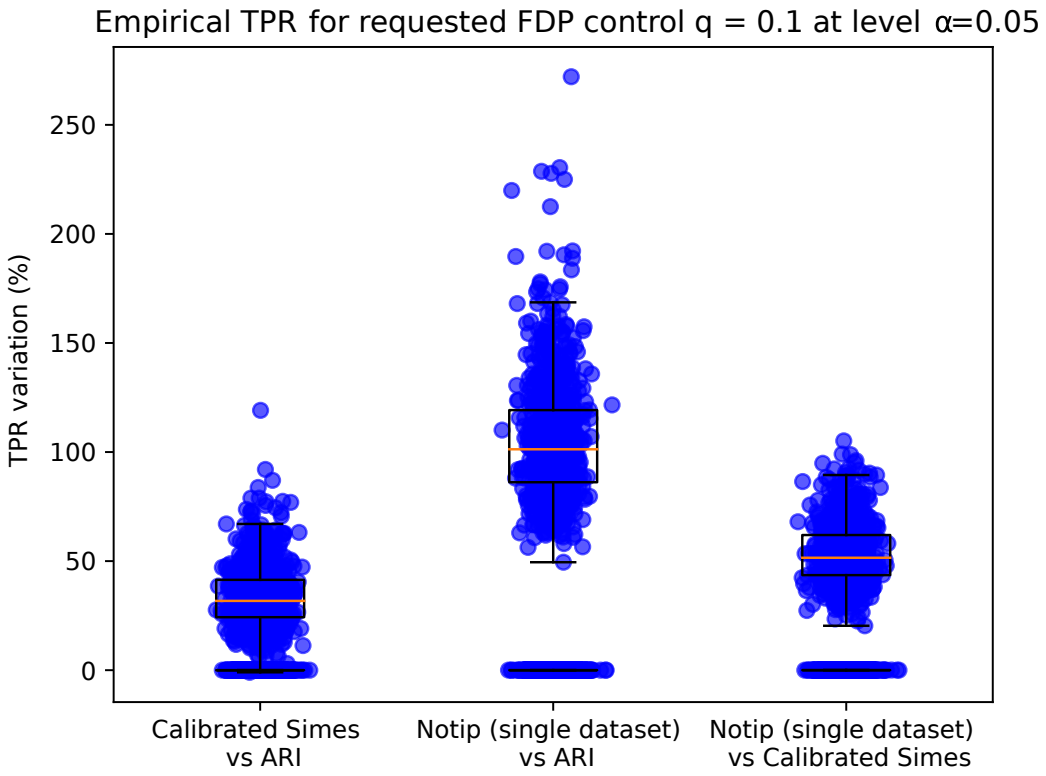


Figure 6.12: **TPR comparison for an FDP budget $q = 0.1$ with risk level $\alpha = 0.05$ using Notip on a single dataset**. We run 1000 simulations and report the empirical TPR for each one. Notice that Notip (single dataset) offers substantial performance gains compared to both ARI (100 % on average) and calibrated Simes (50 % on average).

As seen in Figure 6.12, Notip (single dataset) yields substantial performance gains compared to ARI and calibrated Simes: 50% on average compared to calibrated Simes, and 100% on average compared to ARI. These gains are comparable to those obtained using the classical Notip method on simulated data (see Figure 6.5).

6.4 Discussion

In this chapter, we have proposed the Notip procedure, that allows users to obtain statistical guarantees on the proportion of truly activated voxels in any given cluster. There are at least two ways to perform inference on fMRI data using this procedure. First, one can threshold a statistical map to obtain the largest possible region that satisfies a requested FDP control. Second, users can also obtain an upper bound on the FDP, or, equivalently, a lower bound on the TDP in any cluster of interest (see an example in Section 6.5.3).

This type of analysis is meant to mitigate the arbitrariness of cluster-forming thresholds in cluster-level inference, which remains a popular framework. The underlying observation is that estimates computed on these clusters may be plagued by circularity.

We have introduced a data-driven approach to obtain valid post hoc FDP control, thus achieving this goal. Moreover, controlling the FDP is a substantially more informative

guarantee than controlling the FDR, its expected value. We show that our procedure yields a higher number of detections than existing methods that offer the same statistical guarantees, namely ARI and calibrated Simes. We could go further by applying a step-down procedure as described in Blanchard et al., 2020, but the gains are expected to be marginal (Enjalbert-Courrech and Neuvial, 2022). Also, a noticeable feature of the proposed inference method is that it doesn't require valid p-values to maintain error guarantee.

The gains in detections are maintained across practically all possible training sets, even in cases of poor matching between the training and inference datasets, as seen in Figure 6.9. Figure 6.10 also illustrates the robustness of Notip, this time in the case of a poor match of smoothing parameters between the training and inference data. In this case, the gain in detections obtained by using the learned template is reduced, albeit still non-negligible (30% compared to ARI and 9% compared to calibrated Simes). We found that choosing training contrast pairs that contain a large number of subjects and low signal is optimal for performance. This is coherent with intuition since a large number of subjects and minimal signal allow a more stable and accurate estimation of the distribution of p-values under the null. Therefore, when selecting a template, it is useful to rely on a large-sample dataset with small signal magnitude.

In practice, learning a template ex ante can be inconvenient or simply impossible, for instance when users only have a single dataset at hand. We have shown numerically that it is possible (though not formally supported by the theory) to use Notip on a single dataset. This simplifies the procedure and removes the cumbersome choice of an external dataset to learn the template.

Notip comes with an additional computational cost compared to classical calibration using the Simes template, since we have to learn the template before inference. Generally, this additional cost is acceptable in practice since template learning and inference have the same computation complexity.

We have used $10,000$ permutations for better resolution when learning the template instead of the typical $1,000$ permutations used at the inference step. Learning a template using $B_{train} = 10,000$ permutations with a standard laptop (on a single thread) takes around 7 minutes, while inferring on a contrast pair (using $B_{infer} = 1,000$ takes around 45 seconds). This can be trivially parallelized, as natively done in the implementation we propose.

A current limitation of the proposed method is that it only handles one-sample or two-sample designs. This method could be extended to multivariate linear models in future work.

The idea of learning templates is not specific to fMRI data and could also be used on other types of data on which the calibration procedure is useful such as genomics (Enjalbert-Courrech and Neuvial, 2022).

We have achieved the goal of obtaining valid post hoc FDP control - rather than FDR control, or even weaker guarantees on clusters - while maintaining a satisfactory number of detections. This allows users in the brain imaging community to use more reliable inference methods that provide robust guarantees, avoiding circularity biases. The efforts to build such methods appear to us as important goal for the brain imaging field. The Python code used in this chapter is available at `https://github.com/alexblnn/Notip`. This code relies on the sanssouci package available at `https://github.com/pneuvial/sanssouci.python`.

6.5 Additional experimental results

6.5.1 FDP control on simulated data

In section 4.2 we report the empirical TPR for experiments on simulated data, for which the ground truth is known. We also compute the FDP for each simulation run to verify that, as expected, Notip indeed controls the FDP.

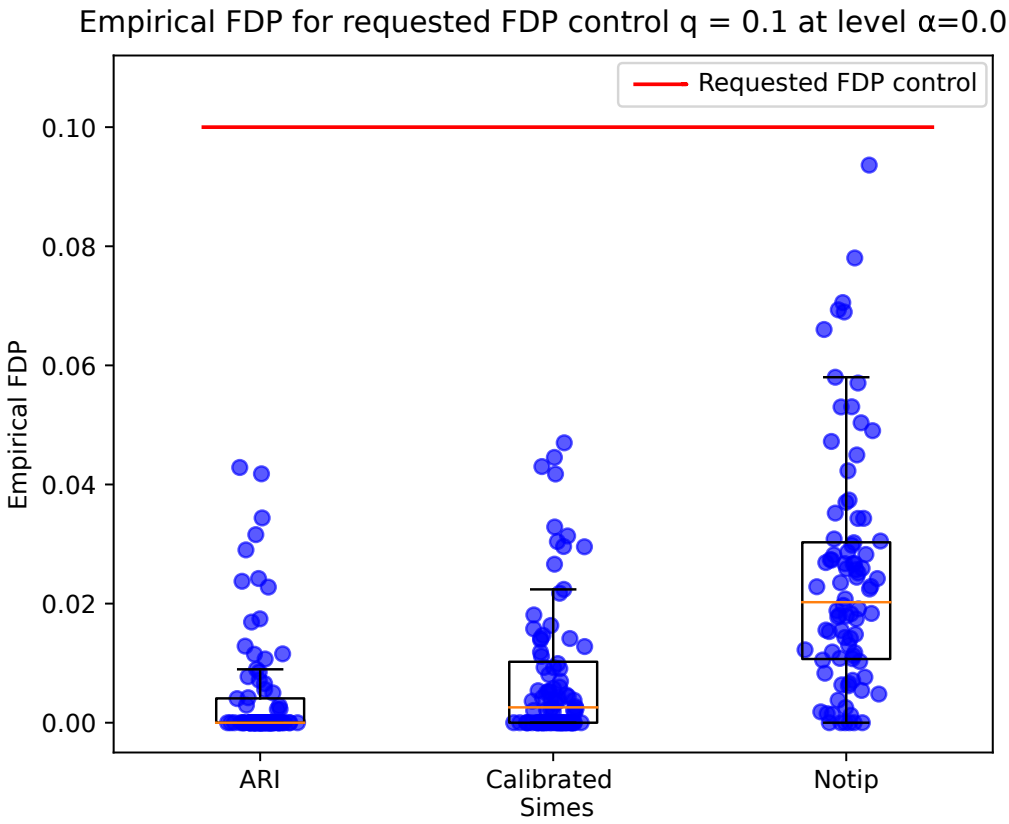


Figure 6.13: **False Discovery Proportion achieved for a FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$. We run 100 simulations and report the empirical FDP for each one. All three methods control the FDP, but Notip is less conservative than ARI and Calibrated Simes.

6.5.2 Variability of Notip

We have observed relatively high variability in number of detections when comparing Notip to ARI and calibrated Simes in Figure 6.3. One may wonder whether this variability is inherent to the Notip procedure or stems from the data. To assess this, we report the empirical TPR of each method (rather than the 3 pairwise comparisons) on simulated data, in the same setup as in Figure 6.5.

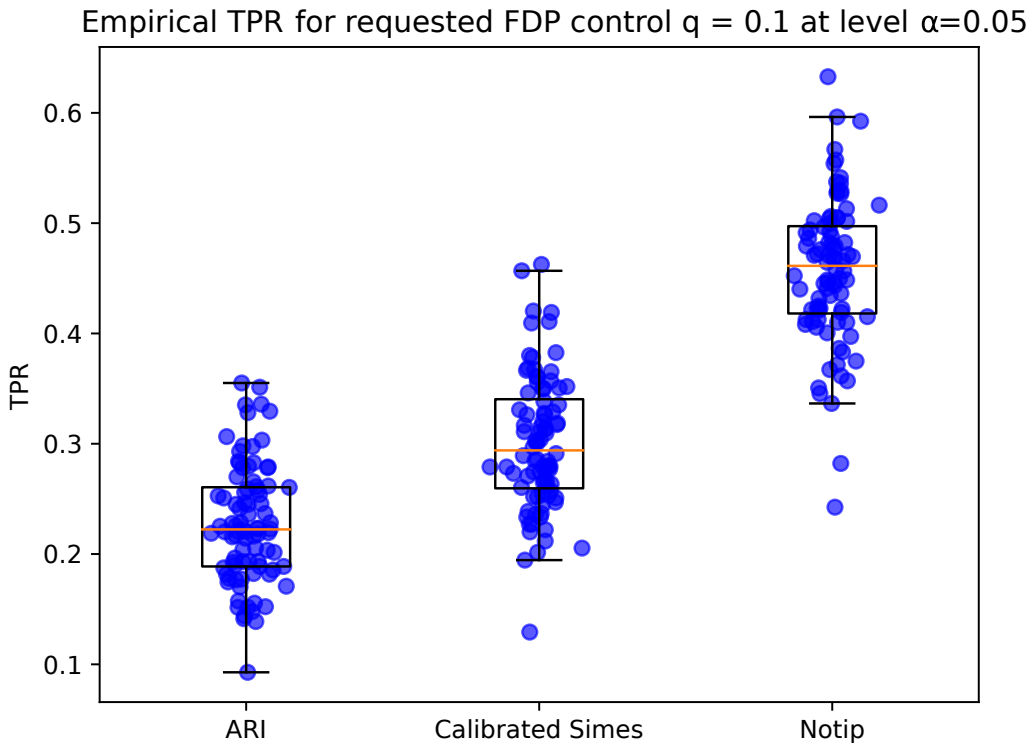


Figure 6.14: **TPR comparison for an FDP budget** $q = 0.1$ **with risk level** $\alpha = 0.05$. We run 100 simulations and report the empirical TPR for each one. Notice that the variability of performance is similar for all three methods.

Figure 6.14 indicates that all three methods exhibit similar performance variability on simulated data. This suggests that the variability observed in Figure 6.3 is due to the data itself rather than to the Notip method.

6.5.3 TDP lower bounds on clusters

Throughout the chapter, we chose to focus on FDP upper bounds - and thus on FDP controlling regions - to make Notip comparable with other methods that control the FDR or the FWER. Since Notip is a post hoc method, it can also be used for inference on data-driven clusters. In this setting, it is natural to formulate the results in terms of TDP lower bounds (obtained as 1 - FDP upper bounds), since users generally want a positive guarantee when inferring on clusters. This is illustrated in Table 6.2. Notice that Notip is able to offer less conservative guarantees on the TDP in all clusters than both ARI and calibrated Simes. In Table 6.3 we retain 3 clusters among the 9 found in Table 6.2 for further study, i.e. changing the cluster-forming threshold to assess its impact on performances of all three methods. In Tables 6.4 ($z > 2.5$) and 6.5 ($z > 3.5$), notice that the same clusters are detected with varying sizes. The TDP guarantees remain less conservative using Notip than both ARI and calibrated Simes when the cluster-forming threshold is either lowered to 2.5 or upped to 3.5.

6.5.4 An example of simulated data

Here is an example of simulated data computed in 2D for clarity. We use 3D images in the experiments to mimick fMRI data. Here, we use a 10×10 2D grid and generate the ground truth, a binary mask that defines the signal. Then, we generate n_{infer} null images and n_{infer} images that comprise signal. Substracting these two sets of images results in a list of n_{infer} one-sample images, as in fMRI experiments. In Figure 6.15 an example of simulated ground truth is displayed, while Figure 6.16 shows an example of simulated one-sample image. Figure 6.16 is a noisy version of the ground truth shown in Figure 6.15.

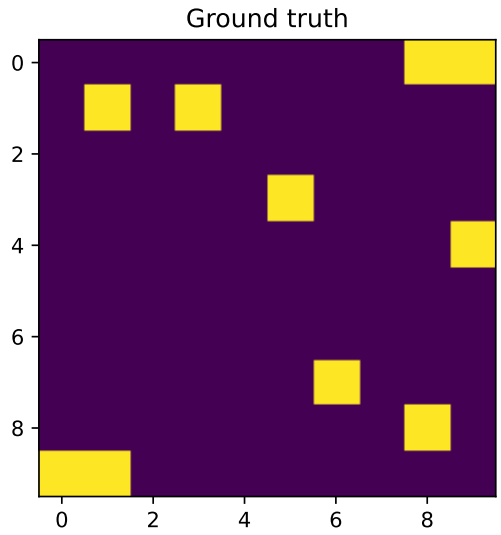


Figure 6.15: **Simulated ground truth.** This binary mask locates the simulated signal on a 2D 10×10 grid. Signal locations have been drawn randomly and account for $(1 - \pi_0)\%$ of the image, the rest of the image being null data.

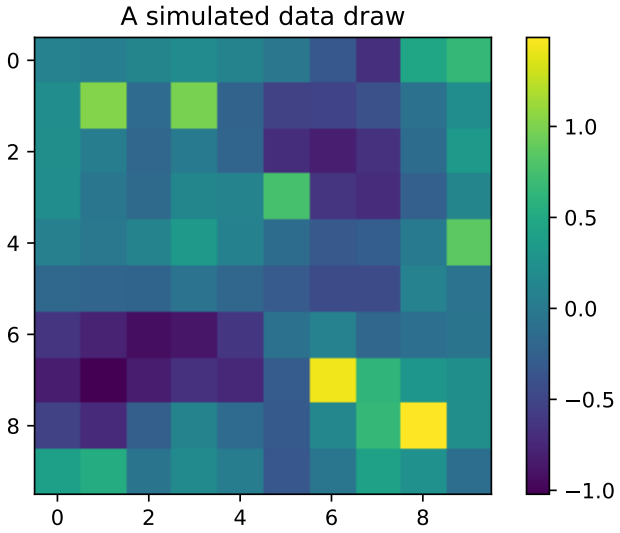


Figure 6.16: **A simulation draw.** This 2D 10×10 grid represents a draw of one-sample image comprising signal at locations determined by the binary mask shown in Figure 6.15. This is a typical example of input data in experiments on simulated data; the goal is then to recover the binary mask using inference methods such as Notip, ARI or calibrated Simes.

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	-33.0	-94.0	-17.0	5.63	7695	0.17	0.24	0.26
1a	-45.0	-79.0	-26.0	4.56				
1b	-48.0	-61.0	-26.0	4.13				
1c	-51.0	-64.0	-35.0	4.08				
2	66.0	2.0	16.0	5.47	14877	0.20	0.33	0.45
2a	69.0	-22.0	10.0	4.67				
2b	69.0	-10.0	13.0	4.59				
2c	69.0	-28.0	13.0	4.43				
3	-12.0	-82.0	-8.0	5.40	14445	0.27	0.38	0.50
3a	30.0	-73.0	-8.0	4.96				
3b	-24.0	-61.0	-11.0	4.91				
3c	30.0	-46.0	-11.0	4.64				
4	-6.0	11.0	52.0	5.30	5238	0.14	0.25	0.29
4a	6.0	8.0	55.0	4.19				
5	45.0	14.0	25.0	5.27	4563	0.24	0.30	0.30
5a	48.0	29.0	13.0	3.36				
6	12.0	-43.0	-26.0	5.08	12555	0.05	0.17	0.35
6a	0.0	-64.0	-14.0	4.43				
6b	3.0	-55.0	-11.0	4.26				
6c	3.0	-16.0	-32.0	4.23				
7	39.0	-73.0	4.0	5.00	6075	0.04	0.09	0.17
7a	39.0	-64.0	16.0	4.44				
7b	30.0	-82.0	10.0	4.42				
7c	27.0	-67.0	34.0	3.63				
8	-63.0	-34.0	16.0	4.95	25812	0.30	0.48	0.66
8a	-63.0	-10.0	13.0	4.90				
8b	-27.0	-19.0	4.0	4.85				
8c	-57.0	-19.0	7.0	4.68				
9	36.0	-94.0	-8.0	4.75	6507	0.08	0.15	0.17
9a	48.0	-70.0	-32.0	3.96				
9b	45.0	-70.0	-23.0	3.92				
9c	33.0	-82.0	-29.0	3.77				

Table 6.2: **Cluster localization ($z > 3$), size, peak statistic and TDP lower bound at risk level $\alpha = 5\%$** using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant. This table can be generated using script https://github.com/alexblnn/Notip/blob/master/scripts/table_2.py.

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	14877	0.20	0.33	0.45
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	14445	0.27	0.38	0.50
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	25812	0.30	0.48	0.66
3a	-63.0	-10.0	13.0	4.90				
3b	-27.0	-19.0	4.0	4.85				
3c	-57.0	-19.0	7.0	4.68				

Table 6.3: **Cluster localization ($z > 3$), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant. Notice that we retained 3 clusters (originally of indices 2, 3 and 8 of Table 6.5.3).

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	28593	0.13	0.18	0.29
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	23355	0.19	0.23	0.35
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	43092	0.19	0.25	0.42
3a	-63.0	-10.0	13.0	4.90				
3b	-27.0	-19.0	4.0	4.85				
3c	-57.0	-19.0	7.0	4.68				

Table 6.4: **Cluster localization ($z > 2.5$), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant.

						True Discovery Proportion		
Cluster ID	X	Y	Z	Peak Stat	Cluster Size (mm3)	ARI	Calibrated Simes	Notip
1	66.0	2.0	16.0	5.47	7425	0.38	0.48	0.69
1a	69.0	-22.0	10.0	4.67				
1b	69.0	-10.0	13.0	4.59				
1c	69.0	-28.0	13.0	4.43				
2	-12.0	-82.0	-8.0	5.40	8397	0.46	0.53	0.73
2a	30.0	-73.0	-8.0	4.96				
2b	-24.0	-61.0	-11.0	4.91				
2c	30.0	-46.0	-11.0	4.64				
3	-63.0	-34.0	16.0	4.95	9585	0.46	0.55	0.76
3a	-63.0	-10.0	13.0	4.90				
3b	-57.0	-19.0	7.0	4.68				
3c	-60.0	-49.0	25.0	4.59				

Table 6.5: **Cluster localization ($z > 3.5$), size, peak statistic and TDP lower bound at risk level** $\alpha = 5\%$ using the three possible templates (ARI, Calibrated Simes and Notip) on contrast pair 'look negative cue vs look negative rating'. Cluster subpeaks are also reported when relevant.

Study	Contrast 1	Contrast 2	$n_{subjects}$
HCP	shapes vs baseline	faces vs baseline	66
HCP	right hand vs baseline	right foot vs baseline	67
HCP	right foot vs baseline	left foot vs baseline	66
HCP	left hand vs baseline	right foot vs baseline	67
HCP	left hand vs baseline	left foot vs baseline	66
HCP	tool vs baseline	face vs baseline	68
HCP	face vs baseline	body vs baseline	68
HCP	tool vs baseline	body vs baseline	68
HCP	body vs baseline	place vs baseline	68
amalric2012mathematicians	equation vs baseline	number vs baseline	29
amalric2012mathematicians	house vs baseline	word vs baseline	37
amalric2012mathematicians	house vs baseline	body vs baseline	27
amalric2012mathematicians	equation vs baseline	word vs baseline	29
amalric2012mathematicians	visual calculation vs baseline	auditory sentences vs baseline	27
amalric2012mathematicians	auditory right motor vs baseline	visual calculation vs baseline	25
cauvet2009muslang	c16 music vs baseline	c02 music vs baseline	35
cauvet2009muslang	c16 language vs baseline	c01 language vs baseline	35
cauvet2009muslang	c02 language vs baseline	c16 language vs baseline	35
cauvet2009muslang	c04 language vs baseline	c16 language vs baseline	35
amalric2012mathematicians	face vs baseline	scramble vs baseline	85
ds107	scramble vs baseline	objects vs baseline	44
ds107	consonant vs baseline	scramble vs baseline	47
ds107	consonant vs baseline	objects vs baseline	44
ds108	reapp negative rating vs baseline	reapp negative cue vs baseline	32
ds108	look negative stim vs baseline	look negative rating vs baseline	34
ds108	reapp negative stim vs baseline	reapp negative rating vs baseline	34
ds109	false photo story vs baseline	false photo question vs baseline	36
ds109	false belief story vs baseline	false photo story vs baseline	36
ds109	false belief question vs baseline	false photo question vs baseline	36
ds109	false belief story vs baseline	false belief question vs baseline	36
ds109	false belief question vs baseline	false photo story vs baseline	36
pinel2007fast	visual right motor vs baseline	vertical checkerboard vs baseline	113
pinel2007fast	auditory right motor vs baseline	visual right motor vs baseline	121
ds107	scramble vs baseline	face vs baseline	85
amalric2012mathematicians	house vs baseline	scramble vs baseline	85
ds107	words vs baseline	face vs baseline	100

Table 6.6: **36 pairs of fMRI contrasts used for experiments.** These contrasts images have been downloaded from Neurovault 1952 collection.

Chapter 7

False Discovery Proportion control for aggregated Knockoffs

Summary. Controlled variable selection is an important analytical step in various scientific fields, such as brain imaging or genomics. In these high-dimensional data settings, considering too many variables leads to poor models and high costs, hence the need for statistical guarantees on false positives. Knockoffs are a popular statistical tool for conditional variable selection in high dimension. However, they control for the expected proportion of false discoveries (FDR) and not their actual proportion (FDP). We present a new method, KOPI, that controls the proportion of false discoveries for Knockoff-based inference. The proposed method also relies on a new type of aggregation to address the undesirable randomness associated with classical Knockoff inference. We demonstrate FDP control and substantial power gains over existing Knockoff-based methods in various simulation settings and achieve good sensitivity/specificity tradeoffs on brain imaging and genomic data.

Contents

7.1 Background

A major caveat of the Knockoffs procedure described in Section 3.2 is the random nature of the Knockoffs generation process: for two runs of the Knockoffs procedure on the same data, different Knockoffs will be built and subsequently different variables may be selected. This undesirable behavior hinders reproducibility. A second caveat is that False Discovery Rate (FDR) control does not imply False Discovery Proportion (FDP) control as shown in Figure 2.1. This leads to potentially unreliable inference: single runs of the method can produce a much higher proportion of False Discoveries than the chosen FDR level.

In this chapter, we propose a novel Knockoff-based inference procedure that addresses both concerns while offering power gains over existing methods, for no significant computation cost. The chapter is organized as follows. We first discuss existing work on Knockoffs aggregation before moving on to the main contribution. Using the symmetry of knockoffs under the null hypothesis, we construct explicit upper bounds on the JER of these statistics, leading to FDP control. We then use the calibration principle of Blanchard et al., 2020, to obtain sharper bounds. Finally, we obtain a robust version of this method using harmonic mean aggregation of the π statistics across multiple Knockoffs draws.

We demonstrate empirical power gains in various simulation settings and show the practical benefits of the proposed method for conditionally important region identification on fMRI and genomic datasets.

7.2 Related work

There has been much effort in the statistical community to achieve derandomized Knockoff-based inference. Ren et al., 2021, introduced the idea of running Model-X Knockoffs (Candès et al., 2018) multiple times and computing for each the proportion of runs for which it was selected. Gimenez and Zou, 2019, explore the idea of sampling multiple Knockoffs simultaneously. This induces a massive computational cost, which is prohibitive compared to methods that can support parallel computing. Nguyen et al., 2020, introduced an aggregation method that relies on viewing Model-X Knockoffs as a Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) on so-called *intermediate p-values*. Such p-values can be computed on different Knockoff runs and aggregated using quantile aggregation (Meinshausen et al., 2009) – then, BH is performed on the aggregated p-values to select variables. This approach relies on the heavy assumption that Knockoff statistics are i.i.d. under the null. Additionally, it is penalized by the conservativeness of the quantile aggregation scheme. Alternative aggregation schemes such as the harmonic mean (Wilson, 2019) can be used but do not yield valid p-values.

Ren and Barber, 2022, introduced an alternative aggregation procedure where Model-X Knockoffs are viewed as an e-BH procedure (Wang and Ramdas, 2022) on well-defined e-values (Vovk and Wang, 2021). Since the mean of two e-values remains an e-value, aggregation is done by averaging e-values across different Knockoffs draws. Then, e-BH is performed on the aggregated e-values to select variables. FDR control on aggregated Knockoffs is achieved without any additional assumption compared to Model-X Knockoffs. However, this method requires the difficult setting of a hyperparameter related to the chosen risk level, which highly impacts power in practice. Other recent developments in Knockoffs include the conditional calibration framework of Luo et al. (2022) which aims at improving the power of Knockoffs-based methods.

There have been a few attempts at controlling other type 1 errors than the FDR using Knockoffs. Janson and Su, 2016, achieves k-FWER control and proposes that FDP control

can be obtained by using a procedure that leverages joint k-FWER control. Recently, Li et al., 2022, introduced such a procedure to reach FDP control based on the k-FWER control introduced in Janson and Su, 2016. In summary, the KOPI approach is the first one that aims at controlling the FDP of knockoffs-based inference for any aggregation scheme, leading to both accurate FDP control and increased sensitivity.

7.3 Main contribution: FDP control for aggregated Knockoffs

7.3.1 Joint distribution of π statistics under the null

To obtain FDP control, we rely on JER control as explained in Section 3.1. The JER associated to the π-statistics is written with the usual notations:

$$\mathrm{JER}(\mathbf{t}) = \mathbb{P}\left(\exists j \in [\![k_{max} \wedge p_0]\!] : \pi_{(j:\mathcal{H}_0)} < t_j\right).$$

In the remainder of this section, we show how to obtain JER control for π statistics. Notice in Equation 3.5 that the JER of a given threshold family only depends on the joint null distribution of the π statistics. As for earlier FDR control (Barber and Candès, 2015) or k-FWER control (Janson and Su, 2016) results, the key idea to obtain JER control for π statistics is to prove that the relevant part of this distribution is in fact known, thanks to the properties of knockoff statistics. We use the same notation as in Janson and Su, 2016. Letting $Z_j = |\{k \in [\![p]\!] : W_k \leq -W_j\}|$ and $\chi_j = sign(W_j)$, the π statistics $(\pi_j)_{j=[\![p]\!]}$ are given by:

$$\pi_j = \frac{1+Z_j}{p} 1_{\{\chi_j=1\}} + 1_{\{\chi_j=-1\}}.$$

For a given $\mathbf{W}$, let $\sigma(\mathbf{W})$ be a permutation of $[\![p]\!]$ that sorts $\mathbf{W}$ by decreasing modulus: $\sigma(\mathbf{W}) = (\sigma_1, \ldots, \sigma_p)$ such that $|W_{\sigma_1}| \geq |W_{\sigma_2}| \cdots \geq |W_{\sigma_p}|$. We start by proving that the Z statistics can be expressed as a function of the vector of χ statistics:

Lemma 1. *For $j \in [\![p]\!]$ such that $\chi_{\sigma_j} = 1$, $Z_{\sigma_j} = \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}$.*

Proof of Lemma 1. Since $\chi_{\sigma_j} = 1$, we have:

$$\begin{aligned}
Z_{\sigma_j} &= |\left\{k \in [\![p]\!] : W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ and } W_{\sigma_k} \leq -W_{\sigma_j}\right\}| \\
&= |\left\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ and } |W_{\sigma_k}| \leq |W_{\sigma_j}|\right\}| \\
&= |\{k \in [\![p]\!] : W_{\sigma_k} < 0 \text{ and } k \leq j\}| \\
&= \sum_{k=1}^{j-1} 1_{\{\chi_{\sigma_k}=-1\}}.
\end{aligned}$$

□

Lemma 1 implies that the distribution of order statistics of $\pi|\sigma(\mathbf{W})$ is entirely determined by that of $\chi|\sigma(\mathbf{W})$. To formalize this, we introduce π^0 statistics.

Definition 12 (π^0 statistics)**.** Let $\chi^0 = (\chi^0_j)_{1\le j\le p}$ be a collection of p i.i.d. Rademacher random variables, that is, for all j, $\mathbb{P}(\chi^0_j = 1) = \mathbb{P}(\chi^0_j = -1) = 1/2$. The associated π^0 statistics are defined for $j \in [\![p]\!]$ by

$$\pi^0_j = \frac{1 + Z^0_j}{p} 1_{\{\chi^0_j = 1\}} + 1_{\{\chi^0_j = -1\}}, \text{ where } Z^0_j = \sum_{k=1}^{j-1} 1_{\{\chi^0_k = -1\}}. \tag{7.1}$$

Theorem 3. *Let* $\mathbf{t}$ *be a threshold family of length* k_{max}*. Then, for* $\pi^0 = (\pi^0_j)_{j\in[\![p]\!]}$ *as in* (7.1)*,*

$$\mathrm{JER}(\mathbf{t}) \le \mathrm{JER}^0(\mathbf{t}) := \mathbb{P}\left(\exists k \in [\![k_{max}]\!] : \pi^0_{(k)} < t_k\right). \tag{7.2}$$

Proof of Theorem 3. Let $k \in [\![k_{max}]\!]$. Since $t_k \le 1$, we have $\pi_{(k:\mathcal{H}_0)} < t_k$ if and only if $N_k \ge k$, where

$$N_k = \left|\left\{j \in \mathcal{H}_0, \chi_j = 1 \text{ and } \frac{1 + Z_j}{p} < t_k\right\}\right|.$$

With the notation of Definition 12, we define the random variable

$$N^0_k = \left|\left\{j \in \mathcal{H}_0, \chi^0_j = 1 \text{ and } \frac{1 + Z^0_j}{p} < t_k\right\}\right|.$$

If $\mathcal{H}_0 = [\![p]\!]$, then Lemma 1 implies that conditional on $\sigma(\mathbf{W})$, N_k and N^0_k have the same distribution. Indeed, the vectors $(W_j)_{j/\chi_j=1}$ and $(Z_j)_{j/\chi_j=1}$ have the same ordering, and conditional on $\sigma(\mathbf{W})$, $(\chi_j)_{j\in\mathcal{H}_0}$ are jointly independent and uniformly distributed on $\{-1, 1\}$ (Lemma 2.1 in Janson and Su, 2016; Barber and Candès, 2015). Using the same argument as in the proof of Lemma 3.1 in Janson and Su (2016), in the case where $\mathcal{H}_0 \subsetneq [\![p]\!]$, false null χ_j will insert -1's into the process on the nulls, implying that N_k is stochastically dominated by N^0_k. Noting that $N^0_k \ge k$ if and only if $\pi^0_{(k)} < t_k$, we obtain that

$$\begin{aligned}\mathbb{P}\left(\exists k \in [\![k_{max} \wedge p_0]\!], \pi_{(k:\mathcal{H}_0)} < t_k | \sigma(\mathbf{W})\right) &\le \mathbb{P}\left(\exists k \in [\![k_{max} \wedge p_0]\!], \pi^0_{(k)} < t_k\right) \\ &\le \mathbb{P}\left(\exists k \in [\![k_{max}]\!], \pi^0_{(k)} < t_k\right).\end{aligned}$$

Taking the expectation with respect to $\sigma(\mathbf{W})$ yields the desired result. □

Theorem 3 is related to Lemma 3.1 of Janson and Su (2016) and Lemma 3.1 of Li et al. (2022), that rely on the sign-flip property of Knockoff statistics under the null (Barber and Candès, 2015). The interest of Theorem 3 is that the upper bound $\mathrm{JER}^0(\mathbf{t})$ only depends on the π^0 statistics and the threshold family $\mathbf{t}$, and not on the original data. Therefore, it can be estimated with arbitrary precision for any given $\mathbf{t}$ using Monte-Carlo simulation, as explained in the next section and described in Algorithm 4:

Algorithm 4: Sampling from the joint distribution of π statistics under the null according to Theorem 3.

Input: B the number of MC draws; p the number of variables
Output: $\mathbf{\Pi}_0 \in [0,1]^{B\times p}$ a matrix of π^0 statistics
$\mathbf{\Pi}_0 \leftarrow \text{zeros}(B, p)$
for $b \in [1, B]$ **do**
  $\boldsymbol{\chi} \leftarrow \text{draw\_random\_vector}(\{-1, 1\}^p)$ `// Draw signs`
  $Z = 0$ `// Initialize count`
  for $j \in [1, p]$ **do**
    if $\boldsymbol{\chi}[j] < 0$ **then**
      $\mathbf{\Pi}_0[b][j] \leftarrow 1$
      $Z \leftarrow Z + 1$ `// Increment` Z
    end
    else
      $\mathbf{\Pi}_0[b][j] \leftarrow \frac{1+Z}{p}$
    end
  end
end
$\mathbf{\Pi}_0 \leftarrow \text{sort\_lines}(\mathbf{\Pi}_0)$ `// Sort samples`
Return $\mathbf{\Pi}_0$

7.3.2 Joint Error Rate control for π statistics via calibration

To approximate the JER upper bound derived in Theorem 3, we draw B Monte-Carlo samples using Algorithm 4. This yields a set of B vectors of π^0 statistics denoted by $\pi^0_b \in \mathbb{R}^p$ for each $b \in [\![B]\!]$. This allows us to evaluate the empirical JER, which estimates the upper bound of interest.

Definition 13 (Empirical JER)**.** For B vectors of π^0 statistics and a threshold family $\mathbf{t}$, the empirical JER is defined as:

$$\widehat{\text{JER}}^0_B(\mathbf{t}) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\exists k \in [\![k_{max}]\!] : \pi^0_{b(k)} < t_k\right\}, \tag{7.3}$$

where for each $b \in [\![B]\!]$, $\pi^0_{b(1)} \leq \cdots \leq \pi^0_{b(p)}$.

Since $\widehat{\text{JER}}^0_B(\mathbf{t})$ can be made arbitrarily close (by choosing B large enough) to $\widehat{\text{JER}}^0(\mathbf{t})$ for any given threshold family $\mathbf{t}$, it remains to choose $\mathbf{t}$ such that $\widehat{\text{JER}}^0(\mathbf{t}) \leq \alpha$ in order to ensure JER control. To this end, we consider a sorted set of candidate threshold families called a *template*:

Definition 14 (Template; Blanchard et al., 2020)**.** A template is a component-wise non-decreasing function $\mathbf{T} : [0,1] \mapsto \mathbb{R}^p$ that maps a parameter $\lambda \in [0,1]$ to a threshold family $\mathbf{T}(\lambda) \in \mathbb{R}^p$.

This definition is naturally extended to the case of templates containing a finite number of threshold families. The template corresponding to B' threshold families is then denoted by $(\mathbf{T}(b'/B'))_{b' \in [\![B']\!]}$.

Once a template is specified, the *calibration* procedure (Blanchard et al., 2020) can be performed; this consists in finding the least conservative threshold family $\mathbf{t}$ amongst the

template that controls the empirical JER at level α. Formally, we consider the threshold family defined $\mathbf{t}_\alpha^B = \mathbf{T}(\lambda_B(\alpha))$, where

$$\lambda_B(\alpha) = \frac{1}{B'} \max \left\{ b' \in [\![B']\!] \quad s.t. \quad \widehat{\mathrm{JER}}_B^0 \left(\mathbf{T} \left(\frac{b'}{B'} \right) \right) \leq \alpha \right\}.$$

As observed by Blain et al. (2022), optimal power is reached when the candidate families match the shape of the distribution of the null statistics. We define a template based on the distribution of the π^0 statistics appearing in Theorem 3. In practice, we draw B' samples from this distribution independently from the B Monte Carlo samples to avoid circularity biases. Since a template has to be component-wise non-decreasing, i.e. the set of candidate threshold families has to be sorted, we extract empirical quantiles from these B' sorted vectors. This yields a template $\mathbf{T}^0$ composed of B' candidate curves that match quantiles of the distribution of π^0 statistics. The $\frac{b'}{B'}$-quantile curve defines the threshold family $\mathbf{T}^0\left(b'/B'\right)$. To prove that this procedure yields JER control, we start by proving a useful Lemma:

Lemma 2. *For any threshold family* $\mathbf{t}$*, we have*

$$\mathrm{JER}^0\left(\mathbf{t}\right) - \widehat{\mathrm{JER}}_B^0(\mathbf{t}) = O_P(1/\sqrt{B})$$

Proof of Lemma 2. Let $Z_B(\mathbf{t}) = \sqrt{B}\left(\mathrm{JER}^0\left(\mathbf{t}\right) - \widehat{\mathrm{JER}}_B^0(\mathbf{t})\right)$. By the Central Limit Theorem, we have

$$Z_B(\mathbf{t}) \xrightarrow[B\to\infty]{d} Z(\mathbf{t}),$$

where $Z(\mathbf{t})$ is a centered Gaussian random variable with variance $\sigma^2(\mathbf{t}) = \mathrm{JER}^0\left(\mathbf{t}\right)(1 - \mathrm{JER}^0\left(\mathbf{t}\right))$. As such, for any $M > 0$, we have

$$\mathbb{P}\left(|Z_B(\mathbf{t})| \geq M\right) \xrightarrow[B\to\infty]{} \mathbb{P}\left(|Z(\mathbf{t})| \geq M\right).$$

Since $\mathrm{JER}^0\left(\mathbf{t}\right) \leq 1$, we have $\sigma^2(\mathbf{t}) \leq 1/4$ for any $\mathbf{t}$, so that $Z(\mathbf{t})$ is stochastically dominated by $\mathcal{N}\left(0, 1/4\right)$, which does not depend on the threshold family $\mathbf{t}$. As such, we have $\mathbb{P}\left(|Z(\mathbf{t})| \geq M\right) = 2\mathbb{P}\left(Z(\mathbf{t}) \geq M\right) \leq 2\overline{\Phi}(2M)$, where $\overline{\Phi}$ denotes the tail function of the standard normal distribution. Since $\overline{\Phi}(x)$ tends to 0 as $x \to +\infty$, we have proved that $Z_B(\mathbf{t}) = O_P(1)$. □

We then obtain the main result:

Theorem 4 (JER control for π-statistics)**.** *Consider the threshold family defined by* $\mathbf{t}_\alpha^B = \mathbf{T}^0(\lambda_B(\alpha))$*. Then, as* $B \to +\infty$*,*

$$\mathrm{JER}(\mathbf{t}_\alpha^B) \leq \alpha + O_P(1/\sqrt{B}).$$

Proof. We treat the case where $\mathbf{t}_\alpha^B$ is well defined for all B, i.e. that there exists a threshold family amongst $\mathbf{T}^0$ controls the empirical JER^0 for B draws. If this is not the case for some B, then $\mathbf{t}_\alpha^B$ is set to the null family and the result holds.

By Theorem 3 we have for all $\mathbf{t}$ that $\mathrm{JER}\left(\mathbf{t}\right) \leq \mathrm{JER}^0\left(\mathbf{t}\right)$. We can write:

$$\begin{aligned}\mathrm{JER}^0\left(\mathbf{t}\right) &= \widehat{\mathrm{JER}}_B^0(\mathbf{t}) + \left(\mathrm{JER}^0\left(\mathbf{t}\right) - \widehat{\mathrm{JER}}_B^0(\mathbf{t})\right) \\ &= \widehat{\mathrm{JER}}_B^0(\mathbf{t}) + O_P(1/\sqrt{B})\end{aligned}$$

by Lemma 2. Applying the above to $\mathbf{t} = \mathbf{t}_\alpha^B$ yields the desired result since $\widehat{\mathrm{JER}}_B^0(\mathbf{t}_\alpha^B) \leq \alpha$ by definition. □

The number B of Monte-Carlo samples in Theorem 4 can be chosen arbitrarily large to obtain JER control, leading to valid FDP bounds via Equation 3.6.

7.3.3 False Discovery Proportion control for aggregated Knockoffs

In the previous section we have seen how to reach FDP control via Knockoffs. As explained above, aggregation is needed to mitigate the randomness of the Knockoff generation process. Therefore, we aim to extend the previous result to the case of aggregated Knockoffs. Let us first define aggregation:

Definition 15. For D draws of Knockoffs, an aggregation procedure is a function $f : \mathbb{R}^D \mapsto \mathbb{R}$ that maps a vector of $(\pi^d)_{d\in[\![D]\!]}$ statistics to an aggregated statistic $\overline{\pi}$.

In practice, since we have p variables, aggregation is performed for each variable, i.e.:

$$\forall j \in [\![p]\!], \quad f(\pi_j^1, \dots, \pi_j^D) = \overline{\pi_j}.$$

Then, inference is performed on the vector of aggregated statistics $(\overline{\pi}_1, \dots, \overline{\pi}_p)$.

For a fixed aggregation scheme f, we can naturally extend the calibration procedure of the preceding section. Instead of drawing a single $B \times p$ matrix of π^0 statistics containing $\pi_b^0 \in \mathbb{R}^p$ for each $b \in [\![B]\!]$, we draw D such matrices. Given $d \in [\![D]\!]$, each matrix contains $\pi_b^{0,d} \in \mathbb{R}^p$ for each $[\![B]\!]$.

Then, for each $b \in [\![B]\!]$, we perform aggregation: $\overline{\pi}_b^0 = f\left((\pi_b^{0,d})_{d\in[\![D]\!]}\right)$. The JER in the aggregated case is defined as:

$$\overline{\mathrm{JER}}(\mathbf{t}) = \mathbb{P}\left(\exists j \in [\![k_{max} \wedge p_0]\!] : \overline{\pi}_{(j:\mathcal{H}_0)} < t_j\right).$$

We obtain the aggregated template following the same procedure, i.e. drawing D templates and aggregating them. For each $b' \in [\![B']\!]$, the aggregated threshold family is written:

$$\overline{\mathbf{T}}\left(\frac{b'}{B'}\right) = f\left(\left(\mathbf{T}^d\left(\frac{b'}{B'}\right)\right)_{d\in[\![D]\!]}\right).$$

We can then write the empirical JER in the aggregated case as:

$$\widehat{\overline{\mathrm{JER}}}\left(\overline{\mathbf{T}}\left(\frac{b'}{B'}\right)\right) = \frac{1}{B}\sum_{b=1}^{B} 1\left\{\exists j \in [\![k_{max}]\!] : \overline{\pi}_{b(j)}^0 < \overline{\mathbf{T}}_j\left(\frac{b'}{B'}\right)\right\}.$$

Calibration can be performed in the same way as in the non-aggregated case. Note that we perform calibration *after* aggregating; therefore, JER control is ensured directly on aggregated statistics and is not a result of aggregating JER controlling families. Importantly, this approach holds without additional assumptions on the aggregation scheme f. We consider the threshold family $\overline{\mathbf{t}}_\alpha^B = \overline{\mathbf{T}}(\lambda_B(\alpha))$, where

$$\lambda_B(\alpha) = \frac{1}{B'}\max\left\{b' \in [\![B']\!] \quad s.t. \quad \widehat{\overline{\mathrm{JER}}}_B^0\left(\overline{\mathbf{T}}\left(\frac{b'}{B'}\right)\right) \le \alpha\right\}.$$

With $\overline{\mathbf{T}}^0$ a template composed of B' candidate curves that match quantiles of the distribution of $\overline{\pi}^0$ statistics, we obtain the following result:

Theorem 5 (JER control for aggregated π-statistics)**.** *Consider the threshold family defined by* $\overline{\mathbf{t}}_\alpha^B = \overline{\mathbf{T}}^0(\lambda_B(\alpha))$. *Then, as* $B \to +\infty$,

$$\overline{\mathrm{JER}}(\overline{\mathbf{t}}_\alpha^B) \le \alpha + O_P(1/\sqrt{B}).$$

Proof. The proof is identical to that of Theorem 4 using the empirical aggregated JER. □

The calibrated aggregated threshold family yields valid FDP upper bounds via Equation 3.6. The proposed **KOPI** (Knockoffs - π) method therefore achieves FDP control on aggregated Knockoffs.

7.4 Experiments

Methods considered. In our implementation of KOPI, we rely on the harmonic mean (Wilson, 2019) as the aggregation scheme f. Additionally, we set $k_{max} = \lfloor p/50 \rfloor$ following the approach of Blain et al., 2022. We also consider both state-of-the-art Knockoffs aggregation schemes: AKO (Aggregation of Multiple Knockoffs, Nguyen et al., 2020) and e-values based aggregation (Ren and Barber, 2022). Additionally, we consider Vanilla Knockoffs, i.e. Candès et al., 2018 and FDP control via Closed Testing (Li et al., 2022). In simulated data experiments, we generate Knockoffs assuming a Gaussian distribution for $\mathbf{X}$, with all variables centered. For methods that support aggregation, we use $D = 50$ Knockoff draws.

7.4.1 Simulated data

Setup. At each simulation run, we generate Gaussian data $\mathbf{X} \in \mathbb{R}^{n \times p}$ with a Toeplitz correlation matrix corresponding to a first-order auto-regressive model with parameter ρ, i.e. $\mathbf{\Sigma}_{i,j} = \rho^{|i-j|}$.

Then, we draw the true support $\boldsymbol{\beta}^* \in \{0, 1\}^p$. The number of non-null coefficients of $\boldsymbol{\beta}^*$ is controlled by the sparsity parameter s_p, i.e. $s_p = \|\boldsymbol{\beta}^*\|_0 / p$. The target variable $\mathbf{y}$ is built using a linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon},$$

with σ controlling the amplitude of the noise: $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2 / (\text{SNR}\|\boldsymbol{\epsilon}\|_2)$, SNR being the signal-to-noise ratio. We choose the central setting $n = 500, p = 500, \rho = 0.5, s_p = 0.1, \text{SNR} = 2$. For each parameter, we explore a range of possible values to benchmark the methods across varied settings.

To select variables using FDP upper bounds, we retain the largest possible set of variables S such that $V(S) \leq q|S|$ (Algorithm 6).

Algorithm 5: Performing calibration on π-statistics. First, we use Theorem 3 to build a suitable template and estimate the JER of each candidate threshold family. Then, we perform calibration to select the least conservative possible threshold family that controls the JER at a given level α.

Input: α the desired FDP coverage; B the number of MC draws for JER estimation; B' the number of candidate threshold families
Output: $\mathbf{t}_\alpha$ the calibrated threshold family at level α
$\mathbf{\Pi}_0 \leftarrow$ draw_null_$\pi(B, p)$ `// Algorithm 4`
$\mathbf{\Pi}_0^{'} \leftarrow$ draw_null_$\pi(B', p)$
for $b' \in [1, B']$ **do**
  $\mathbf{T}[b'] \leftarrow$ quantiles$(\mathbf{\Pi}_0^{'}, \frac{b'}{B'})$ `// Build template`
  $\widehat{\text{JER}}_{b'} \leftarrow$ empirical_jer$(\mathbf{\Pi}_0, \mathbf{T}[b'])$ `// Apply Algorithm 3 for each family`
end
$b'_{cal} \leftarrow \max\{b' \in [1, B'] \text{ s.t. } \widehat{\text{JER}}_{b'} \leq \alpha\}$ `// Perform calibration`
$\mathbf{t}_\alpha \leftarrow \mathbf{T}[b'_{cal}]$
Return $\mathbf{t}_\alpha$

Algorithm 6: Performing inference via Knockoffs and calibration. We compute the largest possible region that satisfies the required FDP level q using the JER controlling family computed via Algorithm 5. The bound $V^{\mathbf{t}_\alpha}$ is computed from π using Equation 3.6.

Input: $\mathbf{X}$ the input data; $\mathbf{y}$ the target variable; q the maximum tolerable FDP; $\mathbf{t}_\alpha$ the calibrated threshold family at level α
Output: $\hat{S}$ the selected variables
$n, p \leftarrow$ shape$(\mathbf{X})$ `// n samples, p variables`
$\tilde{\mathbf{X}} \leftarrow$ sample_Knockoffs$(\mathbf{X})$
$\mathbf{W} \leftarrow LCD(\mathbf{X}, \tilde{\mathbf{X}}, \mathbf{y})$ `// Compute W`
$\boldsymbol{\pi} \leftarrow$ compute_proportion$(\mathbf{W})$ `// Equation (3.8)`
$\hat{S} \leftarrow \max_S\{|S| \quad s.t. \quad \frac{V^{\mathbf{t}_\alpha}(S)}{|S|} \leq q\}$ `// Find largest admissible region`
`//` $V^{\mathbf{t}_\alpha}(S)$ `depends on` $\boldsymbol{\pi}$
Return $\hat{S}$

For each of the N simulations and each method, we compute the empirical FDP and True Positive Proportion (TPP):

$$\widehat{FDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S|} \quad \text{and} \quad \widehat{TPP}(S) = \frac{|S \cap \mathcal{H}_1|}{|\mathcal{H}_1|}.$$

If the FDP is controlled at level α, $|\{k \in [\![N]\!] : \widehat{FDP}(S_k) > q\}| \sim \mathcal{B}(N, \alpha)$. Then, we can compute error bands on the α-level using std $(\mathcal{B}(N, \alpha)/N) = \sqrt{\alpha(1-\alpha)/N}$. The second row of Fig. 7.1 represents the empirical power achieved by each method, which corresponds to the average of TPPs defined above for N runs i.e. Power $= \sum_{k=1}^{N} \widehat{TPP}(S_k)/N$. Fig. 7.1 shows that across all different settings, KOPI retains FDP control. We can also see that FDR control does not imply FDP control, as Vanilla Knockoffs are consistently outside of FDP bound coverage intervals. However, the two existing aggregation schemes (AKO and e-values) that formally guarantee FDR control are generally conservative and achieve FDP control empirically. This is consistent with the findings of Ren and Barber, 2022. The

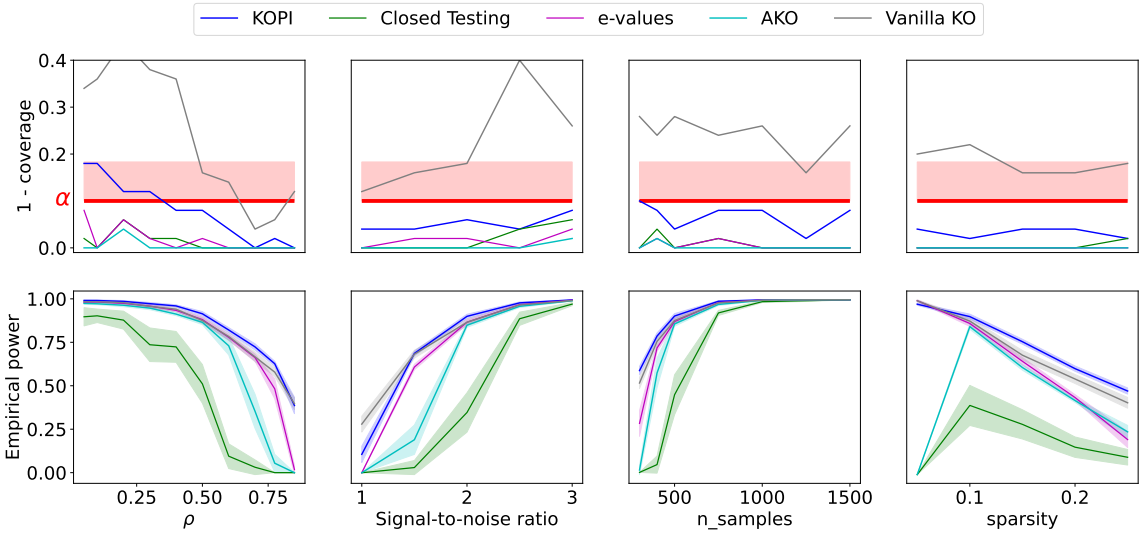


Figure 7.1: **FDP bound coverage at level α and empirical Power for 50 simulation runs and five different methods:** Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation, KOPI and Knockoff inference via Closed Testing. We use $D = 50$ Knockoffs draws and the following simulation settings: $\alpha = 0.1, q = 0.1, p = 500$. Each column represents a varying parameter with the first row displaying FDP coverage and the second row displaying power. The red line and associated error bands represent the acceptable limits for FDP bound coverage. KOPI consistently outperforms all other methods while retaining FDP control.

Closed Testing procedure of Li et al., 2022, achieves FDP control as announced but suffers from a lack of power.

Interestingly, KOPI achieves FDP control while offering power gains compared to FDR-controlling Knockoffs aggregation methods. Yet FDP control is a much stronger guarantee than FDR control, as discussed previously. These gains are especially noticeable in challenging inference settings where most methods exhibit a clear decrease in power or even catastrophic behavior (i.e. zero power).

Moreover, Fig. 7.3 shows that when using $q = 0.05$ rather than $q = 0.1$ as in Fig. 7.1, the robustness of KOPI with regards to difficult inference settings is even more salient. More precisely, for $q = 0.05$, AKO and Closed Testing are always powerless. E-values aggregation yields good power in easier settings such as $\rho \leq 0.6$, SNR ≥ 2.5 or $n > 750$ but exhibits catastrophic behavior in harder settings. Overall, apart from KOPI, only Vanilla Knockoffs exhibit non-zero power, but this method fails to control the FDP as it is intended to control FDR. KOPI preserves FDP control in all settings while yielding superior power compared to all other methods.

7.4.2 Brain data application

The goal of human brain mapping is to associate cognitive tasks with relevant brain regions. This problem is tackled using functional Magnetic Resonance Imaging (fMRI), which consists in recording the blood oxygenation level dependent signal via an MRI scanner. The importance of conditional inference for this problem has been outlined in Weichwald et al., 2015. We use the Human Connectome Project (HCP900) dataset that contains brain images of healthy young adults performing different tasks while inside an MRI scanner. Details about this dataset and empirical results can be found in Section 7.6.2.

While these results demonstrate the face validity of the approach, FDP control and

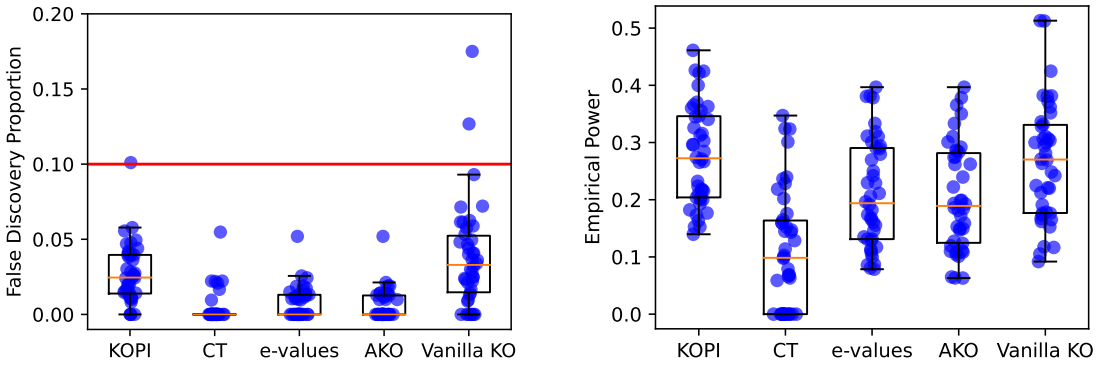


Figure 7.2: **Empirical FDP and power on semi-simulated data for 42 contrast pairs.** We use 7 HCP contrasts C0: "Motor Hand", C1: "Motor Foot", C2: "Gambling", C3: "Relational", C4: "Emotion", C5: "Social", C6: "Working Memory". We consider all 42 possible train/test pairs: the train contrast is used to obtain a ground truth, while the test contrast is used to generate the response. Inference is performed using the 5 methods considered in the chapter and the empirical FDP is reported in the left box plot, while power is reported in the right box plot. Notice (right figure) that KOPI yields superior power compared to all other Knockoffs-based methods while controlling the FDP (left Fig.).

power cannot be evaluated. Therefore, following Nguyen et al., 2022, we consider an additional experiment that consists in using semi-simulated data. We consider a first fMRI dataset $(\mathbf{X}_1, \mathbf{y}_1)$ on which we perform inference using a Lasso estimator; this yields $\boldsymbol{\beta}_1^* \in \mathbb{R}^p$ that we will use as our ground truth. Then, we consider a separate fMRI dataset $(\mathbf{X}_2, \mathbf{y}_2)$ for data generation. The point of using a separate dataset is to avoid circularity between the ground truth definition and the inference procedure. Concretely, we discard the original response vector $\mathbf{y}_2$ for this dataset and build a simulated response $\mathbf{y}_2^{sim}$ using a linear model, with the same notation as previously (we set σ so that $SNR = 4$): $\mathbf{y}_2^{sim} = \mathbf{X}_2\boldsymbol{\beta}_1^* + \sigma\boldsymbol{\epsilon}$.

Then, inference is performed using Knockoffs-based methods on $(\mathbf{X}_2, \mathbf{y}_2^{sim})$. Since we consider $\boldsymbol{\beta}_1^*$ as the ground truth, the FDP and TPP can be computed for each method. As can be seen in Fig. 7.2, KOPI is the most powerful method among those that control the FDP.

7.4.3 Genomic data application

Lymphomatic leukemia mutation classification

Differential gene expression studies aim at identifying genes whose activity differs significantly between two (or more) populations, based on a sample of measurements from individuals from these populations. The activity of a gene is usually quantified by its level of expression in the cell. We consider a microarray data set studied in Bourgon et al., 2010, that consists of expression measurements for biological samples from $n = 79$ individuals with B-cell acute lymphoblastic leukemia (ALL): 37 of these individuals harbor a specific mutation called BCR/ABL, while the remaining 42 do not. Our goal here is to identify, from this sample, genes for which there is a difference in the mean expression level between the mutated and non-mutated populations. We focus on the $p = 90$ genes on chromosome 7 whose individual standard deviation is above 0.5.

The genes selected by different Knockoffs-based methods are summarized in Table 7.1.

Stability selection criteria analogous to Luo et al., 2022; Ren et al., 2021 are displayed. Note that the selection made by KOPI is more robust than that of Vanilla Knockoffs: 4 genes are selected in nearly all runs by KOPI, while none are selected as frequently by Vanilla Knockoffs. Conversely, KOPI only selects 2 genes in less than 50% of all runs compared to 18 for Vanilla Knockoffs. This confirms that error control guarantees of KOPI, together with the stability brought by aggregation, lead to avoiding most spurious/non-reproducible detections. Besides KOPI and Vanilla Knockoffs, all other methods are powerless in all runs.

	KOPI	Vanilla KO	e-values	Closed Testing	AKO
Selected in >90% of runs	4	0	0	0	0
Selected in >50% of runs	6	6	0	0	0
Spurious detections (<50% of runs)	2	18	0	0	0

Table 7.1: **Stability selection criteria for 5 Knockoffs-based methods on "Lymphomatic leukemia mutation" genomic data.** Note that KOPI displays a very stable selection set across all runs with 4 genes present in $> 90\%$ of runs. KOPI also avoids most spurious discoveries, as only 2 genes are selected less than 50% of the time, compared to 18 genes using Vanilla Knockoffs. The 6 genes selected more than 50% of the time by KOPI and Vanilla Knockoffs are the same. All other Knockoffs-based methods are powerless in all runs.

Colon vs Kidney classification

We also considered an additional genomic dataset to reproduce these results with a larger number of samples. The dataset we used is part of **GEMLeR** (Gene Expression Machine Learning Repository, Stiglic and Kokol, 2010), a collection of gene expression datasets that can be used to benchmark ML methods on genomics data.

We chose the "Colon vs Kidney" dataset: this is a binary classification dataset where the goal is to distinguish cancerous tissue from two different organs (Colon and Kidney) using gene expression data. This dataset comprises 546 samples and 10936 genes. To make the problem tractable for Knockoffs-based methods we perform dimensionality reduction to select the 546 genes that have the largest variance. Then, **we run all Knockoffs-based methods 50 times** and report the selected genes.

	KOPI	Vanilla KO	e-values	Closed Testing	AKO
Selected in >90% of runs	21	0	0	0	0
Selected in >50% of runs	22	25	0	0	0
Spurious detections (<50% of runs)	7	34	20	0	0

Table 7.2: **Stability selection criteria for 5 Knockoffs-based methods on "Colon vs Kidney" genomic data.** Note that KOPI displays a very stable selection set across all runs with 21 genes present in $> 90\%$ of runs. KOPI also avoids most spurious discoveries, as only 7 genes are selected less than 50% of the time, compared to 34 genes using Vanilla Knockoffs and 20 using e-values. All other Knockoffs-based methods are powerless in all runs.

The genes selected by different Knockoffs-based methods are summarized in Table 7.2. Stability selection criteria analogous to Luo et al., 2022; Ren et al., 2021 are displayed. Note that the selection made by KOPI is more robust than that of Vanilla Knockoffs: 21

genes are selected in nearly all runs by KOPI, while none are selected as frequently by Vanilla Knockoffs. Conversely, KOPI only selects 7 genes in less than 50% of all runs compared to 34 for Vanilla Knockoffs and 20 for e-values aggregation.

7.5 Discussion

In this chpater, we have proposed a novel method that reaches FDP control on aggregated Knockoffs. It combines the benefits of aggregation, i.e. improving the stability of the inference, in addition to providing a probabilistic control of the FDP, rather than controlling only its expectation, the FDR.

Simulation results support that KOPI indeed controls the FDP. Furthermore, while FDP control is a stricter guarantee than FDR control, KOPI actually offers power gains compared to state-of-the-art aggregation-based Knockoffs methods.

This sensitivity gain is a direct benefit from the JER approach and its adaptivity to arbitrary aggregation schemes. While the latter has been formulated and used so far in mass univariate settings (Blain et al., 2022), the present work presents a first use of this approach in the context of multiple regression.

Moreover, KOPI does not require any assumption on the data at hand or on the law of Knockoff statistics under the null.

The computation time of the proposed approach is comparable to existing aggregation schemes for Knockoffs: sampling π statistics under the null using Algorithm 4 can be done once and for all for a given value of p. JER estimation via Algorithm 3 and calibration can be performed via binary search of complexity $\mathcal{O}(log(B'))$. Finding the rejection set $\hat{S}$ after performing calibration is done in linear time via Enjalbert-Courrech and Neuvial, 2022. In practice, the computation time is the same as for classical knockoff aggregation (Ren et al., 2021) and is in minutes for the brain imaging datasets considered. Avenues for future work include a theoretical analysis of the False Negative Proportion (FNP) (Genovese and Wasserman, 2002) of KOPI and developing a step-down version of the method to further improve power.

We provide a Python package containing the code for KOPI available at `https://github.com/alexblnn/KOPI`.

7.6 Additional simulation results

7.6.1 A harder inference setup

We evaluated the performance of all five methods in the more challenging setting $q = 0.05$ instead of using $q = 0.1$. The results are presented in Fig. 7.3. In this setting, AKO and Closed Testing are always powerless and aggregation via e-values suffers from a lack of power in most cases. Vanilla Knockoffs exhibit satisfactory power but consistently fail to control the FDP. KOPI preserves FDP control and yields acceptable power.

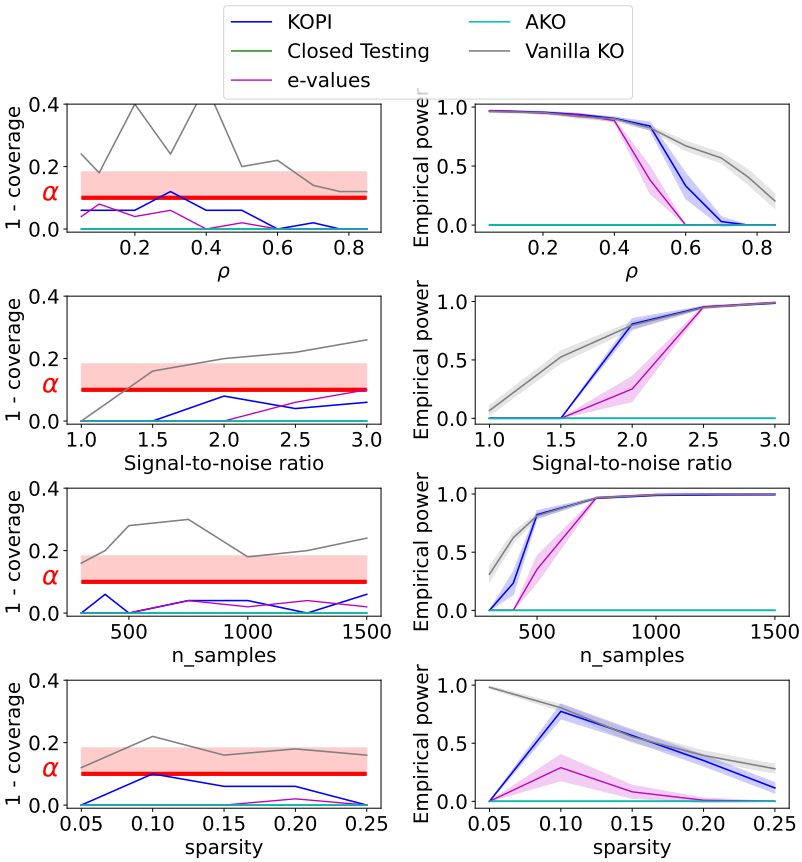


Figure 7.3: **FDP bound coverage at level α and empirical Power for 50 simulation runs and five different methods.** The five methods are Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation, KOPI and Knockoff inference via Closed Testing. We use 50 Knockoffs draws and the following simulation setting $\alpha = 0.1, q = 0.05, p = 500$. Each row represents a varying parameter with the left panel displaying FDP coverage and the right panel displaying power. The red line and associated error bands represent the acceptable limits for FDP bound coverage. Notice that KOPI consistently outperforms all other methods while retaining FDP control.

Impact of aggregation scheme choice

While the theoretical guarantees we obtain hold for all choices of aggregation schemes, these hyperparameter impacts the power of KOPI. To assess this, we use the same simulated data setup as in Figure 7.1 to compare four aggregation schemes: arithmetic mean, geometric mean, harmonic mean and quantile aggregation.

Importantly, we first check that the FDP is controlled for all types of aggregation and in all settings considered by reporting the bound non-coverage. We use three settings of varying difficulty, parametrized by the correlation level ρ and use $\alpha = 0.1, q = 0.1$:

	Harmonic	Arithmetic	Geometric	Quantile aggregation
$\rho = 0.5$	10%	0%	2%	10%
$\rho = 0.6$	2%	0%	0%	4%
$\rho = 0.7$	2%	0%	0%	0%

Table 7.3: **FDP control of KOPI for four aggregation schemes and three different correlation levels.** Note that FDP control is maintained in all scenarios which is coherent with the result obtained in Theorem 5.

The FDP is indeed controlled in all cases since non-coverage never exceeds the chosen level $\alpha = 10\%$ as seen in Table 7.3. This is coherent with the theoretical guarantees we obtain in Theorem 5. We now report the average power to benchmark aggregation schemes:

	Harmonic	Arithmetic	Geometric	Quantile aggregation
$\rho = 0.5$	**0.91**	0.77	0.87	0.90
$\rho = 0.6$	**0.83**	0.58	0.77	**0.83**
$\rho = 0.7$	**0.72**	0.39	0.61	**0.72**

Table 7.4: **Empirical power of KOPI for four aggregation schemes and three different correlation levels.** Note that harmonic mean aggregation consistently outperforms arithmetic aggregation and geometric aggregation. Quantile aggregation performs similarly to harmonic aggregation.

Note that harmonic mean aggregation outperforms arithmetic and geometric mean consistently and performs similarly to quantile aggregation as seen in Table 7.4.

7.6.2 Details and results on HCP data

HCP dataset

We use the HCP900 task-evoked fMRI dataset (Van Essen et al., 2012), in which we take the masked 2 mm resolution z-statistics maps of the 778 subjects from 7 tasks to solve binary regression problems, namely predicting which condition is associated with the brain image: emotion (*emotional face* vs *shape outline*), gambling (*reward* vs *loss*), language (*story* vs *math*), motor hand (*left* vs *right* hand), motor foot (*left* vs *right* foot), relational (*relational* vs *match*) and social (*mental interaction* vs *random interaction*).

We consider the fixed-effect maps (average across right-left and left-right phase encoding schemes) for each condition, yielding one image per subject per condition (which corresponds to two images per subject for each classification problem). Then, for each problem, the number of samples available is 1556 ($= 2 \times 778$) and the number of voxels is 156 374 after gray-matter masking. Dimension reduction was carried out using Ward parcellation scheme to $1k$ clusters, which is known to yield spatially homogeneous regions (Thirion et al., 2014). The signal is then averaged per cluster, yielding a reduced design matrix $\mathbf{X}$ for the problem.

Brain data are non-Gaussian

In the synthetic data experiments we used the Gaussian Knockoff generation process described in Candès et al., 2018. However, fMRI brain maps can be heavily non-Gaussian. In turn, Gaussian Knockoffs cannot satisfy the Knockoffs exchangeability assumption and any statistical control on False Discoveries is rendered spurious.

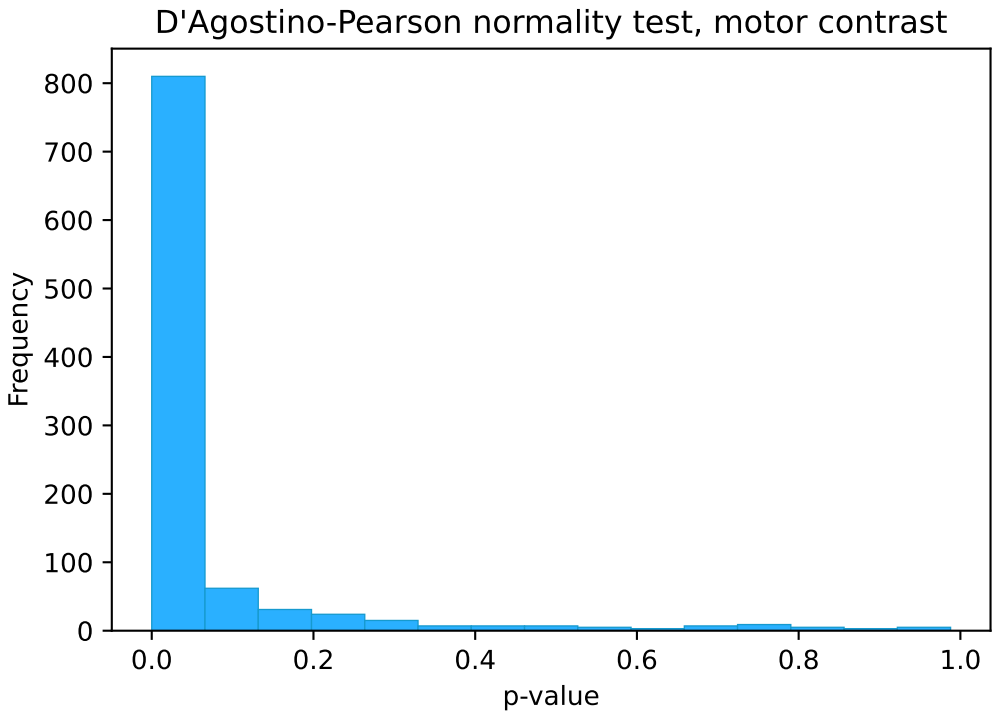


Figure 7.4: **D'Agostino-Pearson normality test for a motor contrast.** We perform a normality test for each cluster amongst the 1000 present for the motor foot contrast. The distribution of the normality test p-values indicates strong non-normality in fMRI data.

To build non-Gaussian Knockoffs, we use a linear variant of the Sequential Conditional Independent Pairs (SCIP) algorithm of Candès et al., 2018:

Algorithm 7: Generating Non-Gaussian Knockoffs using the Sequential Conditional Independent Pairs algorithm of Candès et al., 2018.

for $j \in [1, p]$ **do**
  Fit a Lasso model on $(\mathbf{X_{-j}}, X_j)$
  Compute the residual $\epsilon_j = X_j - \mathbf{X_{-j}}\widehat{\boldsymbol{\beta}}_j$
end
for $j \in [1, p]$ **do**
  Sample $\tilde{X}_j$ from $\mathbf{X_{-j}}\widehat{\boldsymbol{\beta}}_j + \epsilon_{\rho(j)}$ `// ρ is a random ordering of` $[1, p]$
end
Return $\tilde{\mathbf{X}}_{\mathbf{1:p}}$

Additional results

The results corresponding to 7 contrasts of the HCP dataset are presented in Figs 7.5 – Fig 7.11: *foot* contrast of the HCP motor task in Fig 7.5, *hand* contrast of the HCP motor task in Fig 7.6, *relational versus match* contrast of the HCP relational task in Fig 7.7, *gain vs loss* contrast of the HCP gambling task in Fig 7.8, *2-back vs 0-back* contrast of the HCP working memory task in Fig 7.9, *face vs shape* contrast of the HCP Emotional task in Fig 7.10, *interacting vs non-interacting* contrast of the HCP social task in Fig 7.11. These maps display the support of the conditional association test, with a sign that shows whether a region has an upward or downward impact on the decision function.

Overall, many Knockoff-based methods are powerless on all contrasts considered. Only KOPI, Vanilla Knockoffs and e-values aggregation consistently display non trivial solutions. This corresponds to the behavior observed in hard simulation settings in Fig. 7.1, i.e. low SNR and high correlation for instance.

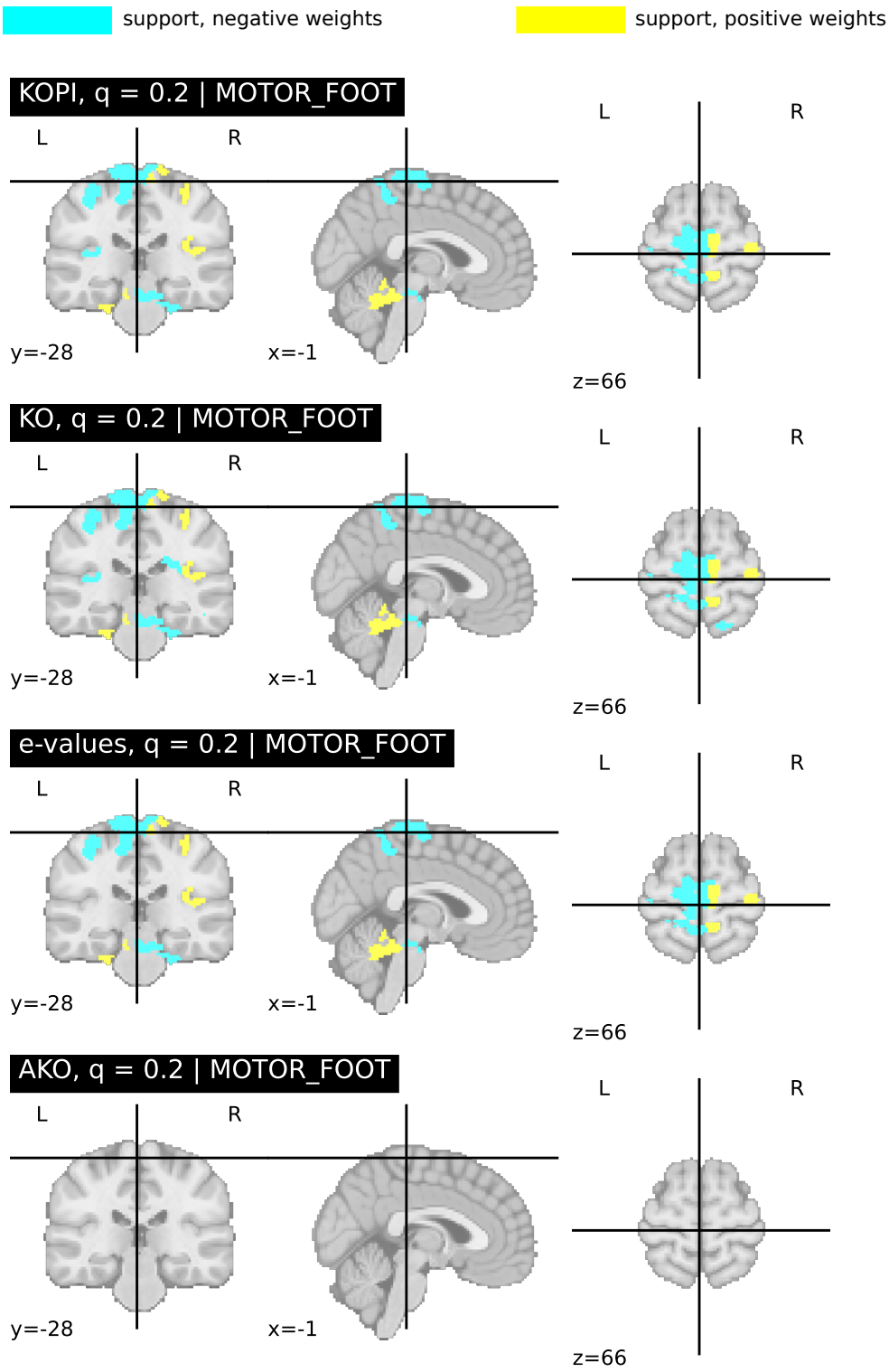


Figure 7.5: **Brain mapping on motor contrast using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs, e-values and KOPI yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws and $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 17 regions, KOPI: 24 regions and e-values: 18 regions.

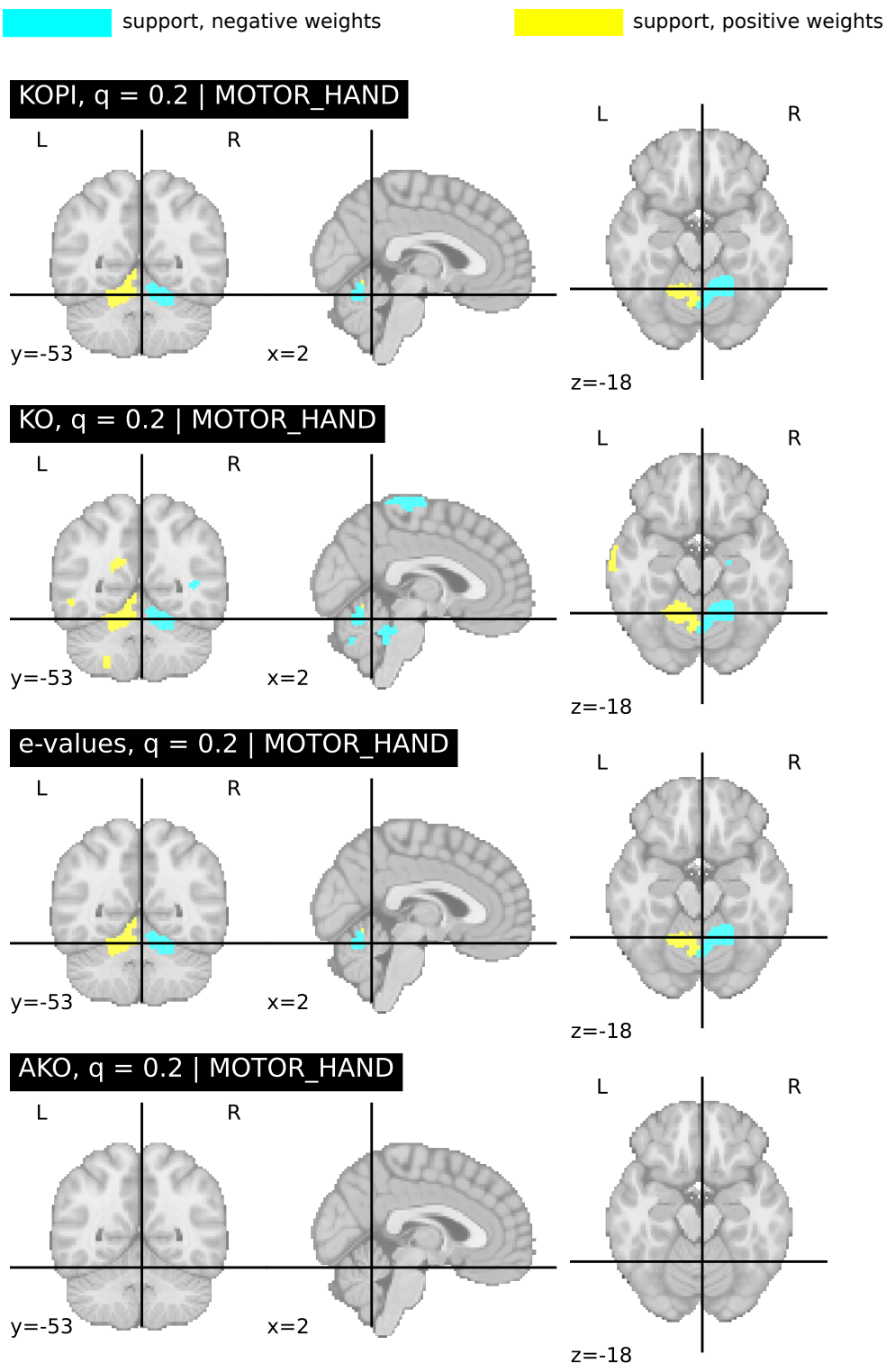


Figure 7.6: **Brain mapping on motor hand contrast using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs, e-values and KOPI yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws and $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 11 regions, KOPI, 10 regions and e-values 11 regions.

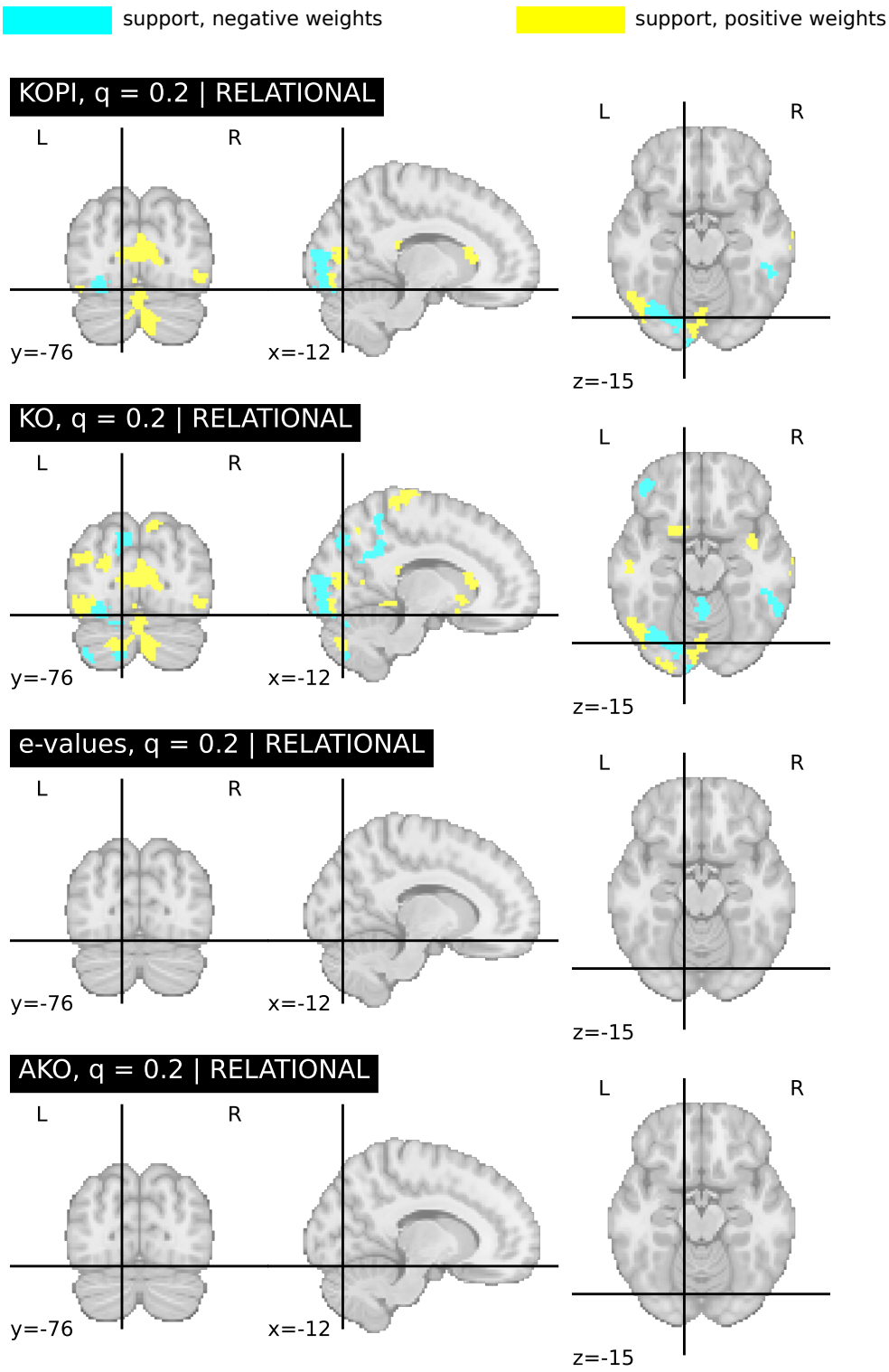


Figure 7.7: **Brain mapping on the HCP Relational task using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs and KOPI yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws, $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 58 regions and KOPI, 24 regions.

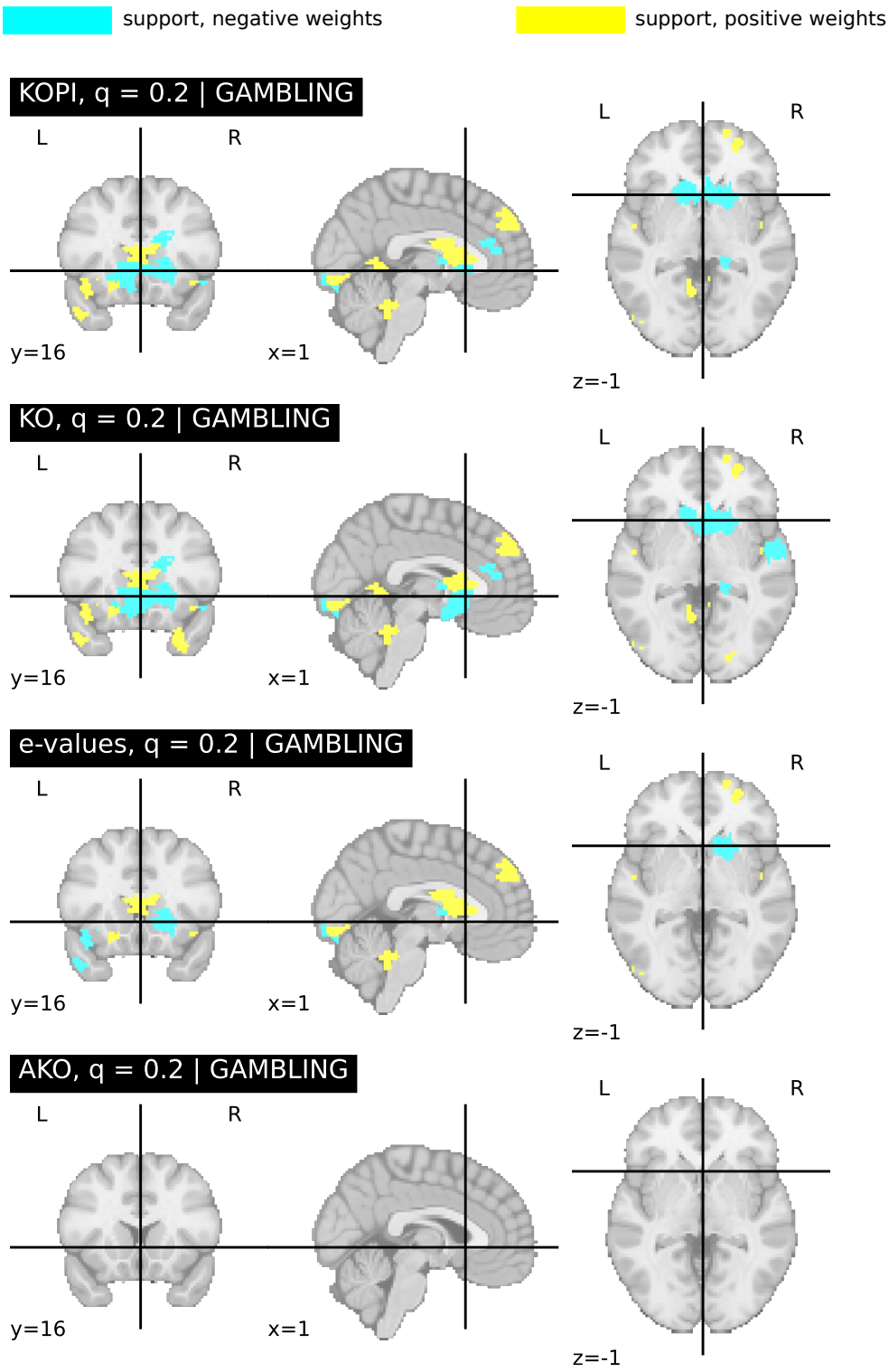


Figure 7.8: **Brain mapping on HCP gambling task using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs, KOPI and e-values aggregation yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws, $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 57 regions, KOPI 57 regions, e-values aggregation, 19 regions.

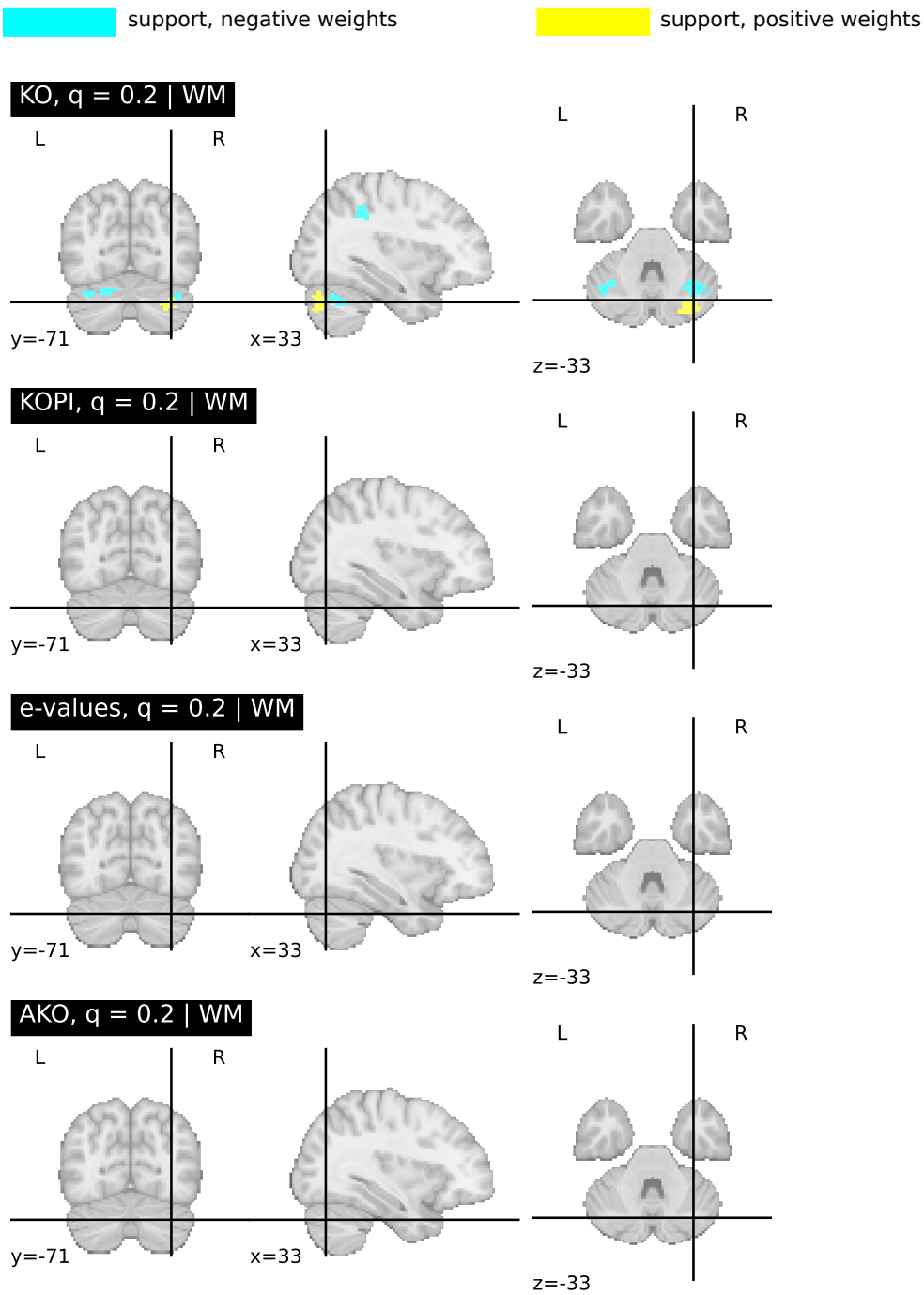


Figure 7.9: **Brain mapping on HCP working memory task using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs yields discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws, $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 8 regions.

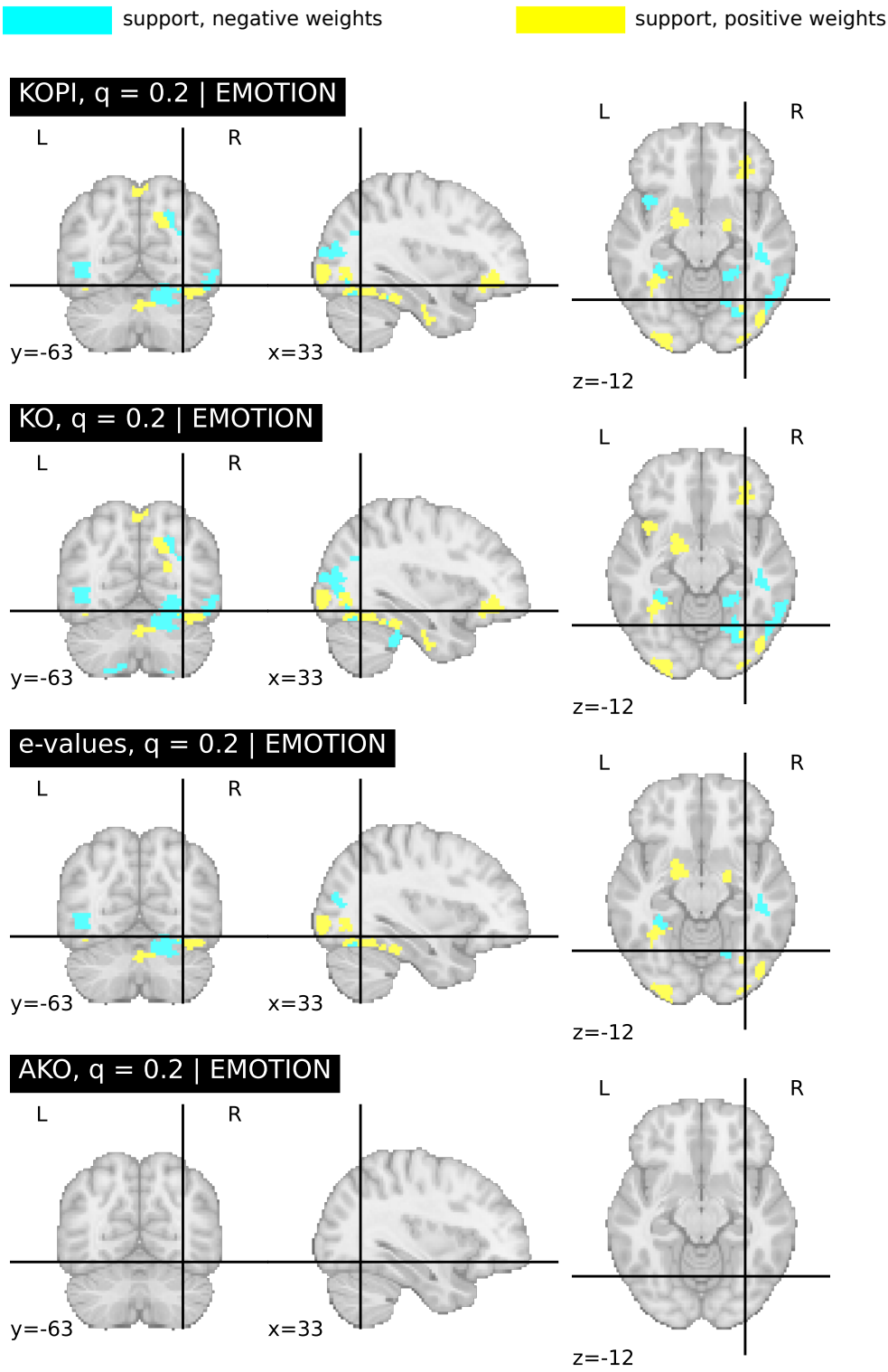


Figure 7.10: **Brain mapping on HCP emotional task using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs, KOPI and e-values aggregation yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws, $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 22 regions, KOPI: 37 regions, e-values aggregation: 20 regions.

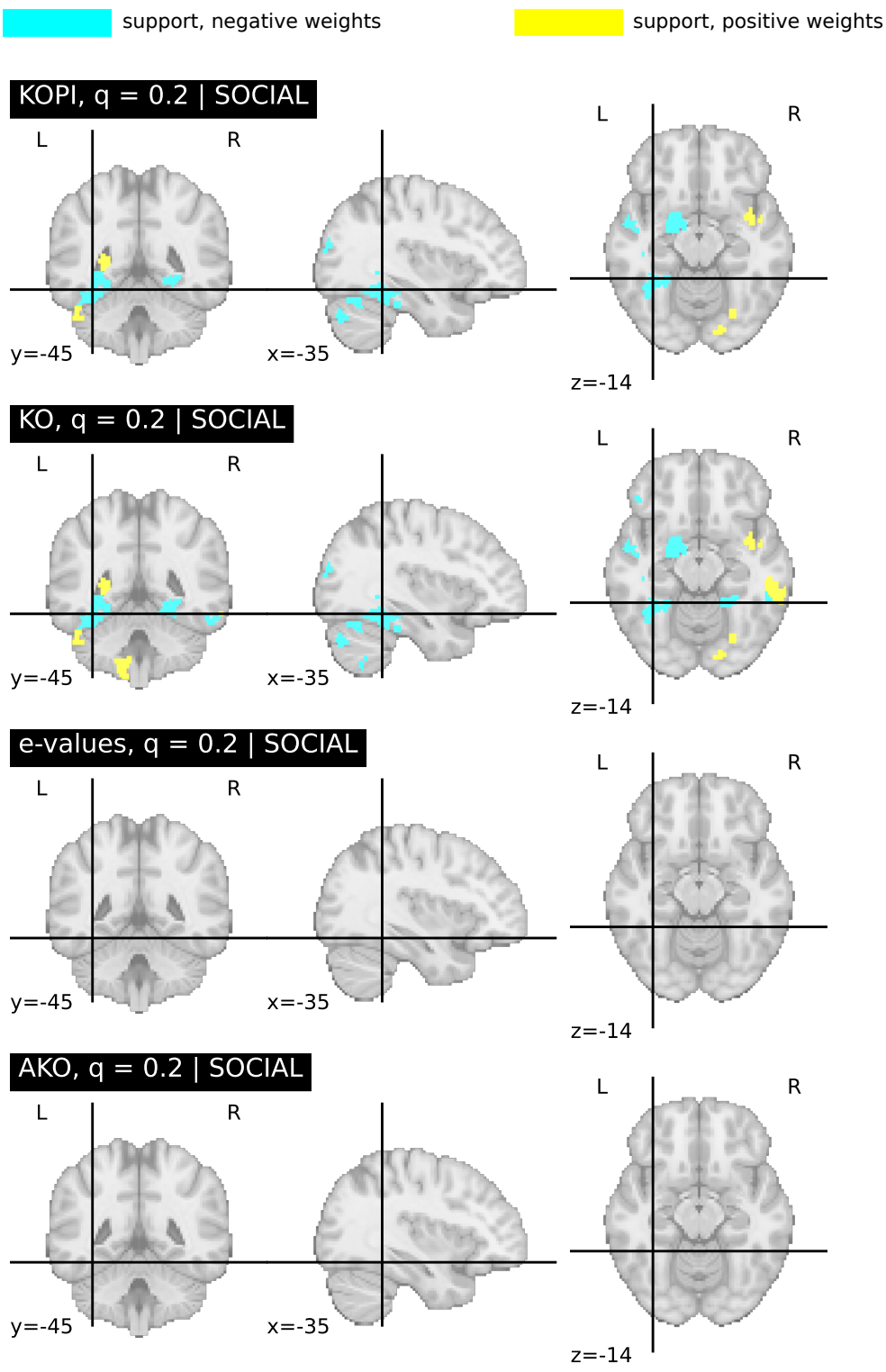


Figure 7.11: **Brain mapping on HCP social task using Knockoffs-based methods.** Among the five methods considered in this chapter –Vanilla Knockoffs, aggregated Knockoffs using e-values, aggregated Knockoffs using quantile-aggregation (AKO), KOPI and Knockoff inference via Closed Testing– only Vanilla Knockoffs and KOPI yield discoveries, plotted above. All other methods are powerless. We use 50 Knockoffs draws, $\alpha = 0.1$ and $q = 0.2$. Each figure represents the region returned by a given method. Vanilla Knockoffs yield 32 regions, KOPI: 27 regions.

Chapter 8

When Knockoffs fail: diagnosing and fixing non-exchangeability of Knockoffs

Summary. Knockoffs are a popular statistical framework that addresses the challenging problem of conditional variable selection in high-dimensional settings with statistical control. Such statistical control is essential for the reliability of inference. However, knockoff guarantees rely on an exchangeability assumption that is difficult to test in practice, and there is little discussion in the literature on how to deal with unfulfilled hypotheses. This assumption is related to the ability to generate data similar to the observed data. To maintain reliable inference, we introduce a diagnostic tool based on Classifier Two-Sample Tests. Using simulations and real data, we show that violations of this assumption occur in common settings for classical Knockoffs generators, especially when the data have a strong dependence structure. We show that the diagnostic tool correctly detects such behavior. To fix knockoff generation, we propose an alternative knockoff construction, which is based on constructing a predictor of each variable based on all others. We also propose a computationally-efficient variant of this algorithm, at the expense of theoretical guarantees. We show empirically that the proposed approach restores error control on simulated data.

Contents

8.1 Background

Error control of the Knockoffs procedure explained in Section 3.2 relies on a critical assumption called *exchangeability*. For exchangeability to hold, the joint distribution of the data must remain unchanged when an original variable is exchanged for its knockoff counterpart. While Knockoffs have shown promise in many applications, this assumption of exchangeability requires careful consideration and assessment, as its validity impacts

the reliability of the entire variable selection process. While the generative processes could ideally be known or derived from first principles, for many practical problems, they are unknown or intractable ; current practice then relies on learning the joint distributions of the observations in order to draw knockoffs from an appropriately perturbed model.

The most popular knockoff generation procedure involves a Gaussian assumption, and thus relies on the knowledge or the accurate estimation of the covariance structure of the covariates, which leads to two potential issues: *i)* the violation of the Gaussian hypotheses and *ii)* inaccuracies in the covariance estimation. This worsens with the number of variables p as the problem becomes harder, especially in high-dimensional regimes where the number of samples n is comparatively small. Some alternative Knockoff generation methods avoid the problem of covariance estimation via e.g. Deep Learning (Romano et al., 2020). However, such methods also suffer from high-dimensional regimes as deep neural networks require massive amounts of data to be properly trained, especially in large feature spaces. Overall, methods that require a large sample size n and a small number of variables p may not be adapted to the Knockoffs framework, which is designed for high-dimensional variable selection. Domain-specific procedures have been developed to tackle some of these issues. For instance, in the field of genomics where variables of interest may be discrete, Knockoffs can be built using Hidden Markov Models (HMM) (Rabiner and Juang, 1986), as demonstrated in the work of Sesia et al., 2019. The Gaussian procedure and existing alternatives are described in 3.2.2.

The aim of this chapter is twofold: First, we propose an effective diagnostic tool that allows practitioners to examine knockoffs along the original data for potential violations. Based on this, we highlight common cases of non-exchangeability when using Gaussian knockoffs and associated error control violations. We describe their consequences, typically a failure to control the false discovery rate at the nominal level. We propose an efficient non-parametric algorithm for constructing knockoffs which comes with theoretical guarantees. We show through simulations that this procedure restores error control in all cases considered.

8.2 An efficient Non-parametric Knockoff generation algorithm

We propose an alternative method to build Knockoffs by learning to predict X_j from X_{-j}, based on the Sequential Conditional Independent Pairs algorithm of Candès et al., 2018:

Algorithm 8: Sequential Conditional Independent Pairs, (SCIP; Candès et al., 2018)

$j = 1$
while $j \leq p$ **do**
  Sample $\tilde{X}_j$ from $\mathcal{L}\left(X_j \mid X_{-j}, \tilde{X}_{1:j-1}\right)$
  $j = j + 1$
end

This algorithm builds provably valid Knockoffs (Candès et al., 2018) without any parametric assumption. However, as explained by the authors, it is hard to run in practice

for two reasons. First, we don't have knowledge of all the conditional distributions in practice. Second, the sequential nature of the algorithm requires computing a new conditional distribution at each step, making computations potentially intractable.

To make this algorithm usable in practice, we propose using a machine learning model f to learn the conditional distributions. A related idea is exploited to quantify variable importance in Random Forests in Chamma et al., 2023. For each variable, f predicts X_j from $X_{-j}, \tilde{X}_{1:j-1}$. Typically, using a Lasso model is a good default choice. In practice, we set $\lambda = \lambda_{max}/100$ with $\lambda_{max} = \frac{1}{n}\left\|\mathbf{X_{-j}}^T X_j\right\|_\infty$. The j-th fitted predictor is denoted f_j. Then, the residual $\widehat{\epsilon}_j = X_j - f_j(\mathbf{X_{-j}})$ is computed. At step j, Knockoffs are built by drawing a residual at random: $\tilde{X}_j$ is chosen as $f_j(\mathbf{X_{-j}}, \mathbf{\tilde{X}_{1:j-1}}) + \widehat{\epsilon}_{\sigma(j)}$ with $\sigma(j)$ a random element of $[\![j]\!]$. Under the assumption that $\mathbf{X}$ is Gaussian and that all residuals ϵ_j are identically distributed, this indeed amounts to sampling from $\mathcal{L}\left(X_j \mid X_{-j}, \tilde{X}_{1:j-1}\right)$. This approach is described in Algorithm 9.

Algorithm 9: Sequential generation of non-parametric Knockoffs by learning to predict X_j from $(\mathbf{X_{-j}}, \mathbf{\tilde{X}_{1:j-1}})$ using a model f.

Input : f
for $j \in [1, p]$ **do**
  Fit a prediction model f_j on $((\mathbf{X_{-j}}, \mathbf{\tilde{X}_{1:j-1}}), X_j)$ `// Typically a Lasso model`
  Compute the residual $\widehat{\epsilon}_j = X_j - f_j((\mathbf{X_{-j}}, \mathbf{\tilde{X}_{1:j-1}}))$
  Sample $\tilde{X}_j = f_j((\mathbf{X_{-j}}, \mathbf{\tilde{X}_{1:j-1}})) + \widehat{\epsilon}_{\sigma(j)}$
end
Return $\mathbf{\tilde{X}_{1:p}}$

In practice, this Algorithm is computationally costly to run. Fitting the models f_j cannot be parallelized as each fit depends on previously built Knockoffs. To make computations tractable, we propose a modified version of this Algorithm, in which samples are drawn from $\mathcal{L}\left(X_j \mid X_{-j}\right)$ instead of $\mathcal{L}\left(X_j \mid X_{-j}, \tilde{X}_{1:j-1}\right)$. In practice, this allows us to fit all models f_j in parallel. Once this is done for all variables, Knockoffs are built by shuffling the residuals: $\tilde{X}_j$ is chosen as $f_j(\mathbf{X_{-j}}) + \widehat{\epsilon}_{\tau(j)}$ with τ a permutation of $[\![p]\!]$.

In words, we remove conditioning on previously built Knockoff variables to make the algorithm easily parallelizable. While this method does not yield theoretically valid Knockoffs, we show empirically that in high-dimensional problems, it outputs very close results to the theoretically grounded sequential approach. The proposed approach is described in Algorithm 10.

The important point is that these approaches circumvents the difficult problem of covariance estimation in high-dimensional settings. Note that using a Lasso model is suitable in regimes where $n \simeq p$ – if $n \gg p$ selecting a more expressive model such as random forests (Breiman, 2001) or gradient boosting trees (Chen et al., 2015) is necessary to retain control.

8.3 When Knockoffs fail: diagnosing non-exchangeability

A direct consequence of non-exchangeability is that Knockoff statistics of null variables $\{X_j \text{ for } j \in H_0\}$ are no longer symmetrical – yet this property is key to achieving error

Algorithm 10: Parallel generation of Non-parametric Knockoffs by learning to predict X_j from $\mathbf{X_{-j}}$ using a model f.

Input : f

for $j \in [1, p]$ **do**
  Fit a prediction model f_j on $(\mathbf{X_{-j}}, X_j)$ `// Typically a Lasso model`
  Compute the residual $\widehat{\epsilon}_j = X_j - f_j(\mathbf{X_{-j}})$
end
for $j \in [1, p]$ **do**
  Sample $\tilde{X}_j = f_j(\mathbf{X_{-j}}) + \widehat{\epsilon}_{\tau(j)}$ `// τ is a permutation of` $[\![p]\!]$
end
Return $\tilde{\mathbf{X}}_{\mathbf{1:p}}$

control via valid Knockoffs as in Candès et al., 2018. Said otherwise, non-exchangeability can make Knockoff importance scores non comparable with real variable importance scores which leads to bias in Knockoff statistics. This is illustrated in Figure 8.1 using a simulated data setup described in Section 8.4.

On the left panel, we display the inverse Cumulative Distribution Function (CDF) of Knockoff statistics of null variables, where Knockoffs come from an oracle, with known covariance. On the central panel, we display the inverse CDF in the same setup but with a covariance estimate that relies on the Graphical Lasso. On the right panel, we display the inverse CDF of statistics obtained using non-parametric Knockoffs built from the data. Knockoff statistics of oracle Knockoffs are nearly perfectly symmetric, as seen on the left panel. Notice that Knockoff statistics of null variables using data-derived Gaussian Knockoffs are not symmetric: they are skewed towards positive values, signalling that real and Knockoff importance scores are not comparable. Non-parametric Knockoffs nearly recover exact symmetry of null Knockoffs statistics.

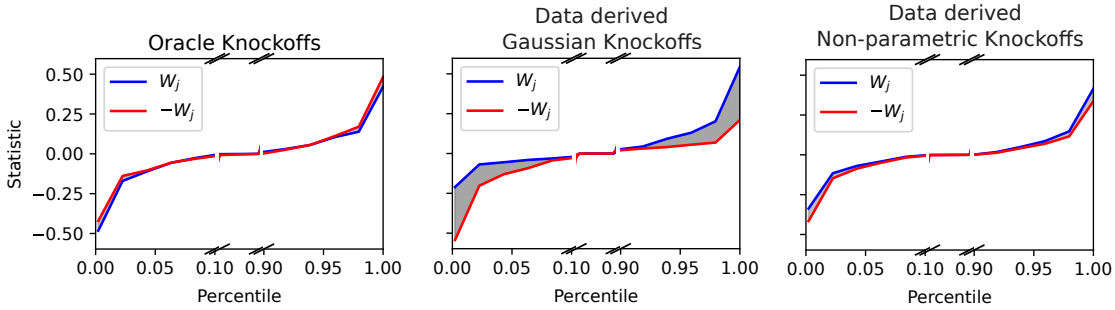


Figure 8.1: **Inverse CDFs of Knockoff statistics of null variables for oracle and data-derived Knockoffs.** On the left panel, we display Knockoff statistics of null variables in an oracle simulation setup where perfectly valid Knockoffs have been built – i.e., using the true covariance of the data, computed using a massive amount of samples. Notice that the inverse CDF is almost symmetrical, which ensures validity of the inference. On the central panel, we display Knockoff statistics of null variables in the same setup but with a covariance estimate that relies on the Graphical Lasso. Notice that null Knockoffs statistics are not symmetrical: they are skewed towards positive values, signalling that real and Knockoff importance scores are not comparable. On the right panel, we display Knockoff statistics of null variables using non-parametric Knockoffs. Notice that symmetry is recovered, which ensures reliable inference. We use $n = 500$ samples and $p = 500$ variables.

In the remainder of this section, we will focus on diagnosing this problem before assessing its consequences on error control in practice.

A sufficient condition for non-exchangeability. Clearly, Knockoffs exchangeability does not hold if $\mathbf{x}$ is non-Gaussian while $\tilde{\mathbf{x}}$ is Gaussian. One can simply take $S = [\![p]\!]$; then $(\mathbf{x}, \tilde{\mathbf{x}})_{swap(S)} = (\tilde{\mathbf{x}}, \mathbf{x}) \overset{d}{\neq} (\mathbf{x}, \tilde{\mathbf{x}})$. More broadly, any Knockoff generation procedure that relies on unfulfilled assumptions on variable distribution or dependence structure fails to replicate the distribution of $\mathbf{x}$. This leads to a natural sufficient condition to diagnose non-exchangeability: if there exists a classifier that is able to accurately distinguish samples from $\mathbf{x}$ versus samples from $\tilde{\mathbf{x}}$, then $\tilde{\mathbf{x}} \neq \mathbf{x}$ in distribution and exchangeability is violated. This idea is related to the C2ST (Classifier Two-Sample Testing) literature (Gretton et al., 2012; Lopez-Paz and Oquab, 2016).

Classifier Two-Sample Testing for Knockoffs. Formally, we wish to test the null hypothesis $H_0 : \tilde{\mathbf{x}} \overset{d}{=} \mathbf{x}$ given n samples of each distribution. Following Section 3 of Lopez-Paz and Oquab, 2016, we proceed by constructing the dataset

$$\mathcal{D} = \left\{\left(x^i, 0\right)\right\}_{i=1}^{n} \cup \left\{\left(\tilde{x}^i, 1\right)\right\}_{i=1}^{n} =: \left\{\left(z^i, l^i\right)\right\}_{i=1}^{2n}.$$

Then, the $2n$ samples of $\mathcal{D}$ are shuffled at random and and split into disjoint training and testing subsets $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$, where $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$ and $n_{\text{te}} := |\mathcal{D}_{\text{te}}|$. Note that in practice, this split is performed several times to mitigate randomness, as in classical cross-validation.

Definition 16 (C2ST statistic, Lopez-Paz and Oquab, 2016)**.** Given $\mathcal{D}_{\text{tr}}$ and $\mathcal{D}_{\text{te}}$, where $\mathcal{D} = \mathcal{D}_{\text{tr}} \cup \mathcal{D}_{\text{te}}$ with $n_{\text{te}} := |\mathcal{D}_{\text{te}}|$ and a binary classifier $g : \mathcal{R}^p \to \{0, 1\}$ trained on $\mathcal{D}_{\text{tr}}$ the C2ST statistic $\hat{t}$ is defined as the classification accuracy on $\mathcal{D}_{\text{te}}$:

$$\hat{t} = \frac{1}{n_{\text{te}}} \sum_{(z^i, l^i) \in \mathcal{D}_{\text{te}}} 1\left[g\left(z^i\right) = l^i\right]$$

Intuitively, if $\tilde{\mathbf{x}} \overset{d}{=} \mathbf{x}$, the test accuracy $\hat{t}$ should remain near 0.5, corresponding to chance-level. Conversely, if $\tilde{\mathbf{x}} \overset{d}{\neq} \mathbf{x}$ and that the binary classifier is able to unveil distributional differences between the two samples, the test classification accuracy $\hat{t}$ should be greater than chance-level. The procedure is summarized in Figure 8.2.

Exchangeability violation by improper sample pairing. Note that the C2ST diagnostic tool only tests that Knockoffs are indeed sampled from the same distribution as the original data. In practice, to have valid Knockoffs, a consistent pairing between original and Knockoffs samples is also needed. Said otherwise, knowing how to sample from the original data distribution is not enough to build valid Knockoffs. We illustrate this point via a simple experiment in the same simulation setup described in Section 8.4. The experiment consists in shuffling 50% of the sample pairings of valid Knockoffs. The middle panel reports the FDP for each simulation run in both scenarios (shuffled and non-shuffled). The left and right sketches schematically illustrate shuffling sample pairings in $2D$. Note that the C2ST diagnostic is invariant to this shuffling.

As seen in Figure 8.3, FDR control is lost when shuffling 50% of the sample pairings of valid Knockoffs. This shows that, beyond equality of the distributions, a proper sample pairing is needed to obtain valid Knockoffs. To test this in practice, one may compute the

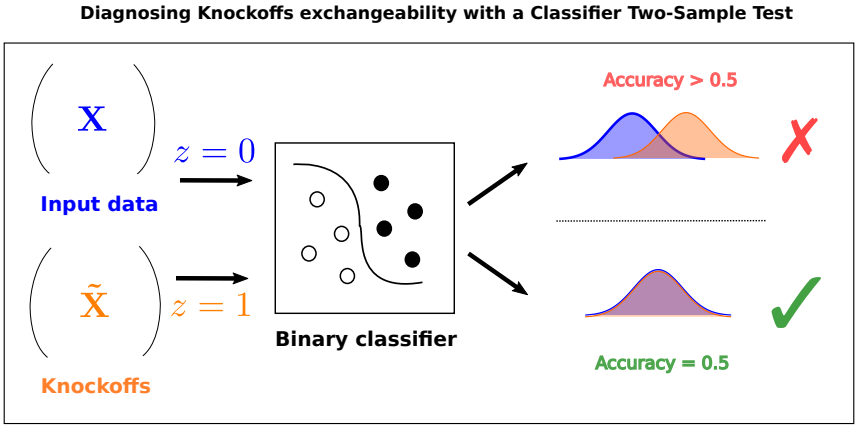


Figure 8.2: **Using a Classifier Two-Sample Test to diagnose exchangeability problems in Knockoffs.** We input both input data and Knockoffs to a binary classifier with two different labels – the goal is to check whether a classifier can distinguish between real variables and Knockoff variables. If the classifier's accuracy is substantially above chance level, exchangeability is violated since the joint distributions of input data and Knockoffs are not equal.

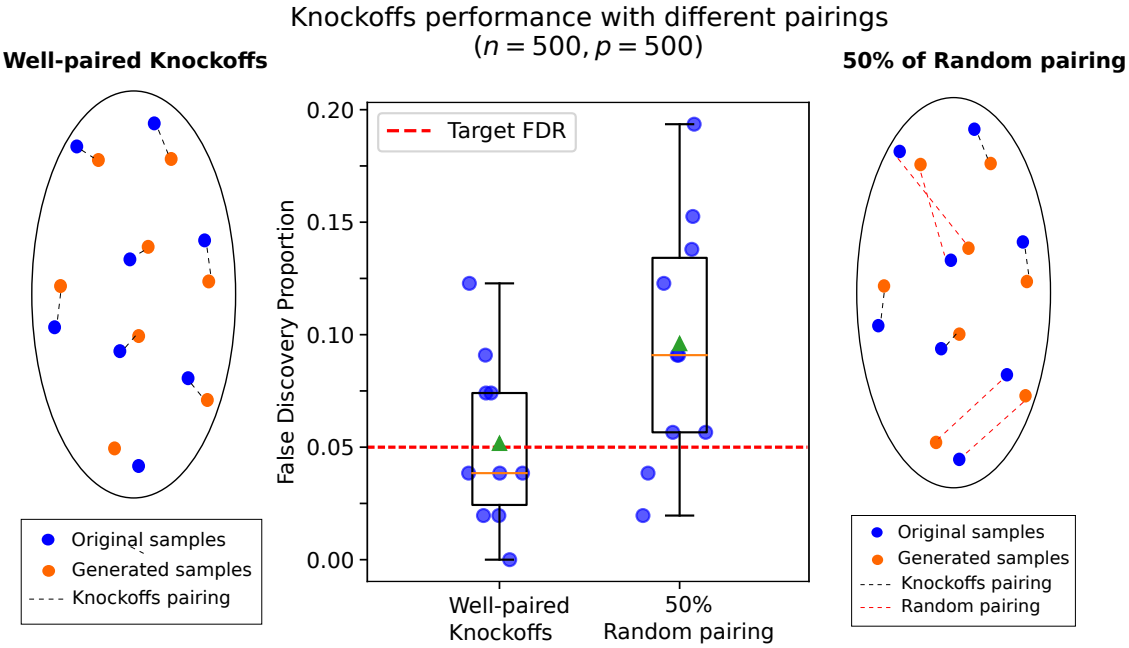


Figure 8.3: **Exchangeability violation by improper sample pairing.** In the left boxplot, we build valid Knockoffs via the Gaussian algorithm and perform inference as in (Candès et al., 2018). In the right boxplot, we use the same Knockoffs but shuffle 50% of the sample pairings before inference. We repeat this experiment 10 times. Note that, coherently with theory, the FDR is indeed controlled in the left boxplot. However, in the right boxplot, error control is lost due to improper sample pairings. The left and right sketches schematically illustrate shuffling sample pairings in $2D$.

optimal assignment between real samples and Knockoff samples e.g. via the Hungarian algorithm (Kuhn, 1955). If the resulting assignment doesn't match the original one, then Knockoffs cannot be exchangeable.

8.4 Experimental results

We aim at showing the practical consequences of exchangeability violations on error control and diagnosis via C2ST statistics.

Simulated data setup. To assess the consequences of varying degrees of non-exchangeability with Gaussian Knockoffs, we design a simulation setup in which we control the difficulty to produce valid Knockoffs. At each simulation run, we first generate i.i.d. Gaussian data with a tensor structure $\mathbf{X} \in \mathbb{R}^{n \times a \times b \times c}$ with $a \times b \times c = p$. The idea is to generate n samples in a 3D feature space of dimensions (a, b, c). Then, we apply an isotropic smoothing kernel of width w across these three dimensions, mimicking a smooth three-dimensional structure, and flatten the data to obtain $\mathbf{X} \in \mathbb{R}^{n \times p}$.

Then, we draw the true support $\boldsymbol{\beta}^* \in \{0, 1\}^p$. The number of non-null coefficients of $\boldsymbol{\beta}^*$ is controlled by the sparsity parameter s_p, i.e. $s_p = \|\boldsymbol{\beta}^*\|_0 / p$. The target variable $\mathbf{y}$ is built using a linear model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \sigma\boldsymbol{\epsilon},$$

where $\sigma = \|\mathbf{X}\boldsymbol{\beta}^*\|_2 / (\mathrm{SNR}\|\boldsymbol{\epsilon}\|_2)$ controls the amplitude of the noise, SNR being the signal-to-noise ratio. We choose the setting $n = 500, p = 500, s_p = 0.1, \mathrm{SNR} = 2$. To obtain $p = 500$, we use $(a, b, c) = (10, 10, 5)$. Note that using $w = 0$ is equivalent to sampling i.i.d. Gaussians since no smoothing is applied in this case. We vary the kernel width w in the interval $[0, 1.25]$ to parametrize the difficulty of the problem. This parameter is closely related to the level of correlation observed in the data; as the kernel width increases, the data becomes increasingly correlated.

In all settings, we use the Graphical Lasso for covariance estimation (Friedman et al., 2008). For each of N simulations, we compute the empirical FDP of the Vanilla Knockoff selection set S and the C2ST statistic averaged across 5-fold cross-validation:

$$\widehat{FDP}(S) = \frac{|S \cap \mathcal{H}_0|}{|S|}, \quad \hat{t} = \frac{1}{n_{\mathrm{te}}} \sum_{(z^i, l^i) \in \mathcal{D}_{\mathrm{te}}} 1\left[g\left(z^i\right) = l^i\right].$$

We first benchmark all existing methods for building Knockoffs for continuous data enumerated in Section 3.2.2 using $w = 0.5$. This value yields local correlations levels that are comparable to those observed in fMRI data.

The left panel of Figure 8.4 shows that all methods except Gaussian and non-parametric Knockoffs fail when p grows larger (using $n = p$). Amongst the state-of-the-art methods, Gaussian Knockoffs are the best candidate as they maintain a C2ST accuracy of at most 0.6. The proposed approach maintains chance-level accuracy in all settings. Note that the data at hand is Gaussian – therefore, the non-exchangeability of Gaussian Knockoffs is necessarily due to the problem of covariance estimation. In contrast, the proposed non-parametric approach circumvents this difficult estimation problem by relying on regressions.

The right panel of Figure 8.4 shows that increasing the number of samples n for a fixed p does not fix the problem, as all methods apart from Non-parametric Knockoffs

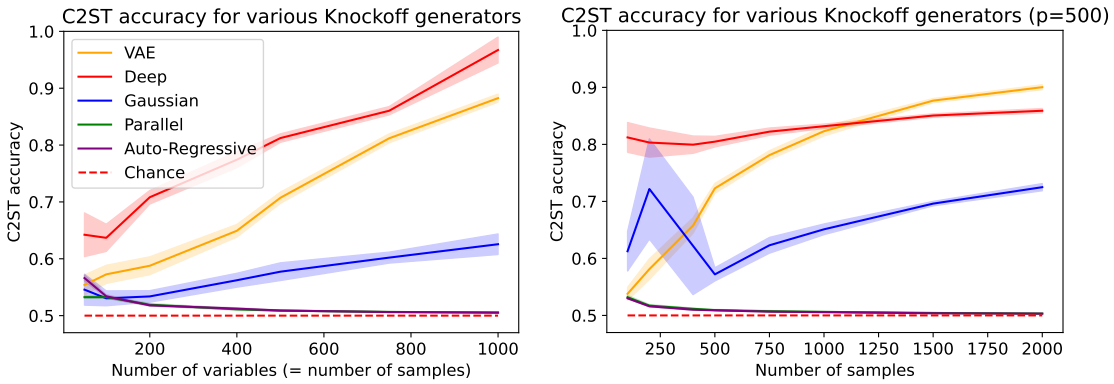


Figure 8.4: **C2ST diagnostic metric for varying number of variables and samples.** In the left panel, the number of variables p grows and $n = p$ at each point. Note that, when $p > 100$ the C2ST accuracy is clearly above chance level for all methods apart from Gaussian Knockoffs and Non-parametric Knockoffs, signaling non-exchangeability in this regime. In the right panel, the number of variables is kept constant with p set to 500. The number of samples varies from 100 to 2000. Apart from Non-parametric Knockoffs, all other methods fail to produce valid Knockoffs, even for large values of n.

fail to provide valid Knockoffs. Interestingly, C2ST accuracy rises as n grows larger for all methods apart from Non-parametric Knockoffs. Intuitively, one might expect that as n grows larger, producing valid Knockoffs becomes easier and therefore that the C2ST accuracy should decrease. A possible interpretation of this result is the presence of two competing effects: while learning features of the underlying joint distribution of the data is indeed easier when n grows, the number of training samples accessible to the classifier also grows as it is equal to $2n$. Therefore, the discriminating power of the classifier also improves when n grows larger. From now on, we discard the VAE and Deep approaches because of their poor empirical performance.

The left panel of Figure 8.5 shows that, for Gaussian Knockoffs, the FDR is controlled only in the easiest settings, i.e. $w \in [0, 0.5]$. For $w > 0.5$ the achieved FDR is substantially above the target FDR, and it grows as the smoothing increases. The $C2ST$ accuracy is clearly above chance level for $w > 0.5$. By contrast, non-parametric Knockoffs maintain FDR control in all settings, while the $C2ST$ accuracy remains near chance level.

Semi-simulated data setup. We now turn to evaluating the proposed approach and diagnostic tool on real data. Following Blain et al., 2023; Nguyen et al., 2022, we use semi-simulated data to evaluate the proposed method with observed $\mathbf{X}$. We use a simulated response $\mathbf{y}$ to be able to compute the FDP. We consider a first functional Magnetic Resonance Imaging (fMRI) dataset $(\mathbf{X}_1, \mathbf{y}_1)$ on which we perform inference using a Lasso estimator; this yields $\boldsymbol{\beta}_1^* \in \mathbb{R}^p$ that we will use as our ground truth. Then, we consider a separate fMRI dataset $(\mathbf{X}_2, \mathbf{y}_2)$ for data generation. The point of using a separate dataset is to avoid circularity between the ground truth definition and the inference procedure. Concretely, we discard the original response vector $\mathbf{y}_2$ for this dataset and build a simulated response $\mathbf{y}_2^{sim}$ using a linear model:

$$\mathbf{y}_2^{sim} = \mathbf{X}_2 \boldsymbol{\beta}_1^* + \sigma \boldsymbol{\epsilon},$$

where we set σ so that $SNR = 4$. We consider 7 binary classifications problems like

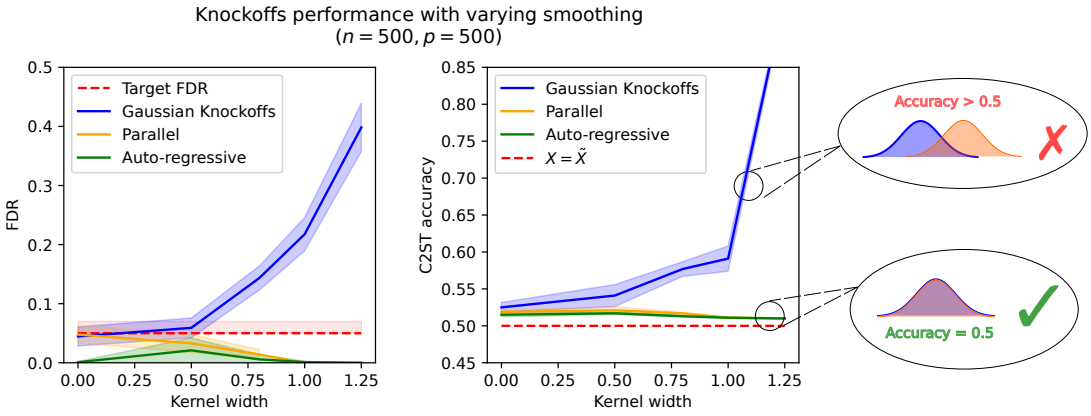


Figure 8.5: **FDP and C2ST diagnostic metric for varying smoothing.** We use the 3D smoothing kernel width to parametrize the correlation present in the data, which in turn tunes the difficulty to produce valid Knockoffs. We run 20 simulations for each kernel width. For Gaussian Knockoffs, the FDR is controlled only in the easiest settings, i.e. $w \in [0, 0.5]$. Note that using $w = 0.5$ represents a common setting encountered on real data: it indeed yields local correlations levels that are comparable to those observed in fMRI data. For $w > 0.5$ the achieved FDR is substantially above the target FDR, and it grows as the smoothing increases. Note that the C2ST accuracy is clearly above chance level for $w > 0.5$, signaling non-exchangeability in this regime. Non-parametric Knockoffs – defined in Section 8.2 – preserve error control in all regimes, which is consistent with C2ST accuracy remaining near chance level.

"gambling" (rewards vs loss) taken from the HCP dataset. For each of these classification problems, the dataset consists in $778 \times 2 = 1556$ samples and 1000 features. These features are obtained by averaging fMRI signals within a Ward parcellation scheme, which is known to yield spatially homogeneous regions (Thirion et al., 2014).

As we use 7 datasets of HCP, we obtain $42 = 7 \times 6$ possible pairs. Since we consider β_1^* as the ground truth, the FDP can be computed. Figure 8.6 shows False Discovery Proportion levels for 42 semi-simulated fMRI datasets based on HCP data for 5 different Knockoff-based inference methods, using either Gaussian Knockoffs or non-parametric Knockoffs[1]. All methods – see Blain et al., 2023 for a detailed comparison – exhibit problematic False Discoveries proportions using default Gaussian Knockoffs (left panel) and Graphical Lasso-based covariance learning. The expeted behavior is recovered using non-parametric Knockoffs (right panel). As one could expect, the C2ST metric obtains high discrimination power (0.95 prediction accuracy) for the Gaussian Knockoffs (signalling non-exchangeability), but is at chance (accuracy of 0.51) for non-parametric Knockoffs. Regarding the pairing condition, the Hungarian algorithm signals that both Gaussian and Non-parametric Knockoffs are optimally paired with original variables.

Computation time of non-parametric Knockoffs. We perform benchmarks to evaluate the computation time needed for the proposed Non-parametric Knockoffs algorithm compared to the Gaussian algorithm. We use the simulation setup described in Section 8.3 and vary the number of variables p. For all values of p, we use $n = p$ samples. We first

[1]In Figure 8.6 we display the result using parallel Knockoffs. AR Knockoffs yield similar results.

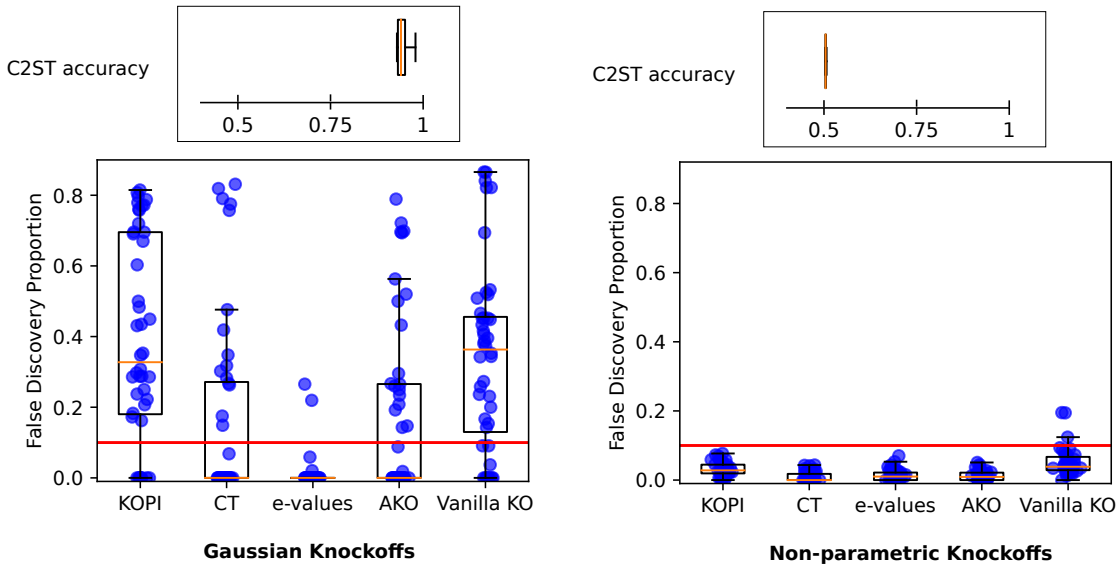


Figure 8.6: **Empirical FDP on semi-simulated data for 42 contrast pairs using Gaussian vs Non-parametric Knockoffs.** We use 7 HCP contrasts C0: "Motor Hand", C1: "Motor Foot", C2: "Gambling", C3: "Relational", C4: "Emotion", C5: "Social", C6: "Working Memory". We consider all 42 possible train/test pairs: the train contrast is used to obtain a ground truth, while the test contrast is used to generate the response. Inference is performed using the 5 methods considered in the chapter and the empirical FDP is reported. Notice that error control of all methods is violated using default Gaussian Knockoffs (left panel) and recovered using Non-parametric Knockoffs (right panel). The C2ST metric is coherent in both cases, with an average accuracy of 0.95 for the Gaussian Knockoffs (signalling non-exchangeability) and 0.51 for the non-parametric Knockoffs.

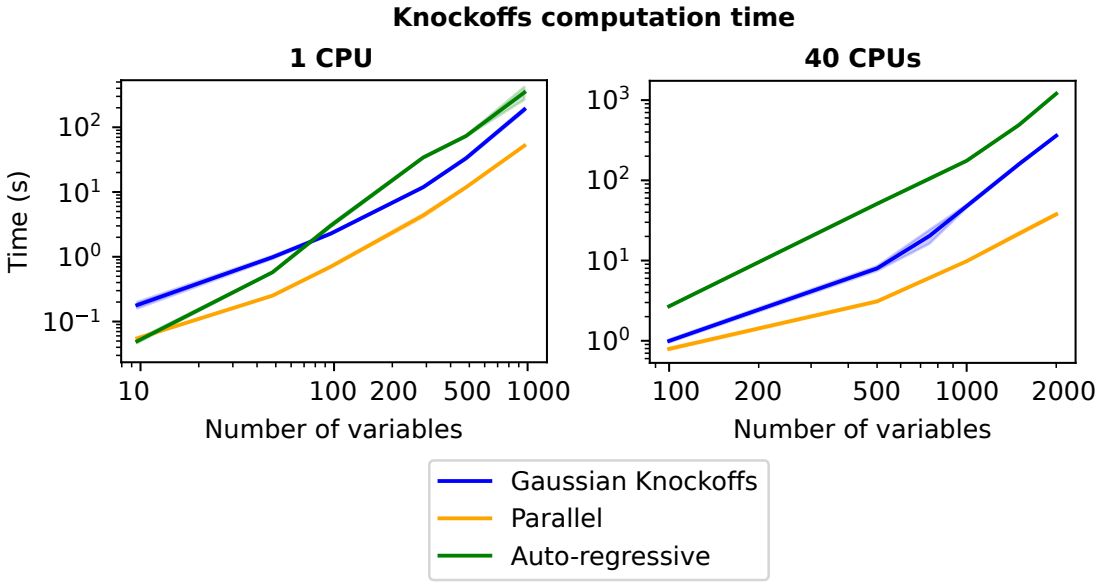


Figure 8.7: **Non-parametric and Gaussian Knockoffs computation time.** We use either one CPU or 40 CPUs and report the computation time for various problem dimensions in log-log scale. Note that using non-parametric (Parallel) Knockoffs on a single CPU yields around a 3× speedup across all problem dimensions compared to Gaussian Knockoffs. Using 40 CPUs, a 5× speedup is achieved for problems under 1000 variables, and a 10× to 15× speedup for problems of up to 2000 variables.

consider using a single CPU and then consider using 40 CPUs. For the Gaussian approach, parallelization is done in the Graphical Lasso covariance estimator via the n_{jobs} argument in scikit-learn (Pedregosa et al., 2011). Note that profiling the code when using a single CPU shows that most of the computation time is spent in covariance estimation. For the Non-parametric approach, the 40 CPUs are used to train Lasso models that predict X_j from $\mathbf{X_{-j}}$ in parallel.

Note that using non-parametric parallel Knockoffs on a single CPU yields around a 3× speedup across all problem dimensions compared to Gaussian Knockoffs. Using 40 CPUs, a 5× speedup is achieved for problems under 1000 variables, and a 10× to 15× speedup for problems of up to 2000 variables. The auto-regressive algorithm is slower than both Gaussian Knockoffs and the parallel version.

8.5 Discussion

Knockoffs are a powerful and efficient method for controlled variable selection. This inference procedure allows performing conditional variable selection in one round of inference, without considering each variable individually as in Conditional Randomization Tests approaches.

The statistical guarantees provided by this method rely on the ability to construct valid Knockoffs - namely, Knockoffs and original variables must be exchangeable so that false positives can be provably controlled. However, constructing valid Knockoffs based only on the available observations is not trivial.

The difficulty of this problem is particularly salient in contexts where *i)* the data at

hand has a strong dependence structure and *ii)* when the number of variables p is large. Note that these are common characteristics of high-dimensional variable selection problems where Knockoffs are relevant, such as fMRI brain mapping or genomic data analyses. In fact, i) and ii) often go together in practice: high dimension is associated with high correlation among measurements. These characteristics make the estimation of the joint distribution of the data increasingly difficult, whether done through covariance estimation (as in the Gaussian algorithm) or via deep learning based techniques. Our first contribution in this context is to show that existing generation procedures are unable to produce valid Knockoffs in such settings.

This chapter also provides insight into the consequences of exchangeability violations: experiments on both real and simulated data show the important impact of departures from exchangeability on the reliability of Knockoff-based methods. These issues can lead to instability in error control for all Knockoff-based methods. To address this challenge in real-world scenarios, we have introduced a diagnostic tool that provides practitioners with the means to identify exchangeability problems. This tool relies on two parts: a classifier two-sample test which aims at detecting distributional differences between original and Knockoff variables, and a procedure that controls the proper pairing between original observations and their knockoff. Note that having a perfect generative model is not enough to obtain valid Knockoffs, as this does not yield samples paired with the original data. We therefore have two necessary (but not sufficient) conditions, for exchangeability.

As part of our efforts to mitigate the exchangeability problems associated with non-Gaussian data or poor covariance estimation, we propose an efficient alternative approach for constructing non-parametric Knockoffs. We prove theoretically that this approach produces valid Knockoffs, provided that the number of samples is large enough. Experiments on simulated data indicate that the Knockoffs we obtain are valid. However, the proposed approach has some limitations. In particular, a sufficient condition for exchangeability is still lacking, so that one may indeed encounter situations where supposedly valid Knockoffs lead to biased FDR guarantees. Another limitation in the non-parametric Knockoff estimation procedure is that the hypothesis that the learner is Bayes optimal leaves open the question of which learner to choose; this choice is obviously important in practice, both for the validity of the Knockoff inference and for the power of the method.

The code for the proposed diagnostic tool and alternative non-parametric Knockoffs algorithm is available at `https://github.com/alexblnn/KnockoffsDiagnostics`. Note that this algorithm is also less costly in terms of computing time than the Gaussian algorithm. Using a single CPU, a 3× speedup is achieved compared to the original algorithm. When leveraging 40 parallel CPUs, a 15× speedup is achieved on fMRI data and other large problems. Details about this benchmark are available in Section 8.4.

Chapter 9

Tight and reliable conformal prediction

Summary. Split Conformal Prediction (SCP) provides a computationally efficient way to construct confidence intervals in prediction problems. Notably, most of the theory built around SCP is focused on the single test point setting. In real-life, inference sets consist of multiple points, which raises the question of coverage guarantees for many points simultaneously. While *on average*, the False Coverage Proportion (FCP) remains controlled, it can fluctuate strongly around its mean, the False Coverage Rate (FCR). We observe that when a dataset is split multiple times, classical SCP may not control the FCP in a majority of the splits. We propose CoJER, a novel method that achieves sharp FCP control in probability for conformal prediction, based on a recent characterization of the distribution of conformal p-values. We show through extensive real data experiments that CoJER provides FCP control while standard SCP does not. Furthermore, CoJER yields shorter intervals than the *state-of-the-art* method for FCP control and only slightly larger intervals than standard SCP.

9.1 Conformal prediction for multiple test points

In most practical applications, the test set for which we want to obtain confidence intervals contains many points. Say that we have m test points $(X_{n+1}, Y_{n+1}), \ldots, (X_{n+m}, Y_{n+m})$. Performing split conformal prediction yields m confidence intervals $\mathcal{C}(\alpha) = \left(\widehat{C}_{i,\alpha}\right)_{i \in [\![m]\!]}$ with $\widehat{C}_{i,\alpha} = \left[\hat{\mu}\left(X_{n+i}\right) \pm S_{(\lceil (n+1)(1-\alpha) \rceil)}\right]$. The marginal guarantee holds, i.e.:

$$\forall i \in [\![m]\!], \quad \mathbb{P}\left\{Y_{n+i} \in \widehat{C}_{i,\alpha}\left(X_{n+i}\right)\right\} \geq 1 - \alpha.$$

To quantify coverage on the set of m points using intervals $\mathcal{I} = (\mathcal{I}_i)_{i \in [\![m]\!]}$, we define the False Coverage Proportion (FCP) and False Coverage Rate (FCR):

$$\mathrm{FCP}(\mathcal{I}) := \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}\left\{Y_{n+i} \notin \mathcal{I}_i\right\}, \qquad \mathrm{FCR}(\mathcal{I}) := \mathbb{E}[\mathrm{FCP}(\mathcal{I})].$$

Clearly, FCR control holds at level α for standard split conformal prediction:

$$\mathrm{FCR}(\mathcal{C}(\alpha)) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{P}\left\{Y_{n+i} \notin \widehat{\mathcal{C}}_{i,\alpha}\right\} \leq \alpha.$$

Interestingly, this doesn't guarantee that $\mathrm{FCP}(\mathcal{C}(\alpha)) \leq \alpha$ with high probability as noted by Gazin et al., 2024. This is analogous to the distinction between False Discoveries Proportion control and False Discovery Rate control highlighted in Figure 2.1.

To ensure precise control, we want to build FCP upper bounds that hold with high probability. For any pre-specified level $\delta > 0$, this amounts to building $\left(\overline{\mathrm{FCP}}_{\alpha,\delta}\right)_{\alpha\in[0,1]}$ such that:

$$\mathbb{P}\left(\forall \alpha \in [0,1], \quad \mathrm{FCP}(\mathcal{C}(\alpha)) \leq \overline{\mathrm{FCP}}_{\alpha,\delta}\right) \geq 1-\delta. \tag{9.1}$$

9.2 Sharp FCP control for conformal prediction

9.2.1 FCP control and conformal p-values

In this work, we intend to obtain JER control for conformal p-values and derive FCP bounds from this control. We first recall a close link between FCP control and the empirical Cumulative Distribution Function (CDF) of conformal p-values. Note that the FCP also solely depends on the joint distribution of the conformal p-values.

Proposition 1 (Empirical CDF of p-values and FCP, Gazin et al., 2024)**.** *Denote by $\widehat{F}_m$ the empirical* CDF *of the joint distribution of $(p_1, \ldots, p_m)$. For any $\alpha \in [0,1]$ denote $\mathcal{C}(\alpha)$ the split conformal intervals. Then:*

$$\mathrm{FCP}(\mathcal{C}(\alpha)) = \widehat{F}_m(\alpha).$$

Using this technical remark, we can obtain a tight link between FCP control and JER control.

Proposition 2 (FCP and JER)**.** *Let $\mathbf{t}$ be an arbitrary threshold family. Denote $j_0(\alpha) = \min\{j \in [\![m]\!] : \alpha \leq t_j\}$. Then:*

$$\mathbb{P}\left(\exists \alpha \in [0,1], \quad \mathrm{FCP}(\mathcal{C}(\alpha)) > \frac{j_0(\alpha)}{m}\right) \leq \mathrm{JER}(\mathbf{t}).$$

Proof. First, we note that $\mathrm{JER}(\mathbf{t})$ may be written in function of $\widehat{F}_m$:

$$\begin{aligned}
\mathrm{JER}(\mathbf{t}) &= \mathbb{P}\left(\exists j \in [\![m]\!] : p_{(j)} < t_j\right) \\
&= \mathbb{P}\left(\exists j \in [\![m]\!] : \sum_{i=1}^{m} \mathbf{1}\left\{p_i \leq t_j\right\} \geq j\right) \\
&= \mathbb{P}\left(\exists j \in [\![m]\!] : \widehat{F}_m(t_j) \geq \frac{j}{m}\right).
\end{aligned}$$

To conclude, we note that if for some $\alpha \in [0,1]$ we have $\mathrm{FCP}(\mathcal{C}(\alpha)) \geq j_0(\alpha)/m$, then by definition of $j_0(\alpha)$ combined with Proposition 1, $j := j_0(\alpha) \in [\![m]\!]$ is such that $\widehat{F}_m(t_j) \geq \widehat{F}_m(\alpha) = \mathrm{FCP}(\mathcal{C}(\alpha)) \geq j/m$. □

Proposition 2 implies that bounds of the form (9.1) may be obtained directly from JER controlling families.

Corollary 6. *Assume that $\mathbf{t}$ controls the* JER *at level $\delta > 0$. Denote $j(\alpha, \delta) = \min\{j \in [\![m]\!] : \alpha \leq t_j\}$. Then*

$$\mathbb{P}\left(\forall \alpha \in [0,1], \quad \mathrm{FCP}(\mathcal{C}(\alpha)) \leq \frac{j(\alpha,\delta)}{m}\right) \geq 1-\delta.$$

9.2.2 Building a JER controlling family

In practice, we need to build a JER controlling family $\mathbf{t}$ to compute FCP bounds. We follow the approach of Blain et al., 2023, which uses similar ideas to reach False Discoveries control in for Knockoff (Candès et al., 2018) inference. Firstly, we need to be able to estimate the JER of any threshold family $\mathbf{t}$. To this end, samples from the joint distribution of ordered conformal p-values are needed. We rely on the main result of Gazin et al., 2024 which characterizes this distribution precisely.

Theorem 7 (Joint distribution of conformal p-values, Gazin et al., 2024)**.** *For any vector $U = (U_1, \dots, U_n) \in [0,1]^n$, define the discrete distribution P^U on $\left\{\frac{\ell}{n+1}, \ell \in [\![n+1]\!]\right\}$ as:*

$$P^U(\{\ell/(n+1)\}) = U_{(\ell)} - U_{(\ell-1)}, \quad \ell \in [\![n+1]\!],$$

where $0 = U_{(0)} \leq U_{(1)} \leq \cdots \leq U_{(n)} \leq U_{(n+1)} = 1$. Then, conformal p-values follow the distribution $P_{n,m}$:

$$P_{n,m} = \mathcal{D}\left(q_i, i \in [\![m]\!]\right), \text{ where}$$

$$\begin{cases} (q_1, \dots, q_m \mid U) \overset{i.i.d.}{\sim} P^U; \\ \text{and } U = (U_1, \dots, U_n) \overset{i.i.d.}{\sim} \mathrm{Unif}([0,1]). \end{cases}$$

The approach of Gazin et al., 2024 consists in using this distribution to obtain a Dvoretzky–Kiefer–Wolfowitz–Massart like inequality (DKWM; Massart, 1990). The original DWKM inequality only holds under independence and therefore cannot be used in this context. The inequality obtained bounds the gap between the empirical CDF and the true CDF with high probability. FCP bounds can in turn be obtained using the CDF formulation of Proposition 1.

For the approach we propose, we derive a straightforward algorithm to sample from the joint distribution of conformal p-values from this Theorem.

Algorithm 11: Sampling from the joint distribution of conformal p-values using Theorem 7.

Input: B the number of MC draws; n the number of calibration points; m the number of test points
Output: $\mathbf{\Pi}_0 \in [0,1]^{B \times m}$ a matrix of simulated p-values
$\mathbf{\Pi}_0 \leftarrow \mathrm{zeros}(B, m)$
for $b \in [1, B]$ **do**
  Sample $[u_1, \dots, u_n]$ from $\mathcal{U}([0,1])^n$
  Sample $[q_1, \dots, q_m]$ from P^U // $P^U(\{\ell/(n+1)\}) = U_{(\ell)} - U_{(\ell-1)}, \quad \ell \in [\![n+1]\!]$
  $\mathbf{\Pi}_0[b] \leftarrow [q_1, \dots, q_m]$
end
Return $\mathbf{\Pi}_0$

We draw B Monte-Carlo samples using Algorithm 11. This yields a set of B vectors of conformal p-values denoted by $p_b \in \mathbb{R}^p$ for each $b \in [\![B]\!]$. This allows us to evaluate the empirical JER, which estimates the actual JER. Using the same procedure and notation as in Chapter 7 we obtain the following result:

Theorem 8 (JER control for conformal p-values)**.** *Consider the threshold family defined by* $\mathbf{t}_\delta^B = \mathbf{T}^0(\lambda_B(\delta))$*. Then, as* $B \to +\infty$*,*

$$\text{JER}(\mathbf{t}_\delta^B) \leq \delta + O_P(1/\sqrt{B}).$$

The number B of Monte-Carlo samples in Theorem 8 can be chosen arbitrarily large to obtain JER control, leading to valid FCP bounds via Corollary 6. Algorithm 5 describes all the steps needed to compute $\mathbf{t}_\delta^B$. The proof of this result is the same as Theorem 4. This bound is fully nonparametric and therefore expected to yield tighter intervals than Gazin et al., 2024's approach. We call the resulting approach **CoJER** (Conformal - JER).

9.3 Aggregated conformal prediction

While conformal prediction coverage guarantees are *distribution free*, the confidence interval output by the method can strongly depend on the chosen model $\hat{\mu}$ in practice. Mitigating the consequences of such modeling decisions motivates the use of aggregation schemes to obtain more stable and generalizable confidence intervals.

Say that we have K models $\hat{\mu}_1, \ldots, \hat{\mu}_K$ fitted on $\mathcal{D}_{train}$. The goal of aggregation is to build a valid confidence interval $\widehat{C}_\alpha$ that takes into account the information provided by each model. Aggregating schemes for conformal prediction have been introduced in Lei et al., 2018; Barber et al., 2021. Lei et al., 2018 propose a Bonferroni-type construction, where the confidence interval of each of the K models is built at level α/K. An union bound argument shows that the intersection of these intervals is valid at level α, therefore yielding FCR control at level α.

Barber et al., 2021, propose a method that relies on a p-value aggregation result which states that twice the arithmetic mean of valid p-values is a valid p-value – see e.g. (Vovk and Wang, 2020). This results in a FCR controlling procedure.

Therefore, existing solutions require the construction of valid aggregated p-values. Using the exact same notation and procedure as the aggregated case of Chapter 7, we develop a nonparametric aggregation procedure that does not require valid aggregated p-values. We obtain the following result:

Theorem 9 (JER control for aggregated conformal p-values)**.** *Consider the threshold family defined by* $\overline{\mathbf{t}}_\delta^B = \overline{\mathbf{T}}^0(\lambda_B(\delta))$*. Then, as* $B \to +\infty$*,*

$$\overline{\text{JER}}(\overline{\mathbf{t}}_\delta^B) \leq \delta + O_P(1/\sqrt{B}).$$

Proof. The proof is identical to that of Theorem 8 using the empirical aggregated JER. □

The calibrated aggregated threshold family yields valid FCP upper bounds via Corollary 6. We therefore achieve a fully nonparametric aggregation scheme for conformal prediction, along with guarantees on the FCP.

9.4 Experiments

Setup. We use 17 OpenML (Vanschoren et al., 2014) datasets from Grinsztajn et al. (2022). Each dataset is randomly split ($n_{split} = 30$ times) into a train, calibration and

test set. The latter is of size m and denoted by $\mathcal{D}^s_{test}$. We fit 5 regression models on the training sets[1]: Random Forest (RF) (Breiman, 2001), Multi-Layer Perceptron (MLP) (Hinton, 1990), Support Vector Regression (SVR) (Platt et al., 1999) K-Nearest Neighbors (KNN; Cover and Hart, 1967) and Lasso (Tibshirani, 1996).

FCP control. We consider three methods for comparison: classical Split Conformal Prediction, the method proposed by Gazin et al., 2024 to obtain FCP control via DKW-type bounds (Massart, 1990) and the proposed approach. We use $\alpha = 0.1$ for all methods. For FCP controlling methods, we set $\delta = 0.1$ and use SCP with the largest level α' such that $\overline{\text{FCP}}_{\alpha',\delta} \leq \alpha$. For each dataset, we compute for each split the empirical FCP for each model and conformal prediction method. Formally, for a given data set, denoting by $\mathcal{C}^s = \left(\widehat{C}^s_i\right)_{i \in \mathcal{D}^s_{test}}$ the confidence intervals obtained for the s-th split for a given method, the associated empirical FCP is given by:

$$\text{FCP}(\mathcal{C}^s) = \frac{1}{m} \sum_{i \in \mathcal{D}^s_{test}} \mathbf{1}\left\{Y_i \notin \widehat{\mathcal{C}}^s_i\right\}.$$

Then for each dataset, we compute the associated empirical coverage as the proportion of splits for which the FCP control event holds:

$$\text{FCP}_{coverage} = \frac{1}{n_{splits}} \sum_{s=1}^{n_{splits}} \mathbf{1}\left\{\text{FCP}(\mathcal{C}^s) < \alpha\right\}.$$

We also compute the interval length of each method for each dataset. We report the relative length to the shortest interval found amongst all methods, averaged across all splits for each dataset. This allows having a comparable metric for interval informativeness across all datasets.

The left panel of Figure 9.1 shows that across all models and datasets, standard Split Conformal does not guarantee FCP control at level α – this is consistent with theory, as Split Conformal prediction only guarantees FCR control. Strikingly, the proportion of splits for which FCP $\leq \alpha$ for Split Conformal can be as low as 35% for certain models and datasets. Both the proposed method and the method of Gazin et al., 2024 control the FCP as expected. Concretely, this means that for all datasets, the proportion of splits for which FCP $\leq \alpha$ is indeed superior to $1 - \delta$.

The right panel of Figure 9.1 shows that SCP yields the shortest intervals in all settings. This is expected, as FCR control is less stringent than FCP control, leading to shorter intervals. Amongst the two FCP controlling methods, the proposed method is less conservative than the method of Gazin et al., 2024. On average across all models and datasets, the proposed method yields intervals that are only $\sim 15\%$ larger than standard SCP. In worst-case scenarios, the proposed method yields intervals $\sim 25\%$ larger than SCP, while intervals obtained using the method of Gazin et al., 2024 are $\sim 80\%$ larger than SCP intervals. Overall, the proposed method yields sharp FCP control at a modest cost in terms of interval length compared to SCP.

[1]All experiments were performed using 40 CPUs, Intel(R) Xeon(R) CPU E5-2660 v2 @ 2.20GHz

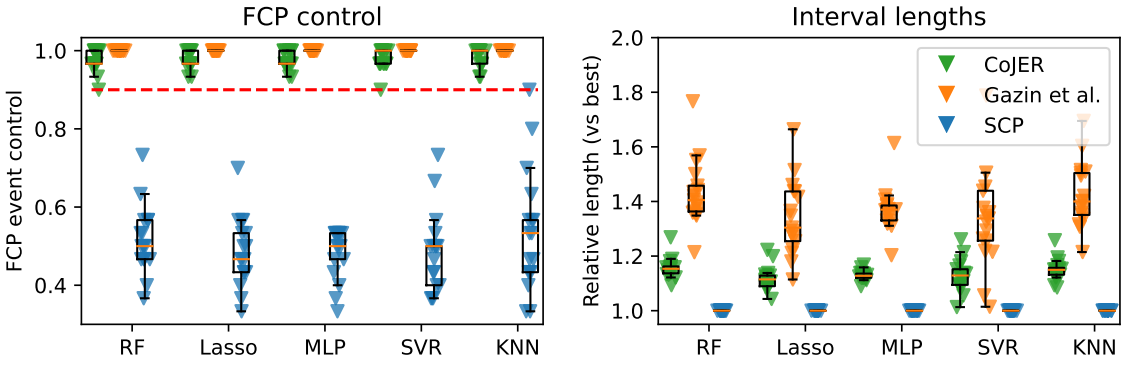


Figure 9.1: **Coverage and relative interval length using 4 models and 17 datasets.** We use 17 OpenML (Vanschoren et al., 2014) datasets from Grinsztajn et al. (2022). Each dataset is split (30 times) into a train, calibration and test set. We fit 5 regression models on the training sets: Random Forest (RF), Multi-Layer Perceptron (MLP), Support Vector Regression (SVR), K-Nearest Neighbors (KNN) and Lasso. Calibration sets are used to compute SCP intervals and conformal p-values. For each method and dataset, we report the FCP coverage, i.e. the proportion of test splits for which the event $FCP \leq \alpha$ was realised. We also report the interval length, relative to the smallest valid interval found amongst all methods. Notice that standard SCP does not guarantee FCP control at level α: for certain datasets and models, FCP event coverage can be as low as 30%. Both the proposed approach and Gazin et al., 2024 obtain FCP control. However, the proposed approach is much less conservative.

Aggregation. We use the five regression models mentioned above and consider three aggregation methods for comparison: the method based on the arithmetic mean of p-value functions proposed by Barber et al., 2021 labeled CV+, the Bonferroni-like construction of Lei et al., 2018 and the proposed method. For the proposed method, we use the harmonic mean as the aggregation scheme.

As in the first experiment, we compute the FCP coverage of each method and the relative interval length. The left panel of Figure 9.2 shows that all three methods control the FCP at level $\delta = 0.1$. While CoJER offers a theoretical guarantee on this control, this is not the case for CV+ and Bonferroni. These two methods likely control the FCP due to excessive conservativeness, as the FCP event is controlled 100% of the time for most datasets using CV+ and Bonferroni intersection.

The right panel of Figure 9.2 shows that CoJER yields the most informative intervals across all datasets. The intervals yielded by the CV+ procedure are $\sim$ 75% larger on average than those of CoJER. The Bonferroni-intersection intervals are $\sim$ 20% larger on average than those of CoJER. Overall, these experiments show that the proposed nonparametric aggregation scheme achieves sharp FCP control while providing tighter intervals that *state-of-the-art* methods.

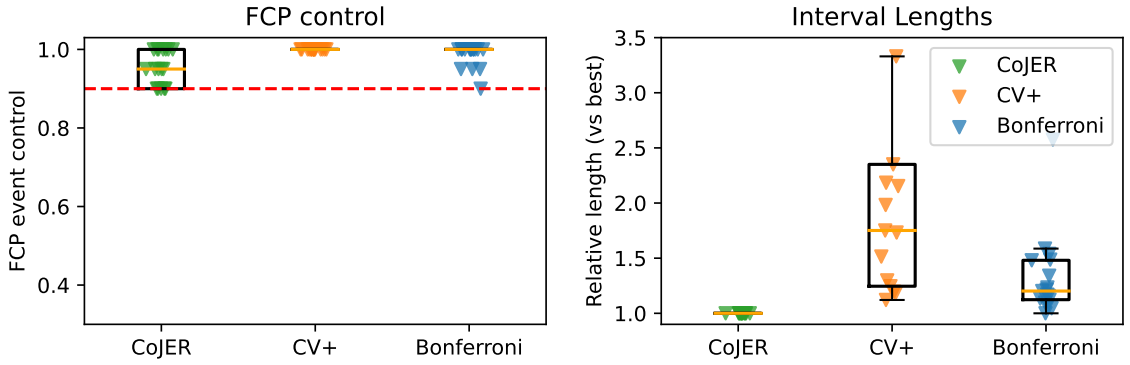


Figure 9.2: **Coverage and relative interval length using 5 models and 17 datasets.** We use 17 OpenML (Vanschoren et al., 2014) datasets from Grinsztajn et al. (2022). Each dataset is split into a train, calibration and test set. We fit 5 regression models on the training sets: Random Forests, Multi-Layer Perceptron, Support Vector Regression, K-Nearest Neighbors and Lasso. Calibration sets are used to compute Split Conformal prediction intervals and conformal p-value functions. These functions are used to compute confidence intervals for the proposed method. For each method and dataset, we report the FCP coverage, i.e. the proportion of test splits for which the event $FCP \leq \alpha$ was realized. We also report the interval length, relative to the smallest valid interval found amongst all methods. The proposed method achieves the expected FCP coverage while providing the most informative intervals.

9.5 Discussion

In this paper, we have proposed a novel method that allows sharp FCP control on conformal prediction. The computational cost of this method is comparable to classical SCP. For given sizes of calibration and test sets, sampling conformal p-values from Algorithm 11 can be done once and for all. Calibration using Algorithm 5 is performed via binary search of complexity $\mathcal{O}(log(B'))$. Computing the empirical JER of a threshold family using Algorithm 3 has a computational complexity of $\mathcal{O}(Bk_{max})$. We provide a Python package containing the code for CoJER available at `https://anonymous.4open.science/r/CoJER-96E5/`.

Additionally, once calibration is performed, the bound of Corollary 6 holds simultaneously for all values of α. In practice, users can try different values of α *post hoc* while retaining valid FCP bounds without needing to relaunch the complete procedure.

We also extend this method to obtain this coverage when aggregating multiple predictors, which provides robustness w.r.t. modeling choices. We show that we obtain tighter intervals than existing aggregation schemes. Notably, valid p-values are not needed to obtain FCP control. Since this control is a direct consequence of JER control on aggregated p-values, it can be obtained for any aggregation scheme f. In particular, our use of the harmonic mean to aggregate p-values leads to valid FCP control, even if the harmonic mean does not yield valid p-values (Chen et al., 2024).

In this chapter, we have focused on the regression setting, while SCP can also be applied in classification tasks. Adapting this method for classification is an interesting prospect. This method could also be extended to other uncertainty quantification frameworks such as

bootstrap or resampling based methods like the jackknife+ (Barber et al., 2021). Since valid p-values are not needed, characterizing the distribution of statistics quantifying uncertainty is sufficient to apply the proposed method.

Another interesting avenue of work is to study the impact of controlling the FCP rather than the FCR on downstream decisions taken using confidence intervals. This type of analysis has been conducted in Vovk and Bendtsen, 2018 in the classical conformal prediction framework and in Perez-Lebel et al., 2024 in the field of model calibration.

Chapter 10

Conclusion

10.1 Summary

In this thesis, we have introduced four contributions to reliable statistical inference for high-dimensional data. First, in Chapter 6 we presented Notip which provides post hoc FDP control with increased power compared to previously existing methods. Post-hoc methods are particularly important in the field of fMRI data analysis, as many inference procedures operate at the cluster level. This can lead to a spatial specificity paradox: detecting large active clusters could indicate the presence of strong signal – yet the information given by cluster-level procedures on the spatial extent of the signal is weaker for large clusters. Such procedures can only guarantee that there is *at least one active voxel* in a given cluster with high probability. In this sense, developing post hoc methods for inference in clusters is an important topic. To promote this type of inference to the neuroimaging community, we have developed an open-source package available at `https://github.com/alexblnn/Notip`.

Second, in Chapter 7, we have introduced KOPI, a novel inference method based on the Knockoffs framework. This approach provides FDP control in probability rather than FDR control. As in Notip, we aim at FDP control in probability rather than in expectation. A starting point of this thesis is to show that these two statistical controls are not equivalent, especially in highly correlated settings as demonstrated in Figure 2.1. To obtain sharp FDP control we use non-parametric methods that rely either on randomization – in Chapter 6 – or on theoretical derivations that allow sampling from the joint distribution of well chosen statistics – in Chapters 7 and 9. In the KOPI procedure, we also leverage multiple Knockoff draws to derandomize inference. We developed an open-source package to allow users to reproduce the chapter's results and to run inference using KOPI which is available at `https://github.com/alexblnn/KOPI`. An important caveat of all Knockoff methods is discussed in Chapter 8: exchangeability violations can occur in common settings using the Gaussian algorithm with catastrophic consequences on error control. To check potential violations, we developed an open-source package that uses C2ST as a diagnostic tool for Knockoffs exchangeability, available at `https://github.com/alexblnn/KnockoffsDiagnostics`. We hope that this incites users of Knockoffs to pay attention to the generation process.

In Chapter 9, we apply the ideas of this thesis to the framework of conformal prediction. Conformal prediction can be viewed as a multiple testing problem across many test points. In this problem, we want to control the proportion of points that are not covered by confidence intervals. Again, controlling the False Coverage Proportion (FCP) does not boil

down to controlling its expectation (FCR). We have shown empirically that standard SCP can yield an empirical FCP greater than the expected level for up to 65% of data splits. To obtain sharp FCP probabilistic bounds, we again rely on a theoretical derivation of the joint distribution of conformal p-values.

10.2 Perspectives

Notip beyond fMRI data. In Chapter 6 we propose Notip and extensively validate the procedure on fMRI data. This procedure is fully nonparametric and can be applied to other types of data. For instance, in genomics data, practitioners face similar challenges as in fMRI analysis. Genomics data presents high local correlation and post hoc analysis of specific genes is of interest (Brzyski et al., 2017). Therefore, performing large-scale experiments using Notip on genomics data is an interesting avenue of work.

Single dataset Notip. Notip relies on an external dataset to build data-driven templates. In practice, choosing – or even obtaining – such a dataset can be cumbersome and costly. We have shown empirically in Section 6.3.6 that using Notip on a single dataset with two different rounds of randomization yields results comparable to the original approach. Obtaining a theoretical result on the validity of Notip used on a single dataset is a desirable goal. This would allow us to recommend definitively this use to practitioners, alleviating the constraint of having an additional dataset.

Notip, calibrated Simes and pARI. In the experimental campaign of Chapter 6, Notip is extensively compared to ARI (Rosenblatt et al., 2018) and to the calibrated Simes template. Using the calibrated Simes template is equivalent to using the original permuted-ARI (pARI, Andreella et al., 2020) procedure. In a recently published comment (Andreella et al., 2024), the authors of pARI propose tuning a hyperparameter to obtain a shifted Simes template: using $\delta > 0$, the shifted Simes template is written $t_k(\lambda) = \frac{(k-\delta)\lambda}{m-\delta}$. The authors argue that this modified procedure using $\delta = 27$ outperforms Notip. Perhaps a similar idea could be used in Notip, i.e. using a $k_{min} > 0$ value in the learned template $\mathbf{t}$ in addition to k_{max}.

Rigorous testing of exchangeability assumptions. In this thesis, we have used two frameworks that required an exchangeability assumption; in the Knockoffs framework, knockoff variables have to be pairwise exchangeable with the original ones for the inference to be valid. In split conformal prediction, calibration and test scores have to be exchangeable for coverage to hold. Both of these assumptions are hard to check in practice. In conformal prediction, much work has been done to extend SCP beyond exchangeability (Gibbs and Candes, 2021; Barber et al., 2023) but only a few papers study the question of exchangeability testing (Vovk, 2023). We studied this at length in Chapter 8 in the context of Knockoffs and proposed a diagnostic tool using classifier two-sample tests. However, this test relies only on checking the violation of a **necessary** condition. A promising avenue of work is to obtain a necessary and sufficient condition that can be tested.

Decision-making using conformal prediction intervals. Conformal prediction produces valid uncertainty quantification via confidence intervals. While we introduced a method that yields error guarantees for multiple test points in Chapter 9, An open question that remains is the use of these confidence intervals to take decisions in practice. In the field of model calibration, Perez-Lebel et al., 2024, study this question through the lens of

maximizing utility functions – perhaps this work could be extended to the framework of conformal prediction.

Chapter 11

Synthèse en Français

Cette thèse de doctorat porte sur le contrôle des fausses découvertes dans les problèmes d'inférence en grande dimension. Elle est divisée en deux parties : dans la première partie, nous introduisons les concepts fondamentaux du test statistique, de la prédiction conforme et de l'analyse des données en neuroimagerie. Dans la seconde partie, nous présentons les nouvelles méthodes développées au cours de cette thèse. Dans la partie consacrée aux bases théoriques, nous introduisons tout d'abord la notion de test statistique au Chapitre 2 et discutons du problème difficile des tests multiples. Nous présentons les principales mesures d'erreur utilisées en tests multiples, telles que le taux d'erreur de famille (Family-wise Error Rate, FWER) et le taux de fausses découvertes (False Discovery Rate, FDR), ainsi que les procédures permettant de contrôler ces erreurs. Nous passons ensuite au Chapitre 3, où nous approfondissons certains outils récents de la littérature sur les tests multiples que nous utilisons dans cette thèse. Notamment, nous introduisons le taux d'erreur conjoint (Joint Error Rate), un cadre général pour contrôler la proportion de fausses découvertes (False Discovery Proportion, FDP), ainsi que la procédure des Knockoffs, une approche novatrice pour le contrôle du taux de fausses découvertes. Dans le Chapitre 4, nous présentons la prédiction conforme, un cadre populaire pour la quantification de l'incertitude dans les problèmes de prédiction. En utilisant les valeurs de conformalité (p-values conformes), nous établissons un lien entre ce cadre et la littérature sur les tests statistiques. Nous abordons ensuite les bases de l'analyse des données en neuroimagerie, depuis l'acquisition et le prétraitement jusqu'à l'analyse statistique, dans le Chapitre 5. L'analyse des données d'IRMf constitue une motivation centrale de cette thèse, la plupart des expériences réalisées dans le cadre de nos contributions étant menées sur des ensembles de données d'IRMf.

Les contributions de cette thèse sont organisées autour de quatre articles, détaillés dans les sections suivantes :

- Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492
- Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*
- Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024a). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*
- Blain, A., Thirion, B., and Neuvial, P. (2024b). Tight and reliable conformal prediction. *Under review*

11.1 Notip : Contrôle non paramétrique de la proportion de découvertes réelles pour l'imagerie cérébrale

Les procédures d'inférence au niveau des clusters sont largement utilisées pour la cartographie cérébrale. Ces méthodes comparent la taille des clusters obtenus par seuillage des cartes cérébrales à une borne supérieure sous l'hypothèse nulle globale, cette borne étant calculée via la théorie des champs aléatoires (Random Field Theory) ou par permutation. Cependant, les garanties fournies par ce type d'inférence – à savoir qu'au moins un voxel est véritablement activé dans le cluster – ne sont pas informatives quant à l'étendue du signal présent. Il est donc nécessaire de disposer de méthodes permettant d'évaluer la quantité de signal au sein des clusters, tout en prenant en compte le fait que ces clusters sont définis à partir des données, ce qui introduit une circularité dans le raisonnement statistique. Cela a motivé l'utilisation d'estimateurs *post hoc* permettant une estimation statistiquement valide de la proportion de voxels activés dans les clusters. Dans le contexte des données d'IRMf, le cadre d'Inférence à Toutes les Résolutions (All-Resolutions Inference), introduit dans Rosenblatt et al., 2018, fournit de telles estimations. Cependant, cette méthode repose sur des familles de seuils paramétriques, ce qui conduit à une inférence conservatrice. Dans le Chapitre 6, nous proposons une approche adaptative aux caractéristiques des données afin d'obtenir un contrôle plus précis des fausses découvertes. Pour cela, nous exploitons des méthodes de randomisation. Nous obtenons ainsi *Notip* (**Non-parametric True Discovery Proportion control**), une méthode puissante et non paramétrique qui fournit des garanties statistiques sur la proportion de voxels activés dans des clusters définis à partir des données. Des expériences numériques montrent des gains substantiels en nombre de détections par rapport aux méthodes de pointe, sur 36 ensembles de données d'IRMf. Nous discutons également des conditions dans lesquelles la méthode proposée apporte un bénéfice.

Travail publié. Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492

11.2 Contrôle de la proportion de fausses découvertes pour les Knockoffs agrégés

La sélection de variables sous contrôle statistique est une étape analytique essentielle dans de nombreux domaines scientifiques, comme l'imagerie cérébrale ou la génomique. Dans ces

contextes de données en grande dimension, inclure trop de variables peut conduire à des modèles peu performants et coûteux, d'où la nécessité d'obtenir des garanties statistiques sur le taux de faux positifs. Les Knockoffs sont un outil statistique populaire pour la sélection de variables conditionnelle en grande dimension. Cependant, ils contrôlent la proportion attendue de fausses découvertes (FDR) mais non leur proportion effective (FDP). Dans le Chapitre 7, nous présentons une nouvelle méthode, KOPI, qui permet de contrôler la proportion de fausses découvertes dans les inférences basées sur les Knockoffs. Cette méthode repose également sur un nouveau type d'agrégation afin de limiter l'aléa associé aux Knockoffs classiques. Nous démontrons un contrôle efficace de la FDP et des gains substantiels en puissance par rapport aux méthodes existantes, aussi bien en simulations que sur des données réelles d'imagerie cérébrale et de génomique.

Travail publié. Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*

11.3 Quand les Knockoffs échouent : diagnostic et correction du non-échangeabilité des Knockoffs

Les Knockoffs constituent un cadre statistique populaire pour la sélection conditionnelle de variables en grande dimension sous contrôle statistique. Un tel contrôle est essentiel pour la fiabilité des inférences. Cependant, les garanties offertes par les Knockoffs reposent sur une hypothèse d'échangeabilité qui est difficile à tester en pratique, et peu d'études discutent des solutions à adopter lorsque cette hypothèse est violée. Nous introduisons un outil diagnostique basé sur les tests de deux échantillons par classifieur (Classifier Two-Sample Tests) au Chapitre 8. Nous montrons, sur des simulations et des données réelles, que cette violation survient fréquemment lorsque la structure de dépendance des données est forte. Nous proposons une alternative non paramétrique et efficace au niveau computationnel, qui permet de restaurer le contrôle des erreurs.

Prépublication. Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024a). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*

11.4 Prédiction conforme rigoureuse et fiable

La prédiction conforme avec séparation (Split Conformal Prediction, SCP) offre un moyen computationnellement efficace de construire des intervalles de confiance dans les problèmes de régression. Notamment, la plupart des travaux théoriques sur SCP se concentrent sur le problème d'un seul point de test. Or, dans les applications réelles, les ensembles d'inférence contiennent plusieurs points, ce qui soulève la question des garanties de couverture pour un ensemble de points simultanément. En moyenne, la proportion de non-couverture (False Coverage Proportion, FCP) reste contrôlée, mais elle peut varier fortement autour de sa moyenne. Nous montrons que lorsque l'on partitionne un ensemble de données plusieurs fois, la SCP classique peut ne pas contrôler la FCP dans jusqu'à 65% des partitions. Dans le Chapitre 9, nous introduisons CoJER, une nouvelle méthode qui assure un contrôle précis de la FCP en probabilité pour la prédiction conforme, en exploitant la connaissance de la distribution des p-valeurs conformes sous échangeabilité. Grâce à des expériences approfondies sur des données réelles, nous montrons que CoJER garantit la couverture

annoncée, contrairement à la SCP standard. De plus, CoJER produit des intervalles plus courts que la méthode *state-of-the-art* et seulement légèrement plus larges que la SCP classique.

En révision. Blain, A., Thirion, B., and Neuvial, P. (2024b). Tight and reliable conformal prediction. *Under review*

Bibliography

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

Adler, R. J. (2010). *The geometry of random fields*. SIAM.

Andreella, A., Hemerik, J., Weeda, W., Finos, L., and Goeman, J. (2020). Permutation-based true discovery proportions for fmri cluster analysis. *arXiv preprint arXiv:2012.00368*.

Andreella, A., Vesely, A., Weeda, W., and Goeman, J. (2024). Selective inference for fmri cluster-wise analysis, issues, and recommendations for critical vector selection: A comment on blain et al. *Imaging Neuroscience*.

Arlot, S., Blanchard, G., and Roquain, E. (2007). Some nonasymptotic results on resampling in high dimension, i: Confidence regions, ii: Multiple tests. *arXiv preprint arXiv:0712.0775*.

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

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). Predictive inference with the jackknife+.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845.

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

Belilovsky, E., Kastner, K., Varoquaux, G., and Blaschko, M. B. (2017). Learning to discover sparse graphical models. In *International Conference on Machine Learning*, pages 440–448. PMLR.

Bender, R. and Lange, S. (2001). Adjusting for multiple testing—when and how? *Journal of clinical epidemiology*, 54(4):343–349.

Benjamini, Y. (2020). Selective inference: The silent killer of replicability.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

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

Bennett, C. M., Wolford, G. L., and Miller, M. B. (2009). The principled control of false positives in neuroimaging. *Social cognitive and affective neuroscience*, 4(4):417–422.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *Annals of Statistics*.

Blain, A., Thirion, B., Grisel, O., and Neuvial, P. (2023). False discovery proportion control for aggregated knockoffs. *NeurIPS 2023*.

Blain, A., Thirion, B., Linhart, J., and Neuvial, P. (2024a). When knockoffs fail: diagnosing and fixing non-exchangeability of knockoffs. *arXiv preprint arXiv:2407.06892*.

Blain, A., Thirion, B., and Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage*, 260:119492.

Blain, A., Thirion, B., and Neuvial, P. (2024b). Tight and reliable conformal prediction. *Under review*.

Blanchard, G., Neuvial, P., and Roquain, E. (2021). On agnostic post hoc approaches to false positive control. In Cui, X., Dickhaus, T., Ding, Y., and Hsu, J. C., editors, *Handbook of Multiple Comparisons*, Chapman & Hall/CRC Handbooks of Modern Statistical Methods. Chapman and Hall/CRC, 1st edition edition.

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

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilita. *Pubblicazioni del R istituto superiore di scienze economiche e commericiali di firenze*, 8:3–62.

Bourgon, R., Gentleman, R., and Huber, W. (2010). Independent filtering increases detection power for high-throughput experiments. *Proceedings of the National Academy of Sciences*, 107(21):9546–9551.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Brzyski, D., Peterson, C. B., Sobczyk, P., Candès, E. J., Bogdan, M., and Sabatti, C. (2017). Controlling the rate of gwas false discoveries. *Genetics*, 205(1):61–75.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., and Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, 14(5):365–376.

Cabeza, R. and Nyberg, L. (2000). Imaging cognition ii: An empirical review of 275 pet and fmri studies. *Journal of Cognitive Neuroscience*, 12(1):1–47.

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

Cauchois, M., Gupta, S., Ali, A., and Duchi, J. C. (2024). Robust validation: Confident predictions even when distributions shift. *Journal of the American Statistical Association*, pages 1–66.

Chamma, A., Engemann, D. A., and Thirion, B. (2023). Statistically valid variable importance assessment through conditional permutations. *NeurIPS 2023*.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., Zhou, T., et al. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4.

Chen, Y., Wang, R., Wang, Y., and Zhu, W. (2024). Sub-uniformity of harmonic mean p-values. *arXiv preprint arXiv:2405.01368*.

Chevalier, J.-A., Nguyen, T.-B., Salmon, J., Varoquaux, G., and Thirion, B. (2021). Decoding with confidence: Statistical control on decoder maps. *NeuroImage*, 234:117921.

Consortium, . G. P. et al. (2010). A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27.

Efron, B. (2012). *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press.

Eklund, A., Nichols, T. E., and Knutsson, H. (2016). Cluster failure: Why fmri inferences for spatial extent have inflated false-positive rates. *Proceedings of the national academy of sciences*, 113(28):7900–7905.

Enjalbert-Courrech, N. and Neuvial, P. (2022). Powerful and interpretable control of false discoveries in two-group differential expression studies. *Bioinformatics*, 38(23):5214–5221.

Fan, J., Han, X., and Gu, W. (2012). Estimating false discovery proportion under arbitrary covariance dependence. *Journal of the American Statistical Association*, 107(499):1019–1035.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

Foygel Barber, R., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2021). The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455–482.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Friston, K. J., Frith, C., Liddle, P., and Frackowiak, R. (1991). Comparing functional (pet) images: the assessment of significant change. *Journal of Cerebral Blood Flow & Metabolism*, 11(4):690–699.

Friston, K. J., Holmes, A. P., Poline, J.-B., Price, C. J., and Frith, C. D. (1996). Detecting activations in pet and fmri: levels of inference and power. *Human brain mapping*, 4(3):170–181.

Fukunaga, K. (2013). *Introduction to statistical pattern recognition*. Elsevier.

Gammerman, A. and Vovk, V. (2007). Hedging predictions in machine learning. *The Computer Journal*, 50(2):151–163.

Gazin, U., Blanchard, G., and Roquain, E. (2024). Transductive conformal inference with adaptive scores. In *International Conference on Artificial Intelligence and Statistics*, pages 1504–1512. PMLR.

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

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics*.

Genovese, C. R., Lazar, N. A., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, 15(4):870–878.

Gibbs, I. and Candes, E. (2021). Adaptive conformal inference under distribution shift. *Advances in Neural Information Processing Systems*, 34:1660–1672.

Gimenez, J. R. and Zou, J. (2019). Improving the stability of the knockoff procedure: Multiple simultaneous knockoffs and entropy maximization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2184–2192. PMLR.

Goeman, J. J. and Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics*, 23(8):980–987.

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

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

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., and Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.

Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in neural information processing systems*, 35:507–520.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., and Drummond, G. B. (2015). The fickle p value generates irreproducible results. *Nature methods*, 12(3):179–185.

Hinton, G. E. (1990). Connectionist learning procedures. In *Machine learning*, pages 555–610. Elsevier.

Hirsch, J., Ruge, M. I., Kim, K. H. S., et al. (2000). An integrated functional magnetic resonance imaging procedure for preoperative mapping of cortical areas associated with tactile, motor, language, and visual functions. *Neurosurgery*, 47(3):711–721.

Hommel, G. (1986). Multiple test procedures for arbitrary dependence structures. *Metrika*, 33(1):321–336.

Howles, S. A., Wiberg, A., Goldsworthy, M., Bayliss, A. L., Gluck, A. K., Ng, M., Grout, E., Tanikawa, C., Kamatani, Y., Terao, C., et al. (2019). Genetic variants of calcium and vitamin d metabolism in kidney stone disease. *Nature Communications*, 10(1):5175.

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

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909.

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

König, G., Molnar, C., Bischl, B., and Grosse-Wentrup, M. (2021). Relative feature importance. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 9318–9325. IEEE.

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

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S., and Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature neuroscience*, 12(5):535.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Ledoit, O. and Wolf, M. (2003). Honey, i shrunk the sample covariance matrix. *UPF economics and business working paper*, (691).

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Lei, J. and Wasserman, L. (2014). Distribution-free prediction bands for non-parametric regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):71–96.

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

Liu, Y. and Zheng, C. (2018). Auto-encoding knockoff generator for fdr controlled variable selection. *arXiv preprint arXiv:1809.10765*.

Logothetis, N. K. (2008). What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878.

Lopez-Paz, D. and Oquab, M. (2016). Revisiting classifier two-sample tests. *arXiv preprint arXiv:1610.06545*.

Luo, Y., Fithian, W., and Lei, L. (2022). Improving knockoffs with conditional calibration.

Mandozzi, J. and Bühlmann, P. (2016). Hierarchical testing in the high-dimensional setting with correlated variables. *Journal of the American Statistical Association*, 111(513):331–343.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283.

Matthews, P. M., Honey, G. D., and Bullmore, E. T. (2006). Applications of fmri in translational medicine and clinical practice. *Nature Reviews Neuroscience*, 7(9):732–744.

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

Meinshausen, N., Meier, L., and Bühlmann, P. (2009). P-values for high-dimensional regression. *Journal of the American Statistical Association*, 104(488):1671–1681.

Mumford, J. A. and Nichols, T. E. (2009). Simple group fmri modeling and inference. *NeuroImage*, 47(4):1469–1475.

Neuvial, P. (2020). *Contributions to statistical inference from genomic data*. Habilitation thesis, Université Toulouse III Paul Sabatier.

Neyman, J. and Pearson, E. S. (1933). Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337.

Nguyen, B. T., Thirion, B., and Arlot, S. (2022). A Conditional Randomization Test for Sparse Logistic Regression in High-Dimension. In *NeurIPS 2022*, volume 35 of *Advances in Neural Information Processing Systems*, New Orleans, United States.

Nguyen, T.-B., Chevalier, J.-A., and Thirion, B. (2019). Ecko: Ensemble of clustered knockoffs for robust multivariate inference on fmri data. In *Information Processing in Medical Imaging: 26th International Conference, IPMI 2019, Hong Kong, China, June 2–7, 2019, Proceedings 26*, pages 454–466. Springer.

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

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical methods in medical research*, 12(5):419–446.

Nichols, T. E. (2012). Multiple testing corrections, nonparametric methods, and random field theory. *Neuroimage*, 62(2):811–815.

Nieuwenhuis, S., Forstmann, B. U., and Wagenmakers, E.-J. (2011). Erroneous analyses of interactions in neuroscience: a problem of significance. *Nature Neuroscience*, 14(9):1105–1107.

Noble, W. S. (2009). How does multiple testing correction work? *Nature biotechnology*, 27(12):1135–1137.

Papadopoulos, H., Proedrou, K., Vovk, V., and Gammerman, A. (2002). Inductive confidence machines for regression. In *Machine learning: ECML 2002: 13th European conference on machine learning Helsinki, Finland, August 19–23, 2002 proceedings 13*, pages 345–356. Springer.

Paus, T. (2010). Population neuroscience: Why and how. *Human Brain Mapping*, 31(6):891–903.

Pearson, K. (1900). X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Penny, W. D., Holmes, A. P., and Friston, K. J. (2003). Random effects analysis. In Frackowiak, R. S., Friston, K. J., Frith, C. D., Dolan, R. J., Price, C. J., Zeki, S., Ashburner, J., and Penny, W. D., editors, *Human Brain Function*, pages 843–850. Academic Press.

Perez-Lebel, A., Varoquaux, G., Koyejo, S., Doutreligne, M., and Morvan, M. L. (2024). Decision from suboptimal classifiers: Regret pre- and post-calibration. *Under Review*.

Platt, J. et al. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of Functional MRI Data Analysis*. Cambridge University Press.

Poline, J.-B. and Mazoyer, B. M. (1993). Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of Cerebral Blood Flow & Metabolism*, 13(3):425–437.

Rabiner, L. and Juang, B. (1986). An introduction to hidden markov models. *ieee assp magazine*, 3(1):4–16.

Ren, Z. and Barber, R. F. (2022). Derandomized knockoffs: leveraging e-values for false discovery rate control. *arXiv preprint arXiv:2205.15461*.

Ren, Z., Wei, Y., and Candès, E. (2021). Derandomizing knockoffs. *Journal of the American Statistical Association*, pages 1–11.

Roche, A., Mériaux, S., Keller, M., and Thirion, B. (2007). Mixed-effect statistics for group analysis in fMRI: a nonparametric maximum likelihood approach. *Neuroimage*, 38(3):501–510.

Romano, Y., Sesia, M., and Candès, E. (2020). Deep knockoffs. *Journal of the American Statistical Association*, 115(532):1861–1872.

Roquain, E. (2015). *Contributions to multiple testing theory for high-dimensional data*. PhD thesis, Université Pierre et Marie Curie.

Rosenblatt, J. D., Finos, L., Weeda, W. D., Solari, A., and Goeman, J. J. (2018). All-resolutions inference for brain imaging. *Neuroimage*, 181:786–796.

Sadinle, M., Lei, J., and Wasserman, L. (2019). Least ambiguous set-valued classifiers with bounded error levels. *Journal of the American Statistical Association*, 114(525):223–234.

Saunders, C., Gammerman, A., and Vovk, V. (1999). Transduction with confidence and credibility. *Sixteenth International Joint Conference on Artificial Intelligence*.

Sesia, M., Sabatti, C., and Candès, E. J. (2019). Gene hunting with hidden markov model knockoffs. *Biometrika*, 106(1):1–18.

Shafer, G. and Vovk, V. (2008). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Simes, R. J. (1986). An improved bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Smith, R. C. (2013). *Uncertainty quantification: theory, implementation, and applications*. SIAM.

Smith, S. M., Jenkinson, M., Woolrich, M. W., et al. (2009). Advances in functional and structural mr image analysis and implementation as fsl. *NeuroImage*, 23:S208–S219.

Smith, S. M. and Nichols, T. E. (2009). Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, 44(1):83–98.

Stein, C., Efron, B., and Morris, C. (1972). *Improving the usual estimator of a normal covariance matrix*. Stanford University. Department of Statistics.

Stiglic, G. and Kokol, P. (2010). Stability of ranked gene lists in large microarray analysis studies. *Journal of biomedicine and biotechnology*, 2010.

Sullivan, G. M. and Feinn, R. (2012). Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Thirion, B., Varoquaux, G., Dohmatob, E., and Poline, J.-B. (2014). Which fMRI clustering gives good brain parcellations? *Frontiers in Neuroscience*, 8(167):13.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. (2019). Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1):1–21. Number: 1 Publisher: Nature Publishing Group.

Van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*.

Van Essen, D. C., Ugurbil, K., Auerbach, E. J., Barch, D. M., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D. A., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L. J., Marcus, D. S., Michalareas, G., Moeller, S., Oostenveld, R., Petersen, S. E., Prior, F. W., Schlaggar, B. L., Smith, S. M., Snyder, A. Z., Xu, J., and Yacoub, E. (2012). The Human Connectome Project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231.

Vanschoren, J., Van Rijn, J. N., Bischl, B., and Torgo, L. (2014). Openml: networked science in machine learning. *ACM SIGKDD Explorations Newsletter*, 15(2):49–60.

Varoquaux, G., Schwartz, Y., Poldrack, R. A., Gauthier, B., Bzdok, D., Poline, J.-B., and Thirion, B. (2018). Atlases of cognition with large-scale human brain mapping. *PLoS computational biology*, 14(11):e1006565.

Vovk, V. (2023). Testing exchangeability in the batch mode with e-values and markov alternatives.

Vovk, V. and Bendtsen, C. (2018). Conformal predictive decision making. In *Conformal and Probabilistic Prediction and Applications*, pages 52–62. PMLR.

Vovk, V., Gammerman, A., and Saunders, C. (1999). Machine-learning applications of algorithmic randomness. *Sixteenth International Conference on Machine Learning (ICML-1999)*.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic learning in a random world*, volume 29. Springer.

Vovk, V. and Wang, R. (2020). Combining p-values via averaging. *Biometrika*, 107(4):791–808.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wang, R. and Ramdas, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852.

Wasserstein, R. L. and Lazar, N. A. (2016). The asa statement on p-values: context, process, and purpose.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59.

Weinstein, A., Su, W. J., Bogdan, M., Barber, R. F., and Candès, E. J. (2020). A power analysis for knockoffs with the lasso coefficient-difference statistic. *arXiv preprint arXiv:2007.15346*.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing: Examples and methods for p-value adjustment*, volume 279. John Wiley & Sons.

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences*, 116(4):1195–1200.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., and Nichols, T. E. (2014). Permutation inference for the general linear model. *Neuroimage*, 92:381–397.

Woo, C.-W., Krishnan, A., and Wager, T. D. (2014). Cluster-extent based thresholding in fmri analyses: pitfalls and recommendations. *NeuroImage*, 91:412–419. 24412399[pmid].

Worsley, K. (2003). Developments in random field theory. *Human brain function*, 2:881–886.

Worsley, K. J. (1994). Local maxima and the expected euler characteristic of excursion sets of χ 2, f and t fields. *Advances in Applied Probability*, 26(1):13–42.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for cbf activation studies in human brain. *Journal of Cerebral Blood Flow & Metabolism*, 12(6):900–918.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fmri time-series revisited—again. *NeuroImage*, 2(3):173–181.

Zaffran, M. (2024). *Post-hoc predictive uncertainty quantification: methods with applications to electricity price forecasting*. PhD thesis, Ecole Polytechnique.

Zhang, C.-H. and Zhang, S. S. (2014). 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.

Zhu, Z., Fan, Y., Kong, Y., Lv, J., and Sun, F. (2021). Deeplink: Deep learning inference using knockoffs with applications to genomics. *Proceedings of the National Academy of Sciences*, 118(36):e2104683118.