DISSERTATION

STATISTICAL INFERENCE ON REPRODUCIBILITY IN HIGH-THROUGHPUT EXPERIMENTS

Submitted by

Austin Ellingworth

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2025

Doctoral Committee:

Advisor: Yawen Guan
Co-Advisor: Wen Zhou

Kayleigh Keller
Piotr Kokoszka
Donald Mykles

ABSTRACT

STATISTICAL INFERENCE ON REPRODUCIBILITY IN HIGH-THROUGHPUT EXPERIMENTS

Results in high-throughput genomics are known to have large variability across independent replicate studies. For this reason, the formal assessment of the agreement of results for many hypotheses across replicate studies has been a burgeoning area of research in statistical genomics. Hypotheses with consistent results are called reproducible, while those without consistency are called irreproducible. The presence of reproducibility in experimental research is critical, as it ensures the validity of findings. In this dissertation, we devise three methods for assessing the reproducibility of results from high-throughput genomic studies, each with advantages under certain settings.

First, we notice that many of the existing approaches to assessing the reproducibility of results from two replicate high-throughput genomics studies either depend on strict parametric assumptions on available summary statistics or fail to properly consider the consistency of reproducible signal across experiments in addition to its strength. Motivated by Philtron et al. (2018), we introduce a function based on the rankings of summary statistics from each experiment to define a notion for reproducibility and identify reproducible hypotheses. The proposed nonparametric statistic takes into account both the signal strength and consistency of results. By examining the geometry of the space of ranks of summary statistics and utilizing the negative association dependence structure of ranks, a novel procedure is introduced for recognizing reproducible findings while controlling the false discovery rate (FDR). This method controls FDR under relatively mild assumptions. The theoretical FDR findings are validated through simulations that also reveal the method to be more powerful than existing procedures. Finally, the procedure is applied to two large-scale TWAS datasets, uncovering reproducible features.

Second, we notice that existing methods for assessing the reproducibility of high-throughput studies ignore the known group structures of genetic features, such as transcripts belonging to the same gene or genes belonging to the same pathway. Motivated by Li et al. (2011) and Liu et al. (2016), we present an empirical Bayesian framework for reproducibility that incorporates this group structure. Additionally, we introduce algorithms for testing reproducibility at the hypothesis and group levels that maintain control of posterior FDR. Next, a data-driven estimation procedure based on the EM algorithm is proposed to enable the application of these algorithms when the parameters it relies on are unknown. In simulation, we show that the inclusion of the group structure in the hypothesis-level procedure leads to superior performance in terms of power and FDR control compared to more naive methods, and that the group-level procedure outperforms methods that rely on aggregation prior to analysis. The proposed procedures enable researchers to integrate known group structure information into the reproducibility problem, yielding higher-quality results.

Finally, while there is a dearth of existing literature for analyzing reproducibility across two replicate studies, there are strikingly few methods that consider cases with more than two studies, and those that exist generally assume the distributions of irreproducible summary statistics are known. Leveraging Kendall's coefficient of concordance, we introduce a rank-based statistic that quantifies the agreement of results for a particular hypothesis without enforcing such strict assumptions. Noticing that in real high-throughput genomic settings, we have many "housekeeping" genes that are unrelated to the disease of interest and thus can be considered as a control set, we utilize conformal inferential and bootstrapping techniques to devise three procedures for calculating approximate p-values from a set of the proposed statistics that can be used to discover reproducible hypotheses at a nominal level of FDR. Simulation studies reveal that the three methods show preferable performance to existing methods in terms of power and FDR control. Applying the methods to single-cell expression data from five COVID-19 studies, we show that the proposed statistic and its procedures can identify genes and gene pathways associated with COVID-19.

ACKNOWLEDGEMENTS

To my co-advisor: Dr. Wen Zhou, thank you for your invaluable guidance, unwavering support, and steady mentorship. I hope to approach each problem in my career with the rigor and intensity of thought that you bring. Your commitment to my development as a researcher and statistician is greatly appreciated.

To my advisor, Dr. Yawen Guan, thank you for stepping up and agreeing to serve as my co-advisor during my final year. I would not be able to complete my degree without your assistance.

To my committee members: Dr. Kayleigh Keller, Dr. Piotr Kokoszka, and Dr. Donald Mykles, thank you for your invaluable feedback. Our discussions and your thought-provoking questions have improved my dissertation greatly.

To my collaborators: Dr. Debashis Ghosh, Dr. Zhigen Zhao, Dr. Yingxin Lin, Dr. Y. X. Rachel Wang, Dr. Xing Tong, and Dr. Hongyu Zhao, thank you for your contributions. The development of the work within this dissertation would not be possible without you.

To the Department of Statistics administrative staff: Alexandria Peitsmeyer, Martin Sweeney, and Karena Alons, thank you for your assistance. I am certain this dissertation would not have been properly submitted without your help. To the Department of Statistics faculty, thank you. I learned much of the content in this dissertation from your instruction and guidance.

To educators: Judy Reinbold, Al Cassidy, Susan Schaub, Kari Koester, Dr. Silas Bergen, Dr. Tisha Hooks, Dr. Brant Deppa, Dr. Chris Malone, Dr. Eric Errthum, and Dr. Felino Pascual, thank you for instilling and encouraging my intellectual curiosity. I continued to pursue schooling because of educators who made learning challenging and rewarding.

To my classmates, colleagues, and friends, both current and former, thank you. A regrettably incomplete list is provided here: Lane Drew, Dr. Seongwon Im, Dr. Elizabeth Lawler, Dr. Xiangdong Meng, Dr. Maddie Rainey, Dr. Gray Stanton, Dr. Justin Van Ee, Dr. Zifeng Zhang, Dr. Connor Gibbs, Julia Campbell, Dr. Mantautas Rimkus, Dr. Nathan Ryder, Dr. Bingying Dai,

Simon Weller, Dr. Troy Wixson, Hannah Butler, Karissa Palmer, Patrick Busch, Dugan Bradley, John Fink, Lucas Haefner, Thomas McDermott, Jacob Peller, Shane Price, and so on.

Lastly, to my parents: B. Jon and Michelle Ellingworth, and siblings: Evan and Emily, Grace, and Ben Ellingworth, thank you. Your unending love is the reason I am who I am. Without your (sometimes irrational) belief in me, I would not be where I am. There are no words to express my appreciation for you all.

DEDICATION

This dissertation is dedicated to B. Jon and Michelle Ellingworth. You believed in me when I didn't believe in myself.

TABLE OF CONTENTS

Chapter 1

Introduction

1.1 Overview

This dissertation examines methods for performing statistical inference on the reproducibility of hypotheses across studies with a particular interest in applications to high-throughput genomic data. Reproducibility analysis, or alternatively replicability analysis, assesses the agreement or consistency of results for a particular hypothesis across multiple replicate studies. A hypothesis is reproducible if its results show agreement across studies and irreproducible if they do not. Over the last 20 years, reproducibility has become a major research area in many fields (MAQC-Consortium, 2006; Aarts et al., 2015; Errington et al., 2021).

High-throughput genomics is one of the fields where reproducibility analysis is increasingly popular (Li et al., 2011; Heller and Yekutieli, 2014; Philtron et al., 2018; Wang et al., 2022; Lyu et al., 2023; Li et al., 2024). High-throughput genomic results are known to be highly variable across studies that should be replicates of each other. Therefore, genetic features with results that are consistent across multiple studies are prioritized, and those with inconsistent results are discarded. The framework considers n hypotheses that are commonly examined in m studies. Methods are designed to discover sets of hypotheses that are reproducible at a specified nominal level of false discovery rate (FDR), defined as the expected proportion of hypotheses in a set that are truly null (Benjamini and Hochberg, 1995) (in our case, irreproducible). What it means for a hypothesis to be considered reproducible or irreproducible is not well established, however. Some approaches rely on parametric assumptions on the observed summary statistics, while others examine properties of the experiment-wise alignment of hypotheses.

Each chapter in this dissertation describes a notion of reproducibility and proposes a procedure to assess that notion. Chapters 2 and 3 consider the setting where the number of replicate studies is limited to two ($m = 2$). Motivated by Philtron et al. (2018), we examine the reproducibility from

a nonparametric lens in Chapter 2 and provide an intuitive definition for irreproducibility using the distribution of ranks of summary statistics within each experiment. We then introduce a rank-based statistic to measure reproducibility that balances the strength of reproducible signal relative to irreproducible signal and the consistency of reproducible signal through an adaptive tuning parameter λ. Using the statistic, we propose a novel testing procedure that learns the distribution of irreproducible rankings from partitions of the observed data. Unlike existing nonparametric methods, we show theoretical control of FDR under a relatively mild setting. In simulations, the procedure shows two distinct advantages compared to existing methods: 1) better estimation of FDR, resulting in more reliable detection of reproducible hypotheses, and 2) more flexibility in detection for different levels of reproducible signal consistency through the data-adaptive λ.

An interesting feature of the high-throughput genomic setting is the group structure that exists. In RNA sequencing data, quantification of abundances occurs at the transcript-level and aggregated to obtain gene-level expressions. Inference on changes in expression or associations with traits occurs at the gene-level. Finally, post hoc identification of pathways, biological processes, or other groups of genes that were most important is typically performed. Existing reproducibility methods ignore this group structure. We aim to integrate group information into the analysis of reproducibility in Chapter 5, where we view the two study reproducibility problem from the empirical Bayesian framework introduced in Liu et al. (2020). This framework considers reproducibility to be the composite of group and hypothesis-within-group level reproducibility terms. Summary statistics from reproducible and irreproducible hypotheses are assumed to come from a Gaussian copula mixture model where underlying signal strength and consistency are inherited from a bivariate normal distribution, as previously used in Li et al. (2011). Using the group structure, we develop testing procedures for group and hypothesis-level reproducibility that control posterior false discovery rates. These methods are more powerful than approaches that are naive to the group structure.

While there are several rank-based procedures for examining the reproducibility of $m = 2$ studies, the nonparametric approach is underexplored when $m > 2$. There are challenges in defining

a sensible notion and extending forms of rank-based statistics for more than two studies. So, when $m > 2$, r out of m reproducibility is typically assessed. That is, existence of signal in at least r out of m studies is required for a hypothesis to be reproducible. Methods that examine this notion tend to examine the order statistics of the available p-values, assumed to be uniform (Wang et al., 2022; Lyu et al., 2023). The uniformity assumption is problematic when unmeasured biases are present across studies from design, platform, etc. Additionally, results depend heavily on the r selected by a practitioner. Consequently, in Chapter 4 we offer a notion of reproducibility that does not depend on the marginal distribution of observed summary statistics and reduces to similar nonparametric definitions when $m = 2$. Noticing that Kendall's W, a statistic measuring the agreement of ratings across judges, has a neat interpretation for reproducibility, we introduce a statistic that separates reproducible and irreproducible hypotheses using W. Because its distribution is difficult to derive in a general setting, we develop three procedures for calculating approximate p-values from the statistic that take advantage of a control set of known irreproducible hypotheses. In simulations, we explore the properties of each method. Finally, we apply one of the procedures to five real datasets examining COVID-19 patients and discover more genes involved in processes important to the regulation of the viral process than existing methods.

1.2 Outline

Each chapter takes the form of a coauthored academic manuscript. The manuscripts are included in full and can be read independently. Some information regarding the general reproducibility problem may appear in multiple chapters, but will be included for the sake of completeness. The chapters of the dissertation are organized as follows.

Chapter 2 is titled "Reproducible or not: a nonparametric procedure to assess reproducibility across high-throughput studies" and is based on work coauthored with Debashis Ghosh, Zhigen Zhao, and Wen Zhou.

Chapter 3 is titled "Assessing reproducibility of high-throughput studies with group structure" and is based on work coauthored with Zhigen Zhao and Wen Zhou.

Chapter 4 is titled "Assessing the reproducibility of results across multiple high-throughput studies using Kendall's W" and is based on work coauthored with Yingxin Lin, Y. X. Rachel Wang, Xin Tong, Wen Zhou, and Hongyu Zhao.

A summary of the dissertation's contributions and discussion of the avenues for continued research are provided in Chapter 5. The supplemental materials for Chapter 2, 3, and 4 can be found in the appendices.

Chapter 2

Reproducible or not: a nonparametric procedure to assess reproducibility across high-throughput studies

2.1 Introduction

2.1.1 Reproducibility and replicability

The ability to reproduce findings in independent studies is critical to contemporary science. Often, intriguing results become the subject of subsequent follow-up studies. Consequently, the quantification of the consistency of results across different studies has emerged as a significant area of research in many fields, including psychology (Aarts et al., 2015), cancer biology (Errington et al., 2021), genomics (MAQC-Consortium, 2006), among others. Outcomes that demonstrate consistency across experiments are considered reproducible or replicable, while those that do not are regarded as irreproducible or non-replicable. The reproducibility of results provides credibility to the scientific discoveries made within an experiment. Conversely, a lack of reproducibility can have detrimental effects on scientific progress.

Reproducibility has become a central concern in high-throughput studies, where thousands of hypotheses are assessed within each experiment, and the degree of consistency among detected signals across experiments serves as an indicator of reproducibility. The evaluation of reproducibility in high-throughput settings has attracted considerable attention across fields, including genomics, wherein researchers examine the genetic expression of thousands of genes or single nucleotide polymorphisms (SNPs). Reproducibility is particularly indispensable in genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS). When an association between a gene or SNP and a particular trait or disease is discovered in a study, it is posited as a

potential candidate for a causal relationship. Validation of such an association across independent studies bolsters the credibility of its causal candidacy.

Statistically, reproducibility for multiple high-throughput studies has not been defined or measured in a unified way. For instance, Hung and Fithian (2020) propose three distinct types of reproducibility: (1) one that investigates the consistency of the effect's sign for each hypothesis, (2) another that gauges reproducibility based on the disparity between effect sizes across two experiments for each hypothesis, and (3) one that examines the proportion of hypotheses exhibiting a diminished effect in a replication study compared to the original. The authors develop intuitive methods to test each of these definitions. However, these approaches require a serial structure for the studies, with one being the original study and another a replication study, designed for only a subset of hypotheses from the original study. Moreover, assessing discrepancies in effect sizes relies on the assumption that effect sizes closely follow a truncated normal distribution, thereby constraining the method's flexibility. The partial conjunction (PC) test (Friston et al., 2005) has often been employed to evaluate reproducibility (Benjamini and Heller, 2008) of n hypotheses across m studies. The PC null posits that a result is non-null in fewer than r out of m studies. Consequently, testing against it can be interpreted as determining whether a result has been replicated in at least r out of m studies. The AdaFilter procedure (Wang et al., 2022) refines the PC test for high-dimensional settings by adaptively filtering the number of hypotheses under consideration, eliminating those least likely to be reproducible, and then applying the PC framework to the remaining hypotheses. The authors introduced the procedure to assess reproducibility for numerous features that control false discovery rate (FDR). However, simulations reveal that the AdaFilter procedure for FDR tends to be overly conservative, resulting in reduced power for detecting weakly replicated signals. Alternatively, Li et al. (2011) consider n common hypotheses across two studies and examines "*the extent to which the ranks of the signals are no longer consistent across replicates*." Within this framework, they develop a method to categorize hypotheses as reproducible or irreproducible at a nominal level of irreproducible discovery rate (IDR) – analogous to FDR – by jointly modeling rankings of summary statistics across studies using a Gaussian copula mixture

model. Motivated by Li et al. (2011), Philtron et al. (2018) assess the reproducibility of individual discovery in multiple testing using the ranks of summary statistics. In contrast to the parametric assumption, they define a hypothesis as irreproducible if the rank of the corresponding summary statistic in one experiment is independent of that in the other experiment. With this definition, they use the maximum rank across studies to determine whether a discovery is reproducible and provide a decision rule at a nominal level of marginal false discovery rate (mFDR). Since the method relies solely on the rank of summary statistics and makes no parametric assumptions regarding the statistics, it can be applied in a wide variety of settings. While Philtron et al. (2018) has demonstrated satisfactory performance numerically, its theoretical properties remain largely unexplored.

2.1.2 Our contributions

Inspired by the suggestion that reproducibility can be defined by the consistency – or lack thereof – of the ranking of summary statistics, we unify the definition of reproducibility in a high-dimensional setting in Section 2.2.1. Then, we introduce the rank-based statistic, $M_{\lambda,i}$, in (2.3) that assesses reproducibility. The proposed statistic yields advantages over the existing literature in a few manners. First, in assessing reproducibility using ranks, no distributional assumptions are made about summary statistics themselves. This allows the $M_{\lambda,i}$ statistic to be applicable in a wide array of different settings. In fact, as long as the ranking structure is maintained, we can assess the reproducibility of results from two experiments that report different types of summary statistics. Ranks are also beneficial because they are robust against a monotone transformation of summary statistics, which is particularly useful when two studies have different sample sizes or when there exists some uniformly applied study effect. Consider, for example, two studies that perform t-tests for the same n hypotheses. In the first study, there is a sample size of 50 for each t-test compared to a sample size of $1,000$ in the second study. The values of the t-statistics for true signal are expected to be larger in the second study compared to the first study because of the greater sample size. Due to this, a method that only examines the p-values or t-statistics across experiments will fail to find many replicated results. In examining ranks, as long as the alignment of the t-statistics pertaining

to true signal remains relatively consistent within the experiment, the $M_{\lambda,i}$ statistic can mitigate the effect of differences in sample size. Additionally, with the inclusion of a λ in the statistic, we introduce flexibility for the user to adjust the importance of signal strength and signal consistency across experiment in reproducibility. This flexibility is not present in other rank-based procedures, resulting in the $M_{\lambda,i}$ statistic separating reproducible and irreproducible hypotheses more reliably in certain cases.

Based on the $M_{\lambda,i}$ statistic, we design a novel procedure for estimating the number of false discoveries in set and decide a threshold for discovering reproducible hypotheses at any nominal level of FDR. The proposed procedure leverages partitions of the geometric rank-space to estimate the distribution of $M_{\lambda,i}$ statistics, as seen in Section 2.2, and with it, the number of false discoveries in any rejection set. In simulations presented in Section 2.6, we see that this proposed procedure yields far more exact FDR control than existing procedures. It also relies on less stringent assumptions, so there are cases in which existing methods fail to control FDR at a desired nominal level, whereas the proposed procedure does control FDR.

Another advantage of using rankings is that sequences of rank statistics are known to be *negatively associated* (Joag-Dev and Proschan, 1983). Negatively associated sequences have several nice theoretical properties. Leveraging those properties allows us to prove that the proposed procedure based on the $M_{\lambda,i}$ statistic asymptotically controls the false discovery rate at any specified level under a relatively modest assumption, Condition 2.3.1. Showing theoretical asymptotic FDR for the proposed procedure is a meaningful contribution, as most other methods with theoretical false discovery control either make stringent parametric assumptions about the observed summary statistics (Hung and Fithian, 2020) or limit the type of summary statistic observed (Wang et al., 2022) and the methods which examine a similar rank-based notion of reproducibility do not show theoretical false discovery control, despite their simulation-based performance.

2.1.3 Related literature

Reproducibility shares a similar framework with meta-analysis (Borenstein et al., 2021). Both types of analysis integrate data or results for the same statistical hypotheses across multiple sources. The goals of the two analyses are quite different, however. In the reproducibility literature, methods are designed specifically to assess the agreement of results across replicate studies (Li et al., 2011). By contrast, in meta-analysis, we are seeking to combine evidence for a hypothesis across multiple studies to make discoveries. While lack of reproducibility will result in higher variation for effect sizes in a meta-analysis, agreement of results is not the primary focus. Here, we focus only on the question of reproducibility and not on combining signal for increased power.

In a high-throughput genomic study setting, the reproducibility problem is closely related to that of simultaneous signal detection (Zhao et al., 2017), also known as colocalization analysis (Nica et al., 2010; Giambartolomei et al., 2014; Hormozdiari et al., 2016) or integrative genomics (Zhao et al., 2014). In the simultaneous signal problem, the primary interest is to identify hypotheses that pertain to true signal across different types of studies, while reproducibility analysis deals with (biological or technical) replicate studies. For example, a common application in the simultaneous signal problem is to integrate results from GWAS analysis and expression quantitative trait loci (eQTL) analysis together to identify genetic features that show associations between both a disease/trait of interest and gene-level expression. Notice, in this problem, the goal is to discover genetic features that show signal in two different study types and thus the hypotheses (SNPs in this case) are related but not identical across the two studies, as is the focus in reproducibility analysis.

In the broader scientific literature, there has been little agreement about the language used for what we define as "reproducibility" in this chapter. We define the term "reproducibility" to mean "the level of agreement between results from replicate experiments across (biological or technical) replicate samples, test sites or experimental or data analytical platforms" as previously defined in Li et al. (2011). A similar concept has sometimes also been defined as "replicability" (Benjamini and Heller, 2008; Heller and Yekutieli, 2014; Hung and Fithian, 2020; Wang et al., 2022). Elsewhere in scientific literature, "reproducibility" is sometimes used concerning the ability

of another researcher to obtain the same computational results as a study starting with the same raw data (Stodden et al., 2013). Here we do not refer to this broader definition of reproducibility.

2.1.4 Organization and notation

The rest of this chapter is organized in the following manner. Section 2.2 introduces our concept of reproducibility, the statistic $M_{\lambda,i}$, and the process for evaluating reproducibility in two experiments. Section 2.4 provides the theoretical FDR guarantees. Section 2.5 covers an estimation procedure for the sparsity parameter π_1 and examines selection criteria for tuning parameters β and λ. Section 2.6 details three simulation studies. Section 2.7 applies the $M_{\lambda,i}$ procedure to two TWAS datasets. Finally, Section 2.8 provides a discussion of the method.

Notations: For clarity, we list some notations used throughout the chapter. The i^{th} hypothesis is denoted $\mathbb{H}_i$ and has an associated summary statistic for the j^{th} study $T_{j,i}$. $R_{j,i}$ is the ranking of $T_{j,i}$ within experiment j. That is, if a *larger* summary statistic denotes higher significance, then if $T_{j,i}$ has the largest magnitude among all hypotheses in experiment j, $R_{j,i} = 1$ and if $T_{j,i'}$ has the smallest magnitude in its experiment, $R_{j,i'} = n$. $R_{j,i}$ is similarly defined if a *smaller* summary denotes higher significance. Throughout the chapter, we assume all $T_{j,i}$ are continuous random variables, and thus there are no ties in magnitude among summary statistics. If ties are present, we recommend random assignment of ranks among hypotheses with the same magnitude. Denote the proportion of the n hypotheses that are reproducible by π_1 and define the set of reproducible indices by $\mathcal{H}_1$ and irreproducible indices $\mathcal{H}_0$. That is,

$$\mathcal{H}_1 = \{g : \mathbb{H}_g \text{ is reproducible}\} \text{ and } \mathcal{H}_0 = \{h : \mathbb{H}_h \text{ is irreproducible}\}. \tag{2.1}$$

The nominal level for FDR control is always denoted by α. Throughout, we call hypotheses with small ranks (i.e. $R_{1,i} = 1$) *highly* ranked.

2.2 Methodology

2.2.1 Notion of reproducibility

As discussed in Section 2.1, there is no unified statistical definition of reproducibility. In this section, we discuss the notion of reproducibility that the proposed method will assess. Instead of examining what is required for a hypothesis to be reproducible, we evaluate what it means for a hypothesis to be irreproducible. The reproducibility of a result is typically characterized by the level of consistency (Jaljuli et al., 2022; Philtron et al., 2018; Wang et al., 2022) or agreement (Li et al., 2011; MAQC-Consortium, 2006) that a result shows across different studies. Implicit in this statement is the idea that if a result does not show consistency, then it is irreproducible. That is, if $\mathbb{H}_i$ is irreproducible, then results across multiple studies are unrelated to each other. Motivated by that observation, we describe the irreproducibility of a hypothesis by the independence of the rankings of its summary statistics across experiments. More precisely, we denote the set of irreproducible hypotheses as $\mathcal{H}_0$ and define irreproducibility by Definition 2.2.1.

Definition 2.2.1. *If $h \in \mathcal{H}_0$, $R_{1,h}$ and $R_{2,h}$ are independent of each other. Additionally, if $h' \in \mathcal{H}_0$, $R_{1,h}$ and $R_{1,h'}$ are identically distributed and $R_{2,h}$ and $R_{2,h'}$ are identically distributed.*

This notion was enforced through parametric assumptions in Li et al. (2011) and explicitly in Philtron et al. (2018). Essentially, Definition 2.2.1 implies that knowledge of the ranking of a summary statistic in one study provides no information about the ranking of the associated summary statistic in another study. This definition is useful for the following reasons. As in Philtron et al. (2018), defining reproducibility on rankings of summary statistics does not limit the type of summary statistic, nor does it enforce strict parametric assumptions on those summary statistics and thus is applicable in settings in which two studies report any type of summary statistic for each hypothesis examined, while many existing methods examine only p-values (Wang et al., 2022) or require the normality of test statistics (Hung and Fithian, 2020). Additionally, as previously discussed in Section 2.1, the rank structure is robust against any monotone transformation of the summary statistics. This mitigates the effect of differing sample sizes and uniformly applied study

effects. Finally, under a mild assumption, this definition can be leveraged to prove the proposed estimates of the distributions of irreproducible rank statistics in each study and the reproducibility statistic, $M_{\lambda,i}$ asymptotically converge to their true respective distributions, as seen in Section 2.4.

2.2.2 Statistic

The notion of reproducibility considered in Section 2.2.1 is evaluated by the dependence of rank statistics $R_{1,i}$ and $R_{2,i}$. For rank statistics to be dependent on each other, a hypothesis must either have a strong true effect relative to the other hypotheses, which we call signal strength, or it must show a more consistent effect across experiment than other hypotheses yielding summary statistics that are similarly positioned, which we call signal consistency. Thus, this notion of reproducibility can be thought of as a combination – or even competition – between these two properties. Consider a high-dimensional reproducibility problem with n common hypotheses across two studies. For hypothesis $\mathbb{H}_i$, we observe a summary statistics (p-value, t-statistic, log-fold change, etc.) for each study $(T_{1,i}, T_{2,i})$ that can be ranked within experiment from most notable (1) to least (n) yielding ranks $(R_{1,i}, R_{2,i})$. Notice, our notion of reproducibility would imply for any $h \in \mathcal{H}_0$, $R_{1,h}$ is independent of $R_{2,h}$. As discussed in Section 2.1, Philtron et al. (2018) considers the maximum rank reproducibility (MaRR) test statistic to be the maximum rank of summary statistic pertaining to a hypothesis across experiments. When there are two studies the procedure is as follows. For each study, summary statistics for every hypothesis are ranked from most to least significant. $\mathbb{H}_i$, is deemed reproducible if $\mathrm{MaRR}_i = \max(R_{1,i}, R_{2,i}) \leq t$ for some critical value t where $R_{1,i}$ and $R_{2,i}$ are the rankings of summary statistics $T_{1,i}$ and $T_{2,i}$ within their respective studies. Heuristically, if a hypothesis is reproducible, it has some level of signal strength and is expected to have highly ranked test statistics in both experiments, yielding a small maximum rank. In a way, the MaRR_i statistic measures the combination of signal strength and consistency because it admits the critical representation

$$\mathrm{MaRR}_i = \max(R_{1,i}, R_{2,i}) = 2^{-1}\Big[\,(R_{1,i} + R_{2,i}) + |R_{1,i} - R_{2,i}|\,\Big]. \tag{2.2}$$

This decomposition has two distinct components that aid in yielding a small MaRR_i statistic in different ways. First, the sum of the ranks $(R_{1,i} + R_{2,i})$ is small if the signal strength of $\mathbb{H}_i$ is strong in both experiments. As a strong signal strength implies the hypothesis is highly ranked in each experiment, resulting in a small sum of ranks. Second, the absolute deviation between $R_{1,i}$ and $R_{2,i}$ $(|R_{1,i} - R_{2,i}|)$ is small when the signal is highly consistent across experiment. As consistency results summary statistics that are similarly ranked across experiment, yielding a small difference in ranks. In practice, we observe that MaRR performs well in identifying reproducible hypotheses with strong signal; however, it cannot detect reproducible hypotheses with weaker signal that are highly consistent across experiment. The equal weighting MaRR imposes on the signal strength and consistency pieces of the decomposition fails to consider differing levels of signal strength and consistency that exist within studies. Recognizing the competition between signal strength and consistency, we propose assessing the reproducibility of $\mathbb{H}_i$ using the test statistic

$$M_{\lambda,i} = R_{1,i} + R_{2,i} + \lambda |R_{1,i} - R_{2,i}| \tag{2.3}$$

where λ is a data-adaptive weighting parameter. Changing values for λ balances the relationship between signal strength and consistency. For example, Figure 2.1 examines the decision boundaries for reproducibility of the original MaRR procedure from Philtron et al. (2018) and the proposed procedure with $\lambda = 5$ on the rank-space $(R_{1,g}, R_{2,g})$. The sum of the ranks portion of $M_{\lambda,i}$ dictates how close the hypothesis is to the origin and the difference in ranks dictates how close a point is to the dashed diagonal. Notice, when $\lambda > 1$, $M_{\lambda,i}$ captures more hypotheses that fall near the diagonal, while sacrificing area near the axes when compared to the baseline MaRR method. We argue this is beneficial because we expect hypotheses with consistent signal that is weaker in strength to fall in the region near the diagonal. By increasing $\lambda > 1$, $M_{\lambda,i}$ can differentiate those reproducible hypotheses from irreproducible hypotheses. Notice, if signal strength has a more important role in the set of reproducible hypotheses than consistency, one can set $\lambda = 1$ and $M_{\lambda,i}$ reduces to the MaRR statistic in the following manner, $M_{\lambda,i} = R_{1,i} + R_1 + 1|R_{1,i} - R_{2,i}| = 2\mathrm{MaRR}_i$. While nothing about the proposed procedure demands $\lambda \geq 1$, in selecting $\lambda < 1$, one supposes

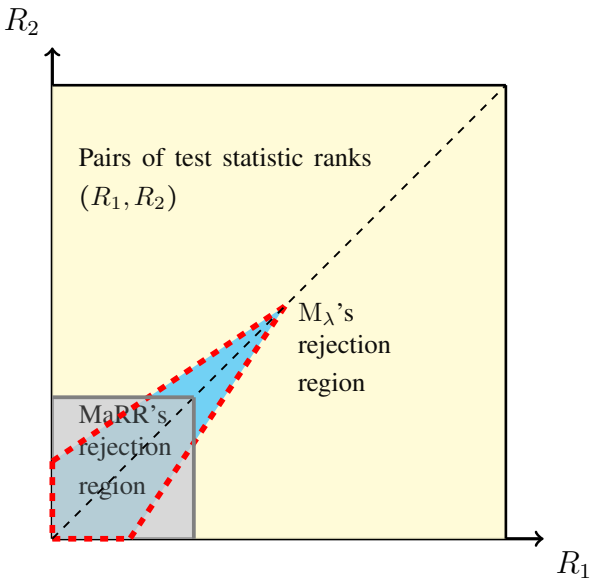


Figure 2.1: Rejection region geometry for MaRR and $M_{\lambda=5,i}$.

hypotheses that live near one axis but far from the origin to be reproducible and de-emphasizes the *agreement* of results for a hypothesis across replicate studies. This is counter to the notion of reproducibility introduced by Li et al. (2011) and thus we consider only $\lambda \geq 1$ throughout the chapter. Now, we describe a novel procedure for discovering reproducible hypotheses with a nominal false discovery rate of α using the $M_{\lambda,i}$ statistic in Section 2.3.

2.3 FDR controlling procedure

In the multiple-testing setting where n hypotheses are being tested simultaneously, the typical goal is to reject a set of hypotheses at a nominal false discovery rate (FDR). That is, we denote V as the number of hypotheses in a rejection set that are truly null and Q as the total number of hypotheses in the rejection set. Then, false discovery proportion (FDP) and FDR, as introduced by Benjamini and Hochberg (1995), are defined by

$$\text{FDP} = \frac{V}{Q \vee 1} \text{ and } \text{FDR} = \mathbb{E}\left[\text{FDP}\right].$$

In the reproducibility framework, with n hypotheses common across $m = 2$ replicate experiments, V represents the number of *irreproducible* hypotheses which are deemed to be reproducible by a procedure and Q is the total number of hypotheses found to be reproducible. Since reproducible hypotheses tend to have small $M_{\lambda,i}$ statistics, we can write the V and Q quantities for sets defined by $M_{\lambda,i}$ statistic as $V_\lambda(x) = \sum_{h \in \mathcal{H}_0} \mathbb{I}[M_{\lambda,h}/n \leq x]$ and $Q_\lambda(x) = \sum_{i=1}^{n} \mathbb{I}[M_{\lambda,i}/n \leq x]$ for any λ and threshold x. Thus, the FDP and FDR using the $M_{\lambda,i}$ statistic is defined as

$$\mathrm{FDP}_\lambda(x) = \frac{V_\lambda(x)}{Q_\lambda(x) \vee 1} \text{ and } \mathrm{FDR}_\lambda(x) = \mathbb{E}\left[\mathrm{FDP}_\lambda(x)\right]. \tag{2.4}$$

Notice, for any fixed λ and x, $Q_\lambda(x)$ is known, but without knowledge of which hypotheses are reproducible, $V_\lambda(x)$ is unknown and thus must be estimated. We propose doing so by noticing $\mathbb{E}[V_\lambda(x)] = n(1 - \pi_1)F_\lambda(x)$ where $F_\lambda(x)$ is the irreproducible distribution of $M_{\lambda,i}/n$ or $\mathbb{P}(M_{\lambda,h}/n \leq x \mid h \in \mathcal{H}_0)$. To estimate that irreducible distribution of $M_{\lambda,i}/n$ – and with, it the FDP – we impose Condition 2.3.1. Then, under Condition 2.3.1, we propose estimates for irreproducible hypotheses rank statistics and $M_{\lambda,i}/n$ statistic.

Condition 2.3.1. *There exists* $\beta_0 \in (\pi_1, 1)$*, such that if* $\max(R_{1,h}/n, R_{2,h}/n) > \beta_0$ *then* $h \in \mathcal{H}_0$.

The existence of a β_0 that satisfies Condition 2.3.1 implies screening regions $\mathcal{R}_1^\beta$ and $\mathcal{R}_2^\beta$ seen in (2.5) will be comprised entirely of irreproducible hypotheses for any $\beta > \beta_0$, which in turn allows the distribution of all irreproducible ranks to be estimated by these screening regions. A common route to showing in controlling false discovery rate is to create a statistic that tests the null hypothesis that is symmetric under the null (Barber and Candés, 2015; Dai et al., 2022; Xing et al., 2023). Then leveraging this symmetry, methods estimate the number of false discoveries by the number of hypotheses that lie in the mirror image of the rejection region. Unfortunately, the $M_{\lambda,i}$ statistic is not symmetric, and thus we cannot follow this exact path.

However, the utility of symmetry in FDR calculation is that it provides a partition of the empirical distribution of test statistics consisting of two spaces: one that is overwhelmingly comprised of null (in our case irreproducible) hypotheses and another comprised of null (irreproducible) and

non-null (reproducible) hypotheses. Then, using the null partition, one can closely approximate the number of null hypotheses that lie in any region of the partition with both types of hypotheses by examining the mirror image of the region. Condition 2.2.1 provides such a partitioning of the distribution of ranks in each experiment. That is, for all $\mathbb{H}_h$ where $R_{1,h}/n > \beta_0$, $\mathbb{H}_h$ is irreproducible and the same is true for $\mathbb{H}_h$ with $R_{2,h}/n > \beta_0$.

We argue that Condition 2.3.1 is relaxed relative to other assumptions for other reproducibility methods. We make no parametric assumptions on the test statistics $T_{1,i}$ and $T_{2,i}$, unlike the copula mixture in Li et al. (2011) or the two-model Bayesian approach from Heller and Yekutieli (2014). Among nonparametric methods, Condition 2.3.1 seems relatively minor. For example, assumption I1 in Philtron et al. (2018) requires complete separation of the ranks of reproducible hypotheses and irreproducible hypotheses. That is, for all $g \in \mathcal{H}_1$, I1 presumes $R_{1,g}/n \leq \pi_1$ and $R_{2,g}/n \leq \pi_1$, and if $h \in \mathcal{H}_0$ then $R_{1,h/n} < \pi_1$ and $R_{2,h}/n > \pi_1$. Notice, the assumption I1 is far stronger than Condition 2.3.1. I1 is actually a special case of Condition 2.3.1. Notice, under I1 if $\max(R_{1,h}/n, R_{2,h}/n) > \pi_1$, then $h \in \mathcal{H}_0$ and Condition 2.3.1 is satisfied with $\beta_0 = \pi_1$.

Intuitively, Condition 2.3.1 can be thought of as a lower bound on the signal strength of reproducible hypotheses. Since it assumes no reproducible rank can take a rank beyond $n\beta_0$ in either experiment there is a lower bound on the amount of signal a reproducible hypothesis must show relative to that of the irreproducible hypotheses. This is sensible because for a result to be reproduced across two studies, we expecti it to show at least some level of signal strength. Now, leveraging Condition 2.3.1, we devise screening partitions of the rank-space and use them to estimate the distribution of scaled irreproducible ranks.

2.3.1 Estimators of irreproducible distributions

To estimate the distributions of ranks $R_{1,h}$ and $R_{2,h}$ for $h \in \mathcal{H}_0$, we first define the β-dependent screening sets $\mathcal{R}_1^\beta$ and $\mathcal{R}_2^\beta$ by

$$\mathcal{R}_1^\beta = \{i : R_{2,i}/n > \beta\} \text{ and } \mathcal{R}_2^\beta = \{i : R_{1,i}/n > \beta\}. \tag{2.5}$$

Through Condition 2.3.1, if $\beta > \beta_0$ then $i \in \mathcal{R}_1^\beta \cup \mathcal{R}_2^\beta$ implies $i \in \mathcal{H}_0$. Figure 2.2 shows an example of the screening regions for one iteration of Setting B from Section 2.6 with $\beta = 0.8$. The screening region for experiment 1, $\mathcal{R}_1^\beta$, is shown along the top of the figure and $\mathcal{R}_2^\beta$ is shown along the right-hand side. Notice, all hypotheses in both regions are irreproducible. Definition 2.2.1 states that if

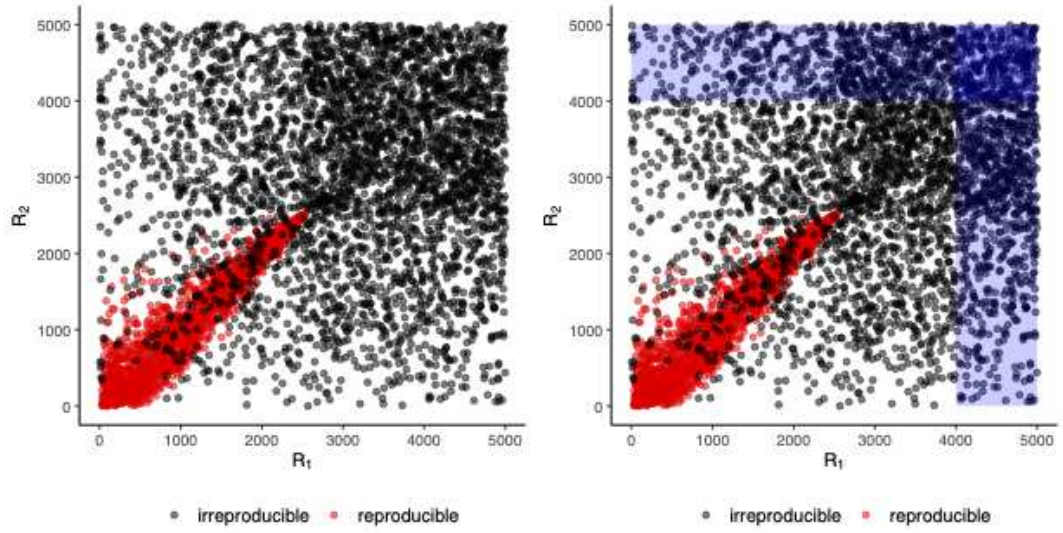


Figure 2.2: One iteration from the Simulation B detailed in Section 2.6 with $a = 0.5$, $b = 3.5$, $c = 0.75$ and $\pi_1 = 0.30$. Reproducible hypotheses are in red and irreproducible in black. The blue shaded areas represent the screening regions $\mathcal{R}_1^\beta$ and $\mathcal{R}_2^\beta$.

$h \in \mathcal{H}_0$ then $R_{1,h}$ is independent of $R_{2,h}$. Thus, for all $h \in \mathcal{H}_0$, $\mathbb{P}(R_{1,h} \leq x \mid h \in \mathcal{R}_1^\beta) = \mathbb{P}(R_{1,h} \leq x)$. For this reason, we use the hypotheses in $\mathcal{R}_1^\beta$ to estimate the distribution of $R_{1,h}/n$ for $h \in \mathcal{H}_0$ across the entire rank-space and the hypotheses in $\mathcal{R}_2^\beta$ to estimate the distribution of $R_{2,h}/n$. That is, the estimates for the distribution of irreproducible scaled ranks, $\widehat{F}_{n,i}^\beta$ for $i \in \{1, 2\}$, are defined by

$$\widehat{F}_{n,1}^\beta(x) = \sum_{\ell \in \mathcal{R}_1^\beta} \frac{\mathbb{I}(R_{1,\ell}/n \leq x)}{\left|\mathcal{R}_1^\beta\right|} \quad \text{and} \quad \widehat{F}_{n,2}^\beta(x) = \sum_{\ell \in \mathcal{R}_2^\beta} \frac{\mathbb{I}(R_{2,\ell}/n \leq x)}{\left|\mathcal{R}_2^\beta\right|}. \tag{2.6}$$

Using Definition 2.2.1 and Condition 2.3.1, one can show Proposition 2.3.1

Proposition 2.3.1. *Under Condition 2.3.1, for any fixed $\beta > \beta_0$, it follows that $\mathbb{E}[\widehat{F}_{n,i}^\beta(t)] = F_1(t)$ for all $t \in (c, \lambda + 1)$ where $c > 0$.*

The result can easily be shown by conditioning on the event $h \in \mathcal{H}_0$ for all $h \in \mathcal{R}_i^\beta$. Next, to estimate the distribution function of $M_{\lambda,h}/n$ for $h \in \mathcal{H}_0$, called $\widehat{F}_{\lambda,n}^\beta$ consider the geometric region $M_{\lambda,h}/n = (R_{1,h}/n + R_{2,h}/n + \lambda|R_{1,h}/n - R_{2,h}/n|) \leq x$ of the scaled rank-space. The space can be visualized in the leftmost panel of Figure 2.3. Notice, it can be found by examining the area of the outer square in the second panel less the triangles in the third and fourth panels of Figure 2.3. To estimate the density of irreproducible $M_{\lambda,h}/n$ statistics that fall in that region, we can integrate

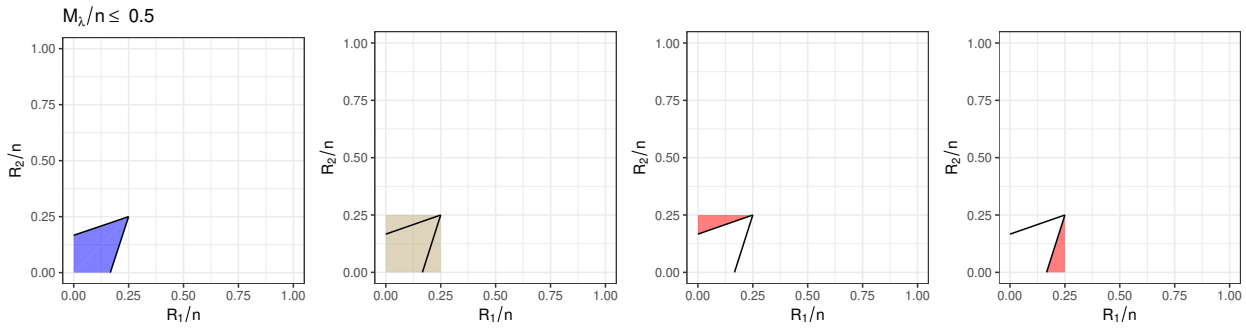


Figure 2.3: Geometric space of $M_{\lambda,i}/n < x$ with $\lambda = 2$.

over the desired region using $\widehat{F}_{1,n}^\beta$ and $\widehat{F}_{2,n}^\beta$. So, our estimate for the distribution function of $M_{\lambda,i}/n$, $\widehat{F}_{\lambda,n}^\beta(x)$ can be calculated by

$$\begin{aligned}\widehat{F}_{\lambda,n}^\beta(x) &= \int_0^{\frac{x}{2}} \int_0^{\frac{x}{2}} d\widehat{F}_{1,n}^\beta(t_1) d\widehat{F}_{2,n}^\beta(t_2) - \int_0^{\frac{x}{2}} \int_{a(t_1)}^{\frac{x}{2}} d\widehat{F}_{2,n}^\beta(t_2) d\widehat{F}_{1,n}^\beta(t_1) \\ &\quad - \int_0^{\frac{x}{2}} \int_{a(t_2)}^{\frac{x}{2}} d\widehat{F}_{1,n}^\beta(t_1) d\widehat{F}_{2,n}^\beta(t_2) \\ &= \sum_{t_1=\frac{1}{n}}^{x/2} \widehat{F}_{2,n}^\beta(a(t_1)) \sum_{i \in \mathcal{R}_1} \frac{\mathbb{I}[R_{1,i} = t_1]}{|\mathcal{R}_1|} + \sum_{t_2=\frac{1}{n}}^{x/2} \widehat{F}_{1,n}^\beta(a(t_2)) \sum_{i \in \mathcal{R}_2} \frac{\mathbb{I}[R_{2,i} = t_2]}{|\mathcal{R}_2|} \\ &\quad - \widehat{F}_{1,n}^\beta\left(\frac{x}{2}\right) \widehat{F}_{2,n}^\beta\left(\frac{x}{2}\right)\end{aligned} \tag{2.7}$$

where $a(t) = (1 - \frac{2}{1+\lambda})t + \frac{x}{1+\lambda}$. Essentially, the first integral can be thought of as the distribution function when $\lambda = 1$, and the next two integrals are the difference in distribution between $\lambda = 1$ and the specified λ. Additionally, since it is a continuous mapping of the random processes $\widehat{F}_{1,n}^\beta$ and $\widehat{F}_{1,n}^\beta$ we can leverage their individual convergence results to show $\widehat{F}_\lambda^\beta$ converges to the

true distribution function of $M_{\lambda,i}/n$, F_λ. The theoretical properties of these estimates are further examined in Section 2.4. Now, using $\widehat{F}^{\beta}_{\lambda,n}$ we define a procedure for declaring features reproducible for a nominal FDR level.

2.3.2 Controlling FDR

Now, using $\widehat{F}^{\beta}_{\lambda}$, we introduce a procedure to estimate FDP and decide a threshold that defines a rejection set at the nominal level of FDR of α. Notice, we can estimate $V_\lambda(x)$ from (2.4), by using $\widehat{F}_{\lambda,n}(x)$ derived in (2.7) as follows.

$$\widehat{V}^{\beta}_{\lambda,\pi_1}(x) = n(1-\pi_1)\widehat{F}^{\beta}_{\lambda,n}(x). \tag{2.8}$$

In Lemmas A.1.2 and A.1.3 in Appendix A.1, we show $\widehat{V}^{\beta}_{\lambda,\pi_1}(x)$ and $V_\lambda(x)$ both converge in probability to $\mathbb{E}[V_\lambda(x)]$, so $\widehat{V}^{\beta}_{\lambda,\pi_1}(x)$ can be used to estimate the true number of false discoveries made at threshold x. Combining (2.8) and the number of total discoveries, $Q_\lambda(x)$, we estimate FDP, as defined in (2.4), at a threshold of x by

$$\widehat{\mathrm{FDP}}^{\beta}_{\lambda}(x) = \frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{Q_\lambda(x) \vee 1} = \frac{n(1-\pi_1)\widehat{F}^{\beta}_{\lambda,n}(x)}{\sum_{i=1}^{n} \mathbb{I}\left[M_{\lambda,i}/n \le x\right] \vee 1}. \tag{2.9}$$

Then, we choose a threshold, $\widehat{t_\alpha}$ for assessing reproducibility at desired FDR level α by

$$\widehat{t_\alpha} = \max_{t\epsilon}\{t : \widehat{\mathrm{FDP}}^{\beta}_{\lambda}(t) \le \alpha\} \tag{2.10}$$

and deem any hypothesis, $\mathbb{H}_i$ reproducible if $M_{\lambda,i}/n \le \widehat{t_\alpha}$. Algorithm 1 summarizes the $M_{\lambda,i}$ procedure for a fixed λ with π_1 known.

2.4 Theoretical guarantees

We first establish uniform convergence of the rank distributions, $\widehat{F}^{\beta}_{n,1}$ and $\widehat{F}^{\beta}_{n,2}$ in Theorem 2.4.1. We then use the uniform convergence of the rank distributions to show that our estimate

Algorithm 1 Step-by-step $M_{\lambda,i}$ procedure (fixed λ and π_1 known).

1: Rank hypotheses by summary statistics, $T_{1,i}$ and $T_{2,i}$, to get $R_{1,i}$ and $R_{2,i}$.
2: Calculate $M_{\lambda,i}$ statistic for each $\mathbb{H}_i$ by (2.3).
3: For fixed $\beta > \beta_0$, create screening regions $\mathcal{R}_1^\beta$ and $\mathcal{R}_2^\beta$ by (2.5) and estimate the marginal rank distributions for $h \in \mathcal{H}_0$ by $\widehat{F}_{n,1}^\beta$ and $\widehat{F}_{n,1}^\beta$ from (2.6).
4: Estimate the marginal distribution of $M_{\lambda,h}/n$ by $\widehat{F}_{\lambda,n}$ by integrating as in (2.7).
5: Estimate FDP at threshold x by $\widehat{\mathrm{FDP}}_\lambda^\beta(x)$ from (2.9) and select critical value for nominal FDR level α by $\widehat{t}_\alpha$ in (2.10).
6: Deem $\mathbb{H}_i$ reproducible at nominal FDR level α if

$$M_{\lambda,i}/n \le \widehat{t}_\alpha.$$

$\widehat{\mathrm{FDP}}_\lambda^\beta$ converges to the true FDP in Theorem 2.4.2 and show that the method proposed in Section 2.2 asymptotically controls FDR at a nominal level α in Theorem 2.4.3.

Theorem 2.4.1. *Under Condition 2.3.1, for fixed $\beta > \beta_0$, the approximate distribution function $\widehat{F}_{n,i}^\beta$ from* (2.6) *satisfies $\sup_{t\in(0,1)} \left|\widehat{F}_{n,i}^\beta(t) - F_i(t)\right| \xrightarrow{p} 0$, for $i = 1, 2$, where $F_i(t)$ is the rank distribution of irreproducible hypotheses, $F_i(x) = \mathbb{P}\left(R_{i,h}/n \le x \mid h \in \mathcal{H}_0\right)$.*

Theorem 2.4.1 ensures the convergence of the estimated rank distributions. An advantage of using rank statistics is that they are *negatively associated* (Joag-Dev and Proschan, 1983). That is, the covariance of any monotone function of ranks is not positive. Negatively associated random variables have many desirable properties (Miao et al., 2014). One such property paramount in the proof of Theorem 2.4.1 is that the Glivenko-Cantelli theorem applies to any sequence of negatively associated random variables (see Lemma 3.6 in Miao et al., 2014).

We then use the estimates, $\widehat{F}_{n,i}^\beta$ to estimate the distribution of irreproducible $M_{\lambda,h}$ statistics by $\widehat{F}_{\lambda,n}$ given by (2.7). Since $\widehat{F}_{\lambda,n}$ is a continuous transformation of $\widehat{F}_{n,i}^\beta$ for $i = 1, 2$, Theorem 2.4.1 guarantees the uniform convergence of $\widehat{F}_{\lambda,n}$ to its corresponding distribution function. We can then show that $\widehat{\mathrm{FDP}}_\lambda^\beta$ converges uniformly to the true FDP, as seen in Theorem 2.4.2.

Theorem 2.4.2. *Under Condition 2.3.1, for a fixed λ and $\beta > \beta_0$, the estimated false discovery proportion $\widehat{\mathrm{FDP}}_\lambda^\beta$ in* (2.9) *satisfies*

$$\sup_{t\in(c,\lambda+1)} \left|\widehat{\mathrm{FDP}}_\lambda^\beta(t) - \mathrm{FDP}_\lambda(t)\right| \xrightarrow{p} 0,$$

for any $c > 0$ where $\mathrm{FDP}_\lambda(t)$ is the true false discovery proportion defined in (2.4).

Theorem 2.4.2 shows the convergence of the estimated $\widehat{\mathrm{FDP}}_\lambda$ to the true FDP. The lower bound c is required to ensure the number of total and false discoveries are non-zero asymptotically. Theorem 2.4.3 uses the uniform convergence of $F_{\lambda,n}$ and V_λ to their respective expectations to show asymptotic FDR control using the threshold $\widehat{t}_\alpha$ from (2.10).

Theorem 2.4.3. *For a fixed λ and nominal level α, assume there exists t_α such that $\mathbb{P}(\mathrm{FDP}_\lambda(t_\alpha) \le \alpha) \to 1$ as $n(1-\beta) \to \infty$. Under Condition 2.3.1, for a fixed λ and $\beta > \beta_0$ with $n(1-\beta) \to \infty$,*

$$\mathrm{FDP}_\lambda(\widehat{t}_\alpha) \le \alpha + o(1) \;\; \textit{and} \;\; \limsup_{n\to\infty} \mathrm{FDR}_\lambda(\widehat{t}_\alpha) \le \alpha,$$

for any nominal FDR level $\alpha \in (0,1)$ where $\widehat{t}_\alpha$ is the critical value determined by (2.10).

The assumption of the existence of t_α such that $\mathbb{P}(\mathrm{FDP}_\lambda(t_\alpha) \le \alpha) \to 1$ bounds $\widehat{t}_\alpha$ from below and guarantees the existence of the asymptotic FDR. Analogous conditions can be found in Dai et al. (2022), Xing et al. (2023), and Storey et al. (2004). The proof of Theorem 2.4.3 follows a similar path to Theorem 6 in Xing et al. (2023) and Proposition 2.2 in Dai et al. (2022). While both of those methods can leverage test statistic symmetry and parametric regression assumptions to show the convergence of estimated false discoveries and empirical false discoveries, our proof is technically more challenging due to the complicated dependence structure of rank statistics. Instead, we leverage the properties of negatively associated rank statistics to show the convergence of the estimated false discoveries and leverage properties of the α-mixing conditions to bound the variance of the true number of false discoveries. Theorem 2.4.3 provides theoretical justification for the method described in Section 2.3.2 under Condition 2.3.1. Specifically, we show asymptotic

false discovery rate control when selecting a threshold for the $M_{\lambda,i}$ statistic in the manner of (2.10). This result is significant, as previous methods for examining a similar notions of reproducibility (Philtron et al., 2018) show FDR control in simulation but lack theoretical justification, while other methods with such justification either make stringent parametric assumptions (Heller and Yekutieli, 2014) or rely on a particular type of summary statistic (Hung and Fithian, 2020; Jaljuli et al., 2022; Wang et al., 2022).

2.5 Practical implementations

2.5.1 Estimation of π_1

The method for estimating FDP described in Section 2.3.2 relied on knowledge of π_1, the true proportion of hypotheses that are reproducible. In practice, however, this quantity is rarely known. We estimate it by $\widehat{\pi}_1$ in a similar way to Philtron et al. (2018). For any t, let $\widehat{S}_n(t)$ be calculated by

$$\widehat{S}_n(t) = n^{-1} \sum_{i=1}^{n} \mathbb{I}\left[\max(R_{1,i}/n, R_{2,i}/n) > t\right].$$

Theorem 1 from Philtron et al. (2018) showed that when ranks from reproducible hypotheses are entirely separated from those of irreproducible hypotheses – or analogously Condition 2.3.1 is met for $\beta_0 = \pi_1$ – then the true asymptotic survival function for $\max(R_{1,h}/n, R_{2,h}/n)$ for any $h \in \mathcal{H}_0$ takes the form

$$S_{\pi_1}(t) = \begin{cases} 1 & t \in (-\infty, \pi_1) \\ 1 - \dfrac{(t-\pi_1)^2}{(1-\pi_1)^2} & t \in [\pi_1, 1] \\ 0 & t \in (1, \infty). \end{cases}$$

They then defined the estimate $\widehat{\pi}_1$ by the value that minimizes the mean square error between the empirical survival function and the asymptotic survival function, or

$$\begin{aligned} \widehat{\pi}_1 &= \underset{\gamma \in \{1/n, 2/n, \cdots, (n-1)/n\}}{\arg\min} \{\mathrm{MSE}_n(\gamma)\} \\ &= \underset{\gamma \in \{1/n, 2/n, \cdots, (n-1)/n\}}{\arg\min} \left\{ (n - n\gamma)^{-1} \sum_{x=n\gamma}^{n} \left(\widehat{S}_n(x/n) - (1-\gamma) S_\gamma(x/n)\right)^2 \right\}. \end{aligned} \tag{2.11}$$

Under the same separation condition previously stated, Philtron et al. (2018) showed $\widehat{\pi}_1 \xrightarrow{p} \pi_1$. Under a relaxed assumption, R1,

R1 $\mathbb{P}(R_{1,g} \leq R_{1,h}) > 1/2$ and $\mathbb{P}(R_{2,g} \leq R_{2,h}) > 1/2$ for any $g \in \mathcal{H}_1$ and $h \in \mathcal{H}_0$

they theoretically justify that $\mathbb{E}(\widehat{\pi}_1) \leq \pi_1$. We favor this estimate of sparsity due to its conservative nature. Notice, if $\widehat{\pi}_1 \leq \pi_1$, then $(1 - \widehat{\pi}_1)\widehat{F}^{\beta}_{\lambda}(x) \geq (1 - \pi_1)\widehat{F}^{\beta}_{\lambda}(x)$, which means the estimated number of false discoveries is larger using $\widehat{\pi}_1$ compared to π_1 and the procedure in Section 2.3.2 will control FDR below a nominal level of α. For this reason, in the practical implementation of this method, we replace the true π_1 with $\widehat{\pi}_1$ from (2.11) and apply the same procedure.

2.5.2 Selection of β

The proposed method also required a user-specified upper bound on the reproducible ranks, denoted β. Notice, all the theoretical properties presented in Section 2.4 assume the selected β value satisfies $\beta > \beta_0$. For this reason, one might suggest a β close to 1 (i.e. $\beta = 0.999$) to ensure $\beta > \beta_0$. Notice, however, β defines the lower boundary of the screening regions, $\mathcal{R}^{\beta}_1$ and $\mathcal{R}^{\beta}_2$, which are used to estimate the rank distributions. So, specifying β too close to 1 yields a large lower bound on the geometry of the screening regions, and the sets $\mathcal{R}^{\beta}_1$ and $\mathcal{R}^{\beta}_2$ will be small in size. Smaller screening regions result in estimating irreproducible rank distributions using less data, introducing more uncertainty in the estimation of rank distributions and FDP.

From Condition 2.3.1, the criteria $\beta > \beta_0$ can be rephrased to any β which satisfies the condition *if $h \in \mathcal{R}^{\beta}_1 \cup \mathcal{R}^{\beta}_2$, then h is irreproducible.* Since it is difficult to devise a method that successfully identifies β_0 exactly without making stringent assumptions about rank distributions of reproducible hypotheses, we devise a method for selecting β such that if $h \in \mathcal{R}^{\beta}_1 \cup \mathcal{R}^{\beta}_2$, then h *looks* irreproducible. To motivate this selection criteria, consider the set $\mathcal{R}^{\beta}_{1,2} = \mathcal{R}^{\beta}_1 \cap \mathcal{R}^{\beta}_2 = \{h : n^{-1}(R_{1,h} \vee R_{2,h}) \geq \beta\}$. By Condition 2.2.1, for any $\beta > \beta_0$, $h \in \mathcal{R}^{\beta}_{1,2}$ implies $h \in \mathcal{H}_0$. Additionally, Definition 2.2.1 in conjunction with Condition 2.3.1 imply that $R_{1,h}$ and $R_{2,h}$ for $h \in \mathcal{R}^{\beta}_{1,2}$ are independent and uniformly distributed on $\{n\beta, n\beta + 1, \cdots, n\}$. With knowledge of these distribu-

tions, we compute the survival functions of the minimum rank statistics across experiment for all $h \in \mathcal{R}^{\beta}_{1,2}$, denoted S^{β}, in Lemma 2.5.1.

Lemma 2.5.1. *Suppose Condition 2.1 holds for β_0. Then, for any $\beta \in \{\beta_0 + n^{-1}, \cdots, 1\}$ and $r \in \{\beta, \beta + n^{-1}, \cdots, 1\}$,*

$$S^{\beta}(r) = \mathbb{P}(n^{-1}(R_{1,h} \wedge R_{2,h}) > r \mid h \in \mathcal{R}^{\beta}_{1,2}) = \left(\frac{1-r}{1-\beta}\right)^2.$$

We propose selecting a β by examining the difference between this true survival function and the empirical survival functions of minimum rank statistics for $h \in \mathcal{R}^{\beta}_{1,2}$, denoted $\widehat{S}^{\beta}$, where

$$\widehat{S}^{\beta}(x) = \frac{\sum_{\ell=1}^{n} \mathbb{I}[n^{-1}(R_{1,\ell} \wedge R_{2,\ell}) > x]}{\sum_{i=1}^{n} \mathbb{I}[n^{-1}(R_{1,\ell} \wedge R_{2,\ell}) > \beta]}.$$

Now, the criterion we use to select a suitable β is as follows.

$$\tilde{\beta} = \min\left\{\beta \in (\pi_1, 1-\delta) : \int_{\beta}^{1} (\widehat{S}^{\beta}(t) - S^{\beta}(t))^2 dt \leq d_{n,\beta}\right\} \tag{2.12}$$

where $d_{n,\beta}$ is a user-specified rate and $1-\delta$ is the upper bound on what we consider to be β_0 could be. The lower bound on the β selection is π_1, as that is the value of β_0 under the total separation of reproducible and irreproducible ranks. For the survival function convergence rate, $d_{n,\beta}$, we recommend using $d_{n,\beta} = (n(1-\beta))^{-3/2}$. Appendix A.2 contains simulation results examining the performance of the proposed method in terms of FDP and power across a broad range of β selections – both smaller and larger than the true β_0. In simulation, we observe that the performance of the proposed method is not sensitive to the selection of β for reasonably large levels of β. It also contains a simulations examining the performance of the method using the proposed β selection procedure for differing rates $d_{n,\beta}$ along with an additional alternative method. We observe preferred performance of the recommended rate of $d_{n,\beta} = (n(1-\beta))^{-3/2}$.

2.5.3 Selection of λ

Finally, to use the $M_{\lambda,i}$ method, we must specify a value of λ. Notice, the theoretical properties of the $M_{\lambda,i}$ procedure presented in Section 2.4 hold for all fixed λ. Thus, control of FDR does not hinge on the selection of a suitable λ. Additionally, through simulations presented in Appendix A.2, we do not observe large differences in the preciseness of the FDR control for different values of λ in simulation. We do, however, observe noticeable differences in the method's ability to detect reproducible hypotheses for different specified values of λ. In simulation, specific values of λ show higher levels of observed power relative to other values under different settings. Thus, the value of λ does not appear to impact FDR control but does impact the power of the proposed method.

Geometrically, the specification of λ controls the shape of the region in the rank-space that is deemed reproducible by the proposed method. Figure 2.4 displays the shapes of reproducible regions for the proposed method when $\lambda \in \{1, 2, 3, 4\}$. In selecting λ from a list of choices, we want to select the λ that corresponds to the region with the most reproducible hypotheses.

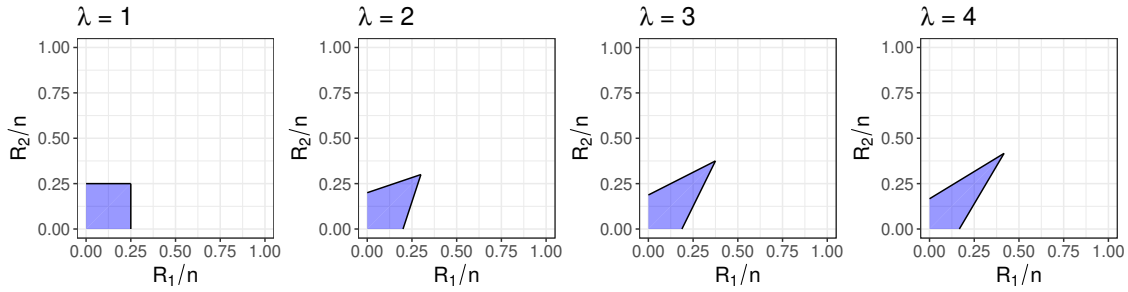


Figure 2.4: Reproducible region geometry for $\lambda \in \{1, 2, 3, 4\}$.

Since we do not observe stark differences in the method's ability to estimate the true FDP across different λ, the reproducible sets from the proposed method with different values of λ will have similar proportions of false discoveries. For that reason, we argue the λ which yields the largest total number of hypotheses deemed reproducible will *tend* to find the most *reproducible* hypotheses reproducible and thus have a higher power than other values λ. So, we recommend selecting the value of λ from a specified list the yields the largest total number of reproducible discoveries.

That is, consider Λ to be a finite set of values to consider for λ (i.e., $\Lambda = \{1, 1.5, 2, 2.5, \cdots, 10\}$ or $\Lambda = \{1, 2, 3, 4, 5\}$) and let $\widehat{t}_\alpha(\lambda)$ be the critical value for a nominal FDR level of α, as defined in (2.10), we recommend selecting a λ value that matches the criterion

$$\lambda = \arg\max_{\ell \in \Lambda} \left| \sum_{i=1}^{n} \mathbb{I}\left[M_{\ell,i}/n \leq \widehat{t}_\alpha(\ell) \right] \right|. \tag{2.13}$$

In Appendix A.2, we examine the performance of this λ specification criteria relative to other values of λ and see strong performance.

2.6 Numerical results

We examine the finite sample performance of the proposed method through three simulation settings: simulation A, B, and C – described in Section 2.6.1. To assess the performance of $M_{\lambda,i}$ compared to existing literature, we compare the proposed method to the maximum rank procedure (MaRR; Philtron et al., 2018) implemented in `R` by the package `marr` employing the default estimation of $\widehat{\pi}_1$; the copula mixture based procedure (IDR; Li et al., 2011) implemented by the `idr` package with initial values of $\mu = 1.75$, $\sigma^2 = 0.5$, $\rho = 0.5$, and $\pi_1 = \widehat{\pi}_1$; the adaptive filtering procedure (AdaFilter; Wang et al., 2022) performed on the p-values from generated summary statistics in each setting via the package `AdaFilter` using type I method of FDR. Additionally, we implement the proposed $M_{\lambda,i}$ procedure with fixed $\lambda \in \{1, 2, 5\}$, $\widehat{\pi}_1$ estimated by Section 2.5 adapted from Philtron et al. (2018), and fixed $\beta = 0.9$ (see Appendix A.2 for sensitivity analysis pertaining to performance for fixed β); and the proposed method with λ selected from $\Lambda = \{1, 1.5, 2, 2.5, 3, 5, 7.5\}$ in the manner described in Section 2.5, $\widehat{\pi}_1$ estimated by Section 2.5, and β selected by the criteria in Section 2.5 with $d_{n,\beta} = \big(n(1-\beta)\big)^{-3/2}$ and $1 - \delta = 0.9$.

Simulations A and B both generate summary statistics for hypotheses from a bivariate Gaussian distribution with differing parameters meant to represent the signal strength and consistency of reproducible hypotheses. Simulation A is proposed to allow for reproducible hypotheses to have differing levels of signal strength and consistency within the same experiment. Simulation B

generates summary statistics from the bivariate Gaussian distribution used to derive the canonical copula mixture in Li et al. (2011) and was also previously considered in Philtron et al. (2018). Simulation C is motivated by the replicating signal simulations considered in Wang et al. (2022) and Deng et al. (2023). In which, the summary statistic provided is a two-sided p-value pertaining to a z-score. Interestingly, this setting allows for hypotheses to demonstrate signal in one experiment but not the other. Figure 2.5 depicts the bivariate density of reproducible ranks from one iteration of each simulation. Notice, reproducible ranks from simulation A closely resemble the shape of the $M_{\lambda,i}$ algorithm's reproducible region shown in Figure 2.4 when λ is 3 or 4, indicating that the inclusion of a larger λ yield better power. Simulations B and C are more varied in shape.

2.6.1 Settings

Simulation A

Simulation A considers a setting with Gaussian summary statistics with differing means and variances across all reproducible hypotheses. We consider $n = 10,000$ hypotheses and π_1 proportion being reproducible. If $h \in \mathcal{H}_1$, then its summary statistics are generated from

$$\begin{bmatrix} T_{1,g} \\ T_{2,g} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_g \\ \mu_g \end{bmatrix}; \begin{bmatrix} \sigma_g^2 & 0 \\ 0 & \sigma_g^2 \end{bmatrix} \right)$$

where μ_g are generated independently by $\text{UNIF}(a, b)$ and $\sigma_g = (1 - \sigma_0)\left(\frac{\mu_g - a}{b - a}\right) + \sigma_0$. If $h \in \mathcal{H}_0$, its summary statistics, $T_{1,h}$ and $T_{1,h}$ are generated independently by $\mathbb{N}(0, 1)$. Heuristically, μ_g represents the signal strength of $\mathbb{H}_g$ and σ_g represents the signal consistency across experiment. Thus a and b represent the upper and lower bounds on the signal strengths of reproducible hypotheses. Note that the function defining σ_g leads to hypotheses with summary statistics strength of $\mu_g = b$ having $\sigma_g = 1$ and summary statistics strength of $\mu_g = a$ having $\sigma_g = \sigma_0$. This setting allows for the existence of reproducible hypotheses with strong signal strength and moderate consistency and reproducible hypotheses with weaker signal strength and strong signal consistency within the

same simulation. We consider all combinations of $a = 0.5$, $b \in \{2.5, 3.5, 5\}$ and $\sigma_0 = 0.01$ for $\pi_1 \in \{0.1, 0.3\}$.

Simulation B

Simulation B considers a setting with Gaussian summary statistics with the same mean, variance, and correlation for all reproducible hypotheses. Again, we consider $n = 10,000$ features, of which π_1 proportion are reproducible. If $g \in \mathcal{H}_1$, then, its summary statistics are generated from

$$\begin{bmatrix} T_{1,g} \\ T_{2,g} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu \\ \mu \end{bmatrix}; \sigma_1^2 \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix} \right).$$

If $h \in \mathcal{H}_0$, its summary statistics, $T_{1,h}$ and $T_{1,h}$ are generated independently by $\mathbb{N}(0, 1)$. The parameter μ represents the strength of signals, as a larger μ will yield highly ranked reproducible hypotheses. The parameters σ^2 and ρ are meant to represent signal consistency across experiment, as smaller σ^2 or ρ will yield more similarly ranked summary statistics for reproducible hypotheses. We consider every combination $\mu = 1.5$, $\sigma_1^2 = 0.5$, and $\rho \in \{0.5, 0.8, 0.95\}$ for $\pi_1 \in \{0.1, 0.3\}$.

Simulation C

Simulation C considers two-sided p-values as the summary statistic of interest. It also considers a setting in which some irreducible hypotheses show spurious signal in one experiment but not the other. That is, we consider $n = 10,000$ hypotheses across two experiments. These hypotheses are characterized by two proportions π_1 and π_{00}. π_1 represents the proportion of hypotheses that are reproducible and thus have true signal in both experiments. π_{00} is the proportion of hypotheses that are truly null in both experiments and thus show no signal in both experiments. The remaining $1-\pi_1-\pi_{00}$ have spurious signal in one experiment but no signal in the other experiment. That is, let $\boldsymbol{\theta}_i = (\theta_{1,i}; \theta_{2,i})$ be indicator variables which indicate whether hypothesis $\mathbb{H}_i$ represents true signal ($\theta_{j,i} = 1$) or not ($\theta_{j,i} = 0$) in each experiment. Then, for π_1 proportion of hypotheses, $\boldsymbol{\theta_i} = (1, 1)$, for $\pi_{0,0}$ proportion of hypotheses, $\boldsymbol{\theta_i} = (0, 0)$, and for the remaining $1 - \pi_1 - \pi_{00}$ hypotheses, $\boldsymbol{\theta_i}$ is sampled independently from $\{(0, 1); (1, 0)\}$. Summary statistics for all hypotheses are two-sided

p-values from z scores in which the z scores are generated by

$$\begin{bmatrix} z_{1,i} \\ z_{2,i} \end{bmatrix} \sim \mathbb{N}\left(\mu_i \boldsymbol{\theta}_i; \begin{bmatrix} 1 & \theta_{1,i}\theta_{2,i}\rho \\ \theta_{1,i}\theta_{2,i}\rho & 1 \end{bmatrix} \right).$$

Where μ_i is sampled independently from $\{\pm\mu_1, \pm\mu_2, \pm\mu_3, \pm\mu_4\}$ and ρ represents the signal consistency for reproducible hypotheses. We consider $\boldsymbol{\mu} = \{\mu_1, \mu_2, \mu_3, \mu_4\} = \{2, 2.5, 3, 3.5\}$ and all $\rho \in \{0, 0.5, 0.9\}$ with $\pi_1 \in \{0.10, 0.30\}$ and $\pi_{00} = 1 - \pi_1 - 0.05$.

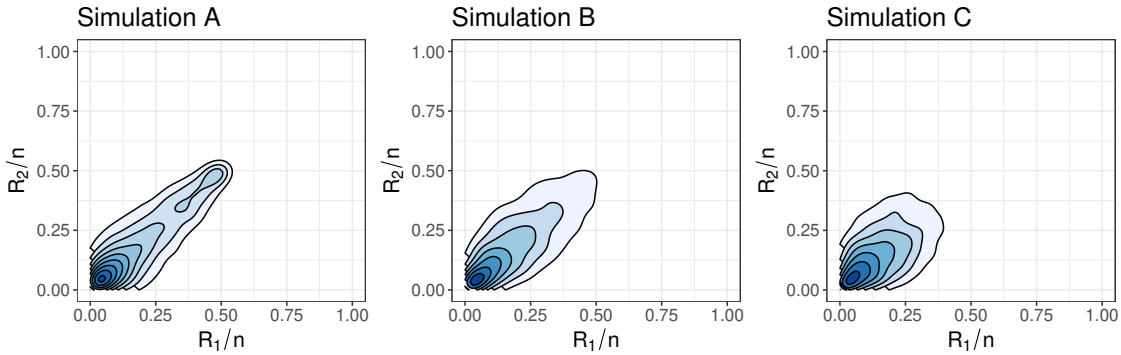


Figure 2.5: Bivariate density of reproducible ranks from one iteration of each simulation. For simulation A, one iteration with $a = 0.5$, $b = 2.5$, and $\sigma_0 = 0.01$ for $\pi_1 = 0.3$ is considered. For simulation B, one iteration with $\mu = 1.5$, $\sigma^2 = 0.5$, and $\rho = 0.8$ for $\pi_1 = 0.3$ is considered. For simulation C, one iteration with $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0.5$ for $\pi_1 = 0.3$ is considered.

2.6.2 Results

Simulation A

Figure 2.6 contains the average FDP and power for 100 iterations of the proposed simulation A. Figure 2.6 supports the theoretical properties examined in Section 2.4, as the average FDP for all implementations of $M_{\lambda,i}$ with a fixed λ and β closely approximates the associated nominal level. Additionally, the novel FDP estimation procedure shows an advantage relative to the original MaRR procedure and IDR, particularly at a nominal level of $\alpha = 0.05$, as the MaRR and IDR procedures tend to display elevated false discovery proportions while the proposed $M_{\lambda,i}$ procedure does not. Figure 2.6 also demonstrates that the proposed procedure with λ and β selected in the

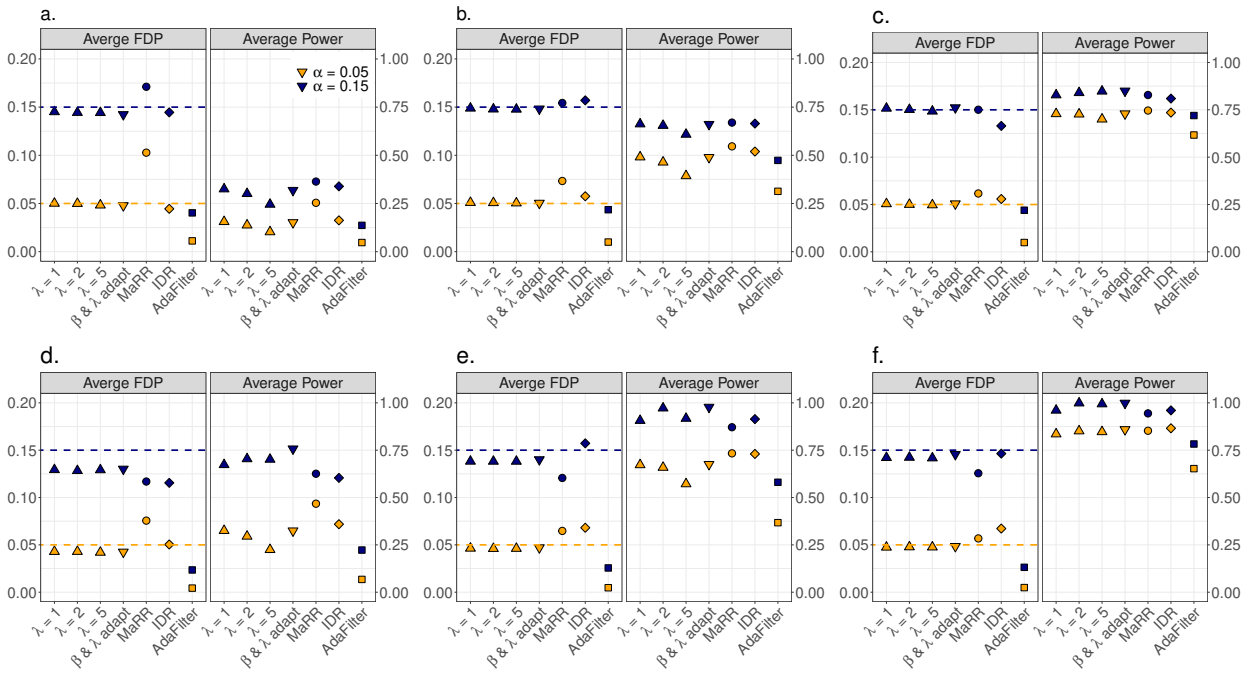


Figure 2.6: Average FDP and power values for 100 iterations of each setting in simulation A at nominal FDR levels $\alpha \in \{0.05, 0.15\}$. The settings are laid out as follows:

a. $a = 0.5$, $b = 2.5$, $\sigma_0 = 0.01$ and $\pi_1 = 0.1$
b. $a = 0.5$, $b = 3.5$, $\sigma_0 = 0.01$, and $\pi_1 = 0.1$
c. $a = 0.5$, $b = 5$, $\sigma_0 = 0.01$, and $\pi_1 = 0.1$
d. $a = 0.5$, $b = 2.5$, $\sigma_0 = 0.01$, and $\pi_1 = 0.3$
e. $a = 0.5$, $b = 3.5$, $\sigma_0 = 0.01$, and $\pi_1 = 0.3$
f. $a = 0.5$, $b = 5$, $\sigma_0 = 0.01$, and $\pi_1 = 0.3$.

Additionally, results for the nominal FDR level $\alpha = 0.05$ are denoted in orange and results $\alpha = 0.15$ in navy. The proposed method with β and λ as described in Section 2.5 selected is represented by $\triangledown$, the proposed method with fixed β and λ by $\triangle$, MaRR by $\bigcirc$, IDR by $\diamondsuit$, and AdaFilter by $\square$.

manner laid out in Section 2.5 demonstrates both FDR control and higher average power compared to the other methods considered. For each combination of a, b, and σ_0, the $M_{\lambda,i}$ procedure with λ and β adaptive selected has an average power that either is the maximum of all methods considered or is within 0.01 of the maximum power. It is interesting to examine the fixed λ values that also perform well under each case. Notice, for more sparse data and smaller levels of α, a fixed $\lambda = 1$ tends to perform approximately as well as the adaptive λ method, signaling that the adaptive procedure is often selecting $\lambda = 1$. However, for larger values of π_1 or α, larger fixed λ values tend to have higher average power, indicating that larger λ values were selected under those cases. Finally, AdaFilter is overly conservative in FDP estimation, leading to sacrificed power.

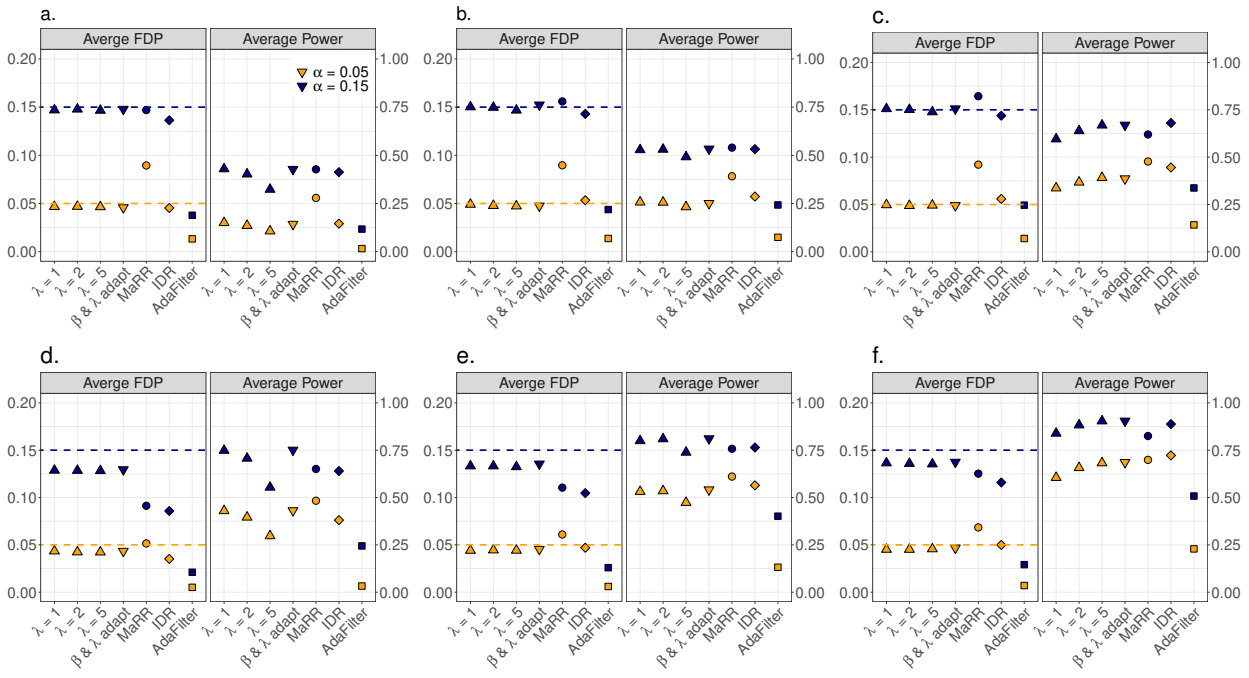


Figure 2.7: Average FDP and power values for 100 iterations of each setting in simulation B at nominal FDR levels $\alpha \in \{0.05, 0.15\}$. The settings are laid out as follows:

a. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.5$ and $\pi_1 = 0.1$
b. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.8$ and $\pi_1 = 0.1$
c. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.95$ and $\pi_1 = 0.1$
d. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.5$ and $\pi_1 = 0.3$
e. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.8$ and $\pi_1 = 0.3$
f. $\mu = 1.5$, $\sigma^2 = 0.5$, $\rho = 0.95$ and $\pi_1 = 0.3$.

Additionally, results for the nominal FDR level $\alpha = 0.05$ are denoted in orange and results $\alpha = 0.15$ in navy. The proposed method with β and λ as described in Section 2.5 selected is represented by $\triangledown$, the proposed method with fixed β and λ by $\triangle$, MaRR by $\bigcirc$, IDR by $\diamondsuit$, and AdaFilter by $\square$.

Simulation B

Figure 2.7 contains the average FDP and power for 100 iterations of the proposed simulation B. Again, Figure 2.7 supports the theoretical findings from Section 2.4 and demonstrates the distinct advantage of the proposed FDR controlling procedure. Additionally, for nearly all cases, the proposed method has at least as high average power as the other considered methods. Notice that this design is the exact model considered in Li et al. (2011) and thus, IDR also enjoys high power levels compared to the other methods. Interestingly, two advantages of the proposed method can be teased out by closely examining Figure 2.7. First, we see the average FDP for the proposed method is closer to its true nominal level than all other methods, demonstrating the precision of the novel FDP estimation method, and second, when ρ is large, we see that the flexibility afforded

to the user by the inclusion of λ yields higher average power. Note, in Figure 2.7 when $\rho = 0.95$, we see that fixed $\lambda = 5$ demonstrates the largest power among the fixed λ options and when $\rho = 0.5$ fixed $\lambda = 1$ demonstrates the largest power. Intuitively, this follows immediately from the notion that larger values of ρ represent across-experiment consistency, and the purpose of increasing λ to capture consistency signal.

Simulation C

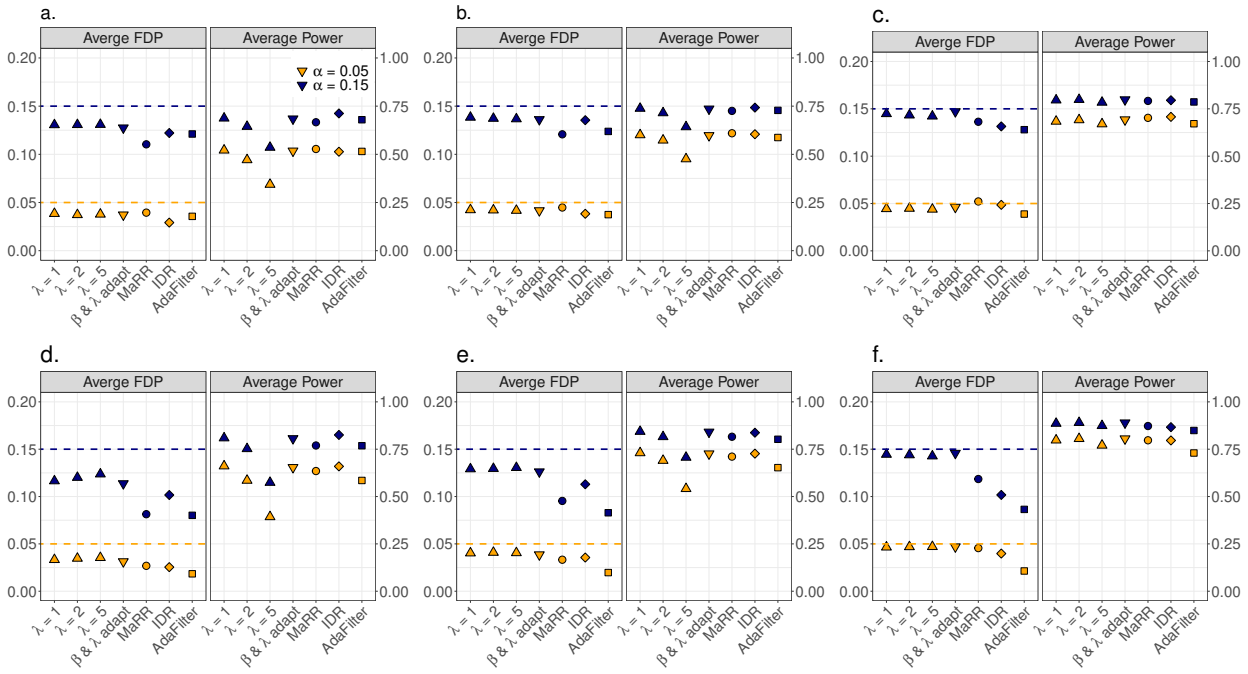


Figure 2.8: Average FDP and power values for 100 iterations of each setting in simulation C at nominal FDR levels $\alpha \in \{0.05, 0.15\}$. The settings are laid out as follows:

a. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0$ and $\pi_1 = 0.1$
b. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0.5$ and $\pi_1 = 0.1$
c. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0.9$ and $\pi_1 = 0.1$
d. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0$ and $\pi_1 = 0.3$
e. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0.5$ and $\pi_1 = 0.3$
f. $\boldsymbol{\mu} = \{2, 2.5, 3, 3.5\}$, $\rho = 0.9$ and $\pi_1 = 0.3$

Additionally, results for the nominal FDR level $\alpha = 0.05$ are denoted in orange and results $\alpha = 0.15$ in navy. The proposed method with β and λ as described in Section 2.5 selected is represented by $\triangledown$, the proposed method with fixed β and λ by $\triangle$, MaRR by $\bigcirc$, IDR by $\diamondsuit$, and AdaFilter by $\square$.

Figure 2.8 contains the average FDP and power for 100 iterations of the proposed simulation C. Notice that the proposed method with λ selected adaptive demonstrates near the largest average

power across all considered values for ρ and α. Under this setting, AdaFilter is far less conservative than the other settings and thus its associated average power values are much closer to the best figures, particularly in the more sparse cases ($\pi_1 = 0.1$). It is notable that under an adaptation of the setting considered in Wang et al. (2022) our proposed method still outperforms the AdaFilter. Again, we note that as ρ increases, we observe that larger values of λ tend to be preferable. This can be observed by examining the power figures for fixed $\lambda = 2$ relative to fixed $\lambda = 1$ across the different levels of ρ.

Across all three settings, the proposed method shows preferable results in both FDR control and increased power relative to the other three methods considered. The novel FDP estimating procedure and resulting thresholding rule yields more exact control of FDP than the other methods, as the proposed method controls FDR at nearly exactly the nominal level. AdaFilter and IDR are both overly conservative and MaRR tends to be anti-conservative for smaller levels of α and overly conservative for larger levels of α. Additionally, the advantage of choosing λ affords higher power on average. Note that, depending on the level of signal consistency, different fixed λ values tend to have higher power. This supports the intuition that λ allows the researcher to have more flexibility to discover reproduced hypotheses.

2.7 Real data application

Genome-wide association studies (GWAS) and transcriptome-wide association studies (TWAS) have been commonly considered in the reproducibility framework (Li et al., 2011; Philtron et al., 2018; Wang et al., 2022; Zhao et al., 2020). TWAS integrates data from GWAS with expression data from specific tissues or cell types that are relevant to a disease or genetic trait of interest. In TWAS, a z-score can be imputed to test the association between a specific genetic feature and a trait. If an association is identified between some genetic feature and a trait, that can be treated as a potential causal link (Wu et al., 2022; Zhang et al., 2020). Reproducibility in TWAS is crucial because multiple independent studies agreeing on the association of a genetic feature and trait provide further credence to the notion of a causal link between that feature and trait.

2.7.1 Real data

We examine the performance of the proposed method in application using TWAS z-scores from two GWAS data sources from the Genotype-Tissue Expression (GTEx) project. The GTEx project, funded by the NIH, was launched in 2010 to create a public catalog of genetic expression across different human tissue types. We will consider data from the UK Biobank (UKB) and the GIANT consortium with genetic expression estimated from skeletal muscle tissue. The data were processed and TWAS z-scores were made available by Zhao et al. (2020). For these studies, the trait of interest was standing height. Associations for a total of 32,362 genes are measured. To be included in our analysis, a gene must have non-zero TWAS z-scores in both experiments. With that inclusion criteria, we examine the reproducibility of 12,517 common genes. The sample size for the UKB data (337k) is larger than that of the GIANT (253K), which results in z-scores smaller in magnitude for the GIANT dataset. Figure 2.9 examines the joint distribution of the 12,517 z-scores we consider and their associated rankings. We apply the $M_{\lambda,i}$ with $\widehat{\pi}_1$, λ, and β selected in the same

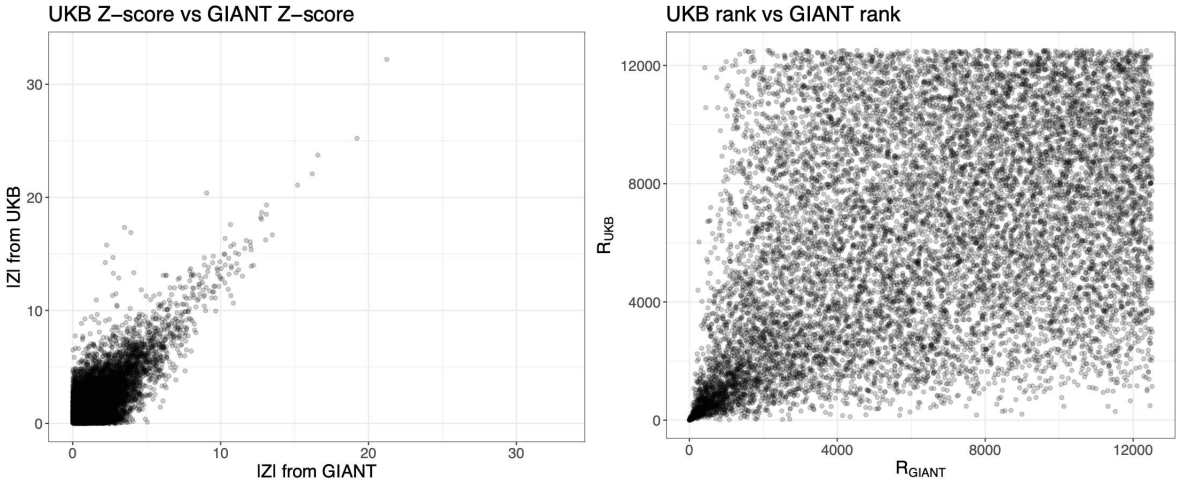


Figure 2.9: Joint distribution of TWAS z-scores and ranks.

manner described in Section 2.6 and discover reproducible genes. We additionally consider MaRR (Philtron et al., 2018), AdaFilter (Wang et al., 2022), and the copula mixture method from Li et al.

(2011) (IDR) with the same implementations described in Section 2.6 and report the percentage of genes deemed reproducible in Section 2.7.2.

2.7.2 Results

The proposed implementations of the $M_{\lambda,i}$ algorithm yield $\widehat{\pi}_1 = 0.1592$, and select $\lambda = 1$ with $\tilde{\beta} = 1 - \delta = 0.9$. The percentage of the 12,517 genes declared reproducible by each method considered is presented in Table 2.1 at nominal levels of $\alpha \in \{0.05, 0.15\}$. **Bold** figures represent the largest percentage for each level of α. Notice, the proposed method using the suggested practical

Table 2.1: Percentage of total features declared reproducible for nominal FDR levels $\alpha \in \{0.05, 0.15\}$.

	$\alpha = 0.05$	$\alpha = 0.15$
$M_{\lambda,i}$; β & λ selected	**15.5%**	**21.5%**
IDR	12.9%	18.3%
AdaFilter	13.9%	18.9%
MaRR	13.8%	18.2%

implementations yields a larger percentage of genes deemed to be reproducible than the existing methods considered. This result is useful, as it results in a larger list of genes with the potential for a causal link to standing height. Compared to the next highest performing method at a nominal false discovery rate of $\alpha = 0.15$ (AdaFilter), the proposed method discovers 389 more reproducible genes. Figure 2.10 shows where the additional reproducible hypotheses are located in both the z-score and rank-space. A smaller sample size for the GIANT study results in z-scores from hypotheses with real signal being smaller in absolute value than their counterparts in the UKB study. Notice then, since the proposed method is a function of ranks, it is less susceptible to the differences in sample sizes across two studies, and thus deems more hypotheses with z-scores smaller in magnitude in the GIANT study to be reproducible.

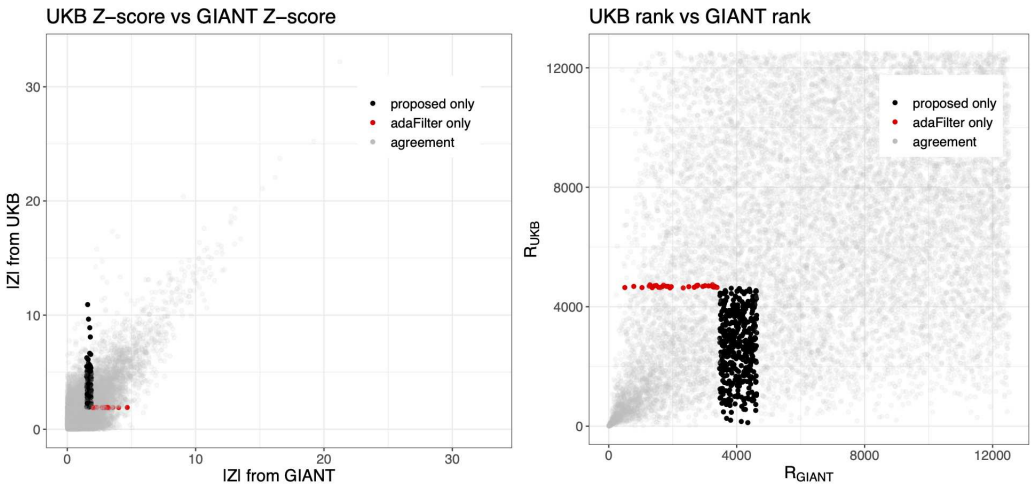


Figure 2.10: Differences in reproducible regions between $M_{\lambda,i}$ procedure and AdaFilter in z-score and rank-space. Points marked in black were deemed reproducible by the proposed method and not AdaFilter and points marked in red were deemed reproducible by AdaFilter but not the proposed method. There was agreement in decision for all other hypotheses.

2.7.3 Gene enrichment analysis

To identify which biological processes were comprised of genes that tend to be more reproducible than the make-up of other processes, we performed gene enrichment analysis with "Gene Ontology: Biological Process" terms as pathways (Ashburner et al., 2000; Aleksander et al., 2023). We perform the gene set enrichment analysis (GSEA) originally proposed by Subramanian et al. (2005), that calculates an enrichment score based on the ranking of a test statistic for a set of genes that is similar to weighted Kolmogorov-Smirnov statistics (Hollander et al., 2013). We used the Gene Ontology: Biological Process terms as organized by the database org.Hs.eg.db (Carlson, 2023) as the gene sets or pathways and the `gseGO` function from the `R` package `clusterProfiler` (Wu et al., 2021a). We consider all pathways that contain at least 10 and at most 800 genes. In this application, we use the proposed $M_{\lambda,i}$ statistic with $\lambda = 1$ as the test statistic of interest so that we can assess which processes tended to have more reproducible genes relative to other processes. 469 biology processes yielded enrichment scores with unadjusted p-values lower than 0.05 and 20 less than 0.001. Table 2.2 contains those with a p-value below 0.001. The table reveals several regulatory and developmental processes.

Table 2.2: GO:BP terms with a p-value under 0.001.

GO:BP terms	
chromatin organization	regulation of cellular senescence
macroautophagy	digestive system development
cellular senescence	short-chain fatty acid metabolic process
skeletal muscle tissue development	receptor recycling
chromosome organization	digestive tract development
chromatin remodeling	branching involved in blood vessel morphogenesis
skeletal muscle organ development	cleavage involved in rRNA processing
regulation of cell development	receptor metabolic process
sister chromatid cohesion	establishment of protein localization to organelle
miRNA metabolic process	cellular response to amino acid starvation

This signals that genes integral to these processes were highly reproducible. For example, Figure 2.11 contains the distribution of $M_{\lambda,i}$ statistics for genes in the processes with the six smallest GSEA p-values compared to the overall distribution of $M_{\lambda,i}$ statistics. Notice that the distributions for these processes are more densely populated at small values, so these genes were highly reproducible relative to the distribution of statistics from all genes. Notably, genes involved in the processes “regulation of cellular senescence" (GO:2000772) and “short-chain fatty acid metabolic process" (GO:0046459) were far more reproducible than genes overall.

2.8 Discussion and conclusion

Inspired by Li et al. (2011) and Philtron et al. (2018), we formalize a notion of reproducibility in a multiple-testing setting. Then, to assess that notion, in Section 2.2, we introduce a rank-based statistic, $M_{\lambda,i}$, and devise a novel procedure to discover hypotheses reproducible hypotheses at the nominal FDR level. We discuss the implementation of the procedure in practice in Section 2.5. Then in Section 2.4, we leverage the properties of the negatively associated sequences to show the procedure admits asymptotic FDR control. To corroborate those theoretical guarantees in the finite sample case, we propose and conduct three simulation studies in Section 2.6 and show preferable performance to existing literature. Finally, we apply the method to two TWAS datasets to discover potential candidates for causal links to the trait of standing height in Section 2.7.

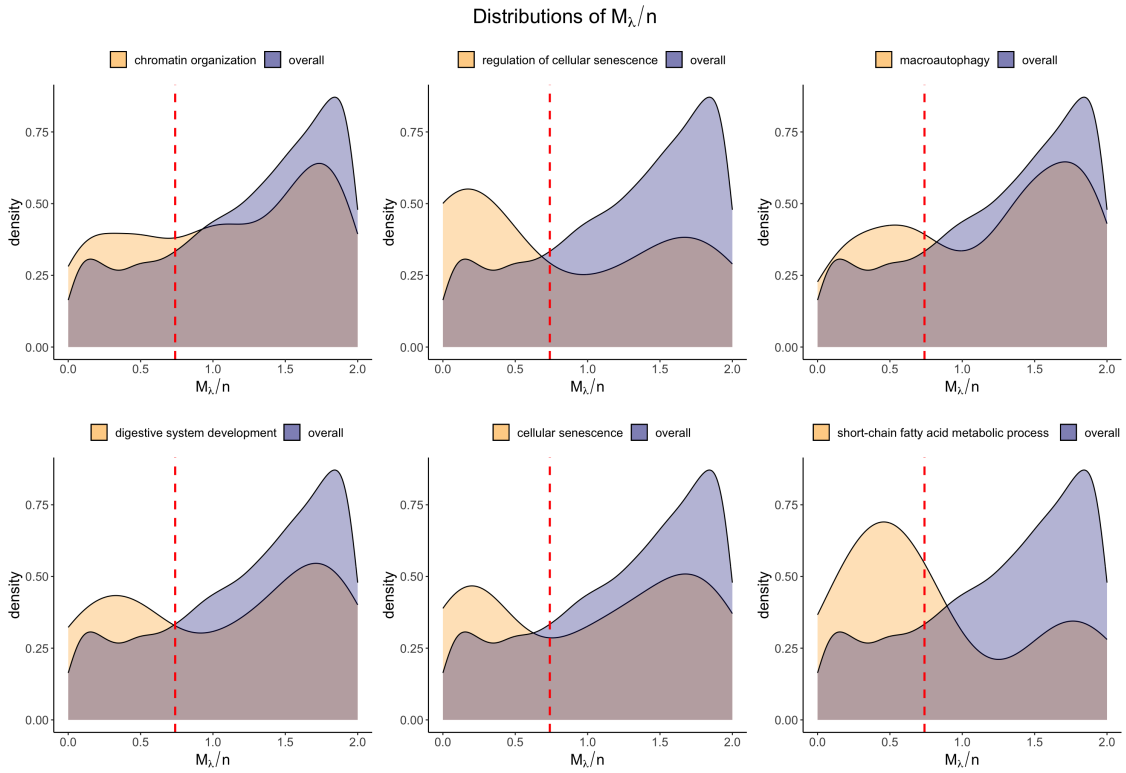


Figure 2.11: Distributions of $M_{\lambda,i}/n$ for the genes in the top six biological processes compared to the distribution of $M_{\lambda,i}/n$ for all genes.

The method represents a meaningful contribution to the reproducibility problem, as it allows practitioners to assess the reproducibility of a large number of hypotheses examined in two studies using only a summary statistic for each hypothesis. Since the procedure is rank-based and does not rely on any parametric assumptions, it can be applied in a wide variety of settings. Compared to existing rank-based methods the $M_{\lambda,i}$ holds a few major advantages. First, under Condition 2.3.1, we show asymptotic false discovery rate control, providing theoretical justification for the method. In practice, the method yields a less conservative control of FDR than existing methods which manifests increased power. Additionally, through the weighting parameter λ, the proposed method has an adaptive rejection region that can be tailored for the type of data observed. This flexibility allows the method to discover more reproducible hypotheses, particularly those with weak but highly consistent signal.

2.8.1 Extension to enforce sign concordance.

In the proposal of the method, we consider reproducibility to be the consistency of the *magnitude* of the signal for a hypothesis. In the case of two-sided tests, this raises an issue as the notion does not consider the *sign* of the signal. That is, we rank the hypotheses in each experiment based solely on their "notability," where hypotheses that are notable in positive and negative directions are not differently ranked. This can lead to hypotheses with a strong positive signal in one study and a strong negative signal in the another study being deemed reproducible using the method. Often it makes sense to consider only hypotheses with consistent signal that has concordant signs to be reproducible and not hypotheses with discordant signs. In those cases, one can adapt the following transformation of summary statistics $T_{1,i}$ and $T_{2,i}$ (assuming $T_{1,i}$ large in magnitude represents notability) to ensure rankings are consistent only when the sign is consistent.

$$T'_{1,i} = |T_{1,i}|,\ T'_{2,i} = \text{sgn}(T_{1,i})T_{2,i}.$$

Then, the $M_{\lambda,i}$ method can be applied using $T'_{1,i}$ and $T'_{2,i}$, ranked from most positive to most negative, to discover reproducible hypotheses that have concordant signs. Incorporating this transformation naturally ensures that hypothesis i will only be highly and consistently ranked in both experiments if 1) the magnitude for the signal is large and similarly aligned in both studies and 2) the sign is consistent.

Chapter 3

Assessing reproducibility of high-throughput studies with group structure

3.1 Introduction

In genomics, high-throughput technologies have made the evaluation of associations between a biological trait of interest and thousands of genes possible. While these technologies have broadened the scope of modern biology, it is known that results for any given gene show large variability across replicate studies due to differences in platform, pipeline, or other experimental design factors.

Consequently, assessing the reproducibility of results pertaining to all genes simultaneously has emerged as an area of research, as it assures the validity of detected links between genetic features and traits of interest. Interestingly, genetic expression data follow a multi-level group structure. Abundances from RNA sequencing data can be measured at the level of the transcript (Pertea et al., 2016). The transcripts that are measured to genes, and genes can be grouped into gene pathways or biological processes (Ashburner et al., 2000; Aleksander et al., 2023) based on shared functions. Results at all of these levels are interesting, as they help researchers further understand the mechanisms of the trait or disease of interest at all biological levels. These group structures are ignored, however, when assessing reproducibility. Usually, inference on the reproducibility of hypotheses is performed without using any group information. Group-level reproducibility is generally analyzed in one of two ways: 1) a post hoc examination of groups or 2) an analysis performed on aggregates of hypotheses within a group. As an example of the first, when examining gene-level reproducibility, it is common to perform gene enrichment analysis on the reproducible genes to identify important pathways and process (Lyu et al., 2023; Li et al., 2024). Here, the pathways act as groups and genes act as the hypotheses, with analysis at the group-level

happening only after hypothesis-level examination has occurred. The second case occurs when gene-level differential expression results are considered. The popular `DESeq2` (Love et al., 2014) and `edgeR` (Chen et al., 2025) pipelines for differential expression analysis require the user to aggregate transcript-level abundance data into gene-level expression data prior to analysis. In this example, the transcripts are the hypotheses and the genes are the groups. Our work aims to devise a method that builds this group structure into the procedures used to measure reproducibility and devise rules for discovering reproducible groups and hypotheses that control false discoveries. In the next section, we explore the existing literature in multiple testing with group structure problems and reproducibility problems.

3.1.1 Existing literature

Multiple testing with group structure

Over the years many methods have been developed for multiple testing of grouped hypotheses within an individual study study. The aim is generally to discover a set of non-null hypotheses that controls the false discovery rate (FDR) at a nominal level. From the frequentist perspective, it is common to examine and combine p-values from the same group to produce group-level statistics and employ multiple testing procedures to test both hypotheses and groups (Hu et al., 2010; Barber and Ramdas, 2016).

In the empirical Bayesian world, the multiple testing problem is answered by assuming null and non-null models and calculating the probability that a hypothesis is null given the observed data – called local false discovery rate (Efron et al., 2001). In cases with group structure, the usual strategy is to combine local FDR scores at the hypothesis and group level and devise testing procedures that control the posterior total false discovery rate based on functions of the local FDR scores (Cai and Sun, 2009; Liu et al., 2016; Sarkar and Zhao, 2022). Of particular interest to this work is Liu et al. (2016), which introduces a mixture model that includes group structure for calculating the local FDR scores both group-wise and within-group. The local FDR for each hypothesis can then be calculated as a composite of the two scores. Using these quantities, they

develop a two-level procedure for testing n hypotheses with control of hypothesis-wise posterior false discovery rate. The procedure uses hypothesis and group-level criteria based on local FDR rates, and a hypothesis is only rejected if it meets both criteria. This procedure leverages the known group structure to yield powerful results.

Reproducibility

In the high-throughput reproducibility (alternatively called replicability) problem, the overarching goal is to identify hypotheses with results that are consistent across m studies and control the false discovery rate below a nominal level. The empirical Bayesian framework is frequently used, with decision rules defined using local FDR score calculated under a specified design (Li et al., 2011; Heller and Yekutieli, 2014; Lyu et al., 2023; Li et al., 2024). A popular method introduced in Li et al. (2011) cleverly avoids making assumptions about the marginal distributions of the observed summary statistics by instead assuming the distributions the underlying signal comes from are a Gaussian mixture model and transformed to the observed summary statistics. The observed data can then be back-transformed to the Gaussian pseudo-data by the observed empirical distributions and local FDR scores can be calculated using the pseudo-data Gaussian model. We adopt this approach later for our estimation procedure in Section 3.4.2.

Alternatively, from the frequentist framework methods either examine reproducibility through assumptions about the available results Hung and Fithian (2020); Wang et al. (2022) or examine the consistency of the alignment of results for hypotheses by functions of rankings Philtron et al. (2018); Ghosh et al. (2021). As an example of the first, Wang et al. (2022) assumes the available results are p-values that are uniform in a certain number of studies for irreproducible hypotheses. Then, they provide filtering and selection criteria to discover reproducible hypotheses. Among rank-based methods, Philtron et al. (2018) consider examining the maximum ranking across two studies as the test statistic and discover hypotheses that are highly ranked in both studies as reproducible.

3.1.2 Our approach and contributions

Our work focuses on extending the empirical Bayesian approach to group structured hypothesis testing from Liu et al. (2020) to the reproducibility problem, where the agreement of signal across two studies is the primary interest. In presenting the two-level procedure for detecting reproducible hypotheses and designing a group-level procedure, we fill a hole in the existing reproducibility problem, where group structure is generally ignored. Currently, for high-throughput genomic settings, hypothesis-level inference is performed ignoring group structures (such as gene pathways or processes), and group-level inference is generally conducted either post hoc or by aggregating before analysis. Through our group structured framework, we can outperform naive methods at the hypothesis-level by using the group information. The framework also allows for group-level inference without needing to initially aggregate data. Additionally, we propose criterion for selecting the tuning parameter η in the hypotheses-level procedure that optimizes results in terms of posterior false nondiscovery rate (FNR). This criterion applies to both the two study reproducibility context and in the general multiple testing problem. To estimate the parameters in the model, we assume summary statistics come from the Gaussian copula mixture model first proposed for the reproducibility problem in Li et al. (2011) and adapt the EM algorithm accordingly, making the method fully data-driven.

3.1.3 Organization and notation

This chapter takes the following form. In Section 3.2, we introduce the group structured setting from Liu et al. (2016) that we adopt for the reproducibility problem and devise procedures for assessing hypothesis and group-level reproducibility under the general setting. Section 3.3 describes the Bernoulli Significant Model that can be used to calculate the local FDR quantities required for the proposed procedures. The estimation procedure used to estimate parameters that are unknown in practice is detailed in Section 3.4.2. Section 3.5 proposes the criterion for selecting an optimal value for a tuning parameter in the method. Then, we assess the performance of the proposed hypothesis and group-level procedures through simulations in Section 3.6. Finally, we provide a

summary of our contributions and discuss future research work regarding this chapter in Section 3.7.

Notations: For clarity, we list some notations used throughout the chapter. Let n represent the total number of hypotheses commonly assessed in 2 studies. These hypotheses can be separated into G non-overlapping groups. n_g represents the size of group g for $g \in \{1, \cdots, G\}$. $\mathbb{H}_{gj}$ is used for the j^{th} hypothesis in group g. $\boldsymbol{T}_{gj} = (T_{gj,1}, T_{gj,2})$ denotes the summary statistics for the j^{th} hypothesis in group g. Additionally, we denote $\boldsymbol{T} = \{\boldsymbol{T}_{11}, \cdots, \boldsymbol{T}_{Gn_G}\}$. The reproducibility status for $\mathbb{H}_{gj}$ is denoted θ_{gj} with $\theta_{gj} = 0$ meaning the hypothesis is irreproducible and $\theta_{gj} = 1$ meaning reproducible.

3.2 Methods for reproducibility with group structure

Suppose there are n hypotheses common across 2 replicate studies that can be divided into G non-overlapping groups of hypotheses. Group g is of size n_g. For each hypothesis, we observe summary statistics in each study that show evidence against a common null hypothesis. Let $\boldsymbol{T}_{gj} = (T_{gj,1}, T_{gj,2})$ represent the summary statistics for the j^{th} hypothesis in group g (denoted $\mathbb{H}_{gj}$), where $j \in \{1, \cdots, n_g\}$ and $g \in \{1, \cdots, G\}$ with $T_{gj,1}$ and $T_{gj,2}$ corresponding to the statistic from study 1 and 2, respectively. We will follow the general group hypothesis testing structure from Liu et al. (2016) and adapt it for the reproducibility setting. We denote θ_{gj} as the reproducibility "status" for $\mathbb{H}_{gj}$. That is, if $\mathbb{H}_{gj}$ is reproducible then $\theta_{gj} = 1$ and if $\mathbb{H}_{gj}$ is irreproducible then $\theta_{gj} = 0$. Hypothesis-level reproducibility is defined by the distribution of the summary statistics. So, if $\theta_{gj} = k$, then $\boldsymbol{T}_{gj} \sim f_k$ for $k \in \{0, 1\}$ and some densities f_0 and f_1. We further assume the reproducibility status for a hypothesis can be decomposed in the following manner.

$$\theta_{gj} = \theta_g * \theta_{j|g} \tag{3.1}$$

where θ_g represents the **group-level** reproducibility status for group g and $\theta_{j|g}$ is the **hypothesis-within-group-level** reproducibility status for hypothesis j in group g conditional on group g being

reproducible. At the group-level, a group of hypotheses is defined to be reproducible ($\theta_g = 1$) by the presence of *at least one reproducible hypothesis* – or there existing some $j \in \{1, \cdots, n_g\}$ such that $\theta_{gj} = 1$ – and irreproducible ($\theta_g = 0$) by the absence of a reproducible hypothesis – or $\theta_{gj} = 0$ for all $j \in \{1, \cdots, n_g\}$. Thus, if group g has any reproducible member hypotheses, then it is reproducible. It can be seen in (3.1) that $\theta_{gj} = 0$ if $\theta_g = 0$ or $\theta_g = 1$ and $\theta_{j|g} = 0$, while $\theta_{gj} = 1$ if and only if $\theta_g = 1$ and $\theta_{j|g} = 1$. Thus, it is clear that $\theta_{j|g}$ dictates whether the hypothesis j, in particular, is reproducible given that it is a member of a reproducible group. The decomposition from (3.1) gives an explicit representation of group structuring often encountered in reproducibility problems across a wide range of applications, including in high-throughput genomics where biologists are often interested in examining genes and particular groups of genes.

Motivated by the structure of the reproducibility status in (3.1), we propose using a reproducibility rejection rules for discovering reproducible groups and hypotheses that follow a similar structure. First, we devise a rejection rule that characterizes reproducibility at the **group-level**, denoted $\delta_g^*(\boldsymbol{T})$. So $\delta_g^*(\boldsymbol{T})$ is a decision rule that tests against the null hypothesis $\theta_g = 0$ and discovers reproducible groups with control of the group-level total posterior false discovery rate (PFDR) at a nominal level of α. Group-level PFDR is defined by the following.

$$g\text{PFDR}(\delta_g^*; \boldsymbol{T}) = \mathbb{E}\left[\frac{\sum_{g=1}^{G} (1 - \theta_g)\delta_g^*(\boldsymbol{T})}{\sum_{g=1}^{G} \delta_g^*(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right]. \tag{3.2}$$

Second, we devise a rejection rule that characterizes reproducibility at the **hypothesis-level**, denoted $\delta_{gj}(\boldsymbol{T})$, that tests against the null hypothesis $\theta_{gj} = 0$ and discovers reproducible hypotheses. Critically, the proposed rejection procedure follows the same structure in (3.1) for θ_{gj}. That is, $\delta_{gj}(\boldsymbol{T})$ has the representation

$$\delta_{gj}(\boldsymbol{T}) = \delta_g(\boldsymbol{T})\delta_{j|g}(\boldsymbol{T}) \tag{3.3}$$

where δ_g is the group-level rejection rule – similar to θ_g from (3.1) and $\delta_{j|g}$ is the rejection rule for hypothesis j in group g given g is reproducible – similar to $\theta_{j|g}$ from (3.1). The proposed rejection rule discovers reproducible hypotheses with control of hypothesis-level PFDR at a given level of

α, with hypothesis-level PFDR defined by

$$h\text{PFDR}(\delta_{gj}; \boldsymbol{T}) = \mathbb{E}\left[\frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} (1 - \theta_{gj}) \delta_{gj}(\boldsymbol{T})}{\sum_{g=1}^{G} \sum_{j=1}^{n_g} \delta_{gj}(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right]. \tag{3.4}$$

Now to build these methods, we introduce the concept of local FDR, both at the group and hypothesis-level. We then leverage the representations of reproducibility status from (3.1) and the rejection region from (3.3) to expand the $h\text{PFDR}$ quantity from (3.4) and define both desired PFDR terms in terms of these local FDR quantities. Finally, we construct δ_g^* and δ_{gj} that assess group and hypothesis-level reproducibility that control of the desired PFDR quantity at a specified nominal level of α, under the ideal oracle setting, where the local FDR quantities are known.

3.2.1 Local FDR quantities

First, remember that the overall hypothesis-level status quantity θ_{gj} can be determined entirely by θ_g and $\theta_{j|g}$. Also assume that each of those two status quantities $(\theta_g, \theta_{j|g})$ are binary random variables for each group or hypothesis. Then, the group-level and the hypothesis-given-group-level for posterior local FDR quantities for each group $g \in \{1, \cdots, G\}$ or hypothesis $j \in \{1, \cdots, n_g\}$ are defined as

$$\text{fdr}_g(\boldsymbol{T}) = \mathbb{P}\left(\theta_g = 0 \mid \boldsymbol{T}\right) \quad \text{and} \quad \text{fdr}_{j|g}(\boldsymbol{T}) = \mathbb{P}\left(\theta_{j|g} = 0 \mid \theta_g = 1, \boldsymbol{T}\right). \tag{3.5}$$

Notice, $\text{fdr}_g(\boldsymbol{T})$ is the probability that group g is irreproducible ($\theta_g = 0$) given the summary statistics and $\text{fdr}_{j|g}(\boldsymbol{T})$ is the probability that hypothesis j in group g is irreproducible ($\theta_{j|g} = 0$) given that group g is reproducible and the summary statistics. We can now expand the representations $g\text{PFDR}(\boldsymbol{T})$ and $h\text{PFDR}(\boldsymbol{T})$ from (3.2) and (3.4) to be in terms of the local FDR quantities from (3.5). Let $\delta^*(\boldsymbol{T})$ be any group-level rejection rule and assume $\sum_{g=1}^{G} \delta_g^*(\boldsymbol{T}) > 0$. Notice then, (3.2) has the form

$$\begin{aligned} g\text{PFDR}(\delta_g^*; \boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G} (1 - \theta_g) \delta_g^*(\boldsymbol{T})}{\sum_{g=1}^{G} \delta_g^*(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right] \\ &= \frac{\sum_g \delta_g^*(\boldsymbol{T}) \text{fdr}_g(\boldsymbol{T})}{\sum_g \delta_g^*(\boldsymbol{T})}. \end{aligned} \tag{3.6}$$

If $\sum_{g=1}^{G} \delta_g^*(\boldsymbol{T}) = 0$, then $\delta_g^*(\boldsymbol{T}) = 0$ for all $g \in \{1, \cdots, G\}$ and thus $g\mathrm{PFDR}(\boldsymbol{T}) = 0$. Now, let δ_{gj} be any hypothesis-level rejection rule and assume $\sum_{g=1}^{G} \sum_{j=1}^{n} \delta_{gj}(\boldsymbol{T}) > 0$. We can represent (3.4) by the equivalent form

$$\begin{aligned} h\mathrm{PFDR}(\delta_{gj}; \boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} (1 - \theta_{gj}) \delta_{gj}(\boldsymbol{T})}{\sum_{g=1}^{G} \sum_{j=1}^{n_g} \delta_{gj}(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right] \\ &= 1 - \frac{\sum_g \left(1 - \mathrm{fdr}_g(\boldsymbol{T})\right) \left[\sum_j \delta_{gj}(\boldsymbol{T}) \left(1 - \mathrm{fdr}_{j|g}(\boldsymbol{T})\right)\right]}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})}. \end{aligned} \tag{3.7}$$

Again, note that $\sum_{g=1}^{G} \sum_{j=1}^{n_g} \delta_{gj}(\boldsymbol{T}) = 0$ implies $\delta_{gj}(\boldsymbol{T}) = 0$ for all g and j and thus $h\mathrm{PFDR}(\boldsymbol{T}) = 0$. With these representations in mind, we propose the following methods, in their oracle forms, to discover reproducible groups and hypotheses at the nominal level of posterior FDR of α.

3.2.2 Oracle proposed procedures

The group and hypothesis-level testing procedures are introduced under the oracle setting before discussing the calculation and estimation of the local FDR scores under an assumed model. So, assume we are in the oracle setting where the local FDR scores $\mathrm{fdr}_g(\boldsymbol{T})$ and $\mathrm{fdr}_{j|g}(\boldsymbol{T})$ are known for all $g \in \{1, \cdots, G\}$ and $j \in \{1, \cdots, n_g\}$. Additionally, throughout the remainder of the paper, we drop the $(\boldsymbol{T})$ from the quantities $\mathrm{fdr}_g(\boldsymbol{T})$ and $\mathrm{fdr}_{j|g}(\boldsymbol{T})$ for notational ease.

Group-level algorithm

With knowledge of the local group-level FDR quantities one can discover reproducible groups at a nominal level of $g\mathrm{PFDR}(\boldsymbol{T})$ of α by Algorithm 2.

Algorithm 2 Group-level reproducibility procedure.

1: Order groups by fdr_g scores from smallest to largest and denote $\mathrm{fdr}_{(1)} \leq \cdots \leq \mathrm{fdr}_{(g)}$ as the order statistic of the local FDR quantities for groups $g \in \{1, \cdots, G\}$.
2: Calculate critical value for a nominal PFDR level of α, l_α by

$$l_\alpha = \max\left\{k \in \{1, \cdots, G\} : \frac{\sum_{g=1}^{k} \mathrm{fdr}_{(g)}}{k} \leq \alpha\right\}. \tag{3.8}$$

3: Deem any group, g, with $\mathrm{fdr}_{(1)} \leq \mathrm{fdr}_g \leq \mathrm{fdr}_{(l_\alpha)}$ to be reproducible.

Theorem 3.2.1 states that the proposed group-level procedure controls $g\mathrm{PFDR}(\boldsymbol{T})$ at the nominal level under the oracle setting.

Theorem 3.2.1. *Assume the quantities* $(\mathrm{fdr}_1, \cdots, \mathrm{fdr}_g)$ *are known. Denote the proposed group-level rejection rule for group* g *by* $\delta_g^*(\boldsymbol{T})$*. For any* $\alpha \in (0, 1)$*, the* $g\mathrm{PFDR}(\boldsymbol{T})$ *using* $\delta_g^*(\boldsymbol{T})$ *is controlled at the level* α *That is,*

$$g\mathrm{PFDR}(\delta_g^*; \boldsymbol{T}) = \mathbb{E}\left[\frac{\sum_{g=1}^{G}(1-\theta_g)\delta_g^*(\boldsymbol{T})}{\sum_{g=1}^{G}\delta_g^*(\boldsymbol{T}) \vee 1}\middle| \boldsymbol{T}\right] \leq \alpha.$$

Proof. The proof is immediate given the form of $g\mathrm{PFDR}$ from (3.6).

Hypothesis-level algorithm

At the hypothesis level, we use a two-fold procedure for discovering reproducible hypotheses at the FDR control level α in a similar manner to the oracle form of the two-level-testing algorithm (TLTA) method devised in Liu et al. (2016). The general shape of the procedure is as follows. It begins by discovering potential reproducible "candidate" hypotheses using hypothesis-level local FDR scores using a fixed $\eta \in (0, 1)$ (see Section 3.5.1 for a discussion of the selection of an optimal η). Then, it selects reproducible groups based on the number of candidate hypotheses and the group-level local FDR quantities. Finally, the intersection the group and hypothesis-level rejection rules are the hypotheses found to be reproducible. Algorithm 3 describes the specifics of the two-fold procedure.

Algorithm 3 Two-fold hypothesis-level reproducibility procedure.

1: Order hypotheses by $\mathrm{fdr}_{j|g}$ scores from smallest to largest within each group. Denote $\mathrm{fdr}_{(1)|g} \leq \cdots \leq \mathrm{fdr}_{(n_g)|g}$ as the order statistic of the local FDR quantities for hypotheses $j \in \{1, \cdots, n_g\}$ in group $g \in \{1, \cdots, G\}$.

2: Given fixed $\eta \in (0, 1)$, calculate the hypothesis-given-group-level critical value $R_g(\eta)$ for each group g by

$$R_g(\eta) = \max\left\{k : \sum_{j=1}^{k} \frac{\mathrm{fdr}_{(j)|g}}{k} \leq \eta\right\}. \tag{3.9}$$

3: Deem any hypothesis j in group g to be a reproducible candidate if $\mathrm{fdr}_{(1)|g} \leq \mathrm{fdr}_{j|g} \leq \mathrm{fdr}_{(R_g(\eta))|g}$. Denote the set of these candidate hypotheses for group g by $C_g(\eta)$.

4: Denote $\phi_g(\eta) = \sum_{j=1}^{R_g(\eta)} \mathrm{fdr}_{(j)|g}$ and $\mathrm{fdr}_g^*(\eta) = 1 - (1 - \phi_g(\eta))\,(1 - \mathrm{fdr}_g)$ for each group g.

5: Order groups from smallest to largest by fdr_g^* scores and denote $\mathrm{fdr}_{(1)}^*(\eta) \leq \cdots \leq \mathrm{fdr}_{(G)}^*(\eta)$ as the order statistics of these values.

6: Calculate the group α-level critical value, $l_\alpha(\eta)$, by

$$l_\alpha(\eta) = \max\left\{k : \frac{\sum_{g=1}^{k} R_{(g)}(\eta)\mathrm{fdr}_{(g)}^*(\eta)}{\sum_{g=1}^{k} R_{(g)}(\eta)} \leq \alpha\right\}$$

where $R_{(1)}(\eta)$ is the $R_g(\eta)$ value for the group with the first order statistic in $\mathrm{fdr}_g^*(\eta)$ values.

7: Deem hypothesis j in group g to be reproducible if $\mathrm{fdr}_{(1)}^*(\eta) \leq \mathrm{fdr}_g^*(\eta) \leq \mathrm{fdr}_{(l_\alpha(\eta))}^*(\eta)$ and $j \in C_g(\eta)$.

Theorem 3.2.2 shows that the proposed hypothesis-level procedure controls $h\mathrm{PFDR}(\boldsymbol{T})$ at the nominal level under the oracle assumption.

Theorem 3.2.2. *Assume the quantities* $\mathrm{fdr}_{j|g}$ *and* fdr_g *are known for all* $g \in \{1, \cdots, G\}$ *and* $j \in \{1, \cdots, n_g\}$. *Denote the proposed two-fold hypothesis-level decision rule for hypothesis* j *in group* g *for a fixed* $\eta \in (0, 1)$ *by* $\delta_{gj}^\eta(\boldsymbol{T})$. *For any* $\alpha \in (0, 1)$, *the* $h\mathrm{PFDR}(\boldsymbol{T})$ *using* $\delta_{gj}^\eta(\boldsymbol{T})$ *is controlled at the level* α. *That is,*

$$h\mathrm{PFDR}(\delta_{gj}^\eta; \boldsymbol{T}) = \mathbb{E}\left[\frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g}(1 - \theta_{gj})\delta_{gj}^\eta(\boldsymbol{T})}{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\delta_{gj}^\eta(\boldsymbol{T}) \vee 1}\middle|\boldsymbol{T}\right] \leq \alpha.$$

Proof. The proof is immediate by construction of δ_{gj}^η, since $h\mathrm{PFDR}(\delta_{gj}^\eta; \boldsymbol{T}) = \frac{\sum_{g=1}^{\ell_\alpha(\eta)} R_{(g)}(\eta)\mathrm{fdr}_{(g)}^*(\eta)}{\sum_{g=1}^{\ell_\alpha(\eta)} R_{(g)}(\eta)}$.

Notice that this algorithm leverages the group structure of reproducibility status by containing criteria for both a hypothesis to show evidence of being reproducible (1-3 in Algorithm 3) and a group as well (4-6 in Algorithm 3).

Remark 3.2.1. *In many applications, including transcript-level expression genomic data, there are some groups (genes) with multiple hypotheses (transcripts) and some groups that contain only a single hypothesis. Notice that the two-stage hypothesis-level rejection rule reduces nicely for single-hypothesis groups. It is clear from the definition of group reproducibility and* (3.5) *that* $\mathrm{fdr}_{1|g} = 0$ *for any group* g *with a single hypothesis. So it follows that* $\mathrm{fdr}^*_g(\eta) = \mathrm{fdr}_g$ *for any* η *and the method finds the single hypothesis to be reproducible if and only if* $\mathrm{fdr}^*_{(1)}(\eta) \leq \mathrm{fdr}_g \leq \mathrm{fdr}^*_{(l_\alpha(\eta))}(\eta)$. *In fact, if all groups contain only one hypothesis, this procedure reduces to the proposed group-level procedure.*

3.3 Group structured model

In this section, we discuss the model for reproducibility statuses $(\theta_g, \theta_{j|g})$ that is assumed to derive closed forms for the local FDR scores fdr_g and $\mathrm{fdr}_{j|g}$.

3.3.1 Bernoulli significant group model

The group structure for $(\boldsymbol{T}_{gj}, \theta_g, \theta_{j|g})$ with $g \in \{1, \cdots, G\}$ and $j \in \{1, \cdots, n_g\}$ is assumed to follow the form of the Bernoulli Significant Group (BSG) discussed in Liu et al. (2016). That is, consider the following setting.

For all $g \in \{1, \cdots, G\}$, $\theta_g \overset{\text{i.i.d.}}{\sim} \mathrm{BERN}(\Pi_1)$.

For all $j \in \{1, \cdots, n_g\}$, the hypothesis-given-group-level reproducibility for irreproducible groups is distributed by the point mass at zero. That is

$$\theta_{j|g} \mid \theta_g = 0 \overset{\text{i.i.d.}}{\sim} \mathrm{BERN}(0)$$

For all $j \in \{1, \cdots, n_g\}$, the hypothesis-given-group-level reproducibility for reproducible groups is distributed by a truncated Bernoulli distribution which avoids reproducible groups with zero reproducible hypotheses. So,

$$\theta_{j|g} \mid \theta_g = 1 \overset{\text{i.i.d.}}{\sim} \text{truncBERN}(\pi_1^1)$$

where truncBERN is the truncated Bernoulli distribution characterized by the following joint point mass function.

$$\theta_{j|g}, \cdots, \theta_{n_g} \mid \theta_g = 1 \sim \frac{\prod_{j=1}^{n_g} (\pi_1^1)^{\theta_{j|g}} (1-\pi_1^1)^{1-\theta_{j|g}} \mathbb{I}\left[\sum_{j=1}^{n_g} \theta_{j|g} \geq 1\right]}{1-(1-\pi_1^1)^{n_g}}. \tag{3.10}$$

Notice that this model ensures that at least one hypothesis within a reproducible group is reproducible, which is in line with the assumed definition of group-level reproducibility.

Lastly, the observed summary statistics for hypothesis j in group g conditional on θ_{gj} have the density

$$\boldsymbol{T}_{gj} \mid \theta_{j|g} \sim (1-\theta_{gj}) f_0(\boldsymbol{t}_{gj}) + \theta_{gj} f_1(\boldsymbol{t}_{gj}) \tag{3.11}$$

where f_0 and f_1 are any continuous density functions for irreproducible and reproducible summary statistics, respectively.

Formulas for $\text{fdr}_{j|g}$ and fdr_g under the BSG Model

The BSG design allows for the quantities fdr_g and $\text{fdr}_{j|g}$ to be have closed formulas in terms of the parameters Π_1, π_1^1, f_0, and f_1. First, let

$$\widetilde{\text{fdr}}_{gj} = \frac{(1-\pi_1^1) f_0(\boldsymbol{T}_{gj})}{f(\boldsymbol{T}_{gj})} \text{ and } \widetilde{\text{fdr}}_g = \prod_{k=1}^{n_g} \widetilde{\text{fdr}}_{gk} \tag{3.12}$$

where $f(\mathbf{t}) = (1-\pi_1^1) f_0(\mathbf{t}) + \pi_1^1 f_1(\mathbf{t})$. Notice, $\widetilde{\text{fdr}}_{gj}$ and $\widetilde{\text{fdr}}_g$ would be the local FDR scores for hypothesis and group if we did not require for any reproducible group $\sum_{j=1}^{n_g} \theta_{j|g} \geq 1$. Liu et al.

(2016) show that the group local FDR score is

$$\text{fdr}_g = \mathbb{P}(\theta_g = 0 \mid \boldsymbol{T}) = \frac{(1-\Pi_1)\widetilde{\text{fdr}}_g}{(1-\Pi_1)\widetilde{\text{fdr}}_g + \pi_1(1-\widetilde{\text{fdr}}_g)} \tag{3.13}$$

where $\pi_1 = \Pi_1\left(\frac{\pi_1^1}{1-(1-\pi_1^1)^{n_g}}\right)$, and the hypothesis-given-group local FDR score is

$$\text{fdr}_{j|g} = \mathbb{P}\left(\theta_{j|g} = 0 \mid \theta_g = 1, \boldsymbol{T}\right) = \frac{\widetilde{\text{fdr}}_{gj} - \widetilde{\text{fdr}}_g}{1-\widetilde{\text{fdr}}_g}. \tag{3.14}$$

Remark 3.3.1. *Again, there is a nice simplification of these quantities in the case of single-hypothesis groups. Notice, that is,*

$$\text{fdr}_g = \frac{(1-\Pi_1)(1-\pi_1^1)f_0(\boldsymbol{T}_{gj})/f(\boldsymbol{T}_{gj})}{(1-\Pi_1)(1-\pi_1^1)f_0(\boldsymbol{T}_{gj})/f(\boldsymbol{T}_{gj}) + \pi_1(1-(1-\pi_1^1)f_0(\boldsymbol{T}_{gj})/f(\boldsymbol{T}_{gj}))}.$$

With knowledge of the parameters $(\Pi_1, \pi_1^1, f_0, f_1)$, one can use the expressions in (3.13) and (3.14) to explicitly calculate the local FDR scores and, in turn, use Algorithms 2 or 3 to discover reproducible groups or hypotheses.

3.4 Estimation

In practice, however, Π_1, π_1^1, and the distributions f_0 and f_1 are unknown and thus must be estimated. To obtain estimates for the local FDR scores, we assume summary statistics come from a bivariate copula mixture model that is commonly used to develop methods (Li et al., 2011) and in generate simulations (Philtron et al., 2018) in the reproducibility literature. With this assumption, we can utilize a version of the EM algorithm to estimate the unknown parameters in the BSG model.

3.4.1 Gaussian copula mixture model

Assume that f_0 and f_1 follow the copula mixture model introduced to assess reproducibility in Li et al. (2011). Specifically, we assume that the dependence structure and relative signal strength

for summary statistics conditional on the reproducibility status (θ_{gj}) is inherited from a bivariate Gaussian distribution. For reproducible hypotheses ($\theta_{gj} = 1$), we expect summary statistics to have more signal strength than irreproducible hypotheses, thus the mean of the bivariate Gaussian that induces the summary statistics is larger for reproducible hypotheses than irreproducible hypotheses. It is also expected that summary statistics for a reproducible hypothesis will be positively associated across study, while irreproducible hypotheses have independent summary statistics. Thus, the correlation coefficient will be non-zero when $\theta_{gj} = 1$ and equal to 0 when $\theta_{gj} = 0$. The assumed copula model that produces the summary statistics is as follows.

Let θ_{gj} have the structure from (3.1) in Section 3.2. Then, conditional on θ_{gj}, the latent Gaussian signals for each hypothesis, denoted $\boldsymbol{z}_{gj} = (z_{gj,1}, z_{gj,2})$, are distributed by

$$\boldsymbol{z}_{gj} \mid \theta_{gj} = k \sim \mathbb{N}\left(\begin{bmatrix}\mu_k \\ \mu_k\end{bmatrix}, \sigma_k^2 \begin{bmatrix}1 & \rho_k \\ \rho_k & 1\end{bmatrix}\right) \text{ for } k \in \{0, 1\} \tag{3.15}$$

with $\mu_0 = \rho_0 = 0$, $\sigma_0^2 = 1$, $\mu_1 > 0$, $0 < \rho_1 \leq 1$, and $\sigma_1^2 > 0$. Then, the observed summary statistics are obtained by transforming $\boldsymbol{z}_{gj}$ so that they are marginally uniform using a mixture of the two bivariate Gaussian distributions and then back-transforming these uniform variables by the unknown distribution of the observed summary statistics, as is described here. Let $\boldsymbol{\beta} = (\Pi_1, \pi_1^1, \mu_1, \sigma_1, \rho_1)$ and denote

$$\begin{aligned} u_{gj,1} &= G(z_{gj,1}; \boldsymbol{\beta}) = \pi_1 \Phi\left(\frac{z_{gj,1} - \mu_1}{\sigma_1}\right) + (1 - \pi_1)\Phi(z_{gj,1}) \\ u_{gj,2} &= G(z_{gj,2}; \boldsymbol{\beta}) = \pi_1 \Phi\left(\frac{z_{gj,2} - \mu_1}{\sigma_1}\right) + (1 - \pi_1)\Phi(z_{gj,2}) \end{aligned} \tag{3.16}$$

where Φ is the standard Gaussian distribution function, and $\pi_1 = \Pi_1\left(\frac{\pi_1^1}{1-(1-\pi_1^1)^{n_g}}\right)$. The actual observed summary statistics are

$$t_{gj,1} = H_1^{-1}(u_{gj,1}) \text{ and } t_{gj,2} = H_2^{-1}(u_{gj,2}) \tag{3.17}$$

for any continuous distribution functions H_1 and H_2. Notice, this allows the observed summary statistics to follow any distributions H_1 and H_2 while still have signal inherited from the bivariate

normal distribution. Additionally, this structure allows f_0 and f_1 from (3.11) to be written in the manner

$$\begin{aligned} f_0(\boldsymbol{t}_{gj}) &= g_0\left(G^{-1}(H_1(\boldsymbol{t}_{gj,1})), G^{-1}(H_2(\boldsymbol{t}_{gj,2}))\right) = g_0(\boldsymbol{z}_{gj}) \\ f_1(\boldsymbol{t}_{gj}) &= g_1\left(G^{-1}(H_1(\boldsymbol{t}_{gj,1})), G^{-1}(H_2(\boldsymbol{t}_{gj,2}))\right) = g_1(\boldsymbol{z}_{gj}) \end{aligned} \tag{3.18}$$

where g_0 is the density function of the $\mathbb{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}\right)$ and g_1 is the density function of the $\mathbb{N}\left(\begin{bmatrix}\mu_1\\\mu_1\end{bmatrix}, \sigma_1^2\begin{bmatrix}1 & \rho_1\\\rho_1 & 1\end{bmatrix}\right)$. In this setting, μ_1 represents the signal strength for reproducible hypotheses compared to irreproducible hypotheses, σ_1^2 represents the variability reproducible hypotheses show relative to irreproducible hypotheses, and ρ_1 represents the consistency of signal across studies for reproducible hypotheses. Now, the model is fully parameterized by $\boldsymbol{\beta} = (\Pi_1, \pi_1^1, \mu_1, \sigma_1, \rho_1)$ along with with the distributions H_0 and H_1. The likelihood function for the data as follows.

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta}) &= \prod_{g=1}^{G}\left\{(1-\Pi_1)\prod_{j=1}^{n_g} f_0(\boldsymbol{t}_{gj}) + \Pi_1\frac{\left[\prod_{j=1}^{n_g} f(\boldsymbol{t}_{gj}) - (1-\pi_1^1)^{n_g}\prod_{j=1}^{n_g} f_0(\boldsymbol{t}_{gj})\right]}{1-(1-\pi_1^1)^{n_g}}\right\} \\ &= \prod_{g=1}^{G}\left\{(1-\Pi_1)\prod_{j=1}^{n_g} f_0(\boldsymbol{t}_{gj}) + \Pi_1\frac{1-\widetilde{\text{fdr}}_g}{1-(1-\pi_1^1)^{n_g}}\prod_{j=1}^{n_g}\left[(1-\pi_1^1)f_0(\boldsymbol{t}_{gj}) + \pi_1^1 f_1(\boldsymbol{t}_{gj})\right]\right\} \end{aligned}$$

3.4.2 Estimation algorithm

We now propose a method to estimate the $\boldsymbol{\beta}$ that does not rely on the specific distributions H_1 and H_2. Imposing no assumptions on the distribution functions H_1 and H_2 except that they are continuous introduces some difficulty in the implementation of standard estimation approaches, as the likelihood function for the actual observed data cannot be calculated. Instead, a common approach to estimate the parameters in the copula model context is using the observed data to calculate scaled ranks, treating these scaled ranks as estimated versions of the $u_{gj,i}$ values from (3.16), and converting these to pseudo-data from the bivariate Gaussian distribution in (3.15) (Oakes, 1994).That is, we let

$$\widehat{u}_{gj,1} = \frac{r_{gj,1}}{n+1} \text{ and } \widehat{u}_{gj,1} = \frac{r_{gj,2}}{n+1} \tag{3.19}$$

where $r_{gj,k}$ is the ranking of the summary statistic observed in study k for the j^{th} hypotheses in group g among all hypotheses. Note that we scale by $n+1$ and not n to avoid infinites when we use the copula mixture to convert these values. Next, for a given $\boldsymbol{\beta}$, one can calculate the pseudo-data by

$$\widehat{z}_{gj,1} = G^{-1}(\widehat{u}_{gj,1}; \boldsymbol{\beta}) \text{ and } \widehat{z}_{gj,2} = G^{-1}(\widehat{u}_{gj,2}; \boldsymbol{\beta}) \tag{3.20}$$

where the inverse of G can be computed analytically since it does not have a closed form. Now, because the unobserved $\boldsymbol{z}_{gj}$ come from a mixture distribution of bivariate Gaussian distributions, the expectation-maximization (EM) (Dempster et al., 1977) approach on the pseudo-data allows us to estimate a new, updated $\boldsymbol{\beta}$. We can then implement an algorithm that alternates between selecting $\boldsymbol{\beta}$ that maximizes the likelihood of the pseudo-data and subsequently updating the pseudo-data using the newly estimated $\boldsymbol{\beta}$. Specifically, Algorithm 4 details our approach to the EM algorithm for estimating the unknown parameters of interest, $\boldsymbol{\beta} = (\Pi_1, \pi_1^1, \mu_1, \sigma_1, \rho_1)$ that avoids making additional parametric assumptions on H_1 and H_2. We adapt the expectation–maximization (EM) algorithm (Dempster et al., 1977) on pseudo-data G^{-1} to fit our group structured model from Section 3.3.

Algorithm 4 $\boldsymbol{\beta}$ estimation algorithm.

1: For each hypothesis, calculate $\widehat{u}_{gj,1}$ and $\widehat{u}_{gj,2}$ as in (3.19).
2: Set initial values $\boldsymbol{\beta}_0 = \left(\Pi_1^0, \pi_1^{1,0}, \mu_1^0, \sigma_1^0, \rho_1^0\right)$.
3: Transform observed data to pseudo-data, $\widehat{z}_{gj,1}$ and $\widehat{z}_{gj,2}$ by (3.20) using $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ where $G^{-1}(\widehat{u}_{gj,k})$.
4: Calculate fdr_g and $\mathrm{fdr}_{j|g}$ using $\widehat{z}_{gj,k}$ for $k \in \{1,2\}$ and $\boldsymbol{\beta} = \boldsymbol{\beta}^0$ as in (3.13) and (3.14) .
5: Update the parameters as follows.

$$\Pi_1^{\text{new}} = 1 - \frac{\sum_g \mathrm{fdr}_g}{G}$$

$$\pi_1^{1,\text{new}} = \frac{\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})}{\sum_g (1-\mathrm{fdr}_g)}$$

$$\mu_1^{\text{new}} = \frac{\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})(\widehat{z}_{gj,1}+\widehat{z}_{gj,2})}{2\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})}$$

$$\sigma_1^{2,\text{new}} = \frac{\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})((\widehat{z}_{gj,1}-\mu_1^0)^2+(\widehat{z}_{gj,2}-\mu_1^0)^2)}{2\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})}$$

$$\rho_1^{\text{new}} = \frac{2\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})(\widehat{z}_{gj,1}-\mu_1^0)(\widehat{z}_{gj,2}-\mu_1^0)}{\sum_g \sum_j (1-\mathrm{fdr}_g)(1-\mathrm{fdr}_{j|g})((\widehat{z}_{gj,1}-\mu_1^0)^2+(\widehat{z}_{gj,2}-\mu_1^0)^2)}$$

and let $\boldsymbol{\beta}' = \left(\Pi_1^{\text{neq}},\ \pi_1^{1,\text{new}},\ \mu_1^{\text{new}},\ \sigma_1^{2,\text{new}},\ \rho_1^{\text{new}}\right)$.
6: Set $\boldsymbol{\beta} = \boldsymbol{\beta}'$ and repeat the process from 3-5 until convergence.

See Appendix B.2 for a more detailed derivation of the EM algorithm employed and Appendix B.3 for an examination of the performance of the procedure in simulation.

3.5 Practical implementations

3.5.1 Selection of η

Notice, the FDR result pertaining to the hypothesis-level procedure from Theorem 3.2.2 holds for any fixed $\eta \in (0,1)$. Naturally, one would desire to select the η which is optimal in some sense. Genovese and Wasserman (2002) introduced false non-discovery rate (FNR) as a counterpart to FDR. Similar to FDR, FNR is the *expectation* of false non-discovery proportion (FNP), where FNP is the proportion of hypotheses not in the rejection set that are truly non-null. That is, in hypothesis-level procedure context the posterior FNR for any decision rule $\delta_{gj}(\boldsymbol{T})$ can be expressed as in (3.21).

$$h\mathrm{PFNR}(\delta_{gj};\boldsymbol{T}) = \mathbb{E}\left[\frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\theta_{gj}\left(1-\delta_{gj}(\boldsymbol{T})\right)}{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(1-\delta_{gj}(\boldsymbol{T})\right)\vee 1}\middle|\boldsymbol{T}\right]. \tag{3.21}$$

In the reproducibility problem, the numerator of fraction in (3.21) represents the number of hypotheses that are truly reproducible that the decision rule does not find to be reproducible and the denominator represents the total number of hypotheses not declared reproducible by the decision rule. In selecting the value of $\eta \in (0,1)$ for the proposed hypothesis-level decision rule, we will focus on minimizing the FNR quantity at a fixed level of FDR, as minimizing this value yields a decision rule with the optimal ability to avoid reproducible non-discoveries. Notice, the FNR quantity in (3.21) has a similar form to the quantity from (3.4) and thus when can be written in as functions of fdr_g and $\mathrm{fdr}_{j|g}$ in the following manner.

$$\begin{aligned} h\mathrm{PFNR}(\delta_{gj};\boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\theta_{gj}\left(1-\delta_{gj}(\boldsymbol{T})\right)}{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\left(1-\delta_{gj}(\boldsymbol{T})\right)\vee 1}\middle|\boldsymbol{T}\right] \\ &= \frac{\sum_g\left(1-\mathrm{fdr}_g\left(\boldsymbol{T}\right)\right)\sum_j\left(1-\mathrm{fdr}_{j|g}\left(\boldsymbol{T}\right)\right)}{\left(n-\sum_g\sum_j\delta_{gj}(\boldsymbol{T})\right)\vee 1} \\ &\quad - \frac{\sum_g\left(1-\mathrm{fdr}_g\left(\boldsymbol{T}\right)\right)\sum_j\left(1-\mathrm{fdr}_{j|g}\left(\boldsymbol{T}\right)\right)\delta_{gj}(\boldsymbol{T})}{\left(n-\sum_g\sum_j\delta_{gj}(\boldsymbol{T})\right)\vee 1}. \end{aligned}$$

Now, this representation allows us to select the η in the proposed δ_{gj}^{η} decision rule that minimizes the hPFNR, avoiding false non-discoveries. That is, we propose selecting η by

$$\eta' = \underset{\eta\in(0,1)}{\arg\min}\{h\mathrm{PFNR}(\delta_{gj}^{\eta};\boldsymbol{T})\} \tag{3.22}$$

and assessing hypothesis-level reproducibility using $\delta_{gj}^{\eta'}$ as laid out in Algorithm 3. An examination of the performance of this selection procedure in simulation can be found in Appendix B.3.

3.6 Simulations

In this section, we compare the performance of the proposed procedures, in both the oracle and fully data-driven forms, with existing methods through extensive sets of simulations. The primary interest is the ability of each method to discover reproducible groups and hypotheses with control of false discovery rate at a nominal level of α. Namely, for hypothesis-level inference, we compare algorithm from Algorithm 3 in the oracle form and fully data-driven form to the original copula

mixture procedure from Li et al. (2011) which we denote IDR, MaRR (Philtron et al., 2018), and adaFilter (Wang et al., 2022). In both the oracle and estimated cases, we select η by the criteria in Section 3.5.1. For group-level inference, we compare the Algorithm in 2 in the oracle and estimated form to applying MaRR (Philtron et al., 2018) to each group by aggregating summary statistics for hypotheses from the same group together and then applying the method to the aggregated summary statistic.

In these simulations, the summary statistics from the two studies are generated by the mixture of bivariate Gaussian distributions described in Section 3.4.1 with the group structure from Section 3.3. The bivariate distribution is commonly considered for simulation in the literature (Li et al., 2011; Philtron et al., 2018; Ghosh et al., 2021). The design allows us to assess the performance of proposed methods under many different circumstances. We describe the specifics of the simulation settings in Section 3.6.1 and detail the results for both group and hypothesis-level procedures in Section 3.6.2.

3.6.1 Simulation settings

Our simulations consider the BSG model structure from Section 3.2 with summary statistics conditional on hypothesis reproducibility status (θ_{gj}), simulated by the distribution in (3.15). That is, we consider n hypotheses commonly assessed in $m = 2$ replicate studies that can be divided into G disjoint groups. The size of group $g \in \{1, \cdots, G\}$ is denoted by n_g. For group g, the group-level reproducibility status, θ_g, is generated by the $\mathrm{BERN}(\Pi_1)$. Then, if group g is irreproducible, the hypothesis-given-group-level reproducibility status for hypothesis $j \in \{1, \cdots, n_g\}$ in group g, $\theta_{j|g}$, comes from a $\mathrm{BERN}(0)$; and if the group is reproducible, then $\theta_{j|g} \sim \mathrm{truncBERN}\,(\pi_1^1)$, as described in (3.10). the overall reproducibility status for a hypothesis is determined by $\theta_{gj} = \theta_g * \theta_{j|g}$. Then, conditional on reproducibility status, we begin by generating the Gaussian signals $\boldsymbol{z}_{gj}$ from the mixture distribution in (3.15), with μ_1, σ_1^2 and ρ_1 representing the signal strength, variability, and across study consistency of reproducible hypotheses. Then, the observed summary statistics,

$\boldsymbol{t}_{gj}$, are the one-sided p-values from the Gaussian signals, $\boldsymbol{z}_{gj}$, calculated using the standard normal distribution.

In these simulations, we set $n = 5,000$ and $G = 500$. The group sizes, n_g for $g \in \{1, \cdots, G\}$, are distributed by $n_g = 1 + n'_g$ with $(n'_1, \cdots, n'_G) \sim \mathrm{M}_G(n; \boldsymbol{p} = (p_1, \cdots, p_G))$ where M_k is the multinomial distribution with k groups. We consider $\boldsymbol{p} = (1/G, \cdots, 1/G)$. To examine differences in performance across differing levels of sparsity, simulations are performed for all combinations of $\Pi_1 \in \{0.3, 0.5, 0.8\}$ and $\pi_1^1 \in \{0.3, 0.5, 0.8\}$. Finally, for the signal strength, variability, and consistency, we set $(\mu_1, \sigma_1^2, \rho_1) = (1.5, 1, 0.7)$ to represent a setting where reproducible hypotheses are not easily separable from irreproducible hypotheses. Results are examined for 50 iterations of each simulation setting.

In applying the proposed hypothesis and group-level procedures to each iteration, we initialize the parameters in $\boldsymbol{\beta}_0$ the following distributions: $\Pi_1^0 \sim \mathrm{UNIF}(0.3, 0.8)$, $\pi_1^{1,0} \sim \mathrm{UNIF}(0.3, 0.8)$, $\mu_1^0 \sim \mathrm{UNIF}(0.5, 3)$, $\sigma_1^{2,0} \sim \mathrm{UNIF}(0.7, 1.5)$, and $\rho_1^0 \sim \mathrm{UNIF}(0.3, 0.9)$. The adaFilter, IDR, and MaRR methods were applied using the `idr`, `adaFilter`, and `marr` packages in `R`, respectively. In implementing the IDR method, we use two initializations: in the first, we use the true value for the parameters and in the second, we use the same initialization used for the fully-data adaptive version of our proposed method.

3.6.2 Simulation results

To compare hypothesis and group-level results, we compare the average observed power and average observed FDP (denoted FDR, as FDR is defined by the expected FDP) at nominal levels of FDR of $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$. Where observed power and FDP for any rejection rule δ_i and reproducibility status θ_i are calculated for each iteration by

$$\text{Power} = \frac{\sum_i \theta_i \delta_i}{\sum_i \theta_i \vee 1} \quad \text{and} \quad \text{FDP} = \frac{\sum_i (1 - \theta_i)\delta_i}{\sum_i \delta_i \vee 1}. \tag{3.23}$$

Figures 3.2 and 3.1 show these average power and FDR values across 50 iterations of each of the simulation settings described in Section 3.6.1 at the hypothesis and group-level. Each point

corresponds to a desired nominal level. The nominal levels of α are marked by the vertical dashed lines. The x-coordinate corresponds to the observed FDR and the y-coordinate corresponds to the observed average power. In general, a point to the left of its corresponding nominal level represents a method that controls FDR at that nominal level. The closer a point is to its corresponding nominal level, the more closely the observed empirical FDR aligns with the nominal level. Additionally, methods with points higher in the vertical direction represent methods that are able to more reliably detect reproducible hypotheses.

Hypothesis-level results

It can be seen in Figure 3.2 that the proposed method in its oracle controls FDR at a level nearly identical to the nominal level and tends to maintain a higher average power level than all other methods considered. The fully estimated version of the proposed method also controls at levels nearly identical to the nominal level with high average power levels, except in the cases with $\pi_1^1 = 0.3$ where the observed FDR is nearly identical to the nominal level for small levels of α but is inflated by roughly 0.03 for $\alpha = 0.20$. The proposed procedures show their largest advantages relative to existing methods within group reproducibility is dense relative to group-wise density. That is, when π_1^1 is large relative to Π_1, the proposed procedures are able to discover reproducible hypotheses far more reliably than the existing methods. This signals that, in these cases, leveraging group information substantially improves performance. When examining existing methods, it can be seen that adaFilter tends to be overly conservative and MaRR tends to be anti-conservative for smaller nominal FDR levels (see $\alpha \in \{0.01, 0.05\}$) and overly conservative for larger levels of nominal FDR (see $\alpha \in \{0.1, 0.2\}$). Additionally, for MaRR, we see the method fail when the reproducible signal is dense (see $\Pi_1 = \pi_1^1 = 0.8$) due to poor estimation of the sparsity parameter needed for the method. For IDR, it is clear that the initialization scheme for their estimation algorithm has a great impact on the method's performance. When the estimation procedure is initialized at the true parameters, we can see that the original IDR method shows nearly exact control of FDR. When initializations are randomly selected as is done for the fully estimable version of the proposed method, however, the original IDR method shows an inability to control FDR at its

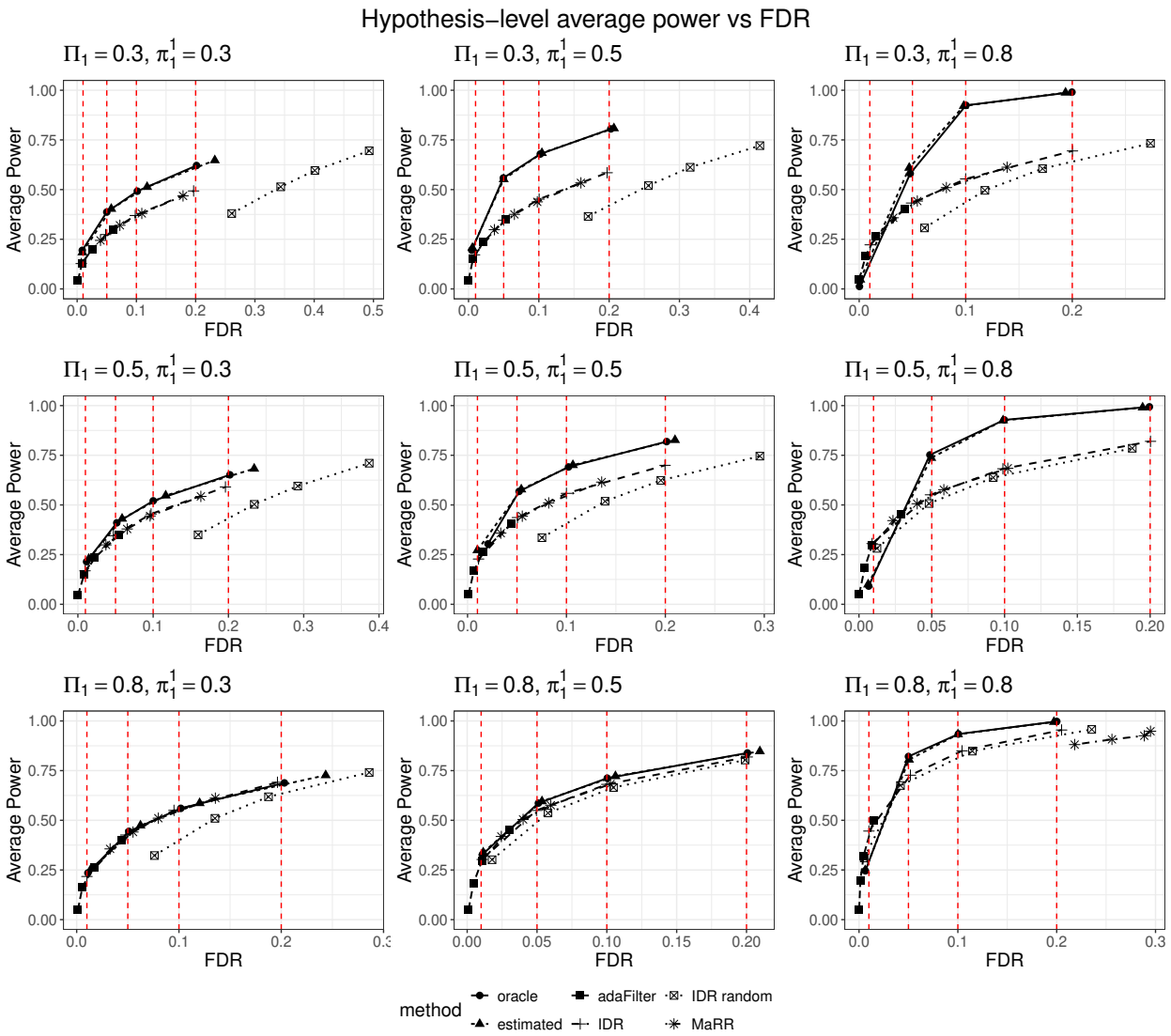


Figure 3.1: Hypothesis-level average power against the average observed FDP from 50 iterations of each setting described in Section 3.6.1 for the proposed methods in oracle (•) and estimated (▲) forms, adaFilter (■), IDR initialized using the true parameters (+) and randomly (⊠), and MaRR (∗). Each point corresponds with a nominal level of FDR of $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ (as marked by the dashed lines).

desired nominal level in nearly all cases. In general, it can be seen that the proposed procedure has advantages in terms of control of FDR and average power when compared to existing methods.

Group-level results

As for the group-level statistical inference, Figure 3.2 shows that the proposed procedure in the oracle and fully estimated versions maintains FDR control at a nearly identical level to that of the nominal FDR level and higher power than that of applying the MaRR procedure to the aggregate

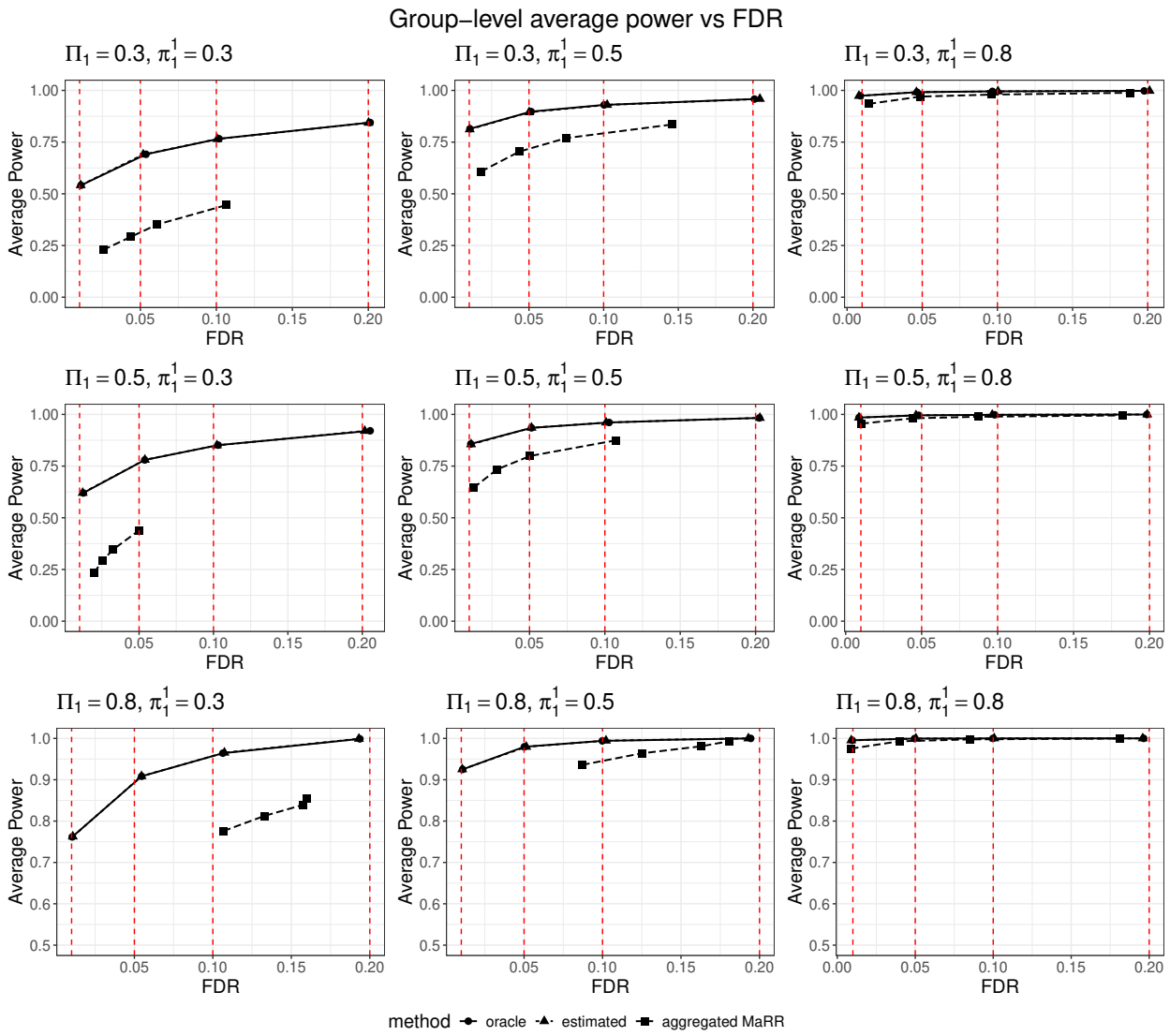


Figure 3.2: Group-level average power against the average observed FDP from 50 iterations of each setting described in Section 3.6.1 for the proposed methods in oracle (•) and estimated (▲) forms, and MaRR on the average summary statistics (■). Each point corresponds with a nominal level of FDR of $\alpha \in \{0.01, 0.05, 0.10, 0.20\}$ (as marked by the dashed lines).

of summary statistics within a group. In particular, the proposed procedures show an advantage when reproducible groups have few reproducible member hypotheses (see $\pi_1^1 = 0.3$). Interestingly, applying the group-level MaRR procedure shows little ability to correctly identify reproducible groups at a nominal level of FDR, as the method is neither consistently conservative nor anti-conservative. In all, the proposed group-level procedure shows distinct advantages compared to

applying reproducible methods to the summary statistics aggregated to the group-level in terms of controlling FDR at the nominal level and average power.

3.7 Discussion

Existing methods to assess the reproducibility of results across high-throughput studies generally ignore the group structure that exists in genetic data. Typically, reproducibility analysis is performed at the hypothesis level without the use of the group information, and any group-level inference is either performed post-hoc – such as pathway analysis – or hypothesis-level results are aggregated to the group-level, and then reproducibility analysis is performed on the aggregates – such as aggregating transcript abundances to gene expressions for differential expression analysis. Both of these approaches are underpowered, as neither adequately leverages the known group structure in their reproducibility analysis. In this work, we extend the empirical Bayesian approach in the multiple testing with group structure problem from Liu et al. (2016) to the reproducibility context, where we examine the agreement of results for n hypothesis examined in two studies. Based on this approach, we propose procedures for discovering sets of reproducible groups and hypotheses that control the total posterior false discovery rate at a nominal level under the oracle the assumption. Blending the BSG model from Liu et al. (2016) with the Gaussian copula mixture model from Li et al. (2011), we adapt the EM algorithm to devise a method for the estimation of the unknown parameters that the models rely on. Additionally, we devise a novel method for selecting the tuning parameter, η, that has optimal performance in terms of posterior false nondiscovery rate. In simulations, we see that the proposed hypothesis-level procedure in the oracle and fully estimable versions outperform existing methods in power while generally controlling the false discovery rate. The results signal that the inclusion of group information can substantially improve reproducibility detection. Additionally, simulations show that the proposed group-level procedure controls the false discovery rate, while the validity of performing group-level inference after aggregation can be heavily dependent on the particular setting.

One limitation of the proposed approach in the high-throughput genomics setting is the assumption that the groups are non-overlapping. Genes often belong to many pathways that are interesting and thus the assumption that the groups are non-overlapping is limiting. The challenges overlapping groups present lie in ensuring consistent group and hypothesis-level inference for hypotheses that are members of multiple groups. Expanding the proposed framework to allow for overlapping groups is an interesting direction for future research.

In conclusion, we present one approach to analyzing the reproducibility of n hypotheses from two studies that come from G non-overlapping groups and show through simulations that it can is beneficial compared to methods naive to the group structure of the data.

Chapter 4

Assessing the reproducibility of results across multiple high-throughput studies using Kendall's W

4.1 Introduction

Advances in high-throughput technologies have allowed single-cell genomic researchers to obtain data for thousands to millions of cells from multiple individuals across disease conditions. This has led to the wide availability of studies on different cohorts that investigate the same disease, providing ample opportunity to uncover signatures across these cohorts and enhance our understanding of disease mechanisms. For example, Lin et al. (2022) compiled 20 studies comprising more than five million cells from 1,000 samples, all investigating COVID-19 patients with varying degrees of severity; and Wang et al. (2024) collected 1,053 samples from 67 Alzheimer's Disease (AD)-related single-cell studies with over seven million cells. In the single-cell sequencing setting, most research integrating these curated resources focuses on removing batch effects to improve cell type identification and profiling (Lücken et al., 2022). However, few efforts have been directed toward leveraging such large-scale datasets to investigate cell type-specific genes with changes in expression *consistently* linked to the disease across multiple studies.

For bulk-sample RNA sequencing datasets, assessing the consistency of gene expression results has become an area of increasing interest (Li et al., 2011; Philtron et al., 2018; Wang et al., 2022; Bogomolov and Heller, 2023). Genes that are consistently up- or down-regulated in the disease population are called reproducible (sometimes replicable), and those that are not are deemed irreproducible (sometimes non-replicable). Identifying these genes and their underlying pathways is essential to scientific progress because individual studies are known to carry large amounts of technical variability, often inherited from differences in platforms, sample preparation, processing, and sequencing depth, that can impact the biological findings of individual studies. Thus, discov-

ering the genes with expression results that have been reproduced across multiple studies allows researchers to narrow their focus on genes with signals that do not depend on experiment-specific effects. Motivated by access to five single-cell datasets studying genetic expression in COVID-19 patients, we aim to examine reproducibility in the single-cell context.

4.1.1 scRNA-seq studies on COVID-19

As of April 2025, the COVID-19 pandemic is responsible for more than 7,000,000 deaths worldwide (World Health Organization, 2025). As such, there has been a collaborative effort across science to better understand the disease. This has motivated research across a wide range of statistical problems, including modeling the dynamics of the COVID-19 pandemic (Hao et al., 2020; Quick et al., 2021), examining gene-level responses at both the bulk-sample (Ellinghaus et al., 2020) and single-cell (Edahiro et al., 2023) RNA sequencing levels, and classifying patients for virus diagnosis (Laddha et al., 2022) or severity (Wang et al., 2023) using machine learning techniques (Alballa and Al-Turaiki, 2021; Wu et al., 2021b; Raman et al., 2023).

Of interest to this work, the increasing number of available single-cell RNA sequencing (scRNA-seq) datasets allows for the identification of genes with consistent cell type-specific expression responses to the virus, as these studies can each be used to examine changes in gene expression in patients with and without COVID-19. Thus, examining the agreement of results across these studies allows us to discover genes with cell type-specific expression responses that are reproduced across independent studies. We devise and apply a method to assess the reproducibility of gene-level expression changes for CD14 Monocyte cells for 14,649 genes commonly available in five scRNA-seq datasets (Wilk et al., 2020; Schulte-Schrepping et al., 2020; Liu et al., 2021; Stephenson et al., 2021; Ren et al., 2021). The five studies are a subset of a larger set of 20 compiled and processed in Lin et al. (2022). In each of the studies, gene-level expression data were available for 18 cell types across varying numbers of patients broadly categorized into three COVID-19 severity categories (healthy, mild/moderate, and severe). We aim to identify genes and pathways that are consistently up- or down-regulated in patients with COVID-19 (in both the mild/moderate

and severe categories) when compared to healthy patients for the CD14 Monocyte cell type. The large number of available scRNA studies enables us to identify high-quality sets of consistently expressed genes, leading to a better understanding of gene and pathway-level responses to COVID-19 within a cell type and avoiding genes that show spurious signals in an individual dataset due to the particulars of that study's design. Existing high-throughput genomic reproducibility methods either rely on knowledge of the distribution of available test statistics in each study or are limited to the case with two replicate studies. Our approach extends the nonparametric notions of reproducibility often used when there are two replicate studies (Philtron et al., 2018) to settings with more than two replicate studies, where p-value combining procedures that rely on strict parametric assumptions are typically used (Wang et al., 2022; Bogomolov and Heller, 2023).

4.1.2 Existing reproducibility methods and our approach

There are many approaches to examining the agreement of many results for hypotheses commonly examined in multiple replicate studies. In fact, there is little agreement on the name of the problem. In the statistics community, the problem is often called either replicability (Heller and Yekutieli, 2014; Bogomolov and Heller, 2018; Wang et al., 2022; Lyu et al., 2023) or reproducibility (Li et al., 2011; Philtron et al., 2018; Li and Zhang, 2018; Ghosh et al., 2021). Throughout the paper, we call it the reproducibility problem.

The general reproducibility setting considers n hypotheses that are commonly assessed in m replicate studies. For every hypothesis, we observe summary statistics from each study that represent the notability of a hypothesis when compared to some null (a p-value, t-statistic, log-fold change score, etc.) with t_{ji} representing this for hypothesis i in experiment j. The aim is to devise a method using the t_{ji} statistics that discovers a set of hypotheses that have been reproduced across the m studies with control of false discovery rate. However, there is a lack of consensus on a formal definition for reproducibility or its counterpart, irreproducibility. The numerous definitions depend on the approach, statistical framework, model, and so on. From the frequentist perspective, approaches can be broadly divided into parametric and nonparametric settings. Among paramet-

ric approaches, one common notion for reproducibility is called r/m reproducibility (Wang et al., 2022; Jaljuli et al., 2022). Under this approach, a hypothesis is found to be reproducible if it is truly non-null in at least r out of m studies and irreproducible if it is non-null in at most $r-1$ studies where $2 \leq r \leq m$. When $m = 2$, this amounts to assessing whether a hypothesis has non-null signal in both experiments. When $m > 2$, the practitioner must specify the level they are interested in a result being reproducible. Often, then, reproducibility is assessed by assuming the t_{ji} are uniform p-values under the null and working with these uniform distributions, for example, Wang et al. (2022) uses the order statistics of p-values to filter and then select reproducible hypotheses. Bogomolov and Heller (2023) provides a more complete examination of these approaches. The nonparametric lens to reproducibility in high-throughput genomics began with the practice of computing Spearman's rank correlation among the top genes across two replicate studies (MAQC-Consortium, 2006). Consistent alignment of the most notably ranked genes signals that the results for these experiments are consistent with each other. Motivated by this thought, methods that have been devised to formally assess reproducibility often define irreproducible hypotheses by the properties of the rankings of summary statistics, such as the independence of rankings (Philtron et al., 2018; Ghosh et al., 2021). There are advantages to examining the reproducibility using rankings. First, it does not make parametric assumptions about the summary statistics available, nor does it require the same type of summary statistic in each study. Second, rankings are invariant to monotone transformations, so any experiment-wise effects can be mitigated, as long as the alignment of hypotheses remains consistent. The bulk of these rank-based notations are defined in the setting with $m = 2$ replicate studies, however, which is inadequate given the increasing availability of scRNA-seq datasets. Alternatively, there has been much work done in the empirical Bayesian setting regarding reproducibility (Heller and Yekutieli, 2014; Lyu et al., 2023; Li et al., 2024). As an example, Li et al. (2011) defines irreproducibility based on an assumed copula mixture model, then develops a procedure to calculate local and total irreproducibility discovery rates (IDR) that are analogous to their false discovery counterparts that can be used to discover reproducible hypotheses.

To fundamentally understand what we are testing against, we first define our notion of irreproducibility. Definition 4.1.1 formally establishes the notion of irreproducibility used to develop the proposed procedures.

Definition 4.1.1. *Denote the summary statistic for hypothesis i in study j by t_{ji} and the set of indices pertaining to irreproducible hypotheses as $\mathcal{H}_0$. Then, $t_{ji} \overset{\text{iid}}{\sim} F_j$ for all $i \in \mathcal{H}_0$ for some distributions F_j with $j \in \{1, 2, \cdots, m\}$ and $(t_{1i}, t_{2i}, \cdots, t_{mi})$ are mutually independent of each other for each $i \in \mathcal{H}_0$.*

Effectively, a hypothesis is irreproducible if knowledge of a result in one study does not provide any information regarding a result in a different study. Our notion fits neatly in the nonparametric space since we make no parametric assumptions about the distribution functions F_j. Interestingly, conditional on a hypothesis being irreproducible, Definition 4.1.1 implies that the rankings of its summary statistics are independent of each other across experiments. The independence of rankings immediately defines irreproducibility for some nonparametric approaches (Philtron et al., 2018; Ghosh et al., 2021) and is implied via assumptions regarding generating distributions in the semi-parametric, empirical Bayesian setting from Li et al. (2011) and its subsequent extensions.

Because it is nonparametric, this notion is useful in a wide range of settings. In high-throughput genomics, where differences in experimental design can yield vast differences in observed effects, this notion allows irreproducible hypotheses to show spurious strong observed signal in some studies despite still having a true null effect, as long as these hypotheses do so at random, which can be advantageous when compared to methods that rely an assumed distribution for irreproducible summary statistics, like methods that assume null p-values are uniformly distributed (Wang et al., 2022; Lyu et al., 2023; Li et al., 2024). The uniformity assumption can be troublesome when unmeasured confounding due to a particular study design causes the distribution of null p-values to be stochastically dominated by the uniform distribution. This notion is also applicable when the type of available summary statistic for each study differs, as F_j need not be the same for $j \in \{1, \cdots, m\}$.

Our approach to measuring this notion of reproducibility uses a statistic calculated using the rankings of summary statistics, similar to existing nonparametric methods. For example, Philtron

et al. (2018) developed the MaRR procedure to discover reproducibility across two studies using the maximum ranking for each hypothesis to discover hypotheses that show signal among the most notable in terms of observed effect magnitude. This approach works well in cases with one-sided alternatives; however, there are issues in enforcing the consistency of the direction of an effect when the alternative is two-sided. Consider the high-throughput genomic case where a gene can be strongly up-regulated in one study and strongly down-regulated in another. In both cases, this gene may be ranked among the notable genes, but we would not say this gene had reproducible results. The same issue can arise with methods that rely on combining two-sided p-values, where it is sometimes recommended to apply the method in both directions at nominal levels of $\alpha/2$ and combine results (Wang et al., 2022; Bogomolov and Heller, 2018). Motivated by the scRNA-seq data, it is of interest for a gene to be either up- or down-regulated in COVID-19; we expand the rank-based procedures to the $m > 2$ case in a manner that prioritizes the consistency of direction of an effect. Instead of ranking summary statistics by the magnitude of the observed effect, we rank summary statistics from the largest negative signal (rank 1) to the largest positive signal (rank n) in each study and our proposed statistic is large for hypotheses that are either consistently among the lowest ranked (consistently strong negative signal) or highest ranked (consistently strong positive signal). Through the full ranking of hypotheses with sign information, we avoid discovering hypotheses that have strong negative signal in some experiments and strong positive signal in other experiments. To devise the proposed testing procedures, we make a critical assumption: knowledge of a set of “control" hypotheses that are known to be irreproducible. Within scRNA-seq experiments, expression is measured for many genes, including “housekeeping" genes. Housekeeping genes are genes that are known to be stably expressed and essential for the functioning of a cell, regardless of perturbations to the cell, such as diseases, traits, and experimental conditions Lin et al. (2019). In the COVID-19 context, these genes will not be consistently differentially expressed in patients with the virus compared to healthy patients, and observed strong signal in an individual study can be assumed to happen at random. For these reasons, housekeeping genes act as a control set for all irreproducible genes. In many different contexts, this assumption can be

overly limiting, as there are not always known sets of irreproducible hypotheses; however, there are in the motivating example. We use the control set to build testing procedures for identifying reproducible hypotheses from irreproducible hypotheses at a nominal level of false discovery rate control.

Reproducibility and meta-analysis

The problem of reproducibility in high-throughput experimentation shares a similar framework to that of meta-analysis, where results for n hypotheses across m experiments are integrated. Despite the similarities in framework, the two problems have vastly different goals. The most common approach to meta-analysis is to integrate information from replicate studies by combining p-values by a function (Fisher, 1925; Wilkinson, 1951; Song and Tseng, 2014). The set of discoveries made by these methods is highly dependent on the alternative hypothesis that a given method considers. Chang et al. (2013) classify meta-analysis methods into three settings based on their alternative hypotheses: 1) non-zero effect in at least one study, 2) non-zero effect in all m studies, and 3) non-zero effect in at least r out of m studies. In each case, however, the null hypothesis is the absence of an effect in all m studies. For this reason, many of these meta-analysis methods allow for discoveries driven solely by a large observed effect in a singular study, so these methods do not measure the *consistency* or *agreement* of results across replicate studies but merely the *existence* of signal (Bogomolov and Heller, 2023). As previously discussed, our notion of irreproducibility allows for hypotheses to show spurious signals in studies, as long as they do so randomly. This allows us to focus on detecting hypotheses that show consistent effects across studies.

Similar to the reproducibility problem, there has also been an effort to develop rank-based methods in the meta-analysis context (Kolde et al., 2012; Hong et al., 2006; Tseng et al., 2012). When applied to the reproducibility context, these nonparametric procedures have similar shortcomings to the p-value combining methods. These methods often make discoveries based solely on one study, ignoring the consistency of results across studies. Consider the following example comparing our proposed reproducibility statistic to that of a popular rank-based meta-analysis method.

An instructive example. Consider $m = 2$ studies and $n = 1000$ common hypotheses where $n_{11} = 300$ have moderately strong and highly consistent signal across the two studies, and $n_{10} = 20$ (and $n_{01} = 20$) hypotheses have a strong signal in the first (or second) experiment but no signal in the other studies. Figure 4.1 is a scatterplot of t-statistics and their associated rankings from one realization of this type of setting. In the reproducibility problem, only the n_{11} hypotheses with

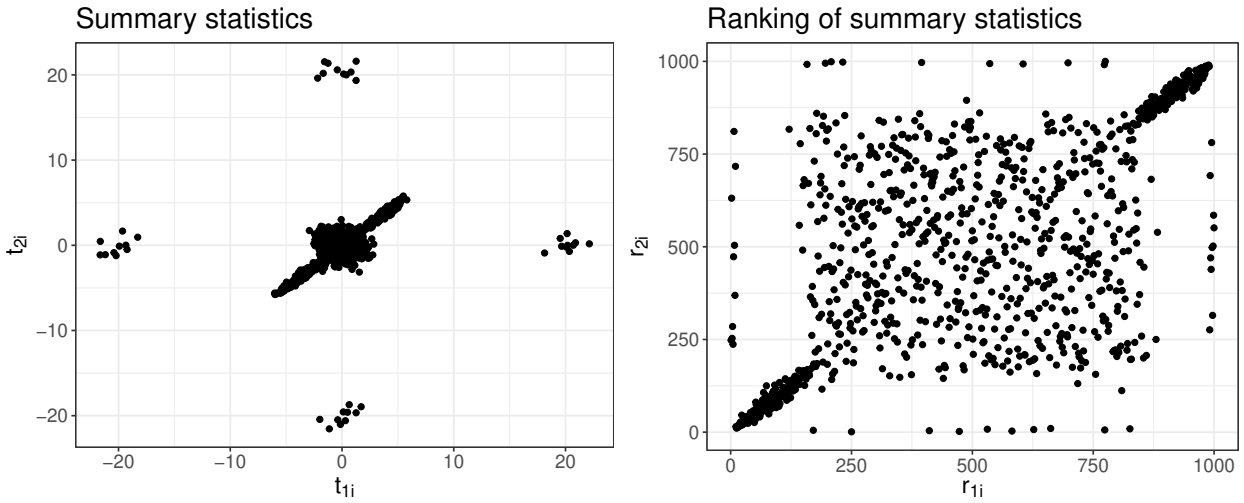


Figure 4.1: A realization of the t-statistics and their associated rankings from the setting described. See Appendix B for the full details regarding the simulation setting that generates these data.

consistent signal should be considered reproducible. With meta-analysis, a typical null hypothesis is the lack of signal in all studies, so the n_{10} and n_{01} hypotheses are also considered non-null. This can be highlighted by examining the hypotheses considered “most significant" using the proposed ΔW_{-i} statistic compared to those from popular rank-based meta method robust rank aggregation (RRA; Kolde et al. (2012)). A brief discussion of the RRA methods can be found in Section 4.4.2. Figure 4.2 shows the top $n_{11}/2 = 150$ hypotheses “most significant" hypotheses in terms of their t_{ji} statistics from the example in Figure 4.1 for the two methods where most significant is measured by the ordering of the statistic used to test against the null for each method. Among the most significant hypotheses for RRA are more than half of the n_{10} and n_{01} hypotheses, while the set for ΔW_{-i} contains only a few of those hypotheses. In this example, 100% of the top 150 hypotheses come from the $n_{11} = 300$ reproducible hypotheses in terms of ΔW_{-i} statistics,

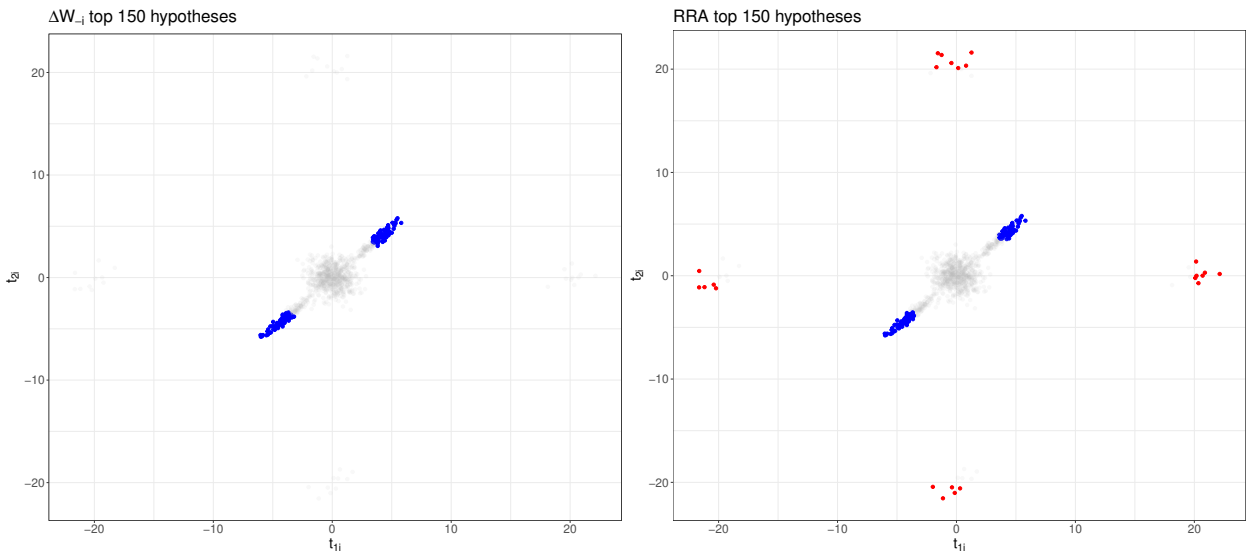


Figure 4.2: Summary statistics for the top $n_{11}/2 = 150$ hypotheses in terms of ΔW_{-i} and RRA. Hypotheses among the n_{10} and n_{01} that have strong signal in only one experiment that are found to be reproducible are shown in red.

while only 84% of the top 150 were truly reproducible for RRA. These hypotheses with strong signal in one experiment that are among the top 150 are shown in red in Figure 4.2. These clearly do not show agreement across experiment. This signals a key difference in the broader goals of reproducibility and meta-analysis, even when comparing nonparametric procedures.

4.1.3 Organization and notation

The remainder of the article is organized as follows. In Section 4.2, we introduce the ΔW_{-i} statistic that measures the reproducibility of hypotheses across m studies and develop three procedures to obtain approximate p-values based on the ΔW_{-i} statistic. The approximate p-values can then be used to discover sets of reproducible hypotheses at a nominal level of false discovery rate. We examine the performance of these three procedures compared to existing literature through comprehensive simulation studies in Section 4.3. In Section 4.4, apply one of the ΔW_{-i} procedures along with existing reproducibility and meta-analysis methods to the five scRNA-seq COVID-19 datasets to discover genes and pathways that are consistently differentially expressed in COVID-19 patients for a particular cell type. A careful comparison of reproducible gene sets for the proposed procedure and other existing methods reveals an interesting advantage for the

proposed procedure. Finally, in Section 4.5 we summarize our contributions and examine avenues for continued research.

Notations: For clarity, we list some notations used throughout the chapter. We consider the m replicate studies that examine the same n hypotheses. $\mathbb{H}_i$ denotes the i^{th} hypothesis. Additionally, we denote the known set of "control" hypothesis indices by $\mathcal{C}_0 \subset \{1, 2, \cdots, n\}$ that is of size $|\mathcal{C}_0| = n_0$, and the remaining set of "test" hypotheses by $\mathcal{D}_t = \{1, 2, \cdots, n\} \setminus \mathcal{C}_0$ of size $|\mathcal{D}_t| = n_1$ with $n = n_0 + n_1$. The sets of all irreproducible hypotheses are denoted by $\mathcal{H}_0$ and $\mathcal{H}_1$. For each hypothesis, we observe summary statistics in each experiment that measure the notability of the hypothesis. We denote the summary statistics for $\mathbb{H}_i$ by $\boldsymbol{t}_i = (t_{1i}, t_{2,i}, \cdots, t_{m,i})$. The ranking of t_{ji} among all summary statistics for study j is r_{ji}.

4.2 Methods

4.2.1 Kendall's W and the ΔW_{-i} statistic

Kendall's W, or Kendall's coefficient of concordance, (Kendall and Smith, 1939) is a nonparametric statistic used to measure the agreement of the rankings of n competitors across m judges. Specifically, assume there are m judges who are tasked with ranking n competitors from best (rank 1) to worst (rank n) by some criteria and denote the ranking of i^{th} competitor by judge j by r_{ji}. Then, Kendall's W is calculated by (4.1).

$$W = \frac{12 \sum_{i=1}^{n} \left(R_i - \bar{R}\right)^2}{m^2(n^3 - n)} \tag{4.1}$$

where $R_i = \sum_{j=1}^{m} r_{ji}$ and $\bar{R} = \frac{1}{n} \sum_{i=1}^{n} R_i$. Notice, when all m judges are in perfect agreement, then $W = 1$, and when there is no agreement in rankings, or no competitors finished higher or lower on average than others, then $W = 0$. Kendall and Gibbons (1990) proposed a hypothesis test that compares $m(n-1)W$ to a χ^2 distribution with $n-1$ degrees of freedom and tests against the null hypothesis that judges produced independent rankings. Permutation testing approaches using W

are also frequently utilized to test against that same null hypothesis (Legendre, 2005). In each of these hypothesis tests, the question of interest is whether there is *any* agreement among judges.

In the high-dimensional reproducibility problem, the Kendall's W framework and its associated hypothesis testing procedures can immediately be extended to assess the agreement of studies overall with hypotheses considered as "competitors" and studies as "judges." In this setting, hypotheses are ranked from the largest negative effect to the largest positive effect (or largest to smallest in the one-tailed case) using some summary statistic, t_{ji}. Notice, the Kendall's W statistic measures the overall agreement of the rankings of hypotheses across the m studies. This measure allows us to test whether there are *any* reproducible hypotheses across studies. Researchers are typically interested in identifying *which* hypotheses are reproducible, as opposed to assessing the reproducibility of studies as a whole, as discussed in Section 4.1.2. Thus, we propose assessing the reproducibility of each hypothesis individually by examining the *change* in Kendall's W from a specified list of hypotheses when the hypothesis of interest is no longer considered. That is, suppose we wish to quantify the reproducibility for $\mathbb{H}_i$ across the m studies relative to the other hypothesis from a larger group of the n hypotheses (i.e., the agreement in ranking across studies for that hypothesis). We propose using the ΔW_{-i} statistic defined in (4.2) to quantify that agreement.

$$\Delta W_{-i} = W - W_{-i} \tag{4.2}$$

where W is the Kendall's W statistic calculated in the manner of (4.1) when considering the entire list of hypotheses and W_{-i} is that quantity calculated using the same list of hypotheses *except* $\mathbb{H}_i$. Heuristically, ΔW_{-i} measures the change in overall agreement level across studies when all hypotheses in the group are considered compared to all hypotheses except the one of interest are considered. It can be thought of as hypothesis i's contribution to the overall agreement of the list of hypotheses. If $\mathbb{H}_i$ shows great agreement across experiments, it will be consistently ranked near the top or bottom hypotheses, and removing it from consideration decreases the overall agreement, yielding a large, positive ΔW_{-i} statistic. If $\mathbb{H}_i$ does not show agreement across experiments, then it will be inconsistently ranked. So removing $\mathbb{H}_i$ from consideration will not impact (or potentially

increase) the overall agreement of results across studies, yielding a near-zero (or negative) ΔW_{-i} statistic. Thus, we expect reproducible hypotheses to have large ΔW_{-i} statistics and irreproducible hypotheses to have a small (or negative) ΔW_{-i} statistic.

4.2.2 Asymptotic distribution of ΔW_{-i} under the global null

In this section, we introduce the global null assumption and show the asymptotic convergence of the ΔW_{-i} statistic under the global null. The global null assumption used to derive the distribution of ΔW_{-i} statistics is as follows.

> **Global null:** All hypotheses are irreproducible. That is, $t_{ji} \overset{\text{iid}}{\sim} F_j$ for all $i \in \{1, 2, \cdots, n\}$ for some distributions F_j with $j \in \{1, 2, \cdots, m\}$ and $(t_{1i}, t_{2i}, \cdots, t_{mi})$ are independent of each other for each $i \in \{1, 2, \cdots, n\}$.

The global null assumption implies the rankings for $\mathbb{H}_i$, $(r_{1i}, r_{2i}, \cdots, r_{mi})$ are each marginally distributed by a discrete uniform distribution on $\{1, 2, \cdots, n\}$ with $(r_{1i}, r_{2i}, \cdots, r_{mi})$ independent of each other, and that rankings for and experiments $(r_{j1}, r_{j2}, \cdots, r_{jn})$ are random permutations of $\{1, 2, \cdots, n\}$ that are independent for each $j \in \{1, 2, \cdots, m\}$. Since the global null assumption fully characterizes the distribution of ranks, we can derive an asymptotically equivalent form for the ΔW_{-i} statistic, as seen in Theorem 4.2.1.

Theorem 4.2.1. *Assume the global null assumption holds for a list of n hypotheses across m replicate studies. Then for $\mathbb{H}_i$,*

$$n\Delta W_{-i} \xrightarrow{D} V \equiv \frac{3(m-1)}{m} + \frac{12}{m^2} \sum_{k=1}^{m} \sum_{j \neq k} U_k(U_j - 1)$$

as $n \to \infty$ where $U_h \overset{\text{iid}}{\sim} \text{UNIF}(0, 1)$ for $h \in \{1, 2, \cdots, m\}$. We leverage the asymptotic form of ΔW_{-i} to approximate p-values based on the ΔW_{-i} statistic for each hypothesis using a “control" set of irreproducible hypotheses in Section 4.2.3.

4.2.3 ΔW_{-i} approximate p-value procedures

Next, propose three procedures that use the ΔW_{-i} statistic to obtain approximate p-values. The general setting for each method is as follows. Let there be n total hypotheses commonly assessed in m studies. For $\mathbb{H}_i$, we observe summary statistics for all studies, denoted $\boldsymbol{t}_i = (t_{1i}, t_{2i}, \cdots, t_{mi})$, that represent both the strength and direction of the observed signal. Additionally, we assume there is a subset of the n hypotheses that are known to be irreproducible based on prior knowledge. We let $\mathcal{C}_0 \subset \{1, 2, \cdots, n\}$ denote the set of indexes for these control hypotheses and $\mathcal{D}_t = \{1, 2, \cdots, n\} \setminus \mathcal{C}_0$ be the indexes of the remaining "test" hypotheses whose reproducibility status is unknown. Let $n_0 = |\mathcal{C}_0|$ and $n_1 = |\mathcal{D}_t|$. The three procedures calculate an approximate p-value for $\mathbb{H}_i$ for every $i \in \mathcal{D}_t$.

Global null procedure

Notice, if $\mathcal{C}_0$ contains all irreproducible hypotheses, then the global null assumption is met for the sets $\mathcal{C}_0$ and $\mathcal{C}_0 \cup \{h\}$ for any irreproducible $h \in \mathcal{D}_t$. In this setting, we can obtain an approximate p value for each $i \in \mathcal{D}_t$ by calculating ΔW_{-i} statistic using the set $\mathcal{C}_0 \cup \{i\}$ and the global null asymptotic result from Theorem 4.2.1 to calculate an approximate p-value. Algorithm 5 describes our proposed procedure to obtain approximate p-values using the global null asymptotic distribution.

Algorithm 5 Global null procedure.

1: Consider $i \in \mathcal{D}_t$, calculate ΔW_{-i} in the manner of (4.2) with $\mathcal{C}_0 \cup \{i\}$ as the overall set of hypotheses.
2: Obtain the approximate p-value by

$$p_i^{\text{gn}} = 1 - F_V\left((n_0 + 1)\Delta W_{-i}\right)$$

where F_V is the cumulative distribution function of the variable V from Theorem 4.2.1.

Notice, if $\mathcal{C}_0$ is not contaminated with any reproducible hypotheses, then by Theorem 4.2.1, for any $i \in \mathcal{H}_0 \cap \mathcal{D}_t$ it holds that

$$\lim_{n_0 \to \infty} \mathbb{P}(n_0 + 1)\Delta W_{-i} < w) = F_V(w). \tag{4.3}$$

Thus, using p_i^{gn} as an approximate p-value is appropriate. The asymptotic convergence result relies on the global null assumption. So this procedure is no longer appropriate when the control set $\mathcal{C}_0$ is contaminated with some reproducible hypotheses. Under this circumstance, the global null hypothesis is no longer met for $\mathcal{C}_0$, and thus the equality in 4.3 no longer holds. Thus, the method is beneficial when the control set is perfectly selected, but it is not valid when the quality of the control set is not guaranteed. We now introduce two additional procedures based on the ΔW_{-i} statistic.

Conformal procedure

Conformal inference (Vovk et al., 2005) is a statistical framework used to quantify variability in predictions from different machine learning algorithms. Bates et al. (2023) use the general conformal setting to calculate valid p-values in the outlier detection problem. Specifically, they used a dataset of n_0 "control" observations, all drawn from the same unknown distribution, P_X. The aim, then, is to use this control set to identify *outlying* observations (those not distributed by P_X) and *inlying* observations (distributed by P_X) from a test set. To do so, Bates et al. (2023) randomly split the control set into two sections: one used to calculate a score, $\widehat{s}(X_i)$, that measures each test point's alignment with P_X and a second portion to calibrate a valid p-value using the empirical distribution of $\widehat{s}(X_i)$ for points in the calibration set.

Many multiple testing procedures are robust under certain p-value dependence structures. One of those p-value dependence structures is *positive regression dependent on a subset* (PRDS).

Definition 4.2.1. *A random vector* $\mathbf{X} = (X_1, X_2, \cdots, X_n)$ *is PRDS on set* $I_0 \subset \{1, 2, \cdots n\}$ *if for any* $i \in I_0$ *and any increasing set A, then* $\mathbb{P}(\mathbf{X} \in A | X_i = x)$ *is increasing in terms of* x.

A major advantage in the multiple testing setting of this conformal approach is that if the following

conditions are met, Bates et al. (2023) show that conformal p-values are PRDS and thus applying the Benjamini-Hochberg (BH) procedure (Benjamini and Hochberg, 1995) to the p-values controls FDR at a desired nominal level of $\pi_0\alpha$ where π_0 is the proportion of hypotheses in $\mathcal{D}_t$ that are irreproducible.

Conformal conditions:

1. All points in the control set are identically distributed.
2. The joint distribution of the control set of points is invariant to ordering or exchanges.
3. All inliers in the test set are exchangeable relative to the control set and the other inliers in the test set.

These assumptions closely align with our definition of irreproducibility from 4.1.1. Now we can leverage the conformal framework to calculate approximate p-values where ΔW_{-i} is used as the score, $\widehat{s}(X_i)$ for each $i \in \mathcal{D}_t$. We use $\mathcal{C}_0$ as the control set of data to identify reproducible (outliers) and irreproducible (inliers) hypotheses in $\mathcal{D}_t$. Algorithm 6 details our conformal procedure for obtaining approximate p-values using the ΔW_{-i} statistics.

Algorithm 6 Conformal testing procedure.

1: Randomly split the set of control hypotheses, $\mathcal{C}_0$ into training and calibration sets, denoted $\mathcal{C}_0^t$ and $\mathcal{C}_0^c$ respectively.
2: For each $\mathbb{H}_i \in \mathcal{D}_t \cup \mathcal{C}_0^c$, calculate ΔW_{-i} in the manner of (4.2) with $\mathcal{C}_0^t \cup \{i\}$ as the overall set of hypotheses.
3: For each $i \in \mathcal{D}_t$, calculate conformal p-values by the form

$$p_i^{\text{con}} = \frac{\sum_{\ell \in \mathcal{C}_0^c} \mathbb{I}\left[\Delta W_{-i} < \Delta W_{-\ell}\right] + \left\lceil U_i \left(1 + \sum_{\ell \in \mathcal{C}_0^c} \mathbb{I}\left[\Delta W_{-i} = \Delta W_{-\ell}\right]\right)\right\rceil}{1 + |\mathcal{C}_0^c|}$$

where $U_i \overset{\text{iid}}{\sim} \text{UNIF}(0,1)$.

The inclusion of the U_i in the conformal p-value comes from the fact that ΔW_{-i} is discrete, as it is a function of ranks and thus there is a non-zero probability of ΔW_{-i} statistics that are tied in

magnitude. These p_i^{con} values are effectively scaled rankings of ΔW_{-i} statistics with ties decided randomly between the hypothesis of interest and the calibration set $\mathcal{C}_0^c$. Proposition 4.2.1 states that if $\mathcal{C}_0$ is well selected, then calculating p_i^{con} in the manner above yields PRDS p-values. Notice, by the definition of irreproducibility from Definition 4.1.1, we know that $\mathbf{t}_i$ for any $i \in \mathcal{H}_0$ are independently distributed by the product of the distributions, F_j for $j = \{1, 2, \cdots, m\}$ and that all $\mathbf{t}_i$ are independent of each other for all $i \in \mathcal{H}_0$. Thus, if $\mathcal{C}_0$ is a random sampling of all irreproducible hypotheses, then the identical and independent conditions on the $\mathbf{t}_i$ values are maintained. Thus, the conformal conditions are met. The identical distribution conformal condition is immediately implied by our definition of irreproducibility. By independence, the $\mathbf{t}_i$'s are also exchangeable, so the second condition is met. Finally, since the identical and independent assumptions hold for all irreproducible hypotheses, they are exchangeable and identically distributed, and the third condition is met. So the p_i^{con} are valid conformal p-values and the implications of Theorem 2 from Bates et al. (2023) are immediately extendable. That is, Proposition 4.2.1 establishes p_i^{con} to be PRDS for irreproducible hypotheses.

Proposition 4.2.1. *Suppose $\mathcal{C}_0$ is a random sampled from $\mathcal{H}_0$. Consider $i \in \{1, 2, \cdots, d\}$ hypotheses in $\mathcal{D}_t$ and assume they are irreproducible. Then, the vector of conformal p-values $(p_1^{\text{con}}, p_2^{\text{con}}, \cdots, p_d^{\text{con}})$ pertaining to these hypotheses are* positive regression dependent on a subset *(PRDS).*

Additionally, if $\mathcal{C}_0$ is a random sample from $\mathcal{H}_0$, the conformal p-values, p_i^{con}, have a distribution that is super-uniform for irreproducible hypotheses. That is,

$$\mathbb{P}(p_i^{\text{con}} \leq u) \leq u \quad \forall u \in (0, 1).$$

Remark 4.2.1. *Benjamini and Yekutieli (2001) show that if test statistics are PRDS on the set of null hypotheses the Benjamini-Hochberg (Benjamini and Hochberg, 1995) method, as discussed in Section 4.2.4, can be applied at a nominal level of α to control FDR at $\pi_0\alpha$. Thus, Proposition 4.2.1 allows us to use the conformal p-values to discover reproducible hypotheses with FDR control.*

Similar to the global null procedure, the theoretical validity of the conformal approach relies on the global null hypothesis holding for the control set. Additionally, although this procedure is quite nice when the control set is large, it has shortcomings when there are few control hypotheses. By construction, the conformal p-values, p_i^{con}, are limited to the set $\left\{\frac{1}{|\mathcal{C}_0^c|+1}, \frac{2}{|\mathcal{C}_0^c|+1}, \cdots, 1\right\}$. When $\mathcal{C}_0^{\text{con}}$ is limited in size, the proposed method is not precise in the calibration of the p-values, which makes it difficult to discover reproducible hypotheses. A limited calibration also yields large variances in results from application to application. Thus, in cases where we have few controls, one might consider an uneven split of the control that results in a larger calibration set and better estimation of the conformal p-values. However, this is not a perfect solution. Reducing the size of the training set $\mathcal{C}_0^t$ results in less separation between the distribution of ΔW_{-i} statistics from reproducible hypotheses and irreproducible hypotheses, since there are fewer hypotheses in the training set to calculate these statistics. Furthermore, when there are few known control hypotheses, regardless of the split used, the size calibration set is still bounded by the total number of control hypotheses. As an alternative, we propose a different procedure that resamples the control data independently to create more sampled control data and break any control set dependence inherited from contamination.

Bootstrap procedure

The final procedure we introduce uses bootstrap resampling of control summary statistics. This procedure is motivated by cases in which the control set is small and contaminated, and thus the global null asymptotic distribution is not valid, and the conformal p-values are not well calibrated. By resampling sets of control summary statistics, we can get as many as $(n_0)^m$ new irreproducible sets of summary statistics. In this framework, we sample new versions of irreproducible summary statistics from $\mathcal{C}_0$ for each experiment independently. Once a new version of irreproducible summary statistics is sampled, it replaces the hypothesis of interest from the test set, and a new version of ΔW_{-i} is calculated. This process is repeated to build a sampled reference distribution of ΔW_{-i} statistics for assuming i was an irreproducible hypothesis and all other hypotheses remained the same. That resampled distribution is then used to calculate the quantile of the observed ΔW_{-i} statistic compared to the reference distribution, which is used as an approximate p-value. Algo-

rithm 7 outlines the procedure that leverages bootstrap sampling to obtain approximate p-values for each hypothesis. This procedure is beneficial for a few reasons. First, because we resample

Algorithm 7 Bootstrap testing procedure.

1: For each $i \in \mathcal{D}_t$, calculate ΔW_{-i} in the manner of (4.2) with $\mathcal{D}_t$ as the overall set of hypotheses.
2: Sample B new vectors of summary statistics, denote each set by $\mathbf{t}_i^b = (t_{1i}^b, t_{2i}^b, \cdots, t_{mi}^b)$ where t_{ji}^b is randomly sampled from $\{t_{jh} : h \in \mathcal{C}_0\}$ independently for each $j \in \{1, 2, \cdots, m\}$. Call this the set of these vectors $\mathcal{B}_0$.
3: Replace the summary statistics from $\mathbb{H}_i$, $\mathbf{t}_i$ by $\mathbf{t}_i^b$ and calculate ΔW_{-i}^b in the manner of (4.2) with $(\mathcal{D}_t \backslash \{i\}) \cup \{b\}$ for all $b \in \mathcal{B}_0$.
4: For each $i \in \mathcal{D}_t$, calculate the approximate p-values by the form

$$p_i^{\text{boot}} = \frac{1 + \sum_{b=1}^{B} \mathbb{I}\left[\Delta W_{-i} < \Delta W_{-i}^b\right]}{1 + B}.$$

from the control set, the bootstrapped irreproducible reference distribution can be built using an adequate number of samples, even when the size of $\mathcal{C}_0$ is limited. Additionally, when the control set is contaminated, the bootstrap resampling mitigates issues caused by the across-study dependence present in the control set. Sampling from each experiment independently ensures the independence of summary statistics across studies for each bootstrapped irreproducible hypothesis, eliminating the dependence structure from contamination of the control set.

Method discussion

Figure 4.3 provides an outline of each of the methods. Notably, the global null and conformal procedures calculate the ΔW_{-i} statistics for each hypothesis in the test set using some subset of the control hypotheses. Their respective approximate p-values examine the misalignment of these ΔW_{-i} statistics with a reference distribution – an asymptotic distribution for the global null and the empirical distribution from the calibration set in the conformal setting. For this reason, the control must contain only irreproducible hypotheses when applying the global null or conformal method. In the global null setting, contamination of the control set results in misalignment ΔW_{-i} statistics for irreproducible hypotheses that are misaligned with the reference asymptotic distribution. For

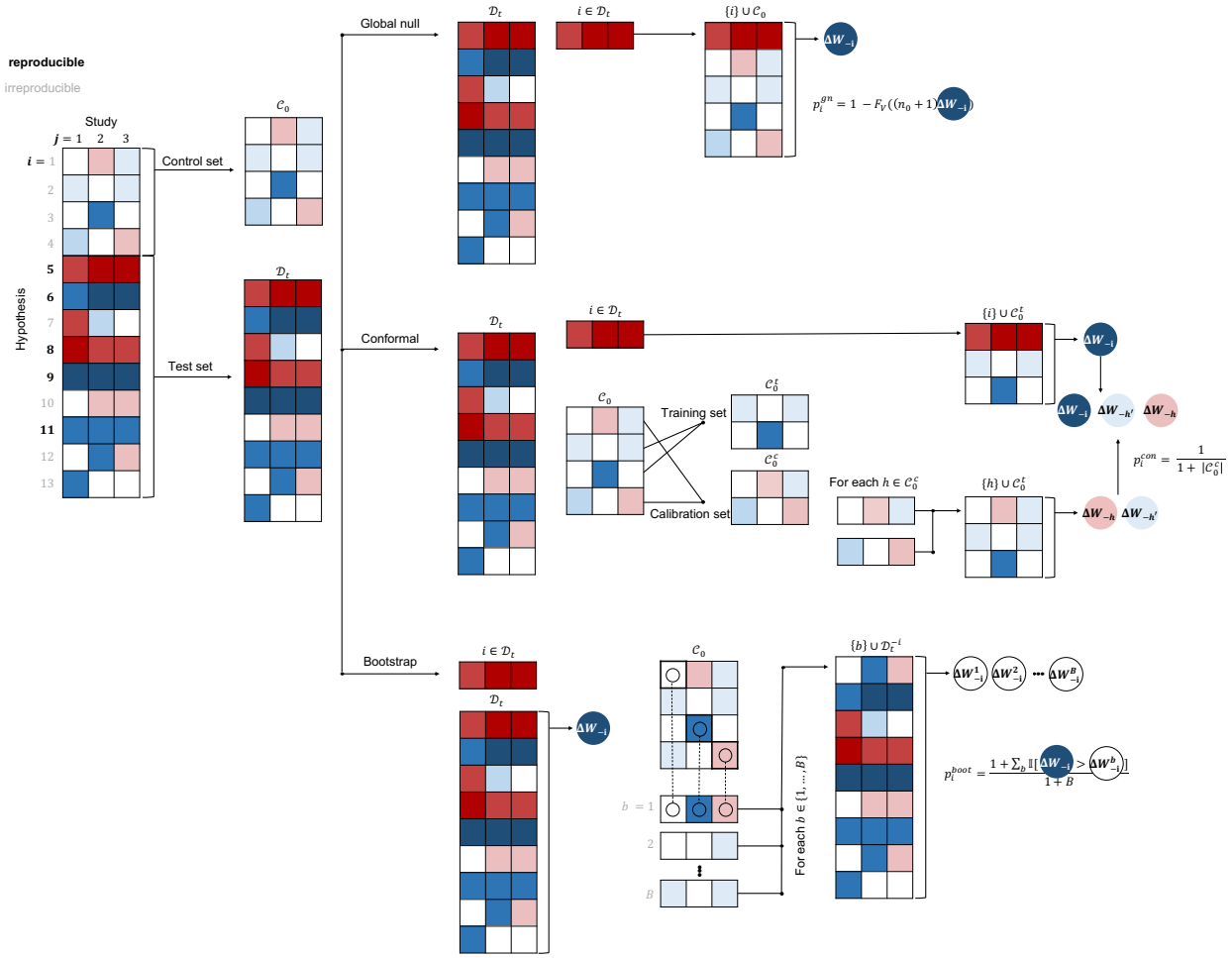


Figure 4.3: Outline of the three proposed ΔW_{-i} procedures. The bolded hypotheses are reproducible. Red represents a negative summary statistic, and blue represents a positive summary statistic. The darkness of each color represents the notability of the observed summary statistic.

the conformal method, contamination of the control set is problematic because the calibration set containing reproducible hypotheses yields a poor reference empirical distribution to calibrate approximate p-values. The bootstrap approach, on the other hand, calculates the ΔW_{-i} statistic for each hypothesis in the test set using the test set and builds the irreproducible reference distribution by sampling summary from the control set independently across study. The method is more robust against contamination, as the dependence of reproducible hypotheses is broken by independent sampling.

If we assume the control set is uncontaminated, then the global null and conformal approaches have theoretical justifications, while a careful examination of the theoretical properties of the boot-

strap procedure under this assumption is left for future work. Specifically, we show the asymptotic convergence of irreproducible ΔW_{-i} statistics to that of the global reference distribution used to obtain the approximate p-values, and for the conformal approach, we leverage results from the outlier detection problem to show that we can apply Benjamini-Hochberg (Benjamini and Hochberg, 1995) to the conformal approximate p-values and control false discovery rate on average. To appeal to these theoretical results, however, it is important to have a large enough control set because the global null result is asymptotic, and the quality of the calibration of the conformal p-value depends on the size of the calibration set.

In practice, the quality of the control set cannot be guaranteed, and the number of control hypotheses is often limited, so we recommend practitioners use the bootstrap procedure. As previously discussed, independently sampling new control hypotheses alleviates both issues. In contexts where the control set is not limited in size and known to be uncontaminated, the global null and conformal approaches might be more appropriate, but in the high-throughput genomic setting, that is not the case.

4.2.4 FDR thresholding

Each of these three procedures can be used to obtain approximate p-values assessing the reproducibility of each hypothesis. To yield a set of reproducible hypotheses with simultaneous control of false discovery, we must apply a multiple-testing correction to the approximate p-values. When evaluating the reproducibility of n hypotheses, the primary concern typically is to discover a subset of hypotheses to be reproducible such that a nominal level of false discovery rate (FDR). That is, we denote V as the number of hypotheses in the reproducible set that are truly irreproducible and Q as the total size of that reproducible set. Then, Benjamini and Hochberg (1995) introduced the notions of false discovery proportion (FDP) and rate (FDR) as

$$\mathrm{FDP} = \frac{V}{Q \vee 1} \text{ and } \mathrm{FDR} = \mathbb{E}\left[\mathrm{FDP}\right].$$

Given a vector of valid p-values for $\mathcal{D}_t$, the most popular manner to control FDR in a multiple testing setting is to use the Benjamini-Hochberg (BH) adjustment (Benjamini and Hochberg, 1995). Which, for a nominal FDR level of α, finds the largest k such that

$$p_{(k)} \leq \frac{k\alpha}{n_1}$$

$p_{(k)}$ is the k^{th} order statistic among all p-values pertaining to hypotheses in $\mathcal{D}_t$. Then, the BH adjustment defines the set of hypotheses to be reproducible by

$$\{i \in \mathcal{D}_t : p_i \leq p_{(k)}\}.$$

To discover a reproducible hypothesis at a nominal FDR level of α, we apply the BH adjustment to the approximate p-values produced by Algorithms 5, 6, and 7.

4.3 Simulations

We conduct simulation studies to examine the performance and robustness of the three proposed methods. The simulation setting generalizes the bivariate Gaussian mixture used to define the copula model in Li et al. (2011) to cases where there exist $m > 2$ replicate experiments. The bivariate version of this setting has frequently been used for nonparametric procedures designed to examine reproducibility for pairs of studies (Philtron et al., 2018; Ghosh et al., 2021). Within the setting, described in detail in Section 4.3.1, reproducible and irreproducible hypotheses are differentiated by parameters that represent the strength of the signal and across-study signal consistency for reproducible hypotheses.

In these simulations we are interested in examining two properties: 1) each of the proposed procedure's abilities to produce approximate p-value that are neither anti-conservative nor overly conservative on average and well-calibrated from application to application; and 2) each of the procedure's ability to detect reproducible hypotheses with relatively high power while controlling FDR at the nominal level. Results for the first properties are discussed in Section 4.3.2, where the

empirical distributions of approximate p-values for irreproducible hypotheses across 50 iterations and their respective mean are examined when the control set is perfectly selected and when the control set is contaminated. In Section 4.3.2 we assess the second property by comparing the distributions of power and FDP for each of the ΔW_{-i} procedures to the AdaFilter procedure (Wang et al., 2022). AdaFilter assesses if a result has been reproduced in at least r out of m by filtering the entire set and then selecting hypotheses based on the order statistics of their p-values. We apply AdaFilter with both $r = m$ and $r = m - 1$ to examine m and $m - 1$ out of m reproducibility in the manner suggested in Wang et al. (2022) that avoids discovering reproducible hypotheses with different signs.

We calculate approximate p-values using a random, 50-50 training-calibration split of the control set, and for the bootstrapping approach, we consider $B = 500$ bootstrap samples from the control. The AdaFilter procedure was implemented using the `adaFilter` package in `R`.

4.3.1 Simulation settings

In each iteration of these simulations, we consider $|\mathcal{D}_t| = n_1$ hypotheses in the test set and $|\mathcal{C}_0| = n_0$ hypotheses in the control set common across m replicate studies. For each hypothesis we observe summary statistics, denoted by $\mathbf{t}_g = (t_{1g}, t_{2g}, \cdots, t_{mg})$. In the test set, π_1 proportion of the n_1 hypotheses are reproducible, and the remaining $(1 - \pi_1)$ proportion are irreproducible. In the control set, π_{00} proportion of the n_0 hypotheses are irreproducible while $(1 - \pi_{00})$ proportion are reproducible, representing the contamination of $\mathcal{C}_0$. For a reproducible hypothesis g, we simulate the associated summary statistics, $\mathbf{t}_g$, by the m-variate Gaussian distribution in (4.4).

$$\mathbf{t}_g \sim \mathbb{N}\left(\mu_g \mathbf{1}_m;\ \sigma^2[\rho \mathbf{1}_m \mathbf{1}_m^\top + (1-\rho)\mathbf{I}_{m\times m}]\right) \tag{4.4}$$

where $\mu_g = \mu$ for half of the reproducible hypotheses and $\mu_g = -\mu$ for the other half of the reproducible hypotheses. Notice that μ represents the strength of the signal, σ_g^2 represents the variability of the signal, and ρ_g represents the consistency of the signal for hypothesis g across replicate studies. If hypothesis h is irreproducible, we simulate the associated summary statistics, $\mathbf{t}_h$, by m

independently by standard Gaussian random variables, as in (4.5).

$$\mathbf{t}_h \sim \mathbb{N}\left(\mathbf{0}_m;\ \mathbf{I}_{m\times m}\right). \tag{4.5}$$

For these simulations, we fix $n_1 = 1000$, $\pi_1 = 0.5$, and $\sigma_1^2 = 1$. We repeat each setting with $n_0 \in \{100, 500\}$ and $m \in \{3, 6\}$ to examine differences in performance between different control set sizes and numbers of experiments. Now, there are three remaining unknown parameters (π_{00}, μ, ρ), each representative of an important capability of a method in detecting reproducibility. π_{00} represents the quality of $\mathcal{C}_0$, as $1 - \pi_{00}$ is the contamination proportion. μ_g and ρ_g represent the general strength and consistency of the signal in all experiments that the reproducible hypotheses demonstrate.

For each simulation, we fix the majority of parameters at a default value and allow one of the parameters to change. The default values selected for each of the unknowns are $\pi_{00} = 1$, $\mu = 1.5$, and $\rho = 0.7$. Note that the default setting has no contamination of the control set. We examine results from 50 iterations of the default setting and 50 iterations of six other settings, each with one of the three unknown parameters changed with the default values for the remaining unknowns. We additionally consider $\pi_{00} \in \{0.98, 0.95\}$ representing 2% and 5% contamination of the control set. As for signal strength of the reproducible hypotheses, we additionally consider $\mu \in \{1, 2\}$ to represent weaker and stronger signal strength for reproducible hypotheses. We also let $\rho \in \{0.3, 0.9\}$ represent weak and strong signal consistency for reproducible hypotheses.

4.3.2 Simulation results

Approximate p-value results

Figure 4.4 displays the empirical distribution function of approximate p-values of all irreproducible hypotheses in the test set for each of the three ΔW_{-i} procedures of 50 iterations of the proposed and their average over 50 iterations. In a. all default values for the parameters are used, and in b. we let the control set be 5% contaminated ($\pi_{00} = 0.95$). Under both settings, we consider control sets of sizes of 100 and 500 to examine how the amount of control data one has impacts

the performance of the methods. When there is no contamination of the control set, the average

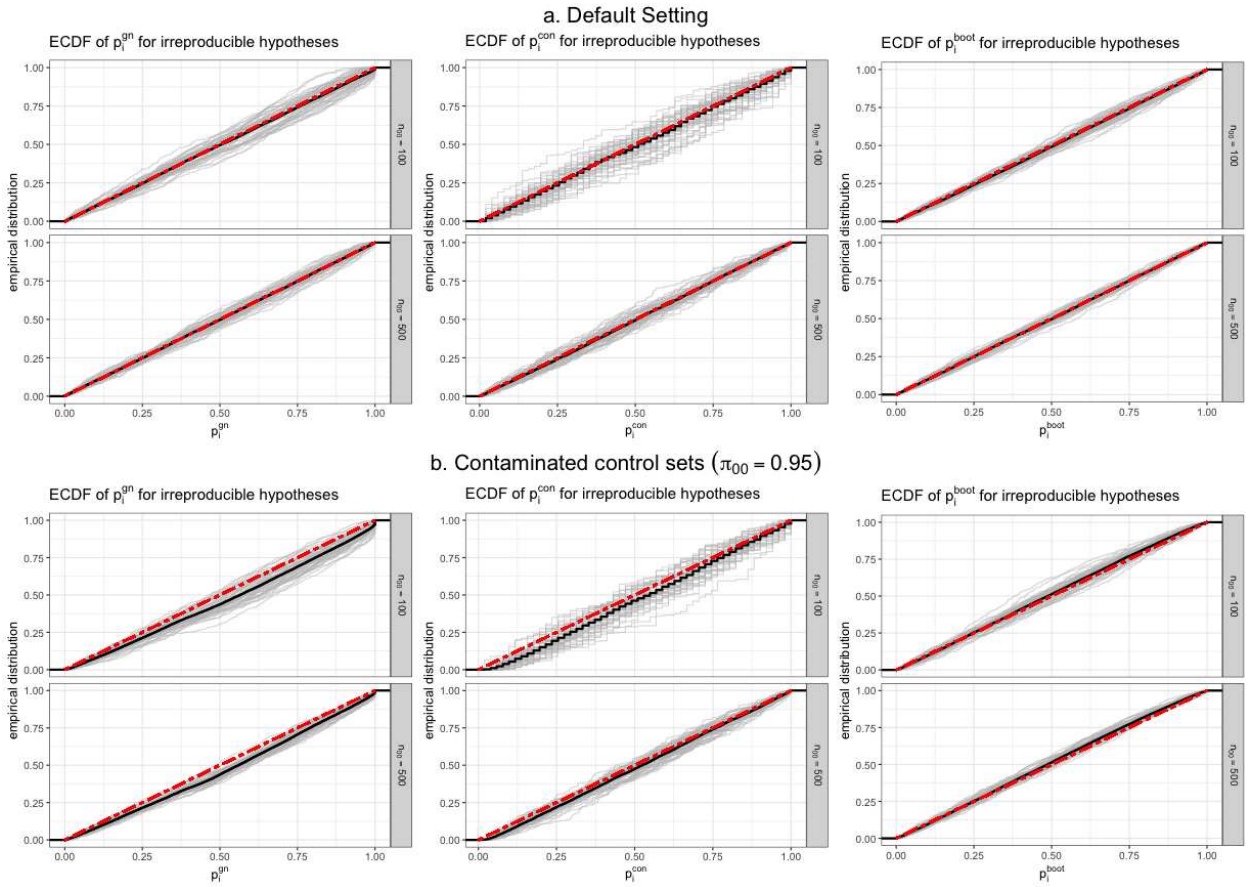


Figure 4.4: Empirical cumulative distribution functions of irreproducible hypotheses' approximate p-values from each of the ΔW_{-i} methods from 50 iterations of proposed $m = 3$ simulation setting (in gray) and their average (in black) for a. the default values for all parameters and b. with 5% contamination of $\mathcal{C}_0$ ($\pi_{00} = 0.95$) and the default values for all other parameters. In both cases, we consider $|\mathcal{C}_0| \in \{100, 500\}$.

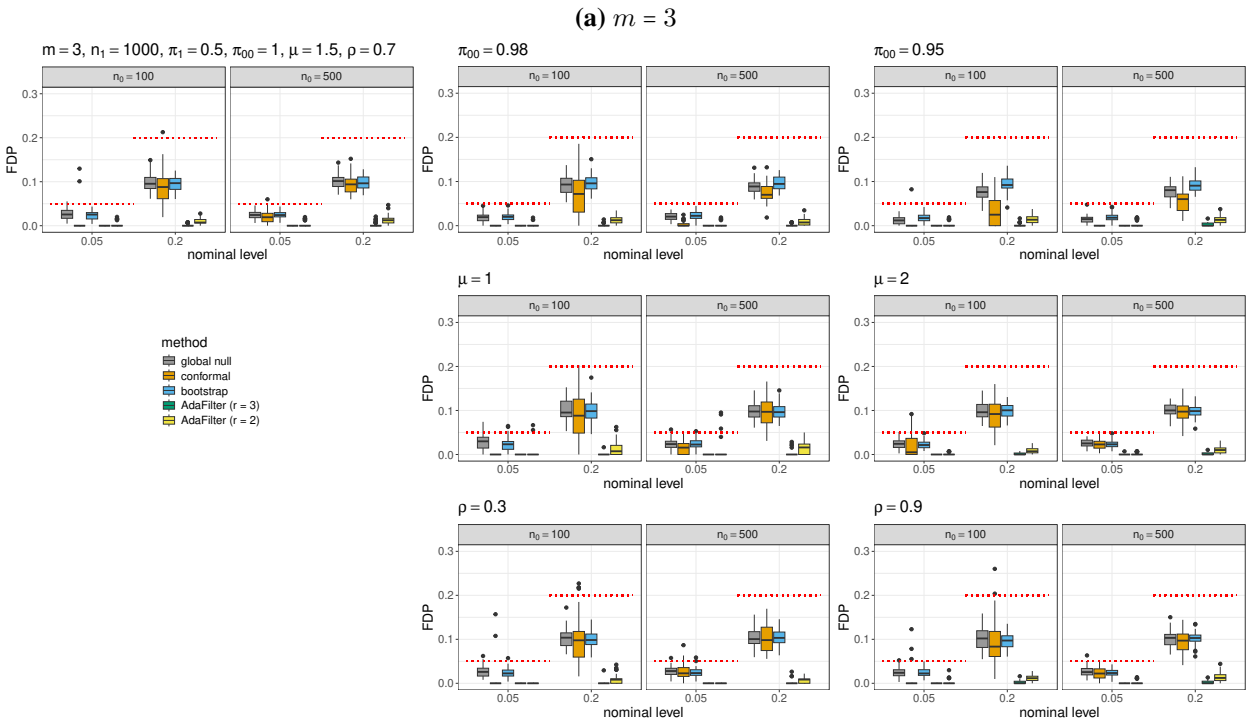
empirical distribution of irreproducible p-values for each of the three methods falls directly on the diagonal. This means that each of the procedures produces approximate p-values that are neither conservative nor anti-conservative on average. That is, under no contamination, these results suggest that for any $t \in (0, 1)$, $\mathbb{E}\left(|\mathcal{D}_t \cap \mathcal{H}_0|^{-1} \sum_{i \in \mathcal{D}_t \cap \mathcal{H}_0} \mathbb{I}[p_i \leq t]\right) \approx t$. When contamination exists, we see that both the global null and conformal approaches tend to have conservative approximate p-values on average, demonstrated by the average empirical distribution line falling below the red, dashed diagonal. As was discussed in Section 4.2, the bootstrap approach produces approximate p-values that are more robust against control set contamination, since dependence inherited from

the contaminated hypotheses is broken by bootstrap sampling summary statistics independently from each experiment.

Additionally, these simulations confirm that when the size of the control set is small, the conformal approximate p-values are not well calibrated. In both a. and b., when $|\mathcal{C}_0| = 100$, we see that there is wide variability in the gray empirical distributions from each iteration. This variability occurs because the p-values are calibrated using only $|\mathcal{C}_0|/2$ hypotheses. This variability is much alleviated when $|\mathcal{C}_0| = 500$. For the other two methods, the change in variability is far less pronounced, as the global null procedure benefits from calculating approximate p-values using an asymptotic distribution and the bootstrap procedure resamples from the control set to create a reference distribution of many irreproducible hypotheses.

Power and FDR results

Figure 4.5 contains boxplots of the FDP from 50 iterations of each simulation setting when the ΔW_{-i} procedures and AdaFilter are applied at nominal levels of FDR of $\alpha \in \{0.05, 0.20\}$. The left panels of each plot are the results when $n_0 = 100$ and the right panels are when $n_0 = 500$. The top left corner is the FDP results when all default values for the parameters of interest are used. The plots in the top row contain results as contamination of the control set is increasing (2% and 5% contamination). The second row contains the results when the strength of the reproducible signal is weaker $(\mu = 1)$ and stronger $(\mu = 2)$ than the default value $(\mu = 1.5)$. The third row shows results when the consistency of the reproducible signal is weaker $(\rho = 0.3)$ and $(\rho = 0.9)$ than the default value $(\rho = 0.7)$. For a method to control FDR at a nominal level of α, the average of the FDP would be below α. Notice that all methods control FDR in simulation across all parameter specifications. All methods are at least moderately conservative in each case. The global null procedure and bootstrap procedure are robust against the size of the control set, as we observe similar FDP boxplots across $n_0 = 100$ and $n_0 = 500$ for both methods. As previously discussed, the conformal procedure tends to be overly conservative when the control set is smaller. This is particularly noticeable at smaller nominal levels of FDR and $m = 3$ replicate experiments. With a small control set, the conformal method shows a wide variability in the observed FDP values,



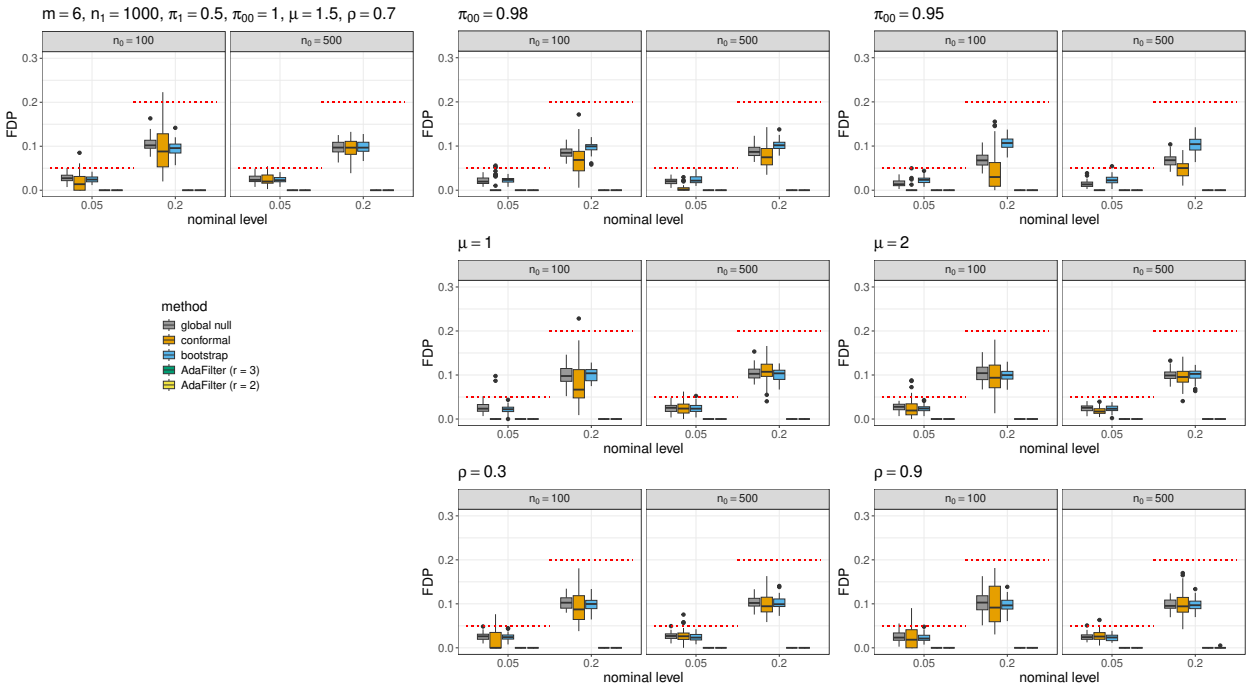


Figure 4.5: FDP values at nominal FDR levels of $\alpha \in \{0.05, 0.20\}$ from 50 iterations of each setting for $m = 3$ in (a) and $m = 6$ in (b) for each of the three ΔW_{-i} procedures and AdaFilter and control set sizes of $|\mathcal{C}_0| \in \{100, 500\}$. The dashed line represents the nominal level of FDR.

but when n_0 increases to $n_0 = 500$, the variability in FDP decreases. Both of these properties were expected, as a small control set implies poor calibration of conformal approximate p-values. Additionally, as expected, the bootstrap approach to the method is robust against contamination of the control set in simulation, while the other two methods become more conservative as π_{00} decreases. This is most notable when there are more replicate experiments. In (b), we see the boxplots for FDP of the global null and conformal procedures decrease along the top row, while the bootstrap does not.

Figure 4.6 shows boxplots of the power from each of the same 50 iterations seen in Figure 4.5. The poor calibration of approximate p-values when there is little data in the control set often yields little power for the conformal method. With a larger control set, the power for the conformal method is more in line with the other two procedures. Performance for the other two methods does not depend on the size of the control set. Outside of the conformal method with a small control set, all of the methods show pronounced power advantages over both AdaFilter procedures under this simulation setting. It is easy to see in Figure 4.5 that AdaFilter is overly conservative under this simulation setting, making the method underpowered compared to the less conservative ΔW_{-i} methods. In part, this is because ΔW_{-i} and other rank-based statistics do not rely on the raw magnitudes of the reproducible signal, but rather the alignment of these hypotheses relative to all other hypotheses, while AdaFilter is based on the magnitudes of p-values for a hypothesis.

Further, comparing the three ΔW_{-i} procedures to each other, we see similar power results when the control set is not contaminated and there are $m = 3$ replicate studies. With $m = 6$ studies, the bootstrap procedure enjoyed a slight advantage compared to the other two methods. Additionally, there were some slight differences between the methods for differing levels of reproducible signal consistency. Notice when the signal for reproducible hypotheses is less consistent ($\rho = 0.3$), the global null and conformal procedures tended to have slightly higher power than the bootstrap procedure, and when the signal for reproducible hypotheses are highly consistent across study ($\rho = 0.9$) the bootstrap procedure shows power advantages compared to the other two methods. As expected, the robustness of the bootstrap procedure against contamination of the control set

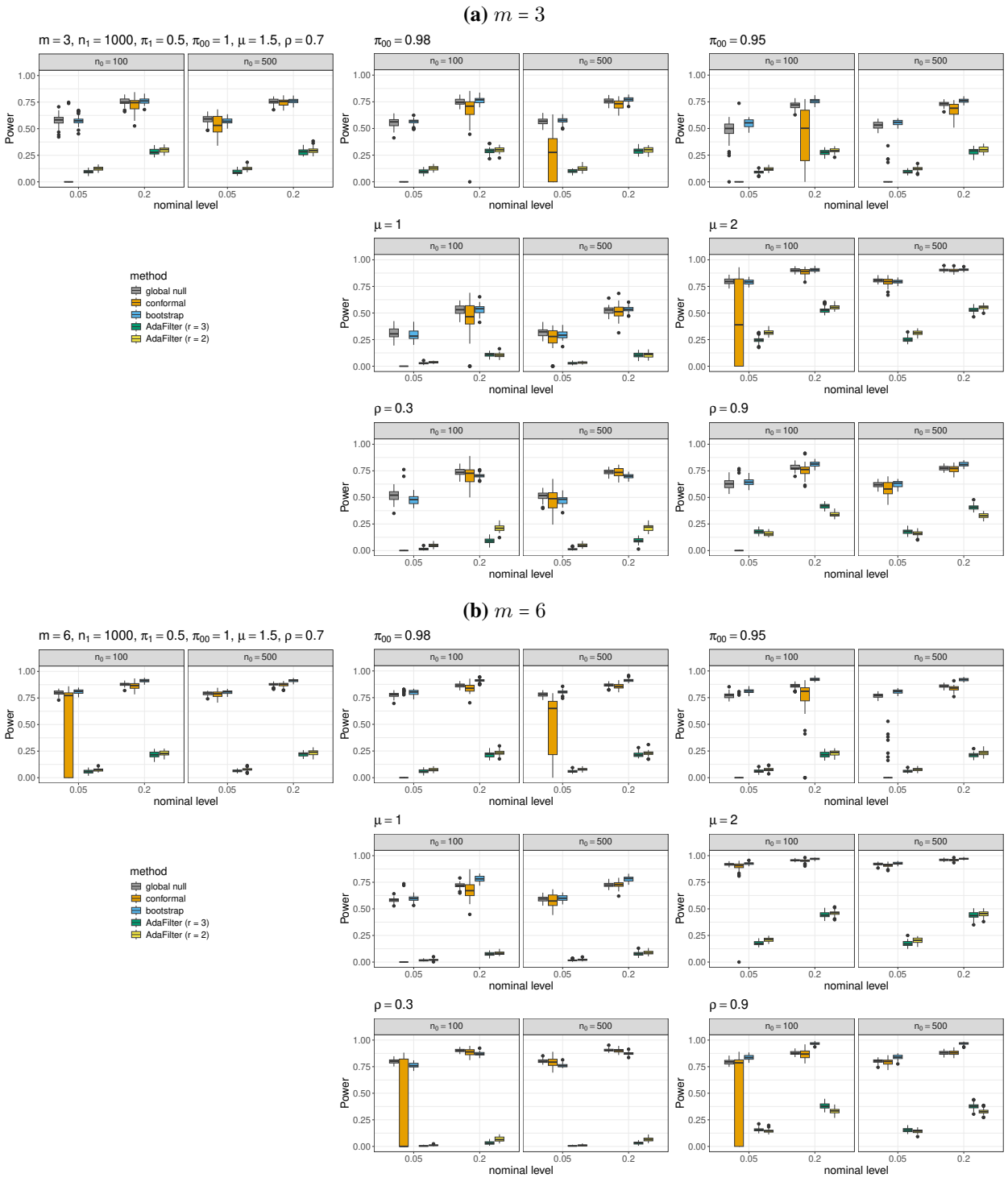


Figure 4.6: Power values at nominal FDR levels of $\alpha \in \{0.05, 0.20\}$ from 50 iterations of each setting for $m = 3$ in (a) and $m = 6$ in (b) for each of the three ΔW_{-i} procedures and AdaFilter and control set sizes of $|\mathcal{C}_0| \in \{100, 500\}$.

resulted in that method dominating the others when the control set was contaminated, particularly when there were more replicate studies.

Ultimately, these results were consistent with the discussions of the advantages and disadvantages of each method from Section 4.2. In practical applications where the control set is limited in size and often contaminated, it is recommended to consider the bootstrap approach, as it enjoys advantages compared to the other two procedures in that setting.

4.4 Application to COVID-19 datasets

We apply the proposed bootstrapping approach to ΔW_{-i} with existing methods in reproducibility and meta-analysis to differential expression results from five independent multi-patient single-cell RNA sequencing (scRNA-seq) datasets, each measured on peripheral blood mononuclear cells (PBMCs) of patients with COVID-19 and control patients. These datasets are a subset of a larger set of scRNA-seq studies examining COVID-19 that were previously processed by Lin et al. (2022) and then integrated by Wang et al. (2023) to classify and predict COVID-19 disease severity. It is often of interest in scRNA-seq studies to understand cell type-specific responses to diseases. Thus, the development and application of differential expression methods to expression data from a particular cell type has become a common interest (Crowell et al., 2020; Squair et al., 2021). Rather than integrating the datasets, our application aims to discover genes – and subsequently gene pathways – that are consistently differentially expressed for a cell type across the five studies by applying the ΔW_{-i} procedure, along with methods from existing literature, and examining the identified genes.

Section 4.4.1 discusses the particular details of these studies, the data processing applied, and the differential expression analyses performed. The details for the reproducibility and meta-analysis methods applied, results of the initial analyses, and a subsequent gene pathway analysis are discussed in Section 4.4.2.

4.4.1 Data processing

In Wang et al. (2023), patients came from one of three COVID-19 severity conditions: Healthy, Mild/Moderate, and Severe/Critical. For the sake of the differential expression analysis, we collapse these classifications into two health conditions: Healthy and COVID-19 (with the Mild/Moderate and Severe/Critical conditions combined). The five studies were selected because they each had an ample number of patients in both conditions, as seen in Table 4.1. In each study, we have an $n_g \times n_c$

Table 4.1: Number of patients in each health condition for the five studies considered.

	Healthy	COVID-19	Total
Liu et al. (2021)	16	54	70
Ren et al. (2021)	20	153	173
Schulte-Schrepping et al. (2020)	51	94	145
Stephenson et al. (2021)	24	102	126
Wilk et al. (2020)	12	43	55

matrix of average gene expressions for each patient. Here, $n_g = 15,955$ represents the number of genes, and $n_c = 18$ is the number of cell types in each study.

The primary differential expression analysis for each study is done using the cell type CD14 Monocyte, as we observed few genes with exactly zero expression and a high level of consistency across study when considering CD14 Monocyte compared to other cell types. CD14 Monocyte cells are also found to be the most predictive of the severity of COVID-19 patients in Lin et al. (2023). Squair et al. (2021) found pseudobulk RNA-seq methods to be preferable to those developed for scRNA data, so we perform differential expression analysis using the size-factor normalized gene-level expression data for CD14 Monocyte cells for each of the five studies using the `limma` package Ritchie et al. (2015) in `R` with condition (Healthy or COVID-19) as the grouping variable to obtain p-values and log-fold change scores for all genes with non-zero expression in at least one patient. The analysis was run independently on each of the five studies. The set of results was consolidated by an inner join, so a gene needed to show non-zero expression in each of the five studies to be considered in the application of reproducibility methods. The joined results contain

p-values and log fold change scores for $14,649$ genes. Figure 4.7 displays the volcano plots for the differential expression analyses for each study. While the scale of the observed log fold change

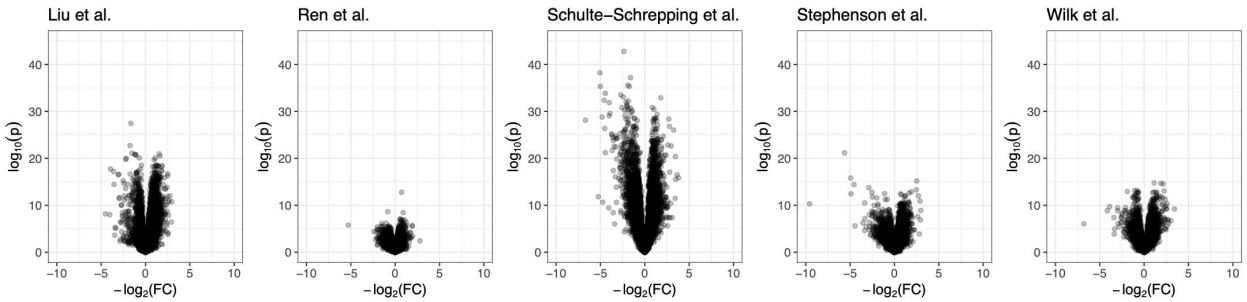


Figure 4.7: Volcano plots from the differential expression analysis for each of the five studies. The x-positioning for each gene is the log fold change score and y-positioning is the $-\log_{10}(p-\text{value})$.

score is relatively similar, it is clear that the observed p-values differ greatly in magnitude across study. In particular, p-values for genes from Liu et al. (2021) and Schulte-Schrepping et al. (2020) show greater magnitude than in the other studies.

The gene-level summary statistic considered for analyzing reproducibility using the ΔW_{-i} procedure was a signed version of the differential expression p-values. That is, let ℓ_{ji} represent and log-fold change score and p_{ji} represent the p-value for gene i in experiment j, then the summary statistic t_{ji} considered in the application of the ΔW_{-i} procedure was as follows

$$t_{ji} = \text{sgn}(\ell_{ji}) \cdot (-\log(p_{ji}))\,. \tag{4.6}$$

We considered this summary statistic because it retains information about both the direction of the observed effect from $\text{sgn}(\ell_{ji})$ and the strength of the observed effect relative to the observed variability from $-\log(p_{ji})$. It is easy to observe that genes with a large observed down-regulating effect will have a large, negative t_{ji} statistic, while those with a large observed up-regulating effect will have a large, positive t_{ji} statistic. Figure 4.8 displays the joint distributions of t_{ji} statistics for each pair of studies in the lower triangle, their associated correlation coefficients in the upper triangle, and the t_{ji} marginal distributions along the diagonal. Note that results from Stephenson

et al. (2021) tended to be the most aligned with results from the other studies, while Ren et al. (2021) tended to be the least aligned with the other studies. Additionally, the marginal distributions of t_{ji} statistics take a similar shape with wide differences in spread. There exists some correlation

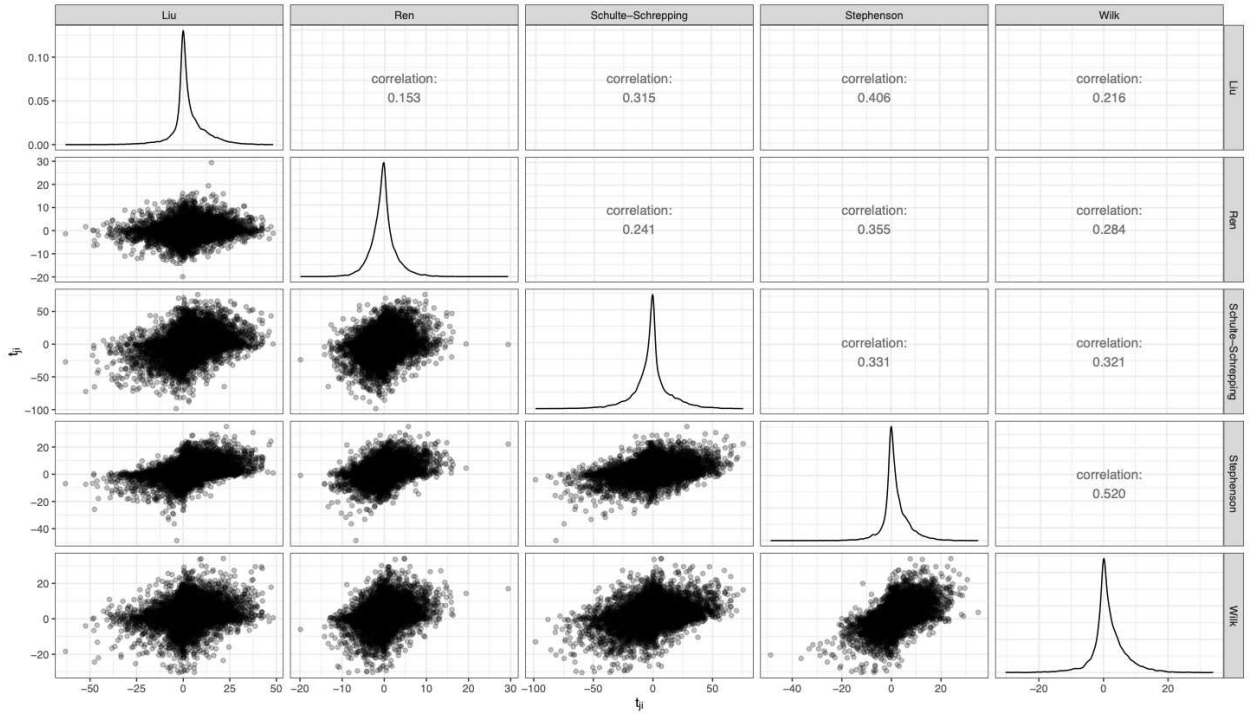


Figure 4.8: The joint distributions of observed t_{ji} statistics for each pair of studies are shown in the bottom triangle with the associated correlation coefficient in the upper triangle. The marginal distribution of t_{ji} statistics for each study is shown along the diagonal.

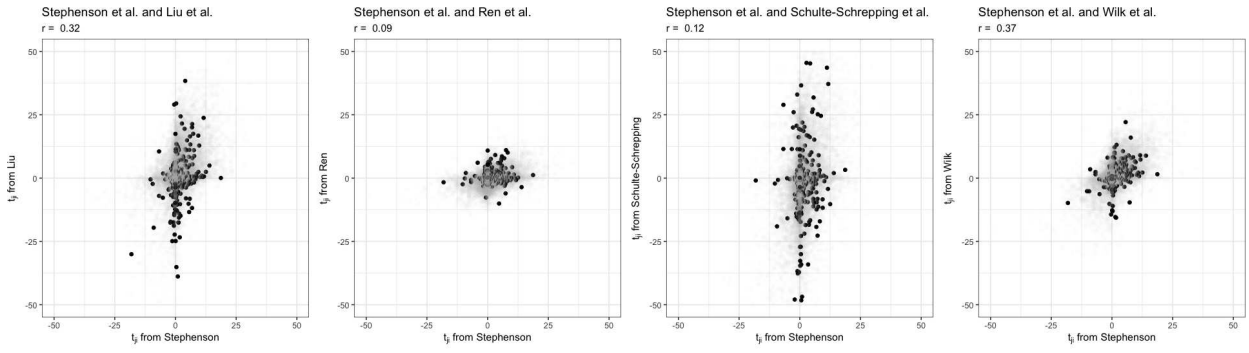


Figure 4.9: Joint distributions of t_{ji} statistics for genes in $\mathcal{C}_0$ from Stephenson et al. (2021) and each of the other studies. The subtitle lists the correlation coefficient for t_{ji} statistics from $\mathcal{C}_0$ for each pair. Stephenson et al. (2021) remained fixed because it had the largest correlation coefficients with the other studies. The joint distributions of all pairs of studies are in Appendix C.3

between the control t_{ji} statistics. This can be observed by examining the correlation coefficients between control t_{ji} statistics from Stephenson et al. (2021) with Liu et al. (2021) ($r = 0.32$) and Wilk et al. (2020) ($r = 0.37$). This observed correlation signals the potential for contamination in $\mathcal{C}_0$. Additionally, it is important to note that the control set is relatively small in size ($|\mathcal{C}_0| = 202$). As previously discussed in Section 4.2 and observed via simulation in Section 4.3, the bootstrap approach to ΔW_{-i} is the most robust of the proposed approaches when $|\mathcal{C}_0|$ is small and the control set is contaminated. For that reason, when examining reproducibility across these studies, we apply only the bootstrap approach to ΔW_{-i}.

4.4.2 Reproducibility results

As discussed in Section 4.4.1, due to the expected contamination of $\mathcal{C}_0$, we employ the proposed bootstrapping approach with $B = 1000$ bootstrap samples to the five COVID-19 datasets to discover reproducible genes. Among replicability/reproducibility methods, we compare the proposed procedure with the AdaFilter method (Wang et al., 2022) using its proposed adaptation to avoid reproducible discoveries with different signs for $r \in \{4, 5\}$ discovering genes that have been reproduced in at least 4 of 5 or all 5 studies, respectively. We also apply a subset of meta-analysis methods from Chang et al. (2013) to emphasize the differences between reproducibility and meta-analysis. Additionally, we apply the robust rank aggregation (RRA) approach proposed in Kolde et al. (2012) that is frequently used in practice to check for signals across multiple replicate studies. A brief description of each method is provided below.

Fisher's: Fisher's method (Fisher, 1925) uses $S_i^F = -2\sum_{j=1}^{m}\log(p_{ji})$. where p_{ji} is the differential expression p-value to test whether gene i is differentially expressed any of the $j \in \{1, 2, \cdots, m\}$ studies.

max$P$: the max$P$ method (Wilkinson, 1951) uses $\max P_i = \max_{j\in\{1,2,\cdots,m\}} p_{ji}$ to test whether gene i shows signal in all m studies.

rOP: The rOP method (Song and Tseng, 2014) generalizes the framework used in $\max P$ to examine r/m reproducibility. The method uses the r^{th} order statistic among the p values

related to the gene i ($p_{(r)i}$) for some $r > m/2$ to test whether a gene is differentially expressed in a "majority" of studies. We consider $r = 4$.

RRA: The robust rank aggregation method (RRA) (Kolde et al., 2012) fits less into the typical p-value combining meta-analysis methods and instead aims to find genes that are highly ranked in many studies. The method uses a binomial model to examine the probability that a gene i is among the top-ranked genes in at least k experiments if it is truly null and uses the minimum of these probabilities across different values of k as a test statistic to find "genes that are highly ranked in many preference lists" while ignoring noninformative studies.

Owen (2009) proposes an adaptation to avoid the aggregation of different signed signals, and similarly, Song and Tseng (2014) presents signed adaptations for both the rOP and max$P$ methods. Thus, for those three methods, we apply the versions that avoid different signed discoveries.

Table 4.2 contains the number of genes deemed reproducible (or significant for the meta-analysis context) for each method at the nominal FDR levels of $\alpha \in \{0.10, 0.20\}$. Notice that

Table 4.2: Total number of genes deemed to be reproducible (or significant) for each method at nominal FDR levels of $\alpha \in \{0.10, 0.20\}$.

	$\alpha = 0.10$	0.20
ΔW_{-i} bootstrap	882	1547
AdaFilter $(r = 5)$	243	328
AdaFilter $(r = 4)$	973	1350
Fisher's	12182	12860
Max$P$	2062	2485
rOP $(r = 4)$	4805	5713
RRA	177	312

each of the three meta-analysis methods that combine p-values discovers the most reproducible genes. This is expected, as these methods only assess the presence or absence of signal in a certain number of studies, as opposed to the presence *and consistency* of signal in the reproducibility context. Interestingly, the rank-based RRA is underpowered relative to all other methods. Among

reproducibility methods, the proposed ΔW_{-i} discovers more reproducible genes than the AdaFilter procedure when $r = 5$ and a similar number of reproducible genes when $r = 4$. Despite a similar number of reproducible genes to AdaFilter, the geometric advantage of ΔW_{-i} compared to other methods can be seen in Figure 4.10, which displays the geometry of these reproducible sets for each method in terms of t_{ji} from Stephenson et al. (2021) and Schulte-Schrepping et al. (2020).

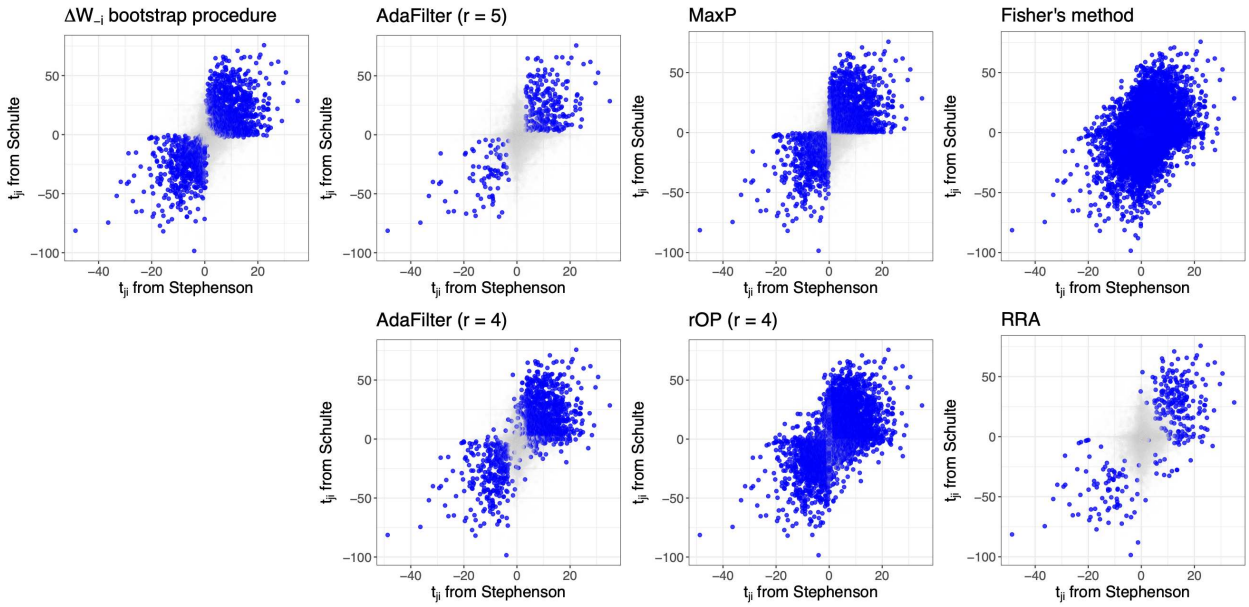


Figure 4.10: Reproducibility (or rejection) regions for a nominal FDR level of $\alpha = 0.20$ for each of the methods considered in terms of t_{ji} statistics from Stephenson et al. (2021) and Schulte-Schrepping et al. (2020). Appendix C.3 contains scatter plots of these regions for all pairs of studies.

Notice that the region that ΔW_{-i} finds to be reproducible is consistent with what one might expect: genes that are consistently down or up-regulated across multiple studies and avoid discovering genes with little signal. This region is preferable to the region discovered by AdaFilter with $5 = 5$, as it is similarly shaped but less restrictive in both the positive and negative directions. AdaFilter with $r = 4$ includes several genes that have t_{ji} statistics in the second and fourth quadrants and does not include nearly as many genes that show consistent but relatively moderately negative signal compared to ΔW_{-i}, which manages to discover those moderately down-regulated genes while avoiding those inconsistent genes. Among meta-analysis methods, Fisher's combining method and rOP are not particularly discerning, and thus their regions are not geometrically inter-

esting in the context of the reproducibility problem. Interestingly, the max$P$ region takes a similar form to that of ΔW_{-i}. A closer examination of differences between these regions can be found in Figure 4.11, where the geometry of the intersections and set differences of the reproducible region for ΔW_{-i} and each other method are inspected. Additionally, Table 4.3 examines the sizes of these intersections and set differences at a nominal FDR level α = 0.20.

Table 4.3: Sizes of the intersections and set differences between the reproducibility (or rejection) regions for ΔW_{-i} and each of the other methods considered at a nominal FDR level of $alpha$ = 0.20. **Both** represents the number of hypotheses found to be reproducible (or significant) by both ΔW_{-i} and the other specified method. $\boldsymbol{\Delta W_{-i}}$ **only** represents the number of genes ΔW_{-i} finds to be reproducible that the other specified method does not. **Other only** similarly represents the number of genes the specified method finds to be reproducible (or significant) that ΔW_{-i} does not.

		Both	ΔW_{-i} only	Other only
ΔW_{-i} bootstrap	AdaFilter (r = 5)	327	1220	1
	AdaFilter (r = 4)	962	585	388
ΔW_{-i} bootstrap	Fisher's	1547	0	11313
	max$P$	1339	208	1146
	rOP (r = 4)	1543	4	4170
	RRA	253	1294	59

In this, we can see that when comparing ΔW_{-i} with max$P$, the meta-analysis method tends to discover more genes that have no signal in both experiments (points near the origin), while ΔW_{-i} deems these irreproducible, as their signal is too weak. Finally, since RRA does not consider the sign of an effect, we see that there is a cluster of genes that demonstrate strong up-regulation in Stephenson et al. (2021) and strong down-regulation in Schulte-Schrepping et al. (2020). This group of genes should not be considered reproducible, as genes react to COVID-19 differently in these two studies.

Gene ontology pathway analysis

After discovering genes that showed consistent differential expression results across the five replicate studies, we now examine the biological processes and gene pathways that were over-represented in the list of reproducible genes for each of the three reproducibility methods (ΔW_{-i}

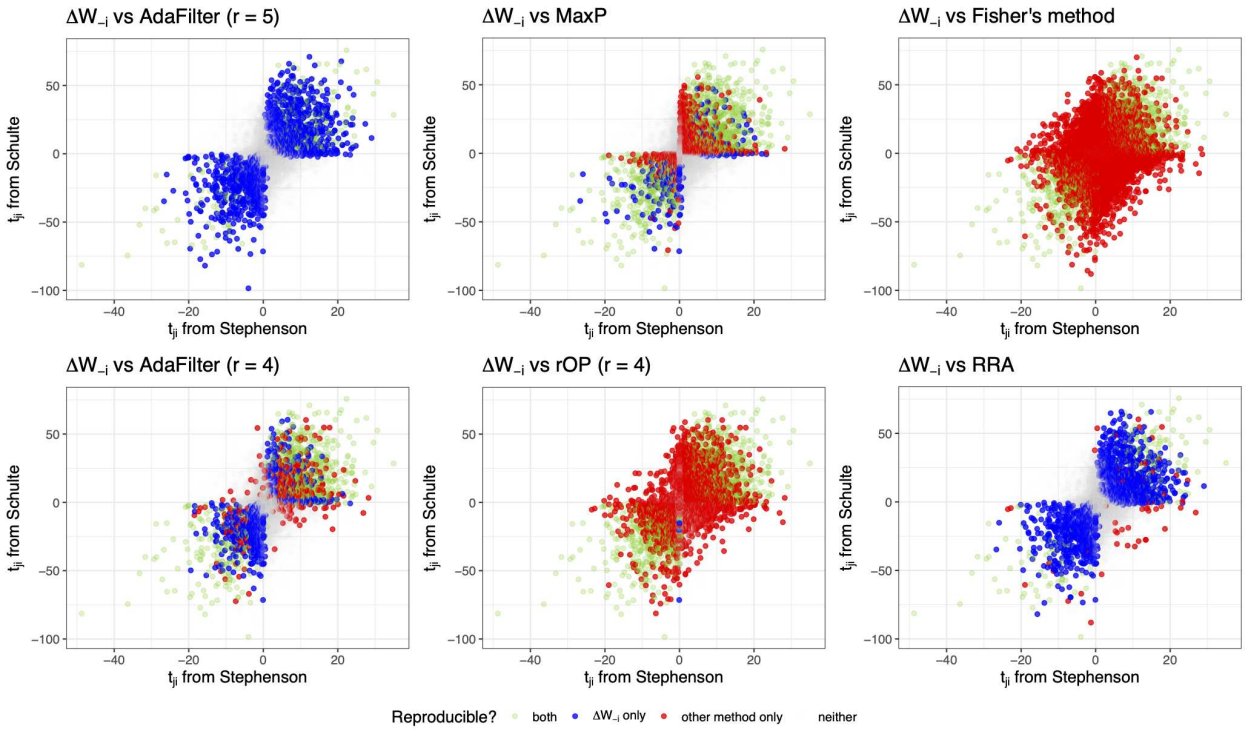


Figure 4.11: Intersection and set differences in reproducibility (or rejection) regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and each of the other methods in the space of t_{ji} statistics from Stephenson et al. (2021) and Schulte-Schrepping et al. (2020). Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only the other method. Appendix C.3 contains the scatterplots with these regions for all pairs of studies.

and AdaFilter with $r \in \{4, 5\}$). To do so, we performed gene ontology enrichment analysis on the biological process (GO:BP) (Ashburner et al., 2000; Aleksander et al., 2023) labels using the `clusterProfiler` package (Wu et al., 2021a) in `R`. The ten most prominent pathways in the reproducibility set for different methods are found in Figure 4.12.

We notice that both ΔW_{-i} and AdaFilter reproducible sets are enriched with antigen processing-related gene sets that are found to be associated with COVID-19 disease severity in previous studies (Wack, 2021). Interestingly, we notice that many pathways related to the viral process – such as regulation of the viral process, life cycle, and entry into the host cell – are among the most represented in the ΔW_{-i} reproducible set, but not discovered as highly represented by either of the AdaFilter regions. Furthermore, Figure 4.13 compares the rankings of pathways that include the

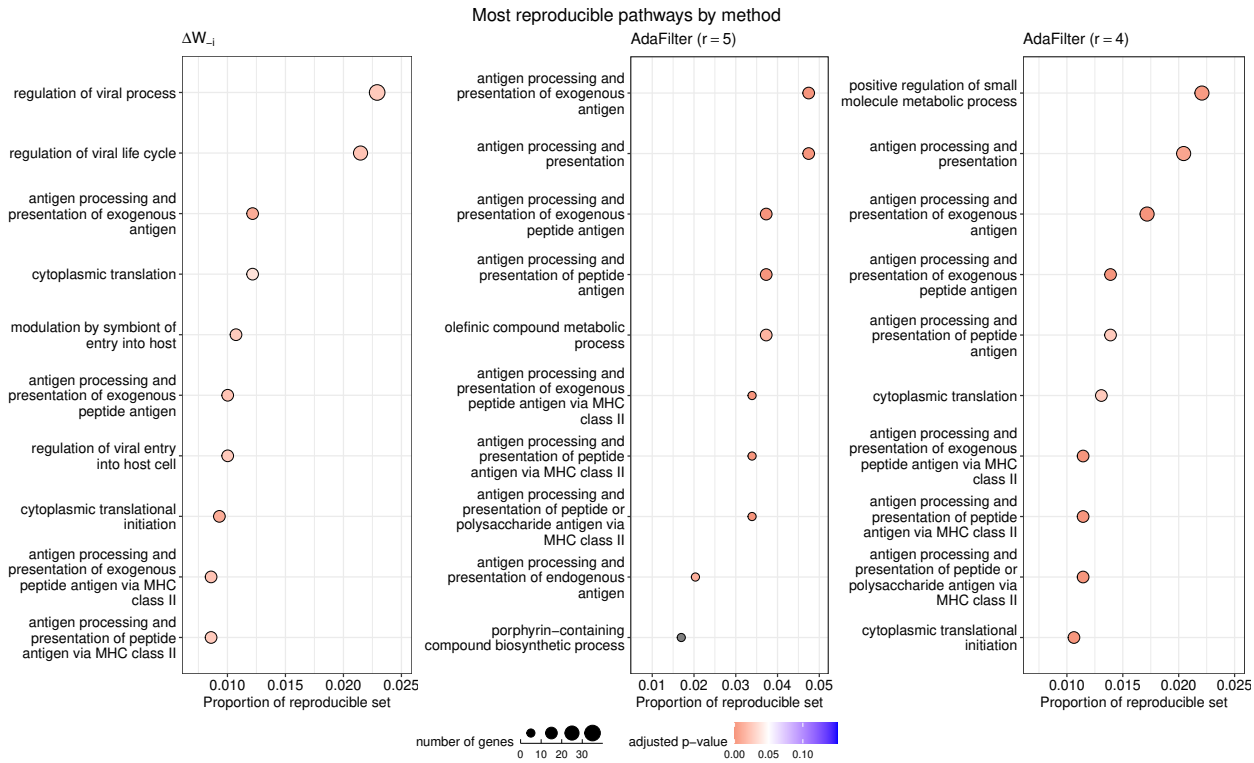


Figure 4.12: Most overrepresented pathways by proportion reproducibility genes at a nominal FDR level of $\alpha = 0.20$ that come from a particular biological process.

word “viral" among the top 200 most overrepresented biological processes identified by ΔW_{-i} and AdaFilter with $r = 4$.

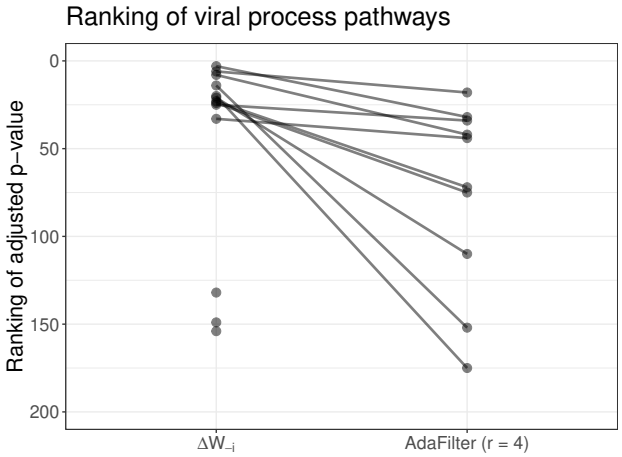


Figure 4.13: Ranking of adjusted enrichment p-values for pathways related to the viral process among the top 200 pathways identified by ΔW_{-i} and AdaFilter with $(r = 4)$. Examination of the top 1000 and all pathways can be found in Appendix C.3.

Notably, these “viral" processes are more highly ranked in terms of representation in the ΔW_{-i} reproducible set than in the AdaFilter set. This further demonstrates that the observed geometric advantage of the proposed method yields the discovery of more reproducible genes from biologically important pathways.

4.5 Discussion

By adapting Kendall's W, we introduce a rank-based statistic, ΔW_{-i}, that measures the agreement results for a hypothesis across multiple high-dimensional replicate studies. We use this statistic to devise three procedures that use a set of control hypotheses to approximate p-values and discover reproducible hypotheses at a specified nominal level of false discovery rate. To calculate approximate p-values, the first procedure uses an asymptotically equivalent form of the ΔW_{-i} statistic under a global null assumption, the second borrows from the conformal framework in the machine learning outlier detection problem, and the third builds a reference distribution by bootstrap sampling the control set. Unlike procedures based solely on the magnitude of p-values, the proposed methods are designed to scrutinize the agreement of both the magnitude and direction of observed effects for a hypothesis. In the high-throughput genomic context, this allows the procedures to identify genes that are consistently up- or down-regulated and avoid those with discordant effects across experiments without needing to apply the method multiple times. Additionally, since the ΔW_{-i} is rank-based, it shows superior power to methods that rely on combining p-values when the observed effects for a gene are highly consistent but only moderately strong. Through careful examination of the properties of each method, we compare the conditions that make each of the proposed methods preferable in application. We show preferable power and more exacting FDR control relative to p-value-based methods across a wide range of different settings via simulation. We then apply the method to five scRNA-seq datasets, examining the links between genes and the COVID-19 virus, and show advantages relative to existing methods in discovering genes that show consistent links to COVID-19. Specifically, the proposed method more reliably discovers genes from biological processes related to the viral process.

We provide theoretical justifications for the global null and conformal procedures under the stringent assumption that the control set is uncontaminated with reproducible genes. Specifically, we show that the conformal procedure controls FDR at a nominal level on average if the data in the control set are treated as *r*andom. A possible direction for future work is to examine the theoretical properties of this method when limited to one control set, as is the case for a practitioner. Examination of the theory for the bootstrap approach is also a future direction.

Chapter 5

Conclusion and discussion

5.1 Overview

In high-throughput genomics, the discovery of reproducible genetic features has become an important concern. Gene-trait associations that have been reproduced in multiple studies are more likely to be indicative of real causal links, rather than a product of a particular experimental design. In the high-throughput genomic context, this dissertation examined methods to identify reproducible hypotheses at a nominal level of false discovery rate. Particularly, we aimed to contribute to the reproducibility body of knowledge in three projects, each examining an interesting problem. The first two projects deal with reproducibility.

In Chapter 2, we formalized a nonparametric notion of irreproducibility. To assess our notion of reproducibility, we introduced a statistic with an accompanying procedure that's data-driven, flexible, and theoretically justified under relatively minor assumptions. Motivated by Philtron et al. (2018), the statistic was a function of ranks of summary statistics, so it did not rely on access to a particular type of summary statistics and was impervious to monotone experimental effects. Additionally, the newly proposed tuning parameter allowed practitioners the flexibility to prioritize the importance of signal consistency relative to strength depending on the context of the problem. The novel procedure for assessing reproducibility operates by partitioning the two study rank-space and learning the distribution of irreproducible hypotheses from the observed data. We showed the method controlled false discovery rate in theory under less stringent assumptions than elsewhere in the literature, a result typically lacking in existing nonparametric procedures Philtron et al. (2018); Ghosh et al. (2021). Through simulations, we observed that the method more closely approximates the desired level of false discovery rate, leading to increased power. We then analyzed two real TWAS analyses and uncovered reproducible genes.

Next, we noticed that the group structure present in high-throughput studies was not being adequately addressed in the reproducibility problem. In Chapter 3, we expanded the empirical Bayesian multiple testing with group structure framework from Liu et al. (2016) to analyze reproducibility across two studies. We proposed procedures for analyzing group and hypothesis-level reproducibility under an oracle assumption and adapted the EM algorithm to estimate the model parameters. Through simulations, we showed that by including group structure information, the proposed methods were more powerful than methods naive to existing groups while maintaining control of FDR.

Finally, the third problem was motivated by access to five scRNA-seq datasets examining expression for the same lists of genes in patients with the COVID-19 virus. The methods from the previous two chapters would be inadequate, as they consider only the case with two replicate studies. So, in Chapter 4 we expanded the non-parametric notion of reproducibility to the general cases with $m \geq 2$ replicate studies. We then designed a statistic based on Kendall's coefficient of concordance to measure the reproducibility of hypotheses in these cases. The statistic prioritized both the agreement of the magnitude and the sign of the observed effect, as opposed to just the magnitude. This statistic could separate reproducible hypotheses from irreproducible hypotheses; however, there was no neat distributional form for the statistic for irreproducible hypotheses. We then used sets of control hypotheses to design three procedures for discovering reproducible hypotheses. The first was based on an asymptotic result under a global null assumption, the second borrowed the conformal framework from the machine learning outlier problem, and the third employed bootstrap sampling techniques. Via simulation, we examined the advantages of the three methods and compared their performance to existing methods. Finally, we applied one of the methods to the five COVID-19 datasets along with existing methods and noticed two distinct advantages: 1) the geometry of the region found to be reproducible was more aligned with the idea of a result being reproducible, and 2) the method was able to more reliably discover genes from pathways known to be related to the viral process.

5.2 Future work

There are many avenues for continued research based on the content of the dissertation. As mentioned in Section 3.7, the requirement that groups are non-overlapping in Chapter 3 is limiting in some cases, including high-throughput genomics where genes can belong to many different groups or pathways. As such, expanding the Bernoulli Significant Model and the discussed framework to cases with overlapping groups is of particular interest. The challenge lies in redesigning the hypothesis and group-level procedures so that information for a hypothesis can be shared across all groups that it is a member hypothesis, and by the dependence of group-level hypothesis tests that are inherited from shared members. Additionally, we are interested in extending this framework to cases where there are additional hypothesis-level covariates (such as sequencing depth in the genomic setting) other than group structure. One approach to including additional covariates considers estimating the density of non-null signal (in our case Π_1 and π_1^1) as a function of the covariate (Cai et al., 2022). Under this design, the probability that a group or hypothesis is reproducible is a function of the covariate of interest, thus allowing us to gain power from the use of additional relevant information.

Extensions to the work in Chapter 4 include examining the theoretical properties of the bootstrapping procedure and considering a different procedure to control the false discovery rate. We propose using the Benjamini-Hochberg (BH) method Benjamini and Hochberg (1995). As seen in our simulations, applying BH is known to be conservative when true signals are dense within the test set. Over the years, there have been alternative mechanisms for controlling false discovery rate, including using sparsity in the estimation of false discovery rate (Storey, 2002; Storey et al., 2004), leveraging the symmetry of test statistics under the null hypothesis (Barber and Candés, 2015; Xing et al., 2023), among others. It might be interesting to examine other criteria for controlling false discovery rates that are slightly less conservative in practice. The assumptions made in many directly these approaches make them challenging to immediately leverage in our case. Additionally, as an alternative to using a control set of irreproducible hypotheses, one could make assumptions about irreproducible summary statistics and generate synthetic versions of irrepro-

ducible hypotheses to use as a reference distribution for irreproducible ΔW_{-i} statistics, similar to the popular knockoff procedures (Barber and Candés, 2015; Barber and Candès, 2016; Liu et al., 2020). The challenge and limitation of this approach is in proposing a model to generate the synthetic irreproducible summary statistics.

Bibliography

Aarts, A., Anderson, J., Anderson, C., Attridge, P., Attwood, A., Axt, J., Babel, M., Bahník, v., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F., and Pen, M. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251).

Alballa, N. and Al-Turaiki, I. (2021). Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. *Informatics in Medicine Unlocked*, 24:100564.

Aleksander, S., Balhoff, J., Carbon, S., Cherry, J., Drabkin, H., Ebert, D., Feuermann, M., Gaudet, P., Harris, N., Hill, D., Lee, R., Mi, H., Moxon, S., Mungall, C., Muruganugan, A., Mushayahama, T., Sternberg, P., Thomas, P., Auken, K., and Westerfield, M. (2023). The gene ontology knowledgebase in 2023. *Genetics*, 224.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., and Cherry, J. (2000). Gene ontology: Tool for the unification of biology. *The Gene Ontology Consortium. Nat Genet*, 25:25–29.

Barber, R. and Candès, E. (2016). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, 47.

Barber, R. and Ramdas, A. (2016). The p -filter: Multilayer false discovery rate control for grouped hypotheses. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79.

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

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

Beare, B. K. (2009). A generalization of hoeffding's lemma, and a new class of covariance inequalities. *Statistics & Probability Letters*, 79(5):637–642.

Benjamini, Y. and Heller, R. (2008). Screening for partial conjunction hypotheses. *Biometrics*, 64:1215–22.

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. *The Annals of Statistics*, 29(4):1165 – 1188.

Bogomolov, M. and Heller, R. (2018). Replicability across multiple studies. *Biometrika*, 105(3):505 – 516.

Bogomolov, M. and Heller, R. (2023). Replicability across multiple studies. *Statistical Science*, 38(4):602 – 620.

Borenstein, M., Hedges, L. V., Higgins, J. P., and Rothstein, H. R. (2021). *Introduction to Meta-Analysis*. John Wiley & Sons.

Cai, T. T. and Sun, W. (2009). Simultaneous testing of grouped hypotheses: Finding needles in multiple haystacks. *Journal of the American Statistical Association*, 104:1467–1481.

Cai, T. T., Sun, W., and Xia, Y. (2022). Laws: A locally adaptive weighting and screening approach to spatial multiple testing. *Journal of the American Statistical Association*, 117(538):1370 – 1283.

Carlson, M. (2023). *org.Hs.eg.db: Genome wide annotation for Human*. `R` package version 3.17.0.

Chang, L.-C., Lin, H.-M., Sibille, E., and Tseng, G. (2013). Meta-analysis methods for combining multiple expression profiles: Comparisons, statistical characterization and an application guideline. *BMC Bioinformatics*, 14:368.

Chen, Y., Chen, L., Lun, A., Baldoni, P., and Smyth, G. (2025). edgeR v4: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *Nucleic Acids Research*, 53.

Crowell, H., Soneson, C., Germain, P.-L., Calini, D., Collin, L., Raposo, C., Malhotra, D., and Robinson, M. (2020). muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nature Communications*, 11.

Dai, C., Lin, B., Xing, X., and Liu, J. S. (2022). False discovery rate control via data splitting. *Journal of the American Statistical Association*, pages 1–18.

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

Deng, L., He, K., and Zhang, X. (2023). Joint mirror procedure: Controlling false discovery rate for identifying simultaneous signals. *arXiv*.

Dubhashi, D. and Ranjan, D. (1998). Balls and bins: A study in negative dependence. *Random Structures & Algorithms, v.13, 99-124 (1998)*, 13.

Edahiro, R., Shirai, Y., Takeshima, Y., Sakakibara, S., Yamaguchi, Y., Murakami, T., Morita, T., Kato, Y., Liu, Y.-C., Motooka, D., Naito, Y., Takuwa, A., Sugihara, F., Tanaka, K., Wing, J., Sonehara, K., Tomofuji, Y., Namkoong, H., Tanaka, H., and Okada, Y. (2023). Single-cell analyses and host genetics highlight the role of innate immune cells in COVID-19 severity. *Nature genetics*, 55:753 – 767.

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

Ellinghaus, D., Degenhardt, F., Bujanda, L., Buti, M., Albillos, A., Invernizzi, P., Fernández, J., Prati, D., Baselli, G., Asselta, R., Grimsrud, M., Milani, C., Aziz, F., Kässens, J., May,

S., Wendorff, M., Wienbrandt, L., Uellendahl-Werth, F., Zheng, T., and Arning, N. (2020). Genomewide association study of severe COVID-19 with respiratory failure. *New England Journal of Medicine*, 383:1522 – 1534.

Errington, T., Mathur, M., Soderberg, C., Denis, A., Perfito, N., Iorns, E., and Nosek, B. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, 10.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Edinburgh: Oliver and Boyd.

Friston, K., Penny, W., and Glaser, D. (2005). Conjunction revisited. *NeuroImage*, 25(3):661 – 667.

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.

Ghosh, T., Philtron, D., Zhang, W., Kechris, K., and Ghosh, D. (2021). Reproducibility of mass spectrometry based metabolomics data. *BMC Bioinformatics*, 22(1):423.

Giambartolomei, C., Vukcevic, D., Schadt, E., Franke, L., Hingorani, A., Wallace, C., and Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics*, 10:e1004383.

Hao, X., Cheng, S., Wu, D., Wu, T., Lin, X., and Wang, C. (2020). Reconstruction of the full transmission dynamics of COVID-19 in wuhan. *Nature*, 584:1–7.

Heller, R. and Yekutieli, D. (2014). Replicability analysis for genome-wide association studies. *The Annals of Applied Statistics*, 8(1):481–494.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30.

Hollander, M., Wolfe, D. A., and Chicken, E. (2013). *Nonparametric Statisitical Methods*. John Wiley & Sons.

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., and Chory, J. (2006). RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22):2825 – 2827.

Hormozdiari, F., Bunt, M., Segrè, A., Li, X., Joo, J., Bilow, M., Sul, J. H., Sankararaman, S., Pasaniuc, B., and Eskin, E. (2016). Colocalization of GWAS and eQTL signals detects target genes. *American Journal of Human Genetics*, 99.

Hu, J., Zhao, H., and Zhou, H. (2010). False discovery rate control with groups. *Journal of the American Statistical Association*, 105:1215–1227.

Hung, K. and Fithian, W. (2020). Statistical methods for replicability assessment. *The Annals of Applied Statistics*, 14(3):1063 –1087.

Jaljuli, I., Benjamini, Y., Shenhav, L., Panagiotou, O., and Heller, R. (2022). Quantifying replicability and consistency in systematic reviews. *Statistics in Biopharmaceutical Research*, 15(2):372 – 385.

Joag-Dev, K. and Proschan, F. (1983). Negative association of random variables with applications. *The Annals of Statistics*, pages 286–295.

Kendall, M. G. and Gibbons, J. D. (1990). *Rank Correlation Methods*. Oxford University Press, New York, 5th edition.

Kendall, M. G. and Smith, B. B. (1939). The problem of m rankings. *Annals of Mathematical Statistics*, 10(3):275 – 287.

Kolde, R., Laur, S., Adler, P., and Vilo, J. (2012). Robust rank aggregation for gene list integration and meta-analysis. *Bioinformatics (Oxford, England)*, 28(4):573 – 580.

Laddha, S., Sami, M., Alghamdi, M., Kumar, V., Kaur, M., Alrashidi, M., Almuhaimeed, A., Alshehri, A., Abdullah, M., and Alkhazi, I. (2022). COVID-19 diagnosis and classification using radiological imaging and deep learning techniques: a comparative study. *Diagnostics*, 12.

Legendre, P. (2005). Species associations: The Kendall coefficient of concordance revisited. *Journal of Agriculutral, Biological, and Environmental Statistics*, 10(2):226 – 245.

Li, Q., Brown, J. B., Huang, H., and Bickel, P. J. (2011). Measuring reproducibility of high-throughput experiments. *Journal of the American Statistical Association*, 5(3):1752 – 1779.

Li, Q. and Zhang, F. (2018). A regression framework for assessing covariate effects on the reproducibility of high-throughput experiments. *Biometrics*, 74(3):803 – 813.

Li, Y., Chen, R., Zhang, X., and Cao, H. (2024). STAREG: Statistical replicability analysis of high throughput experiments with applications to spatial transcriptomic studies. *PLoS Genetics*, 20(10).

Lin, Y., Cao, Y., Willie, E., Patrick, E., and Yang, J. (2023). Atlas-scale single-cell multi-sample multi-condition data integration using scMerge2. *Nature Communications*, 14.

Lin, Y., Ghazanfar, S., Strbenac, D., Wang, A., Patrick, E., Lin, D., Speed, T., Yang, J., and Yang, P. (2019). Evaluating stably expressed genes in single cells. *GigaScience*, 8.

Lin, Y., Loo, L., Tran, A., Lin, D., Moreno, C., Hesselson, D., Neely, G., and Yang, J. (2022). Scalable workflow for characterization of cell-cell communication in COVID-19 patients. *PLOS Computational Biology*, 18(10):e1010495.

Liu, C., Martins, A., Lau, W., Rachmaninoff, N., Chen, J., Imberti, L., Mostaghimi, D., Fink, D., Burbelo, P., Dobbs, K., Delmonte, O., Bansal, N., Failla, L., Sottini, A., quiros roldan, E., Han, K., Sellers, B., Cheung, F., Sparks, R., and Tsang, J. (2021). Time-resolved systems immunology reveals a late juncture linked to fatal COVID-19. *Cell*, 184(7):1836 – 1857.

Liu, W., Ke, Y., Liu, J., and Li, R. (2020). Model-free feature screening and FDR control with knockoff features. *Journal of the American Statistical Association*, 117:1–43.

Liu, Y., Sarkar, S. K., and Zhao, Z. (2016). A new approach to multiple testing of grouped hypotheses. *Journal of Statistical Planning and Inference*, 179.

Love, M., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15:550.

Lyu, P., Yan, L., Wen, X., and Cao, H. (2023). JUMP: replicability analysis of high-throughput experiments with applications to spatial transcriptomic studies. *Bioinformatics (Oxford, England)*, 39.

Lücken, M., Büttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Mueller, M. F., Strobl, D., Zappia, L., Dugas, M., Colomé-Tatché, M., and Theis, F. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nature Methods*, 19(1):41 – 50.

MAQC-Consortium (2006). The microarray quality control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology*, 24(9):1151 – 1161.

Miao, Y., Xu, W., Chen, S., and Adler, A. (2014). Some limit theorems for negatively associated random variables. *Proceedings-Mathematical Sciences*, 124:447–456.

Nica, A., Montgomery, S., Dimas, A., Stranger, B., Beazley, C., Barroso, I., and Dermitzakis, E. (2010). Candidate causal regulatory effects by integration of expression qtls with complex trait genetic associations. *PLoS Genetics*, 6:e1000895.

Oakes, D. (1994). Multivariate survival distributions. *Journal of Nonparametric Statistics*, 3(3-4):343–354.

Owen, A. (2009). Karl Pearson's meta-analysis revisited. *Annals of Statistics*, 37(6B):3867 – 3892.

Pertea, M., Kim, D., Pertea, G., Leek, J., and Salzberg, S. (2016). Transcript-level expression analysis of rna-seq experiments with hisat, stringtie and ballgown. *Nature Protocols*, 11:1650–1667.

Philtron, D., Lyu, Y., Li, Q., and Ghosh, D. (2018). Maximum rank reproducibility: a nonparametric approach to assessing reproducibility in replicate experiments. *Journal of the American Statistical Association*, 113(523):1028–1039.

Quick, C., Dey, R., and Lin, X. (2021). Regression models for understanding COVID-19 epidemic dynamics with incomplete data. *Journal of the American Statistical Association*, 116:1561–1577.

Raman, G., Ashraf, B., Demir, Y., Kershaw, C., Cheruku, S., Atis, M., Atis, A., Atar, M., Chen, W., Ibrahim, F., Bat, T., and Mete, M. (2023). Machine learning prediction for COVID-19 disease severity at hospital admission. *BMC Medical Informatics and Decision Making*, 23.

Ren, X., Wen, W., Fan, X., Hou, W., Su, B., Cai, P., Li, J., Liu, Y., Tang, F., Zhang, F., Yang, Y., He, J., Ma, W., He, J., Wang, P., Cao, Q., Chen, F., Chen, Y., Cheng, X., and Zhang, Z. (2021). COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell*, 184:5838.

Ritchie, M., Phipson, B., Wu, D., Hu, Y., Law, C., Shi, W., and Smyth, G. (2015). LIMMA powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47.

Sarkar, S. K. and Zhao, Z. (2022). Local false discovery rate based methods for multiple testing of one-way classified hypotheses. *Electronic Journal of Statistics*, 16(2):6043 – 6085.

Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., De Domenico, E., Wendisch, D., Grasshoff, M., Kapellos, T., Beckstette, M., Pecht, T., Saglam, A., Dietrich, O., Mei, H., and Sander, L. (2020). Severe COVID-19 is marked by a dysregulated myeloid cell compartment. *Cell*, 182(6):1419 – 1440.

Song, C. and Tseng, G. (2014). Hypothesis setting and order statistic for robust genomic meta-analysis. *The Annals of Applied Statistics*, 8:777 – 800.

Squair, J., Gautier, M., Kathe, C., Anderson, M., James, N., Hutson, T., Hudelle, R., Qaiser, T., Matson, K., Barraud, Q., Levine, A., La Manno, G., Skinnider, M., and Courtine, G. (2021). Confronting false discoveries in single-cell differential expression. *Nature Communications*, 12:5692.

Stephenson, E., Reynolds, G., Botting, R., Calero-Nieto, F., Morgan, M., Tuong, Z., Bach, K., Sungnak, W., Worlock, K., Yoshida, M., Kumasaka, N., Kania, K., Engelbert, J., Olabi, B., Spegarova, J., Wilson, N., Mende, N., Jardine, L., Gardner, L., and Ansaripour, A. (2021). Single-cell multi-omics analysis of the immune response in COVID-19. *Nature Medicine*, 27:904–916.

Stodden, V., Guo, P., and Ma, Z. (2013). Toward reproducible computational research: An empirical analysis of data and code policy adoption by journals. *PloS One*, 8:e67111.

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

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

Su, C., Zhao, L., and Wang, Y. (1997). Moment inequalities and weak convergence for negatively associated sequences. *Science in China Series A: Mathematics*, 40(2):172–182.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

Tseng, G., Ghosh, D., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic acids research*, 40(9):3785 – 3799.

Volkonskii, V. and Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory of Probability & Its Applications*, 4(2):178–197.

Vovk, V., Gammerman, A., and Shafer, G. (2005). *Algorithmic Learning in a Random World*. Springer.

Wack, A. (2021). Monocyte and dendritic cell defects in COVID-19. *Nature Cell Biology*, 23:1–2.

Wang, C., Acosta, D., McNutt, M., Bian, J., Ma, A., Fu, H., and Ma, Q. (2024). A single-cell and spatial RNA-seq database for Alzheimer's disease (ssREAD). *Nature Communications*, 15:4710.

Wang, J., Gui, L., Su, W. J., Sabatti, C., and Owen, A. B. (2022). Detecting multiple replicating signals using adaptive filtering procedures. *The Annals of Statistics*, 50(4):1890–1909.

Wang, L., Wang, Y., Li, J., and Tong, X. (2023). Hierarchical neyman-pearson classification for prioritizing severe disease categories in COVID-19 patient data. *Journal of the American Statistical Association*, 119(545):39 – 51.

Wilk, A., Rustagi, A., Zhao, N., Roque, J., Martínez-Colón, G., McKechnie, J., Ivison, G., Ranganath, T., Vergara, R., Hollis, T., Simpson, L., Grant, P., Subramanian, A., Rogers, A., and Blish, C. (2020). A single-cell atlas of the peripheral immune response in patients with severe COVID-19. *Nature Medicine*, 26(7):1070 – 1076.

Wilkinson, B. (1951). A statistical consideration in psychological research. *Psychological Bulletin*, 48(3):156 – 158.

World Health Organization, a. (2025). WHO COVID-19 dashboard. `https://data.who.int/dashboards/covid19`. Accessed: 2025-04-29.

Wu, D., Li, X., Tanaka, R., Wood, J., Tibbs-Cortes, L., Magallanes-Lundback, M., Bornowski, N., Hamilton, J., Vaillancourt, B., Diepenbrock, C., Li, X., Deason, N., Schoenbaum, G., Yu, J., Buell, C. R., DellaPenna, D., and Gore, M. (2022). Combining GWAS and TWAS to identify candidate causal genes for tocochromanol levels in maize grain. *Genetics*, 221.

Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., Fu, X., Liu, S., Bo, X., and Yu, G. (2021a). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *The Innovation*, 2(3):10041.

Wu, Y.-H., Gao, S.-H., Mei, J., Xu, J., Fan, D.-P., Zhang, R.-G., and Cheng, M.-M. (2021b). JCS: An explainable COVID-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, 30:3113–3126.

Xing, X., Zhao, Z., and Liu, J. S. (2023). Controlling false discovery rate using Gaussian mirrors. *Journal of the American Statistical Association*, 118(541):222–241.

Zhang, Y., Quick, C., Yu, K., Barbeira, A., Consorium, T. G., Luca, F., Pique-Regi, R., Im, H. K., and Wen, X. (2020). PTWAS: investigating tissue-relevant causal molecular mechanisms of complex traits using probabilistic TWAS analysis. *Genome Biology*, 21(232).

Zhao, S. D., Cai, T. T., and Li, H. (2014). More powerful genetic association testing via a new statistical framework for integrative genomics. *Biometrics*, 70(4):881 – 890.

Zhao, S. D., Cai, T. T., P., C. T., Margulies, K. B., and Li, H. (2017). Sparse simultaneous signal detection for identifying genetically controlled disease genes. *Journal of the American Statistical Association*, 112(519):1032 – 1046.

Zhao, Y., Sampson, M., and Wen, X. (2020). Quantify and control reproducibility in high-throughput experiments. *Nature Methods*, 17(12):1207–1213.

Appendix A

Supplemental materials for "*Reproducible or not: a nonparametric procedure to assess reproducibility across high-throughput studies*"

These supplemental materials contain technical results and additional simulations for Chapter 2. Appendix A.1 presents the proofs to all theoretical results presented throughout the paper. Appendix A.2 describes and conducts additional simulations assessing individual parts of the proposed method. Namely, Appendix A.2.1 examines the performance of the proposed sparsity estimate $\widehat{\pi}_1$, Appendix A.2.2 examines different methods to select a suitable β, Appendix A.2.3 discusses the criteria for the selection of λ to increase power. Appendix A.3 contains the proofs of theoretical results introduced throughout Appendix A.2.

Notations: Throughout this appendix, we assume there are n hypotheses common across two studies with π_1 proportion reproducible. We denote the i^{th} hypothesis by $\mathbb{H}_i$ and define the sets irreproducible indices $(\mathcal{H}_0)$ and reproducible indices $(\mathcal{H}_1)$ by

$$\mathcal{H}_0 = \{i : \mathbb{H}_i \text{ is irreproduible}\} \text{ and } \mathcal{H}_1 = \{i : \mathbb{H}_i \text{ is reproduible}\}.$$

For each $\mathbb{H}_i$ we observe a summary statistics for each study denoted $(T_{1,i}, T_{2,i})$ which can be ranked within study from most (rank 1) to least (rank n) notable within experiment with ranks denoted by $(R_{1,i}, R_{1,i})$. We use $\xrightarrow{p}$ to represent convergence in probability, and for a function f let

$$||f||_\infty = \inf\{C \geq : |f(x)| \leq C \text{ for almost every } x\}.$$

A.1 Proofs

A.1.1 Auxiliary results

We first proof some useful Lemmas.

Lemma A.1.1. *If $h, h' \in \mathcal{H}_0$, then for any $x, y \in \{1, \cdots, n\}$ with $y \neq x$,*

$$\mathbb{P}(R_{1,h'} = y \mid R_{1,h} = x) = \left(\frac{n_0}{n_0 - 1}\right)\mathbb{P}(R_{1,h'} = y)$$

where $n_0 = n(1 - \pi_1)$.

Proof. Denote $\mathbf{R}_1^1 = \{R_{1,g} : g \in \mathcal{H}_1\}$. Notice, from the identically distributed statement in the definition of irreproducibility, it follows that

$$\mathbb{P}(R_{1,h} = x \mid \mathbf{R}_1^1 = \mathbf{A}) = \mathbb{P}(R_{1,h'} = x \mid \mathbf{R}_1^1 = \mathbf{A}) = \frac{1}{n_0} \tag{A.1}$$

for all $x \in \{1, \cdots, n\} \backslash \mathbf{A}$ and if $h^* \in \mathcal{H}_0$ also,

$$\mathbb{P}(R_{1,h'} = y \mid R_{1,h} = x \cap \mathbf{R}_1^1 = \mathbf{A}) = \mathbb{P}(R_{1,h^*} = x \mid R_{1,h} = x \cap \mathbf{R}_1^1 = \mathbf{A}) = \frac{1}{n_0 - 1} \tag{A.2}$$

for all $y \in \{1, \cdots, n\} \backslash \mathbf{A}$ such that $y \neq x$. This follows because only one hypothesis can be each rank. Notice, for any $y \neq x$

$$\begin{aligned}
\mathbb{P}(R_{1,h'} = y \mid R_{1,h} = x) &= \sum_{\mathbf{A} \in \mathbb{Z}^{n\pi_1}} \mathbb{P}(R_{1,h'} = y \mid R_{1,h} = x \cap \mathbf{R}_1^1 = A) \\
(\text{from (A.2)}) &= \frac{1}{n_0 - 1} \sum_{\mathbf{A} \in \mathbb{Z}^{n\pi_1}} \mathbb{I}[y \in \{1, \cdots, n\} \backslash \mathbf{A}] \mathbb{P}(\mathbf{R}_1^1 = \mathbf{A}) \\
&= \left(\frac{n_0}{n_0 - 1}\right) \sum_{\mathbf{A} \in \mathbb{Z}^{n\pi_1}} \frac{\mathbb{I}[y \in \{1, \cdots, n\} \backslash \mathbf{A}]}{n_0} \mathbb{P}(\mathbf{R}_1^1 = \mathbf{A}) \\
(\text{from (A.1)}) &= \left(\frac{n_0}{n_0 - 1}\right) \sum_{\mathbf{A} \in \mathbb{Z}^{n\pi_1}} \mathbb{P}(R_{1,h'} = y \mid \mathbf{R}_1^1 = A) \mathbb{P}(\mathbf{R}_1^1 = \mathbf{A}) \\
&= \left(\frac{n_0}{n_0 - 1}\right) \mathbb{P}(R_{1,h'} = y).
\end{aligned}$$

□

Proposition A.1.1. *For $c \in (0, \pi]$ and any $t \in (0, \lambda + 1)$, it holds*

$$\mathbb{V}ar\left(\sum_{i\in\mathcal{H}_0} \mathbb{I}[M_{\lambda,i}/n \le t]\right) \le cn_0.$$

Proof. Notice,

$$\begin{aligned}
&\mathbb{V}\text{ar}\left(\sum_{i\in\mathcal{H}_0} \mathbb{I}\left[\frac{M_{\lambda,i}}{n} \le t\right]\right) \\
=&\sum_{i\in\mathcal{H}_0} \mathbb{V}\text{ar}\left(\mathbb{I}\left[\frac{M_{\lambda,i}}{n} \le t\right]\right) + \sum_{i\neq j\in\mathcal{H}_0} \mathbb{C}\text{ov}\left(\mathbb{I}\left[\frac{M_{\lambda,i}}{n} \le t\right], \mathbb{I}\left[\frac{M_{\lambda,j}}{n} \le t\right]\right) \\
=&n_0\mathbb{V}\text{ar}\left(\mathbb{I}\left[\frac{M_{\lambda,h}}{n} \le t\right]\right) + n_0(n_0-1)\mathbb{C}\text{ov}\left(\mathbb{I}\left[\frac{M_{\lambda,h}}{n} \le t\right], \mathbb{I}\left[\frac{M_{\lambda,h'}}{n} \le t\right]\right) \\
\le&0.25n_0 + n_0(n_0-1)\mathbb{C}\text{ov}\left(\mathbb{I}\left[\frac{M_{\lambda,h}}{n} \le t\right], \mathbb{I}\left[\frac{M_{\lambda,h'}}{n} \le t\right]\right),
\end{aligned}$$

where the last inequality holds since $\mathbb{I}\left[\frac{M_{\lambda,h}}{n} \le t\right]$ is a Bernoulli random variable. Now, to control the covariance for arbitrary $h, h' \in \mathcal{H}_0$, we use the Hoeffding type covariance inequality (Volkonskii and Rozanov (1959), or Equation (5.5) in Beare (2009)), for a sequence of random variables $\{X_i\}$ defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. For any nonempty finite sets of integer, S and T such that $\min T - \max S \ge r$, and any Borel measurable functions $f : [a, b]^{|S|} \to \mathbb{R}$ and $g : [a, b]^{|T|} \to \mathbb{R}$, we have

$$|\mathbb{C}\text{ov}(f(X_s : s \in S), g(X_t : t \in T))| \le 4||f||_\infty||g||_\infty\alpha_r \tag{A.3}$$

where α_r is defined as the mixing coefficient,

$$\alpha_r = \sup_{S,T} \sup_{A\in\mathcal{F}_S, B\in\mathcal{F}_T} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|$$

with $\mathcal{F}_T$ and $\mathcal{F}_S$ being the sub σ fields generated by X_i for $i \in S$ or $i \in T$ respectively. We define sequence $\{X_1, X_2, X_3, X_4\} = \{R_{1,h}/n, R_{2,h}/n, R_{1,h'}/n, R_{2,h'}/n\}$ and define the function $f_\lambda(x, y) = \mathbb{I}[x + y + \lambda|x - y| \le t]$.

Notice, the desired covariance can be written and bounded by (A.3) as follows

$$\begin{aligned}\mathbb{C}\mathrm{ov}\left(\mathbb{I}\left[M_{\lambda,h}/n \le t\right], \mathbb{I}\left[M_{\lambda,h'}/n \le t\right]\right) &= \mathbb{C}\mathrm{ov}\left(f_\lambda(X_1, X_2), f_\lambda(X_3, X_4)\right) \\ &= \mathbb{C}\mathrm{ov}\left(f_\lambda(X_i : i \in \{1,2\}), f_\lambda(X_j : X_j \in \{3,4\})\right) \\ &\le 4\|f_\lambda\|_\infty^2 \alpha_1 = 4\alpha_1.\end{aligned}$$

Further, examine

$$\begin{aligned}\alpha_1 &= \sup_{S\subseteq\{1,2,3\};T\subseteq\{S+1,\cdots,4\}} \sup_{A\in\mathcal{F}_S;B\in\mathcal{F}_T} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \\ &= \sup_{A,B\subseteq\{1,\cdots,n\}^2} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)| \\ &= \sup_{A,B\subseteq\{1,\cdots,n\}^2} \Bigg| \sum_{(r_1,r_2)\in A} \sum_{(q_1,q_2)\in B} \mathbb{P}\left(X_1 = \frac{r_1}{n}, X_2 = \frac{r_2}{n}, X_3 = \frac{q_1}{n}, X_4 = \frac{q_2}{n}\right) \\ &\qquad - \mathbb{P}\left(X_1 = \frac{r_1}{n}, X_2 = \frac{r_2}{n}\right)\mathbb{P}\left(X_3 = \frac{q_1}{n}, X_4 = \frac{q_2}{n}\right)\Bigg| \\ &= \sup_{A,B\subseteq\{1,\cdots,n\}^2} \Bigg| \sum_{(r_1,r_2)\in A} \sum_{(q_1,q_2)\in B} \mathbb{P}\left(R_{1,h} = r_1, R_{2,h} = r_2, R_{1,h'} = q_1, R_{2,h'} = q_2\right) \\ &\qquad - \mathbb{P}\left(R_{1,h} = r_1, R_{2,h} = r_2\right)\mathbb{P}\left(R_{1,h'} = q_1, R_{2,h'} = q_2\right)\Bigg| \\ &= \sup_{A,B\subseteq\{1,\cdots,n\}^2} \Bigg| \sum_{(r_1,r_2)\in A} \sum_{(q_1,q_2)\in B} \mathbb{P}\left(R_{1,h} = r_1, R_{1,h'} = q_1\right)\mathbb{P}\left(R_{2,h} = r_2, R_{2,h'} = q_2\right) \\ &\qquad - \mathbb{P}\left(R_{1,h} = r_1\right)\mathbb{P}\left(R_{2,h} = r_2\right)\mathbb{P}\left(R_{1,h'} = q_1\right)\mathbb{P}\left(R_{2,h'} = q_2\right)\Bigg|\end{aligned}$$

(Definition 2.2.1)

$$
\begin{aligned}
&\leq \sup_{A,B\subseteq\{1,\cdots,n\}^2} \Bigg| \sum_{(r_1,r_2)\in A} \sum_{(q_1,q_2)\in B} \left(\frac{n_0}{n_0-1}\right)^2 \\
&\qquad \times \mathbb{P}\left(R_{1,h}=r_1\right)\mathbb{P}\left(R_{1,h'}=q_1\right)\mathbb{P}\left(R_{2,h}=r_2\right)\mathbb{P}\left(R_{2,h'}=q_2\right) \\
&\qquad - \mathbb{P}\left(R_{1,h}=r_1\right)\mathbb{P}\left(R_{2,h}=r_2\right)\mathbb{P}\left(R_{1,h'}=q_1\right)\mathbb{P}\left(R_{2,h'}=q_2\right)\Bigg| \\
&\qquad + \left|\left(\frac{n_0}{n_0-1}\right)^2 \sum_{(k_1,k_2)\in A\cap B} \mathbb{P}(R_{1,h}=k_1)^2\mathbb{P}(R_{2,h}=k_2)^2\right| \\
&\qquad + \left|\left(\frac{n_0}{n_0-1}\right)^2 \sum_{\substack{(k_1,j_2)\in A\\(k_1,i_2)\in B}} \mathbb{P}(R_{1,h}=k_1)^2\mathbb{P}(R_{2,h}=j_2)\mathbb{P}(R_{2,h'}=i_2)\right| \\
&\qquad + \left|\left(\frac{n_0}{n_0-1}\right)^2 \sum_{\substack{(j_1,k_2)\in A\\(i_1,k_2)\in B}} \mathbb{P}(R_{2,h}=k_2)^2\mathbb{P}(R_{1,h}=j_1)\mathbb{P}(R_{1,h'}=i_1)\right| \\
&\leq \sup_{A,B\subseteq\{1,\cdots,n\}^2} \left[\left(\frac{n_0}{n_0-1}\right)^2-1\right]\mathbb{P}\left(A\right)\mathbb{P}\left(B\right)+\left(\frac{n_0}{n_0-1}\right)^2\left(\frac{n^2+2n^3}{n_0^4}\right) \\
&\leq \left[\left(\frac{n_0}{n_0-1}\right)^2-1\right]+\left(\frac{n_0}{n_0-1}\right)^2\left[\frac{1}{(1-\pi_1)^2n_0^2)}+\frac{2}{(1-\pi_1)^3n_0}\right] \\
&\leq \frac{C_1}{n_0},
\end{aligned}
\tag{Lemma A.1.1}
$$

where $C_1 = 8 + (1-\pi_1)^{-2} + 2(1-\pi_1)^{-3}$. Hence, it leads to

$$
\mathbb{V}\mathrm{ar}\left(\sum_{i,j\in\mathcal{H}_0} \mathbb{I}\left[\frac{M_{\lambda,i}}{n}\leq t\right]\right) \leq 0.25n_0 + n_0(n_0-1)\frac{4C_1}{n_0} \leq (4C_1+0.25)n_0.
$$

□

Lemma A.1.2. *Under Condition 2.3.1, for any fixed $\beta > \beta_0 > \pi_1$, it holds*

$$
\sup_x \left|(n-n\pi_1)^{-1}\widehat{V}^{\beta}_{\lambda,\pi_1}(x) - (n-n\pi_1)^{-1}\mathbb{E}[V_\lambda(x)]\right| \xrightarrow{p} 0
$$

as $n(1-\beta)\to\infty$.

Proof. Notice,

$$\begin{aligned}\sup_{x\in(0,\lambda+1)}\left|\frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{n(1-\pi_1)}-\frac{\mathbb{E}[V_\lambda(x)]}{n(1-\pi_1)}\right| &= \sup_{x\in(0,\lambda+1)}\left|\frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{n(1-\pi_1)}-\frac{\mathbb{E}\left(\sum_{h\in\mathcal{H}_0}\mathbb{I}[M_{\lambda,h}/n\le x]\right)}{n(1-\pi_1)}\right| \\ \text{Definition 2.2.1} &= \sup_{x\in(0,\lambda+1)}\left|\widehat{F}^{\beta}_{\lambda,n}(x)-\mathbb{P}(M_{\lambda,h}/n\le x\mid h\in\mathcal{H}_0)\right| \\ &= \sup_{x\in(0,\lambda+1)}\left|\widehat{F}^{\beta}_{\lambda,n}(x)-F_\lambda(x)\right|.\end{aligned} \tag{A.4}$$

Remember, $\widehat{F}^{\beta}_{\lambda,n}$ is defined as

$$\begin{aligned}\widehat{F}^{\beta}_{\lambda,n}(x) = \int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} d\widehat{F}^{\beta}_{1,n}(t_1)d\widehat{F}^{\beta}_{2,n}(t_2) - \int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} d\widehat{F}^{\beta}_{2,n}(t_2)d\widehat{F}^{\beta}_{1,n}(t_1) \\ - \int_0^{\frac{x}{2}}\int_{a(t_2)}^{\frac{x}{2}} d\widehat{F}^{\beta}_{1,n}(t_1)d\widehat{F}^{\beta}_{2,n}(t_2)\end{aligned}$$

and due to the independence requirement in Definition 2.2.1, $F_\lambda(x)$ can be decomposed in the same manner by

$$\begin{aligned}F_\lambda(x) = \int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} dF_1(t_1)dF_2(t_2) - \int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} dF_2(t_2)dF_1(t_1) \\ - \int_0^{\frac{x}{2}}\int_{a(t_2)}^{\frac{x}{2}} dF_1(t_1)dF_2(t_2).\end{aligned}$$

So, the last expression in (A.4) can be bounded as follows

$$\begin{aligned}&\sup_{x\in(0,\lambda+1)}\left|\widehat{F}^{\beta}_{\lambda,n}(x)-F_\lambda(x)\right| \\ &\quad\le \sup_{x\in(0,\lambda+1)}\left|\int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} d\widehat{F}^{\beta}_{1,n}(t_1)d\widehat{F}^{\beta}_{2,n}(t_2) - \int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} dF_1(t_1)dF_2(t_2)\right| \\ &\qquad + \left|\int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} d\widehat{F}^{\beta}_{2,n}(t_2)d\widehat{F}^{\beta}_{1,n}(t_1) - \int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} dF_2(t_2)dF_1(t_1)\right| \\ &\qquad + \left|\int_0^{\frac{x}{2}}\int_{a(t_2)}^{\frac{x}{2}} d\widehat{F}^{\beta}_{1,n}(t_1)d\widehat{F}^{\beta}_{2,n}(t_2) - \int_0^{\frac{x}{2}}\int_{a(t_2)}^{\frac{x}{2}} dF_1(t_1)dF_2(t_2)\right|.\end{aligned} \tag{A.5}$$

Further, we can examine each term in (A.5). Starting with the first term, we have

$$\begin{aligned}&\sup_{x\in(0,\lambda+1)}\left|\int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} d\widehat{F}^{\beta}_{1,n}(t_1)d\widehat{F}^{\beta}_{2,n}(t_2) - \int_0^{\frac{x}{2}}\int_0^{\frac{x}{2}} dF_1(t_1)dF_2(t_2)\right| \\ &= \sup_{x\in(0,\lambda+1)} |\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \xrightarrow{p} 0.\end{aligned} \tag{A.6}$$

Next, the second and third terms are controlled in the manner

$$
\begin{aligned}
&\sup_{x\in(0,\lambda+1)}\left|\int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} d\widehat{F}^{\beta}_{2,n}(t_2)d\widehat{F}^{\beta}_{1,n}(t_1) - \int_0^{\frac{x}{2}}\int_{a(t_1)}^{\frac{x}{2}} dF_2(t_2)dF_1(t_1)\right| \\
&= \sup_{x\in(0,\lambda+1)}\left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(x/2) - \widehat{F}^{\beta}_{2,n}(a(t_1))d\widehat{F}^{\beta}_{1,n}(t_1) - \int_0^{\frac{x}{2}} F_2(x/2) - F_2(a(t_1))dF_1(t_1)\right| \\
&\leq \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(a(t_1))d\widehat{F}^{\beta}_{1,n}(t_1) - \int_0^{\frac{x}{2}} F_2(a(t_1))dF_1(t_1)\right| \\
&= \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(a(t_1))d(\widehat{F}^{\beta}_{1,n}(t_1) - F_1(t_1) + F_1(t_1)) - \int_0^{\frac{x}{2}} F_2(a(t_1))dF_1(t_1)\right| \\
&\leq \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(a(t_1))dF_1(t_1) - \int_0^{\frac{x}{2}} F_2(a(t_1))dF_1(t_1)\right| \\
&\qquad + \left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(a(t_1))d(\widehat{F}^{\beta}_{1,n}(t_1) - F_1(t_1))\right| \\
&\leq \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \sup_{x\in(0,\lambda+1)}\left|\int_0^{\frac{x}{2}} \widehat{F}^{\beta}_{2,n}(a(t_1)) - F_2(a(t_1))dF_1(t_1)\right| \\
&\qquad + \sup_{x\in(0,\lambda+1)}\left|\widehat{F}^{\beta}_{1,n}(x/2) - F_1(x/2))\right| \\
&\leq \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \sup_{x\in(0,\lambda+1)}\left|\sup_{t\in(0,1)}|\widehat{F}^{\beta}_{2,n}(t) - F_2(t)|\int_0^{\frac{x}{2}} dF_1(t_1)\right| \\
&\qquad + \sup_{x\in(0,\lambda+1)}\left|\widehat{F}^{\beta}_{1,n}(x/2) - F_1(x/2))\right| \\
&\leq \sup_{x\in(0,\lambda+1)}|\widehat{F}^{\beta}_{1,n}(x/2)\widehat{F}^{\beta}_{2,n}(x/2) - F_1(x/2)F_2(x/2)| \\
&\qquad + \sup_{t\in(0,1)}\left|\widehat{F}^{\beta}_{2,n}(t) - F_2(t)\right| + \sup_{x\in(0,1)}\left|\widehat{F}^{\beta}_{1,n}(x) - F_1(x)\right| \\
&\xrightarrow{p} 0 + 0 + 0.
\end{aligned}
\tag{A.7}
$$

The results in (A.6) and (A.6) prove Lemma A.1.2. □

Lemma A.1.3. *Suppose* $0 < c \le \pi_1$ *for arbitrary* c*. It holds*

$$\sup_{x\in\mathcal{R}} \left| n^{-1}V_\lambda(x) - n^{-1}\mathbb{E}[V_\lambda(x)] \right| \xrightarrow{p} 0$$

as $n(1-\pi_1) \to \infty$.

Proof. For any $\epsilon \in (0,1)$, for $N_\epsilon = \lceil 1/\epsilon \rceil$ let $0 = a_0^n > a_1^n > \cdots > a_{N_\epsilon}^n = \lambda{+}1$ such that $n^{-1}\mathbb{E}[V_\lambda(a_{k-1}^n)] - n^{-1}\mathbb{E}[V_\lambda(a_k^n)] \le \epsilon/2$ for $k \in \{1, 2, \cdots, N_\epsilon\}$. Since $\mathbb{E}[V_\lambda(x)]$ converges to continuous functions, we know that for n large enough there exists sequence $\{a_i^n\}$ which satisfies the above condition.

So it follows, since $V_\lambda(x)$ and $\mathbb{E}(V_\lambda(x))$ are monotone, increasing functions,

$$\begin{aligned}
\mathbb{P}\left(\sup_{x\in\mathbb{R}} n^{-1}V_\lambda(x) - n^{-1}\mathbb{E}[V_\lambda(x)] > \epsilon\right) &\le \mathbb{P}\left(\bigcup_{k=1}^{N_\epsilon} \sup_{x\in[a_k^n, a_{k-1}^n)} n^{-1}V_\lambda(x) - n^{-1}\mathbb{E}[V_\lambda(x)] > \epsilon\right) \\
&\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\sup_{x\in[a_k^n, a_{k-1}^n)} n^{-1}V_\lambda(x) - n^{-1}\mathbb{E}[V_\lambda(x)] > \epsilon\right) \\
(\text{Monotonicity}) &\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\sup_{x\in[a_k^n, a_{k-1}^n)} n^{-1}V_\lambda(a_k^p) - n^{-1}\mathbb{E}[V_\lambda(a_{k-1}^p)]\right) \\
&\le \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\sup_{x\in[a_k^n, a_{k-1}^n)} n^{-1}V_\lambda(a_k^p) - n^{-1}\mathbb{E}[V_\lambda(a_k^p)] + \frac{\epsilon}{2} > \epsilon\right) \\
&= \sum_{k=1}^{N_\epsilon} \mathbb{P}\left(\sup_{x\in[a_k^n, a_{k-1}^n)} n^{-1}V_\lambda(a_k^p) - n^{-1}\mathbb{E}[V_\lambda(a_k^p)] > \frac{\epsilon}{2}\right) \\
(\text{Chebyshev's inequality}) &\le \frac{4C_1 N_\epsilon}{n_0\epsilon^2} \to 0 \text{ as } n_0 \to \infty,
\end{aligned}$$

where the last bound holds from of Proposition A.1.1. By the same argument, one can show the following.

$$\mathbb{P}\left(\inf_{x\in\mathbb{R}} n^{-1}V_\lambda(x) - n^{-1}\mathbb{E}[V_\lambda(x)] \le -\epsilon\right) \le \frac{4C_1 N_\epsilon}{n_0\epsilon^2}.$$

Thus, we have shown Lemma A.1.3. □

A.1.2 Proof of Proposition 2.3.1

Notice, for all x, and $\beta > \beta_0$

$$\begin{aligned}
\mathbb{E}[\widehat{F}^{\beta}_{n,1}(x)] &= \mathbb{E}\left[\sum_{\ell \in \mathcal{R}^{\beta}_1} \frac{\mathbb{I}\left(R_{1,\ell}/n \le x\right)}{\left|\mathcal{R}^{\beta}_1\right|}\right] \\
&= \sum_{\ell \in \mathcal{R}^{\beta}_1} \mathbb{E}\left[\frac{\mathbb{I}\left(R_{1,\ell}/n \le x\right)}{\left|\mathcal{R}^{\beta}_1\right|}\right] \\
&= \mathbb{P}(R_{1,h}/n \le x \mid h \in \mathcal{R}^{\beta}_1) \\
&= \mathbb{P}(R_{1,h}/n \le x \mid h \in \mathcal{R}^{\beta}_1 \cap h \in \mathcal{H}_0)\mathbb{P}(h \in \mathcal{H}_0 \mid h \in \mathcal{R}^{\beta}_1) \\
(\text{Condition 2.3.1}) &= \mathbb{P}(R_{1,h}/n \le x \mid R_{2,h}/n > \beta \cap h \in \mathcal{H}_0) \\
&= \mathbb{P}(R_{1,h}/n \le x \mid h \in \mathcal{H}_0) \\
&= F_1(x).
\end{aligned}$$

The proof for $\widehat{F}^{\beta}_{n,2}(x)$ is similar. □

A.1.3 Proof of Theorem 2.4.1

Proof. Suppose Condition 2.1 is met for some $\beta_0 > \pi_1$. Fix β such that $1 > c \ge \beta > \beta_0$. Define the random sequence $\{X_i\}_{i=1}^{n(1-\beta)} = \{R_{1,h}/n : h \in \mathcal{R}^{\beta}_1\}$. Notice, $\{X_i\}$ is a sequence of scaled ranks, which are negatively associated (Joag-Dev and Proschan (1983); Su et al. (1997); Miao et al. (2014)). That is, for any f, g which are coordinate wise non-decreasing and any disjoint subsets of $\{1, 2, \cdots n(1-\beta)\}$ A and B

$$\mathbb{C}\text{ov}\left[f(X_i : i \in A), f(X_j : j \in B)\right] \le 0.$$

Additionally, the marginal distribution function for each X_i can be written as follows

$$\begin{aligned}
F_{X_i}(t) &= \mathbb{P}(R_{1,h}/n \le t \mid h \in \mathcal{R}^{\beta}_1) \\
(\text{Condition 2.1}) &= \mathbb{P}(R_{1,h}/n \le t \mid h \in \mathcal{R}^{\beta}_1 \cap h \in \mathcal{H}_0)
\end{aligned}$$

$$= \mathbb{P}(R_{1,h}/n \le t \mid R_{2,h}/n \ge \beta \cap h \in \mathcal{H}_0)$$

$$(R_{1,h} \perp R_{2,h}) = \mathbb{P}(R_{1,h}/n \le t \mid h \in \mathcal{H}_0)$$

$$= F_1(t).$$

Notice, $\widehat{F}^{\beta}_{1,n}$ is the empirical distribution function for F_{X_i}. Thus Lemma 3.6 in Miao et al. (2014) gives us the Glivenko-Cantelli lemma for negatively associated sequences. That is,

$$\sup_{t\in\mathbb{R}} \left|\widehat{F}^{\beta}_{1,n}(t) - F_1(t)\right| \xrightarrow{p} 0$$

as $n(1-\beta) \to \infty$. The proof for $\widehat{F}^{\beta}_{2,n}$ follows in the exact same manner. □

A.1.4 Proof of Theorem 2.4.2

Proof. Notice,

$$\begin{aligned}
\sup_{x\in(c,\lambda+1)} \left|\widehat{\mathrm{FDP}}^{\beta}_{\lambda}(x) - \mathrm{FDP}(x)\right| &= \sup_{x\in(c,\lambda+1)} \left|\frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{Q_\lambda(x)\vee 1} - \frac{V_\lambda(x)}{Q_\lambda(x)\vee 1}\right| \\
&= \sup_{x\in(0,\lambda+1)} \left|\frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{Q_\lambda(x)\vee 1} - \frac{\mathbb{E}[V_\lambda(x)]}{Q_\lambda(x)\vee 1} + \frac{\mathbb{E}[V_\lambda(x)]}{Q_\lambda(x)\vee 1} - \frac{V_\lambda(x)}{Q_\lambda(x)\vee 1}\right| \\
&\le \sup_{x\in(c,\lambda+1)} \left|\frac{\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{Q_\lambda(x)\vee 1} - \frac{\mathbb{E}[V_\lambda(x)]}{Q_\lambda(x)\vee 1}\right| \\
&\qquad + \sup_{x\in(c,\lambda+1)} \left|\frac{\mathbb{E}[V_\lambda(x)]}{Q_\lambda(x)\vee 1} - \frac{V_\lambda(x)}{Q_\lambda(x)\vee 1}\right| \\
&\quad (\text{Lemmas A.1.2 and A.1.3}) \xrightarrow{p} 0 + 0
\end{aligned}$$

as $n(1-\beta) \to \infty$. □

A.1.5 Proof of Theorem 2.4.3

For notation sake, we denote

$$\widehat{\mathrm{FDP}}^{\beta}_{\lambda}(x) = \frac{n^{-1}\widehat{V}^{\beta}_{\lambda,\pi_1}(x)}{n^{-1}Q_\lambda(x)\vee 1}, \quad \mathrm{FDP}_\lambda(x) = \frac{n^{-1}V_\lambda(x)}{n^{-1}Q_\lambda(x)\vee 1}, \quad \overline{\mathrm{FDP}}_\lambda(x) = \frac{n^{-1}\mathbb{E}[V_\lambda(x)]}{n^{-1}Q_\lambda(x)\vee 1}. \tag{A.8}$$

First, we show for any $\epsilon \in (0, \alpha)$, that

$$\mathbb{P}(\widehat{t}_\alpha \geq t_{\alpha-\epsilon}) \geq 1 - \epsilon$$

there $t_{\alpha-\epsilon}$ satisfies $\mathbb{P}(\mathrm{FDP}_\lambda(t_{\alpha-\epsilon}) \leq \alpha - \epsilon) \to 1$ as $n(1-\beta_1) \to \infty$. From Theorem 2.4.2, and using the definition of $\widehat{t}_\alpha$, it follows that

$$\mathbb{P}(\widehat{t}_\alpha \geq t_{\alpha-\epsilon}) \geq \mathbb{P}(\widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(t_{\alpha-\epsilon}) \leq \alpha) \geq \mathbb{P}(|\widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(t_{\alpha-\epsilon)} - \mathrm{FDP}_\lambda(t_{\alpha-\epsilon})|) \leq \epsilon) \geq 1 - \epsilon.$$

$$\begin{aligned}
\limsup_{n\to\infty} \mathrm{FDR}_\lambda(\widehat{t}_\alpha) &= \limsup_{n\to\infty} \mathbb{E}[\mathrm{FDP}_\lambda(\widehat{t}_\alpha)] \\
&\leq \limsup_{n\to\infty} \mathbb{E}[\mathrm{FDP}_\lambda(\widehat{t}_\alpha)|\widehat{t}_\alpha \geq t_{\alpha-\epsilon}]\mathbb{P}(\widehat{t}_\alpha \geq t_{\alpha-\epsilon}) + \epsilon \\
&\leq \limsup_{n\to\infty} \mathbb{E}[\mathrm{FDP}_\lambda(\widehat{t}_\alpha) - \overline{\mathrm{FDP}}_\lambda(\widehat{t}_\alpha)|\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] \\
&\qquad + \mathbb{E}[\overline{\mathrm{FDP}}_\lambda(\widehat{t}_\alpha) - \widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(\widehat{t}_\alpha)|\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] \\
&\qquad + \mathbb{E}[\widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(\widehat{t}_\alpha)|\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] + \epsilon \\
&\leq \limsup_{n\to\infty} \mathbb{E}[|\mathrm{FDP}_\lambda(\widehat{t}_\alpha) - \overline{\mathrm{FDP}}_\lambda(\widehat{t}_\alpha)||\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] \\
&\quad + \limsup_{n\to\infty} \mathbb{E}[|\overline{\mathrm{FDP}}_\lambda(\widehat{t}_\alpha) - \widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(\widehat{t}_\alpha)||\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] \\
&\quad + \limsup_{n\to\infty} \mathbb{E}[\widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(\widehat{t}_\alpha)|\widehat{t}_\alpha \geq t_{\alpha-\epsilon}] + \epsilon \\
&\leq \limsup_{n\to\infty} \mathbb{E}[\sup_{t\in(t_{\alpha-\epsilon},\lambda+1)} |\mathrm{FDP}_\lambda(t) - \overline{\mathrm{FDP}}_\lambda(t)|] \\
&\quad + \limsup_{n\to\infty} \mathbb{E}[\sup_{t\in(t_{\alpha-\epsilon},\lambda+1)} |\overline{\mathrm{FDP}}_\lambda(t) - \widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(t)|] \\
&\quad + \limsup_{n\to\infty} \mathbb{E}[\sup_{t\in(t_{\alpha-\epsilon},\lambda+1)} \widehat{\overline{\mathrm{FDP}}}^{\beta}_{\lambda}(\widehat{t}_\alpha)] + \epsilon \\
&\leq \alpha + \epsilon,
\end{aligned}$$

where the last inequality holds by the definition of $\widehat{t}_\alpha$, Lemmas A.1.2 and A.1.3 and the dominated convergence theorem.

□

A.2 Additional simulations

A.2.1 Estimation of π_1 simulations

In this section, we propose and conduct simulations that examine the conservative nature of the estimated sparsity parameter $\widehat{\pi}_1$ under each of the three simulation frameworks. Figure A.1 shows the distributions of $\widehat{\pi}_1$ estimates from 100 iterations of simulation A (with $n = 10000$ and $a = 0.5$, $b = 3.5$, and $\sigma_0 = 0.01$), simulation B (with $n = 10000$ and $\mu \in \{1.5, 2\}$, $\sigma^2 = 0.5$, and $\rho = 0.8$), and simulation C (with $n = 10000$ and $\{\mu_1, \mu_2, \mu_3, \mu_4\} = \{2.5, 3, 3.5, 4\}$ and $\rho = 0$) for true $\pi \in \{0.1, 0.3, 0.5\}$ respectively. Notice, across settings and iterations $\widehat{\pi}_1 \leq \pi_1$, and thus replacing

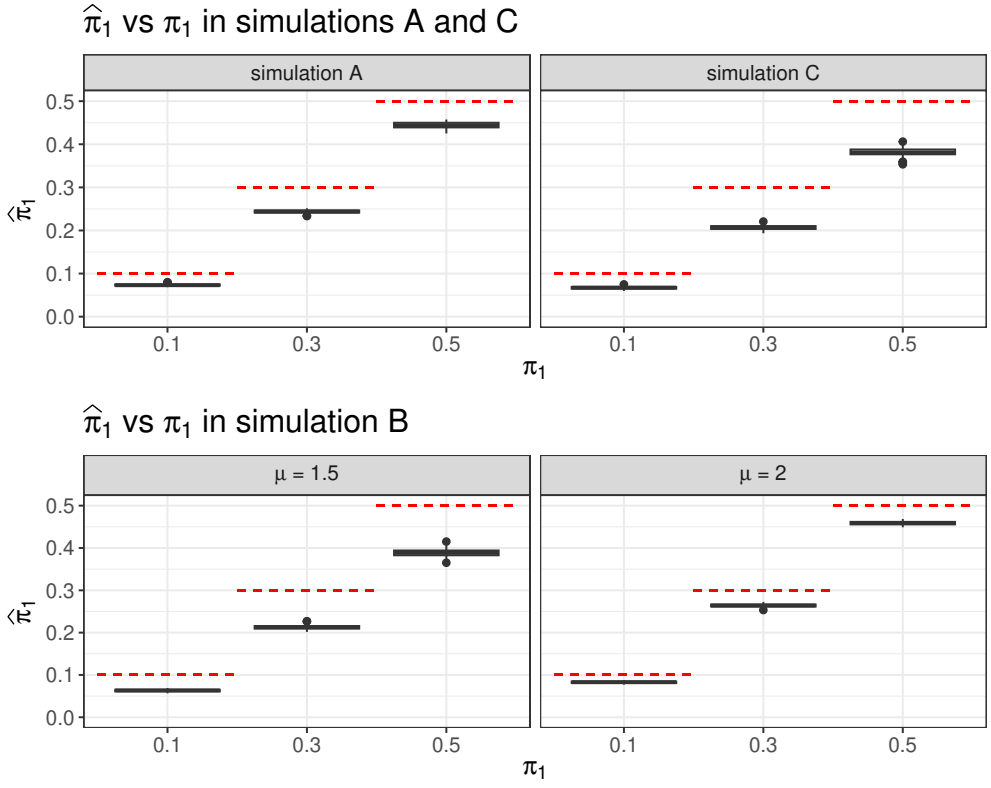


Figure A.1: Estimated $\widehat{\pi}_1$ over 100 iterations of simulations A, B, and C with true $\pi_1 \in \{0.1, 0.3, 0.5\}$.

π_1 with $\widehat{\pi}_1$ in the estimation of FDP will yield slightly conservative estimation, on average. This supports the notion that under realistic settings, $\mathbb{E}[\widehat{\pi}_1] \leq \pi_1$.

Additionally, when examining results from simulation, it can be seen that the estimate is generally less conservative for $\mu = 2$ than when $\mu = 1.5$ signaling that the larger the signal strength that all reproducible hypotheses demonstrate, the less biased this estimate is. Additionally, to examine the theoretical result presented in Philtron et al. (2018), Figure A.2 shows the distribution of $\widehat{\pi}_1$ under the ideal assumption from Philtron et al. (2018). That is, for each iteration ranks from reproducible hypotheses are entirely separated from those of irreproducible hypotheses. This condition was induced by considering simulation B with μ set sufficiently large. Here we see $\widehat{\pi}_1$ appears unbiased for the true π_1 across all three levels of sparsity considered.

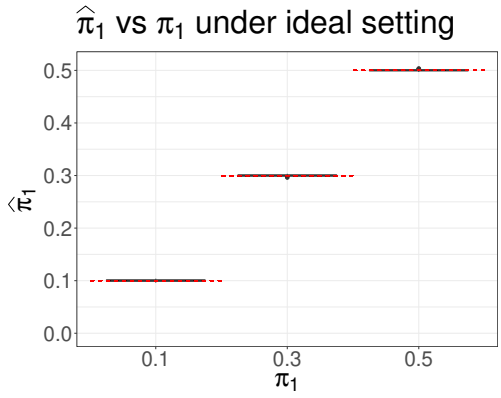


Figure A.2: Estimated $\widehat{\pi}_1$ over 100 iterations with the ideal assumption of rank separation enforced.

A.2.2 Performance of method across β simulations

In this section we compare the performance of the proposed method across different fixed values of β and a few different β selection procedures. For this simulation, we consider the simulation B setting outlined in Section 2.6.1 with $n = 10,000$, $\pi_1 = 0.3$, $\mu = 1.5$, $\sigma^2 = 0.5$, and $\rho = 0.8$. Figure A.3 contains the FDP and power from 100 iterations of this simulation for all fixed $\beta \in \{\pi_1, 0.5, 0.6, 0.7, 0.8, 0.9, 0.95\}$ and the true value of β_0 for each iteration. Additionally, we consider selecting β adaptively in the manner described in Section 2.5 with $d_{n,\beta} \in \left\{\frac{-\log(1/\log(n))(1-\pi_1)}{2n(1-\beta)}, \left(\frac{1}{n(1-\beta)}\right)^{3/2}\right\}$. The first rate we consider is motivated by the bound on the rate by which the survival function of the minimum rank statistic, $\widehat{S}^\beta$ converges to the true survival function, S^β. That is, Theorem A.2.1 provides an upper bound on the rate by which $\widehat{S}^\beta(t)$ converges to $S^\beta(t)$ for any fixed β and t.

Theorem A.2.1. *Suppose Condition 2.1 is met. Denote $\mathcal{R}^\beta_{1,2} = \mathcal{R}^\beta_1 \cap \mathcal{R}^\beta_1$. Let $\epsilon > 0$. For any $\beta > \beta_0$, and any, fixed, t*

$$\mathbb{P}\left(\left|\widehat{S}^\beta(t) - S^\beta(t)\right| > \sqrt{\frac{-\log(\epsilon/4)}{2C_\beta(n,\epsilon)}}\right) < \epsilon \tag{A.9}$$

where $C_\beta(n,\epsilon) = \frac{n(1-\beta)^2}{1-\pi_1} - \sqrt{\frac{-\log(\epsilon/2)n(1-\pi_1)}{2}}$.

To derive the first rate (denoted d^1), we consider this bound with $\epsilon = \frac{4}{log(n)}$ and drop the $\sqrt{\frac{-\log(\epsilon/2)n(1-\pi_1)}{2}}$ term for simplicity sake. The second rate (denoted $d^{(3/2)}$) is set such that the selected β tends to be slightly larger than that from the first rate. This rate is set small enough such that we observe less conservative FDP control. In general, the rate $d_{n,\beta}$ is user-specified, however, this is our recommended rate.

Lastly, we consider an alternative manner to select β which minimizes the difference between the two halves of the screening regions (denoted split). That is β can be selected by comparing the distributions of ranks across different regions of $\mathcal{R}^\beta_1$ and $\mathcal{R}^\beta_2$. Notice, Condition 2.3.1 ensures that for all $\beta > \beta_0$, if $h, h' \in \mathcal{R}^\beta_1 \cup \mathcal{R}^\beta_1$ $R_{1,h}$ is independent of $R_{2,h}$ and thus for any $\beta > \beta_0$, the following

equalities all hold

$$\begin{aligned}\mathbb{P}\left(R_{1,h}/n \le t \mid h \in \mathcal{H}_0\right) &= \mathbb{P}\left(R_{1,h}/n \le t \mid h \in \mathcal{R}_1^{\beta}\right) \\ &= \mathbb{P}\left(R_{1,h}/n \le t \mid R_{2,h} \in (\beta, 1)\right) \\ &= \mathbb{P}\left(R_{1,h}/n \le t \mid R_{2,h} \in (\beta, (\beta+1)/2]\right) \\ &= \mathbb{P}\left(R_{1,h}/n \le t \mid R_{2,h} \in ((\beta+1)/2, 1]\right)\end{aligned} \tag{A.10}$$

for any $t \in (0, 1]$ and the analogous equalities for $R_{2,h}$ hold as well. For any $\beta < \beta_0$, $h \in \mathcal{R}_1^{\beta} \cup \mathcal{R}_2^{\beta}$ does not necessarily imply that h is reproducible so the equalities in (A.10), in turn, do not hold. Thus, to select a β, one could consider comparing the sample estimated versions of $\mathbb{P}\left(R_{1,h}/n \le t \mid R_{2,h} \in (\beta, (\beta+1)/2]\right)$ and $\mathbb{P}\left(R_{1,h}/n \le t \mid R_{2,h} \in ((\beta+1)/2, 1]\right)$,

$$\widehat{F}_{n,1}^{\beta:(\beta+1)/2}(x) = \sum_{\ell \in \mathcal{R}_1^{\beta} \setminus \mathcal{R}_1^{\frac{\beta+1}{2}}} \frac{\mathbb{I}(R_{1,\ell}/n \le x)}{\left|\mathcal{R}_1^{\beta} \setminus \mathcal{R}_1^{\frac{\beta+1}{2}}\right|} \text{ and } \widehat{F}_{n,1}^{(\beta+1)/2}(x) = \sum_{\ell \in \mathcal{R}_1^{(\beta+1)/2}} \frac{\mathbb{I}(R_{1,\ell}/n \le x)}{\left|\mathcal{R}_1^{(\beta+1)/2}\right|}.$$

From (A.10), we know for all $\beta > \beta_0$, the functions that $\widehat{F}_1^{\beta:(\beta+1)/2}$ and $\widehat{F}_1^{(\beta+1)/2}$, and the functions $\widehat{F}_2^{\beta:(\beta+1)/2}$ and $\widehat{F}_2^{(\beta+1)/2}$ are estimating are equal to each other and for $\beta < \beta_0$, they are not. Thus, it is natural to select a β which minimizes the squared distance between those estimates, as that is the value that yields the smallest difference in the distribution in ranks across the different areas of the screening region $\mathcal{R}_j^{\beta}$. Thus, an alternative criterion for selecting β is as follows

$$\beta^{\text{split}} = \max_{j \in \{1,2\}} \left[\arg\min_{\beta \in (\pi_1, 1)} \left(n^{-1} \sum_{i=1}^{n} \left(\widehat{F}_j^{\beta:(\beta+1)/2}(i/n) - \widehat{F}_j^{\beta:(\beta+1)/2}(i/n) \right)^2 \right) \right]. \tag{A.11}$$

From Figure A.3 it can be seen that that FDP control is not sensitive to a selection of β. All levels of β demonstrate empirical distributions of FDP centered below the nominal target level. For smaller fixed β, in particular $\beta = \pi_1 = 0.3$, we see the proposed method is overly conservative. This is a result of the screening regions $\mathcal{R}_1^{\beta}$ and $\mathcal{R}_2^{\beta}$ contain a large number of reproducible hypotheses' rankings and do not do an adequate job of estimating the irreproducible rank distributions. However, as β increases, we see the method quickly move towards its associated level. Note

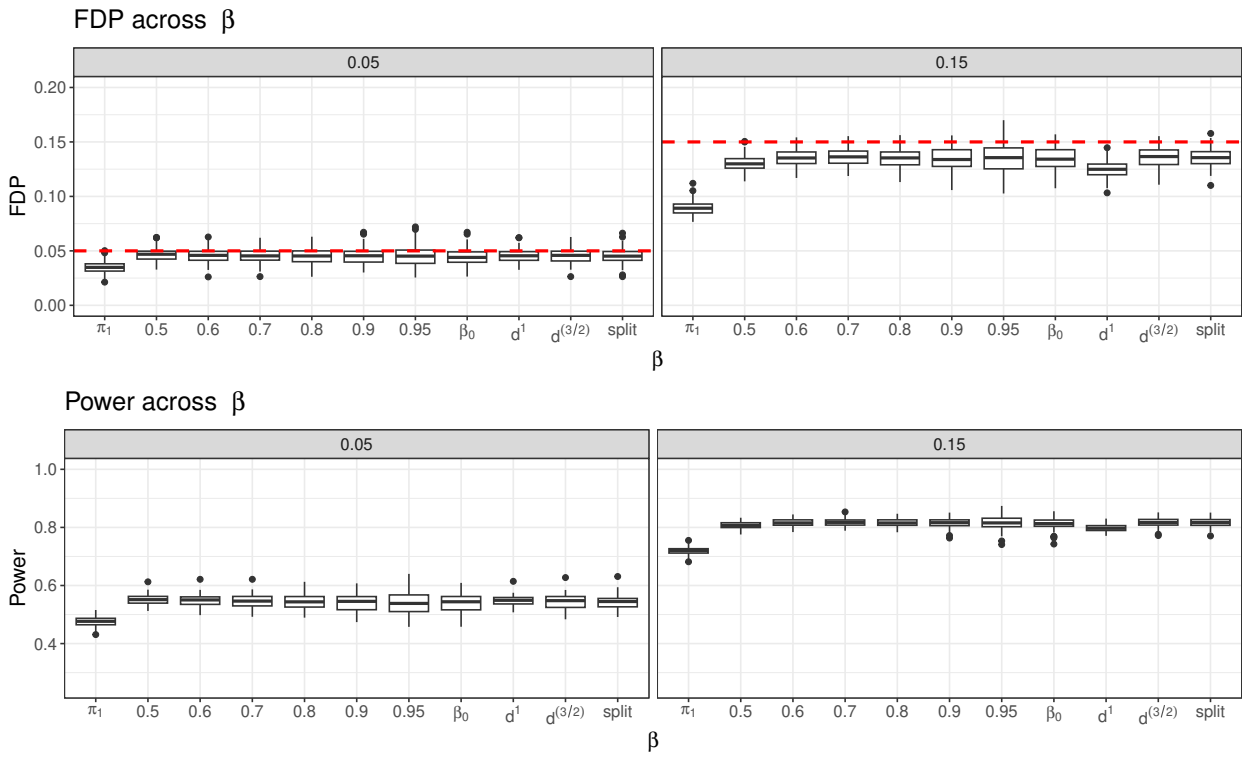


Figure A.3: FDP and power from 100 iteration of simulation B with $\mu = 1.5$, $\sigma^2 = 0.5$, and $\rho = 0.8$ with $\pi_1 = 0.30$.

we see very little difference in the empirical FDP distributions for any $\beta \geq 0.5$. This indicates the proposed method is robust against the selection of β. It can be noted that for extremely large β, in particular $\beta = 0.95$, we see values FDP and power are more variable from iteration to iteration. This is expected, as using a larger value for β results in estimating irreproducible rank distributions using fewer hypotheses.

Additionally, it can be seen that selecting β in the proposed manner from Section 2.5 with the rate of $d_{n,\beta} = (\frac{1}{n(1-\beta)})^{3/2}$ and selecting β using the splitting method perform quite robustly. The proposed method from Section 2.5 with the rate of $d_{n,\beta} = \frac{-\log(1/\log(n))(1-\pi_1)}{2n(1-\beta)}$ appears to select a β too small and thus too conservative of FDP control. It is our recommendation to consider the former rate over the latter or use a fixed, large β value, like $\beta = 0.9$.

A.2.3 Selection of λ simulations

In these simulations we evaluate the procedure used to select λ presented in Section 2.5. To do so, we consider simulating summary statistics by the bivariate Gaussian model presented in simulation B with $n = 10,000$ features with $\pi_1 = 0.3$ proportion reproducible. We consider $\mu = 1.5$, $\sigma^2 = 0.5$ and each $\rho \in \{0.5, 0.8, 0.95\}$. We then apply the proposed method using the $M_{\lambda,i}$ statistics with a fixed $\beta = 0.9$ and $\widehat{\pi}_1$ estimated in the manner presented in Section 2.5 to obtain a reproducible set with nominal FDR level of $\alpha = 0.05$ for each $\lambda \in \{1, 1.5, 2, 2.5, 3, 5, 7.5\}$. The λ selected for each iteration is the one with the largest reproducible set.

FDP control. Figure A.4 contains the boxplot of the true FDP values from the 100 iterations with a nominal level of $\alpha = 0.05$ across all of the λ values along with the FDP of the proposed method with λ selected adaptively for each iteration. We observe that the distribution of FDP values is centered below the nominal target goal for all λ and the selected λ. Additionally, we note that the distribution of FDP values does not appear to depend on the specific value of λ, as the estimation of irreproducible rank distributions does not rely on λ. This observation is critical in the λ selection criteria, as it ensures that selecting λ with the largest reproducible set does not yield a particular λ with a much larger proportion of false rejections.

Power. Here we examine the power properties of the proposed λ selection criteria. Figure A.5 contains the distribution selected λ values for 100 iterations compared to the distribution of λ values with the highest true power among $\lambda \in \{1, 1.5, 2, 2.5, 3, 5, 7.5\}$ for 100 iteration. The intuition of the λ selection procedure is that in selecting the value with the largest reproducible set, we are also selecting the λ which yields the most hypotheses correctly deemed reproducible (i.e. the λ selected also is the λ with the largest power among all values considered). The distributions of the selected λ compared to the value with the largest true power appear nearly identical. In fact, across all 300 iterations, the proposed λ selection procedure selects the value with the highest power 94.67% of the time (100% when $\rho = 0.5$, 91% when $\rho = 0.8$, and 93% when $\rho = 0.95$). Interestingly, the best value of λ increases as the level of signal consistency increases (ρ increases).

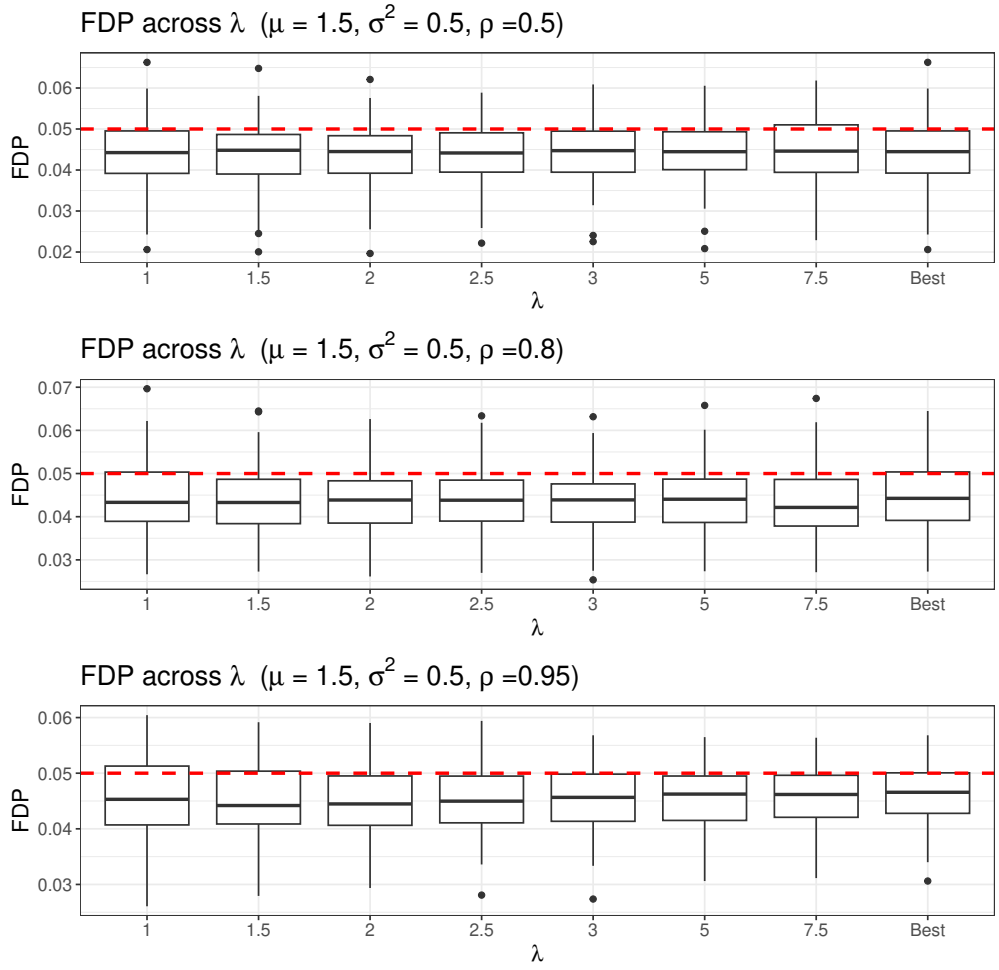


Figure A.4: FDP over 100 iterations of simulation B with $\rho \in \{0.5, 0.8, 0.95\}$ for $\lambda \in \{1, 1.5, 2, 2.5, 3, 5, 7.5\}$ and λ selected by the proposed procedure.

This supports the assertion that the inclusion of λ allows users more flexibility in detecting highly consistent reproducible hypotheses when considering larger values of λ

Additionally, when the procedure does not select the λ value with the largest true power, it is very near the largest power. Figure A.6 contains the difference in true power between the selected λ value and the value with the largest power for all iterations where those values are different. Notice that the difference in power in these iterations is always less than 0.003 indicating that the proposed selection criteria yields a selection of λ results within a negligible difference of the best possible λ in terms of power.

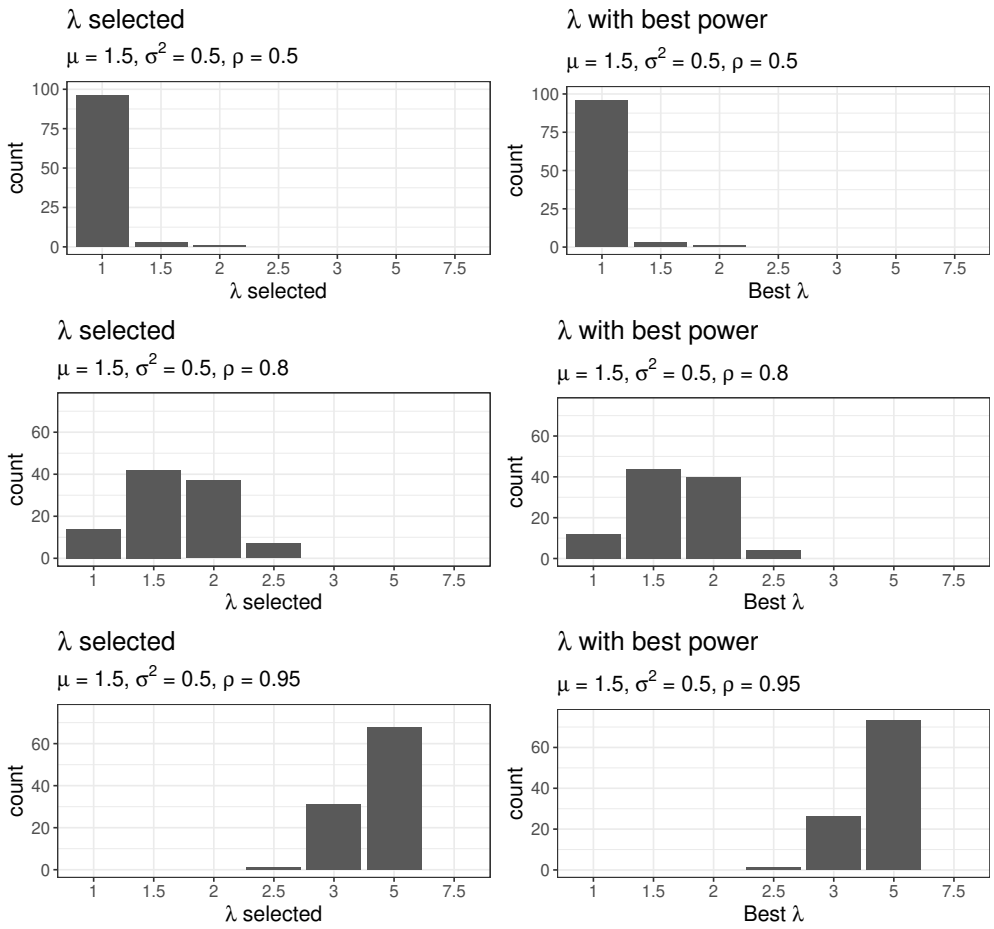


Figure A.5: Selected λ over 100 iterations of simulation B with $\rho \in \{0.5, 0.8, 0.95\}$.

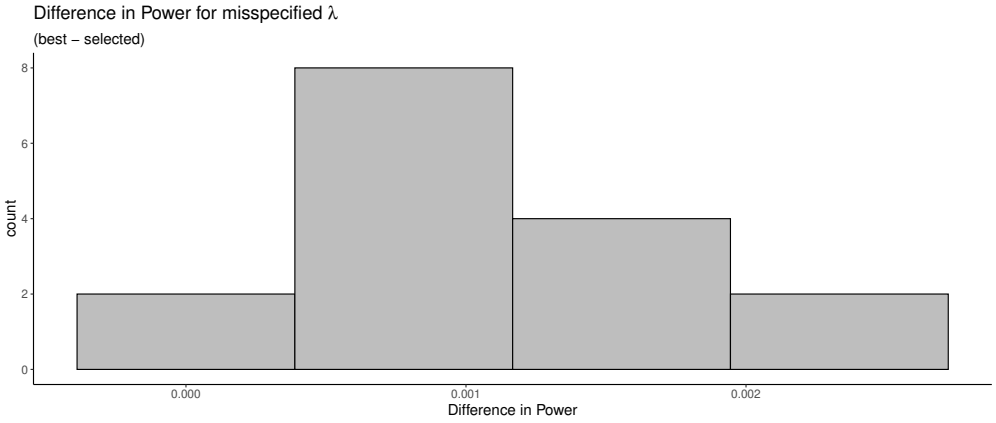


Figure A.6: Difference in power between selected λ and best possible λ over 100 iterations of simulation B with $\rho \in \{0.5, 0.8, 0.95\}$.

A.3 Theory for Appendix A.2

Proof of Lemma 2.5.1

Proof. First, notice, for any $r > \beta_0$, the identically distributed requirement on Condition 2.1 implies that for any $h \in \mathcal{H}_0$

$$\mathbb{P}(R_{1,h} = r) = \frac{1}{1-\pi_1} \text{ and } \mathbb{P}(R_{2,h} = r) = \frac{1}{1-\pi_1}. \tag{A.12}$$

Further, then, for any $\beta > \beta_0$, we have

Notice since $\beta > \beta_0$, Condition 2.1 enforces that for all $h \in \mathcal{R}^{\beta}_{1,2}$, $h \in \mathcal{H}_0$. Also, for all $h \in \mathcal{H}$ ranks are identically distributed, we can extend (A.12) to show

$$\mathbb{P}(R_{1,h} = r \mid R_{1,h} > n\beta) = \frac{1}{1-\beta} \text{ and } \mathbb{P}(R_{2,h} = r \mid R_{2,h} > n\beta) = \frac{1}{1-\beta}. \tag{A.13}$$

Now, using (A.13) we can show

$$\begin{aligned}
S^{\beta}(r) &= \mathbb{P}(n^{-1}(R_{1,h} \wedge R_{2,h}) > r \mid h \in \mathcal{R}^{\beta}_{1,2}) \\
&= \mathbb{P}(n^{-1}(R_{1,h} \wedge R_{2,h}) > r \mid h \in \mathcal{R}^{\beta}_{1,2} \cap h \in \mathcal{H}_0) \\
&= \mathbb{P}(n^{-1}(R_{1,h} \wedge R_{2,h}) > r \mid n^{-1}(R_{1,h} \wedge R_{2,h}) > \beta \cap h \in \mathcal{H}_0) \\
&= \mathbb{P}(R_{1,h}/n > r \cap R_{2,h}/n > r \mid (R_{1,h}/n > \beta \cap R_{2,h}/n > \beta) \cap h \in \mathcal{H}_0) \\
&= \mathbb{P}(R_{1,h}/n > r \mid R_{1,h}/n > \beta \cap h \in \mathcal{H}_0)\mathbb{P}(R_{2,h}/n > r \mid R_{2,h}/n > \beta \cap h \in \mathcal{H}_0) \\
&= \sum_{x_1 > r}^{n} \mathbb{P}(R_{1,h} = x_1 \mid R_{1,h} > n\beta \cap h \in \mathcal{H}_0) \sum_{x_2 > nr}^{n} \mathbb{P}(R_{2,h} = x_1 \mid R_{2,h} > n\beta \cap h \in \mathcal{H}_0) \\
&= \left(\frac{n(1-r)}{n(1-\beta)}\right)\left(\frac{n(1-r)}{n(1-\beta)}\right) \\
&= \left(\frac{1-r}{1-\beta}\right)^2.
\end{aligned}$$

□

To prove, Theorem A.2.1, we first require the technical Lemma A.3.1

Lemma A.3.1

Lemma A.3.1. *Suppose Condition 2.1 holds for β_0. Denote $\mathcal{R}_1^\beta \cap \mathcal{R}_2^\beta$ by $\mathcal{R}_{1,2}^\beta$. Then, for fixed $\beta > \beta_0$, it holds that*

$$\mathbb{P}\left(|\mathcal{R}_{1,2}^\beta| \geq \frac{n(1-\beta)^2}{1-\pi_1} - t\right) \geq 1 - \exp\left(\frac{2t^2}{n(1-\pi_1)}\right)$$

or alternatively,

$$\mathbb{P}\left(|\mathcal{R}_{1,2}^\beta| \geq \frac{n(1-\beta)^2}{1-\pi_1} - \sqrt{-\frac{\log(\delta)n(1-\pi_1)}{2}}\right) \geq 1 - \delta.$$

Proof. First, notice, for any $r > \beta_0$, the identically distributed requirement on Condition 2.1 implies that for any $h \in \mathcal{H}_0$

$$\mathbb{P}(R_{1,h} = r) = \frac{1}{1-\pi_1} \text{ and } \mathbb{P}(R_{2,h} = r) = \frac{1}{1-\pi_1}. \tag{A.14}$$

Now, note the following decomposition for $|\mathcal{R}_{1,2}^\beta|$ when $\beta > \beta_0$,

$$\begin{aligned}
|\mathcal{R}_{1,2}^\beta| &= \sum_{i=1}^{n} \mathbb{I}\left[i \in \mathcal{R}_1^\beta \cap \mathcal{R}_2^\beta\right] \\
&= \sum_{h\in\mathcal{H}_0} \mathbb{I}\left[h \in \mathcal{R}_1^\beta \cap \mathcal{R}_2^\beta\right] + \sum_{g\in\mathcal{H}_1} \mathbb{I}\left[g \in \mathcal{R}_1^\beta \cap \mathcal{R}_2^\beta\right] \\
(\text{Condition 2.1}) \quad &= \sum_{h\in\mathcal{H}_0} \mathbb{I}\left[R_{1,h} > n\beta \cap R_{2,h} > n\beta\right] + 0 \\
&= \sum_{h\in\mathcal{H}_0} \mathbb{I}[R_{1,h}/n > \beta]\mathbb{I}[R_{2,h}/n > \beta]
\end{aligned} \tag{A.15}$$

and

$$\begin{aligned}
\mathbb{E}|\mathcal{R}_{1,2}^\beta| &= \mathbb{E}\left(\sum_{h\in\mathcal{H}_0} \mathbb{I}[R_{1,h}/n > \beta]\mathbb{I}[R_{2,h}/n > \beta]\right) \\
&= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\mathbb{I}[R_{1,h}/n > \beta]\mathbb{I}[R_{2,h}/n > \beta]\right) \\
(\perp \text{ from Condition 2.1}) \quad &= \sum_{h\in\mathcal{H}_0} \mathbb{E}\left(\mathbb{I}[R_{1,h}/n > \beta]\right)\mathbb{E}\left(\mathbb{I}[R_{2,h}/n > \beta]\right) \\
&= n(1-\pi_1)\mathbb{P}(R_{1,h} > n\beta)\mathbb{P}(R_{2,h} > n\beta) \\
(\text{from (A.12)}) \quad &= \frac{n(1-\beta)^2}{(1-\pi_1)}.
\end{aligned} \tag{A.16}$$

Additionally, notice the sequences

$$\{X_{1,h}\}_{h\in\mathcal{H}_0} = \{R_{1,h}/n : h \in \mathcal{H}_0\} \text{ and } \{X_{1,h}\}_{h\in\mathcal{H}_0} = \{R_{2,h}/n : h \in \mathcal{H}_0\}$$

are negatively associated random sequences and independent of each other by Condition 2.1, so, by Property P_7 in Joag-Dev and Proschan (1983), $\{X_{j,h}\}_{j\in\{1,2\},h\in\mathcal{H}_0}$ is a negatively associated sequence. Now, for a fixed β, Let

$$h(x,y) = \mathbb{I}[x > \beta]\mathbb{I}[y > \beta]$$

and notice $h(x,y)$ is a coordinate-wise increasing function. Let

$$\{Z_h\}_{h\in\mathcal{H}_0} = \{h(X_{1,h}, X_{2,h})\}.$$

Notice, from (A.15) $\sum_{h\in\mathcal{H}_0} Z_h = |R^{\beta}_{1,2}|$. Since $\{Z_h\}_{h\in\mathcal{H}_0}$ is an increasing function on disjoint sets of negatively associated random variables, it is a negatively associated random sequence through Property P_6 from Joag-Dev and Proschan (1983). Further, Dubhashi and Ranjan (1998) shows the Chernoff-Hoeffding bounds for the sum of bounded random variables from Hoeffding (1963) holds for negatively associated random variables. Leveraging these bounds, we have

$$\begin{aligned}\mathbb{P}\left(|\mathcal{R}^{\beta}_{1,2}| \geq \frac{n(1-\beta)^2}{1-\pi_1} - t\right) &= \mathbb{P}\left(\sum_{h\in\mathcal{H}_0} Z_h - \mathbb{E}\left[\sum_{h\in\mathcal{H}_0} Z_h\right] \geq -t\right) \\ (\text{Chernoff} - \text{Hoeffding bound}) \quad &\geq 1 - \exp\left(\frac{-2t^2}{\sum_{h\in\mathcal{H}_0} 1}\right) = 1 - \exp\left(\frac{-2t^2}{n(1-\pi_1)}\right).\end{aligned} \tag{A.17}$$

Let $t = \sqrt{-\frac{\log(\delta)n(1-\pi_1)}{2}}$. Then,

$$\begin{aligned}\mathbb{P}\left(|\mathcal{R}^{\beta}_{1,2}| \geq \frac{n(1-\beta)^2}{1-\pi_1} - \sqrt{-\frac{\log(\delta)n(1-\pi_1)}{2}}\right) &\geq 1 - \exp\left(\frac{\log(\delta)n(1-\pi_1)}{n(1-\pi_1)}\right) \\ &= 1 - \delta.\end{aligned} \tag{A.18}$$

□

Proof of Theorem A.2.1

Proof. Define the events

$$A_{\epsilon_1} = \left\{ |\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2|\mathcal{R}^{\beta}_{1,2}|}} \right\} \text{ and } B_{\epsilon_2} = \left\{ |\mathcal{R}^{\beta}_{1,2}| > C_{\beta}(n, \epsilon/2) \right\}.$$

For A_{ϵ_1}, notice the sequence $\{X_{1,j}\} = \{R_{1,j}/n : j \in \mathcal{R}^{\beta}_1\}$ and $\{X_{2,j}\} = \{R_{2,j}/n : j \in \mathcal{R}^{\beta}_2\}$ are negatively associated. By Property P_7 in Joag-Dev and Proschan (1983), the union of independent sets of negatively associated random variables are also negatively associated. Thus, the sequence $\{X_{i,j}\} = \{X_{1,j}\} \cup \{X_{2,j}\}$ is negatively associated. Additionally, notice for any constant c $f(x, y) = \mathbb{I}[\min(x, y) > c]$ is a monotone function of x and y and by Property P_6 in Joag-Dev and Proschan (1983)), we know any concordant monotone functions defined on disjoint subsets of a set of negatively associated random variables are negatively associated. Thus, the set $\{Z_j\} = \{\mathbb{I}[\min(X_{1,j}, X_{2,j}) > c] : j \in \mathcal{R}^{\beta}_{1,2}\}$ is a negatively associated sequence, bounded on $[0, 1]$. Dubhashi and Ranjan (1998) show the Chernoff-Hoeffding (Hoeffding, 1963) inequalities for sums of bounded random variables hold. Thus it follows that

$$\begin{aligned} \mathbb{P}(A_{\epsilon}) &= \mathbb{P}\left(\left| \widehat{S}^{\beta}(t) - S^{\beta}(t) \right| > \sqrt{\frac{-\log(\epsilon/4)}{2|\mathcal{R}^{\beta}_{1,2}|}} \right) \\ &= \mathbb{P}\left(\left| \sum_j Z_j - \mathbb{E} \sum_j Z_j \right| > \sqrt{\frac{-\log(\epsilon/4)|\mathcal{R}^{\beta}_{1,2}|}{2}} \right) \\ &\leq 2 \exp\left(\frac{\log(\epsilon/4)|\mathcal{R}^{\beta}_{1,2}|}{\sum_{j \in R^{\beta}_{1,2}} 1} \right) = \epsilon/2 \end{aligned} \tag{A.19}$$

Additionally, notice from Lemma A.3.1, we get

$$\mathbb{P}(B_{\epsilon/2}) = \mathbb{P}(|\mathcal{R}_{1,2}| > C(n, \epsilon/2)) > 1 - \epsilon/2. \tag{A.20}$$

Finally, notice

$$
\begin{aligned}
&\mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C_{\beta}(n,\epsilon/2)}}\right) \\
&\quad = \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C_{\beta}(n,\epsilon/2)}} \bigcap B_{\epsilon/2}\right) \\
&\qquad\quad + \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C_{\beta}(n,\epsilon/2)}} \bigcap B^{c}_{\epsilon/2}\right) \\
&\quad \leq \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C_{\beta}(n,\epsilon/2)}} \bigcap B_{\epsilon/2}\right) + \mathbb{P}(B^{c}_{\epsilon/2}) \\
&\quad \leq \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C(n,\epsilon/2)}} \bigcap B_{\epsilon/2}\right) + \epsilon/2 \qquad \text{(A.20)} \\
&\quad = \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2C_{\beta}(n,\epsilon/2)}} \bigcap |\mathcal{R}^{\beta}_{1,2}| > C_{\beta}(n,\epsilon/2)\right) + \epsilon/2 \\
&\quad \leq \mathbb{P}\left(|\widehat{S}^{\beta}(t) - S^{\beta}(t)| > \sqrt{\frac{-\log(\epsilon/4)}{2|\mathcal{R}^{\beta}_{1,2}|}} \bigcap |\mathcal{R}^{\beta}_{1,2}| > C_{\beta}(n,\epsilon/2)\right) + \epsilon/2 \\
&\quad = \mathbb{P}\left(A_{\epsilon/2} \cap B_{\epsilon/2}\right) + \epsilon/2 \\
&\quad \leq \mathbb{P}(A_{\epsilon/21}) + \epsilon/2 \\
&\quad \leq \epsilon/2 + \epsilon/2 = \epsilon. \qquad \text{(A.19)}
\end{aligned}
$$

□

Appendix B

Supplemental materials for "*Assessing Reproducibility of High-Throughput Studies with Group Structure*"

These supplemental materials contain technical results and additional simulations for Chapter 3. Appendix B.1 examines the posterior FDR and FNR equalities from Section 4. A derivation of the EM estimation procedure from 4 is found in Appendix B.2. Finally, additional simulations are in Appendix B.3. Appendix B.3.1 examines the proposed estimation procedure and Appendix B.3.2 examines the η selection criteria from Section 3.5.1.

Notations: Throughout this appendix, we assume there are n hypotheses common across two studies that can be divided into G groups. θ_{gj}, θ_g, and $\theta_{g|j}$ are the hypothesis-level, group-level, hypothesis-within-group-level reproducibility statuses. Π_1 and π_1^1 are the parameters from the BSG model described in Section 3.3 and μ_1, σ_1^2 and ρ_1 are the parameters in the copula mixture discussed in Section 3.4.1.

B.1 Posterior FDR and FNR quantities

The equality for the group-level posterior false discovery rate, $g\text{PFDR}(\delta_g^*; \boldsymbol{T})$, from (3.6) can be derived in the following manner.

$$\begin{aligned} g\text{PFDR}(\delta_g^*; \boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G}(1-\theta_g)\delta_g^*(\boldsymbol{T})}{\sum_{g=1}^{G}\delta_g^*(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right] \\ &= 1 - \frac{\sum_g \delta_g^*(\boldsymbol{T})\mathbb{E}[\theta_g \mid \boldsymbol{T}]}{\sum_g \delta_g^*(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \delta_g^*(\boldsymbol{T})\mathbb{P}(\theta_g = 1 \mid \boldsymbol{T})}{\sum_g \delta_g^*(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \delta_g^*(\boldsymbol{T})[1 - \text{fdr}_g(\boldsymbol{T})]}{\sum_g \delta_g^*(\boldsymbol{T})} \\ &= \frac{\sum_g \delta_g^*(\boldsymbol{T})\text{fdr}_g(\boldsymbol{T})}{\sum_g \delta_g^*(\boldsymbol{T})}. \end{aligned}$$

Next, the hypothesis-level posterior false discovery rate, $h\text{PFDR}(\delta_{gj}; \boldsymbol{T})$, from (3.7) can be found by

$$\begin{aligned} h\text{PFDR}(\delta_{gj}; \boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G}\sum_{j=1}^{n_g}(1-\theta_{gj})\delta_{gj}(\boldsymbol{T})}{\sum_{g=1}^{G}\sum_{j=1}^{n_g}\delta_{gj}(\boldsymbol{T}) \vee 1} \middle| \boldsymbol{T} \right] \\ &= 1 - \frac{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})\mathbb{E}[\theta_{gj} \mid \boldsymbol{T}]}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})\mathbb{P}(\theta_{gj} = 1 \mid \boldsymbol{T})}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})\mathbb{P}(\theta_g\theta_{j|g} = 1 \mid \boldsymbol{T})}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})\mathbb{P}(\theta_g = 1 \cap \theta_{j|g} = 1 \mid \boldsymbol{T})}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})\mathbb{P}(\theta_g = 1 \mid \boldsymbol{T})\mathbb{P}(\theta_{j|g} = 1 \mid \theta_g = 1, \boldsymbol{T})}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \left(1 - \mathbb{P}(\theta_g = 0 \mid \boldsymbol{T})\right)\left[\sum_j \delta_{gj}(\boldsymbol{T})\left(1 - \mathbb{P}(\theta_{j|g} = 0 \mid \theta_g = 1, \boldsymbol{T})\right)\right]}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})} \\ &= 1 - \frac{\sum_g \left(1 - \text{fdr}_g(\boldsymbol{T})\right)\left[\sum_j \delta_{gj}(\boldsymbol{T})\left(1 - \text{fdr}_{j|g}(\boldsymbol{T})\right)\right]}{\sum_g \sum_j \delta_{gj}(\boldsymbol{T})}. \end{aligned}$$

Finally, the hypothesis-level posterior false nondiscovery rate, $h\mathrm{PFNR}(\delta_{gj}; \boldsymbol{T})$, from (3.5.1) can be derived as follows.

$$
\begin{aligned}
h\mathrm{PFNR}(\delta_{gj}; \boldsymbol{T}) &= \mathbb{E}\left[\frac{\sum_{g=1}^{G} \sum_{j=1}^{n_g} \theta_{gj} \left(1 - \delta_{gj}(\boldsymbol{T})\right)}{\sum_{g=1}^{G} \sum_{j=1}^{n_g} \left(1 - \delta_{gj}(\boldsymbol{T})\right) \vee 1} \middle| \boldsymbol{T} \right] \\
&= \frac{\sum_g \sum_j \left(1 - \delta_{gj}(\boldsymbol{T})\right) \mathbb{E}[\theta_{gj} \mid \boldsymbol{T}]}{\sum_g \sum_j \left(1 - \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&= \frac{\sum_g \sum_j \mathbb{E}[\theta_{gj} \mid \boldsymbol{T}]}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} - \frac{\sum_g \sum_j \mathbb{E}[\theta_{gj} \mid \boldsymbol{T}] \delta_{gj}(\boldsymbol{T})}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&= \frac{\sum_g \sum_j \mathbb{P}\left(\theta_{gj} = 1 \mid \boldsymbol{T}\right)}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} - \frac{\sum_g \sum_j \mathbb{P}\left(\theta_{gj} = 1 \mid \boldsymbol{T}\right) \delta_{gj}(\boldsymbol{T})}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&= \frac{\sum_g \sum_j \mathbb{P}\left(\theta_g = 1 \mid \boldsymbol{T}\right) \mathbb{P}\left(\theta_{j|g} = 1 \mid \theta_g = 1, \boldsymbol{T}\right)}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&\qquad - \frac{\sum_g \sum_j \mathbb{P}\left(\theta_g = 1 \mid \boldsymbol{T}\right) \mathbb{P}\left(\theta_{j|g} = 1 \mid \theta_g = 1, \boldsymbol{T}\right) \delta_{gj}(\boldsymbol{T})}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&= \frac{\sum_g \left(1 - \mathrm{fdr}_g\left(\boldsymbol{T}\right)\right) \sum_j \left(1 - \mathrm{fdr}_{j|g}\left(\boldsymbol{T}\right)\right)}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1} \\
&\qquad - \frac{\sum_g \left(1 - \mathrm{fdr}_g\left(\boldsymbol{T}\right)\right) \sum_j \left(1 - \mathrm{fdr}_{j|g}\left(\boldsymbol{T}\right)\right) \delta_{gj}(\boldsymbol{T})}{\left(n - \sum_g \sum_j \delta_{gj}(\boldsymbol{T})\right) \vee 1}.
\end{aligned}
$$

B.2 Derivation of estimation procedure

The complete likelihood with the data, $(\theta_g, \theta_{j|g})$ for all $g \in \{1, \cdots, G\}$ and $j \in \{1, \cdots, n_g\}$z can be written as in (B.1).

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\theta}) = \prod_{g=1}^{G} \Bigg\{ & \left[(1 - \Pi_1) \prod_{j=1}^{n_g} f_0(\boldsymbol{t}_{gj}) \right]^{(1-\theta_g)} \\ & \times \left[\Pi_1 \left(\frac{1 - \widetilde{\text{fdr}}_g}{1 - (1 - \pi_1^1)^{n_g}} \right) \prod_{j=1}^{n_g} \left((1 - \pi_1^1) f_0(\boldsymbol{t}_{gj}) \right)^{(1-\theta_{j|g})} \left(\pi_1^1 f_1(\boldsymbol{t}_{gj}) \right)^{\theta_{j|g}} \right]^{\theta_g} \Bigg\}\end{aligned} \tag{B.1}$$

So the complete log-likelihood is

$$\begin{aligned}\ell(\boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\theta}) &= \log(\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\theta})) \\ &= \sum_g (1 - \theta_g) \log(1 - \Pi_1) + \sum_g \sum_j (1 - \theta_g) \log(f_0(\boldsymbol{t}_{gj})) \\ &\quad + \sum_g \theta_g \log(\Pi_1) + \sum_g \theta_g \log(1 - \widetilde{\text{fdr}}_g) - \sum_g \theta_g \log(1 - (1 - \pi_1^1)^{n_g}) \\ &\quad + \sum_g \sum_j \theta_g (1 - \theta_{j|g}) \log(1 - \pi_1^1) + \sum_g \sum_j \theta_g (1 - \theta_{j|g}) \log(f_0(\boldsymbol{t}_{gj})) \\ &\quad + \sum_g \sum_j \theta_g \theta_{j|g} \log(\pi_1^1) + \sum_g \sum_j \theta_g \theta_{j|g} \log(f_1(\boldsymbol{t}_{gj}))\end{aligned} \tag{B.2}$$

and its expectation with respect to $\boldsymbol{\theta}$ when $\boldsymbol{\beta} = \boldsymbol{\beta}'$ is

$$
\begin{aligned}
Q(\boldsymbol{\beta}, \boldsymbol{\beta}') &= \mathbb{E}_{\boldsymbol{\theta}}[\ell(\boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\theta})] \\
&= \sum_g \log(1 - \Pi_1)\mathbb{E}[(1 - \theta_g)] + \sum_g \sum_j \log(f_0(\boldsymbol{t}_{gj}))\mathbb{E}[(1 - \theta_g)] \\
&\quad + \sum_g \log(\Pi_1)\mathbb{E}[\theta_g] + \sum_g \log(1 - \widetilde{\mathrm{fdr}}_g)\mathbb{E}[\theta_g] - \sum_g \log(1 - (1 - \pi_1^1)^{n_g})\mathbb{E}[\theta_g] \\
&\quad + \sum_g \sum_j \log(1 - \pi_1^1)\mathbb{E}[\theta_g(1 - \theta_{j|g})] + \sum_g \sum_j \log(f_0(\boldsymbol{t}_{gj}))\mathbb{E}[\theta_g(1 - \theta_{j|g})] \\
&\quad + \sum_g \sum_j \log(\pi_1^1)\mathbb{E}[\theta_g\theta_{j|g}] + \sum_g \sum_j \log(f_1(\boldsymbol{t}_{gj}))\mathbb{E}[\theta_g\theta_{j|g}] \\
&= \sum_g \log(1 - \Pi_1)\mathbb{P}(\theta_g = 0|\boldsymbol{t}, \boldsymbol{\beta}') + \sum_g \sum_j \log(f_0(\boldsymbol{t}_{gj}))\mathbb{P}(\theta_g = 0|\boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad + \sum_g \log(\Pi_1)\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}') + \sum_g \log(1 - \widetilde{\mathrm{fdr}}_g)\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad - \sum_g \log(1 - (1 - \pi_1^1)^{n_g})\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(1 - \pi_1^1)\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}')\mathbb{P}(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(f_0(\boldsymbol{t}_{gj}))\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}')\mathbb{P}(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(\pi_1^1)\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}')\mathbb{P}(\theta_{j|g} = 1|\theta_g = 1, \boldsymbol{t}, \boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(f_1(\boldsymbol{t}_{gj}))\mathbb{P}(\theta_g = 1|\boldsymbol{t}, \boldsymbol{\beta}')\mathbb{P}(\theta_{j|g} = 0|\theta_g = 1, \boldsymbol{t}, \boldsymbol{\beta}') \\
&= \sum_g \log(1 - \Pi_1)\mathrm{fdr}_g(\boldsymbol{\beta}') + \sum_g \sum_j \log(g_0(\boldsymbol{z}_{gj}))\mathrm{fdr}_g(\boldsymbol{\beta}') \\
&\quad + \sum_g \log(\Pi_1)(1 - \mathrm{fdr}_g(\boldsymbol{\beta}')) + \sum_g \log(1 - \widetilde{\mathrm{fdr}}_g)(1 - \mathrm{fdr}_g(\boldsymbol{\beta}')) \\
&\quad - \sum_g \log(1 - (1 - \pi_1^1)^{n_g})(1 - \mathrm{fdr}_g(\boldsymbol{\beta}')) \\
&\quad + \sum_g \sum_j \log(1 - \pi_1^1)(1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))\mathrm{fdr}_{j|g}(\boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(f_0(\boldsymbol{t}_{gj}))(1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))\mathrm{fdr}_{j|g}(\boldsymbol{\beta}') \\
&\quad + \sum_g \sum_j \log(\pi_1^1)(1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}')) \\
&\quad + \sum_g \sum_j \log(g_1(\boldsymbol{z}_{gj}))(1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}'))
\end{aligned}
\tag{B.3}
$$

The final equality follows from (3.18), $\mathbb{P}(\theta_g = 0 \mid \boldsymbol{t}, \boldsymbol{\beta}') = \mathrm{fdr}_g(\boldsymbol{\beta}')$, and $\mathbb{P}(\theta_{j|g} = 0 \mid \theta_g = 1, \boldsymbol{t}, \boldsymbol{\beta}') = \mathrm{fdr}_{j|g}(\boldsymbol{\beta}')$. Now, the MLE estimates for Π_1 and π_1^1 are

$$\Pi_1^{\text{new}} = 1 - \frac{\sum_g \mathrm{fdr}_g(\boldsymbol{\beta}')}{G} \text{ and } \pi_1^{1,\text{new}} = \frac{\sum_g \sum_j (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}'))}{\sum_g (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))}. \tag{B.4}$$

Notice, for $(\mu_1, \sigma_1^2, \rho_1)$, we can examine the final term in (B.3), since these parameters are contained in g_1,

$$g_1(\boldsymbol{z}_{gj}; \boldsymbol{\beta}) = \frac{1}{2\pi\sigma_1^2\sqrt{1-\rho_1}} \exp\left(-\frac{(z_{gj,1} - \mu_1)^2 - 2\rho_1(z_{gj,1} - \mu_1)(z_{gj,2} - \mu_1) + (z_{gj,2} - \mu_1)^2}{2\sigma_1^2(1-\rho_1^2)} \right)$$

We solve the MLE for the μ_1 with the pseudo-data, $\widehat{\boldsymbol{z}}_{gj}$, filled in for the latent $\boldsymbol{z}_{gj}$ by examining the last taking a derivative with respect to μ_1.

$$\mu_1^{\text{new}} = \frac{\sum_g \sum_j (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}'))(\widehat{z}_{gj,1} + \widehat{z}_{gj,2})}{2\sum_g \sum_j (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}'))}. \tag{B.5}$$

For σ_1^2 and ρ_1, taking the derivative with respect to each and solving the resulting system of equations yields the following.

$$\sigma_1^{2,\text{new}} = \frac{\sum_g \sum_j (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}(\boldsymbol{\beta}')_{j|g})((\widehat{z}_{gj,1} - \mu_1^0)^2 + (\widehat{z}_{gj,2} - \mu_1^0)^2)}{2\sum_g \sum_j (1 - \mathrm{fdr}_g(\boldsymbol{\beta}'))(1 - \mathrm{fdr}_{j|g}(\boldsymbol{\beta}'))} \tag{B.6}$$

$$\rho_1^{\text{new}} = \frac{2\sum_g \sum_j (1 - \mathrm{fdr}_g)(1 - \mathrm{fdr}_{j|g})(\widehat{z}_{gj,1} - \mu_1^0)(\widehat{z}_{gj,2} - \mu_1^0)}{\sum_g \sum_j (1 - \mathrm{fdr}_g)(1 - \mathrm{fdr}_{j|g})((\widehat{z}_{gj,1} - \mu_1^0)^2 + (\widehat{z}_{gj,2} - \mu_1^0)^2)} \tag{B.7}$$

B.3 Additional simulations

B.3.1 Estimation performance

Figure B.1 shows the final estimates for $\boldsymbol{\beta} = (\Pi_1, \pi_1^1, \mu, \rho, \sigma^2)$ produced by the proposed adapted EM algorithm scheme, Algorithm 4, from 50 iterations of each of the settings described in Section 3.6.1. In general, the algorithm produces estimates that are close to the true parameters,

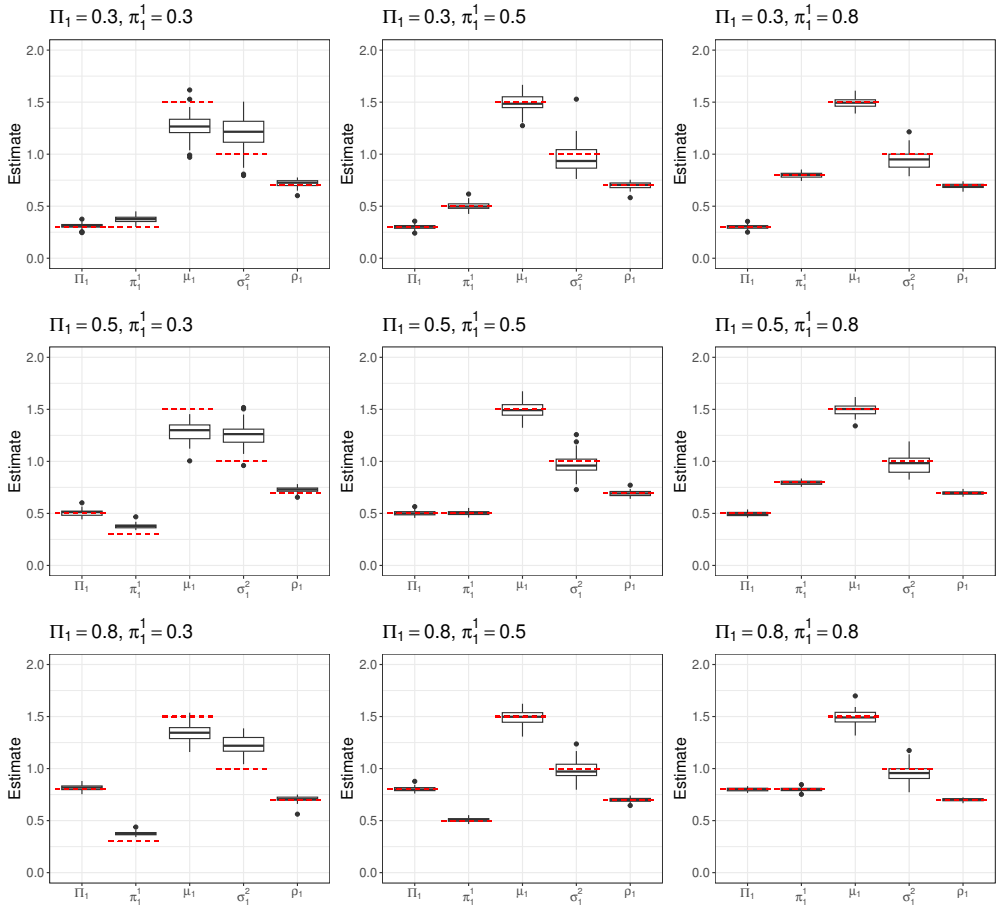


Figure B.1: Estimates for $\boldsymbol{\beta} = (\Pi_1, \pi_1^1, \mu, \rho, \sigma^2)$ using the adapted EM algorithm presented in Algorithm 4 from 50 iterations of each settings described in Section 3.6.1. The dashed line in each figure represents the parameter's true value.

except in the settings with $\pi_1^1 = 0.3$. In this case, a slight overestimation of π_1^1 leads to a slight underestimation of μ_1 and overestimation of σ_1^2. This follows from an overestimation of π_1^1 implying the local hypothesis-within-group local FDR scores, $\text{fdr}_{j|g}$ calculated using $\boldsymbol{\beta}'$ are slightly too

small, thus impacting calculations of μ_1^{new} and $\sigma^{2,\text{new}}$ because they are reliant on those local FDR scores. It can be seen along the left column in Figure 3.1 that under these settings, the estimated application of the method remains relatively close to the nominal level of FDR control and more powerful than the existing literature, particularly for smaller nominal levels of false discovery rate control. For $\alpha = 0.20$, this misestimation yields observed FDP values that are slightly larger than the nominal level.

B.3.2 Selection of η

To examine the performance of the proposed procedure for selecting a value for η that minimizes the $h\text{PFNR}$ in the hypothesis-level testing procedure as is proposed in Section 3.2, we compare that selection criteria to set levels of η. Specifically, we consider 100 iterations of each simulation setting described in Section 3.6.1 and apply the proposed hypothesis-level procedure at a nominal FDR level of $\alpha = 0.10$ in its oracle form for fixed values of $\eta \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.5\}$ and with η selected by the data-driven criteria.

Figure B.2 examines the distributions of observed FDP and Power and FNP from these 100 iterations. The left-hand side of each setting displays the observed FDP values for the method at each η, as defined in (3.23), and the horizontal, dashed line represents the nominal FDR level of $\alpha = 0.10$. Since FDR is defined as the expectation of FDP, the middle of the distributions of the observed FDP levels coinciding with the dashed line would signal an alignment of the method's observed FDR with the specified nominal level. The right-hand sign displays both the observed power, as defined in (3.23), and observed FNP for the proposed method at each η. The observed FNP for a general δ_i and θ_i is calculated as follows.

$$\text{FNP} = \frac{\sum_i \theta_i(1-\delta_i)}{\sum_i(1-\delta_i)}.$$

Notice, the η selected by minimizing $h\text{PFNR}$ shows lower observed FNP levels and higher observed power levels compared to using fixed $\eta \in \{0.01, 0.05, 0.10, 0.20, 0.30, 0.50\}$ while also controlling FDR nearly exactly at the nominal level of $\alpha = 0.10$. Among the fixed levels, allowing

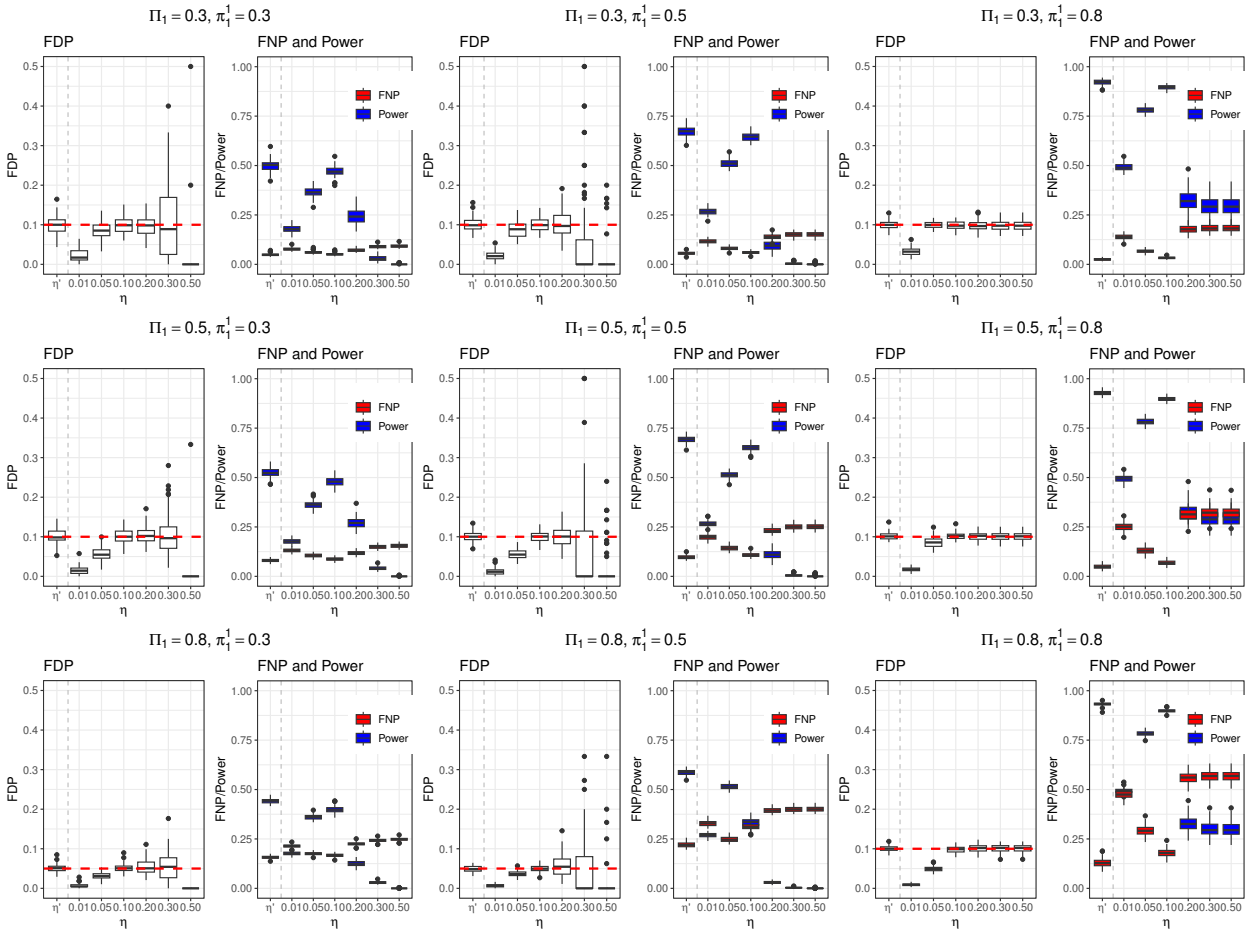


Figure B.2: Observed levels of FDP (left) and Power and FNP (right) from the proposed procure applied at the nominal FDR level of $\alpha = 0.10$ to 100 iterations of each simulation setting described in Section 3.6.1 with η selected by the data-driven criteria and for all fixed $\eta \in \{0.01, 0.05, 0.10, 0.20\}$. "Selected" represents the results for η selected by the criteria in Section 3.5.1. The horizontal, dashed line in the FDP figure represents the nominal level.

$\eta = \alpha = 0.10$ showed the lowest observed FNP levels and highest observed power levels. It does not, however, outperform the selection proposed selection criteria. Under certain settings, for example, when $\Pi_1 = 0.8$ and $\pi_1^1 \in \{0.3, 0.5\}$, selecting $\eta = \alpha = 0.1$ results in the hypothesis-level method being overly conservative relative to the selection criteria from Section 3.5.1. Figure B.3 further examines the distributions of the η that was selected by the proposed criteria. It is interesting to note that the optimal η in terms of $h\text{PFNR}$ for these 100 iterations tended to be between roughly 0.12 and 0.16, larger than the nominal level of $\alpha = 0.10$. This implies that when integrating group information, it is preferable to be more aggressive at the hypothesis level because η operates

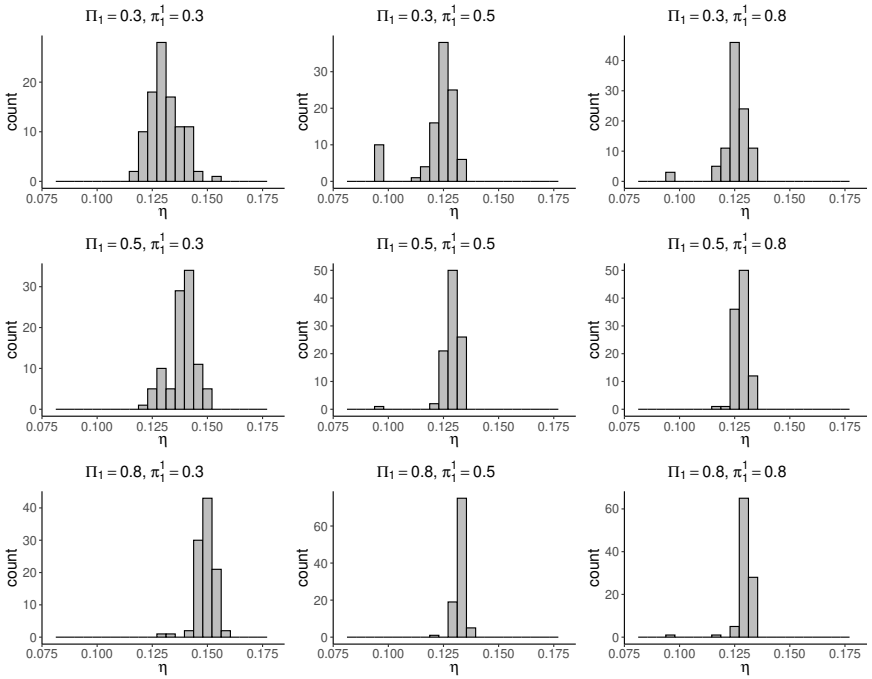


Figure B.3: The selected η that minimized $h\mathrm{PFNR}$ for the nominal FDR level of $\alpha = 0.10$ for 100 iterations of each simulation setting described in Section 3.6.1.

as the criteria for marking a hypothesis within a group as a candidate to be reproducible. There were some slight differences in distributions of the selected η across the levels of Π_1 and π_1^1. In particular, when there are fewer reproducible groups, or when Π_1 is smaller, the selected η tended to be slightly larger cases with more reproducible groups.

Figures B.4 and B.5 are the same as Figure B.2 but for $\alpha = 0.05$ and 0.20. In the same manner, as before, we see that any selection of η yields control of FDR at or below the specified nominal level and that selecting the η that minimizes $h\mathrm{PFNR}$ shows tends to have the smallest FNP and largest power in simulation. These simulations support the appropriateness of proposed η selection criteria from Section 3.5.1.

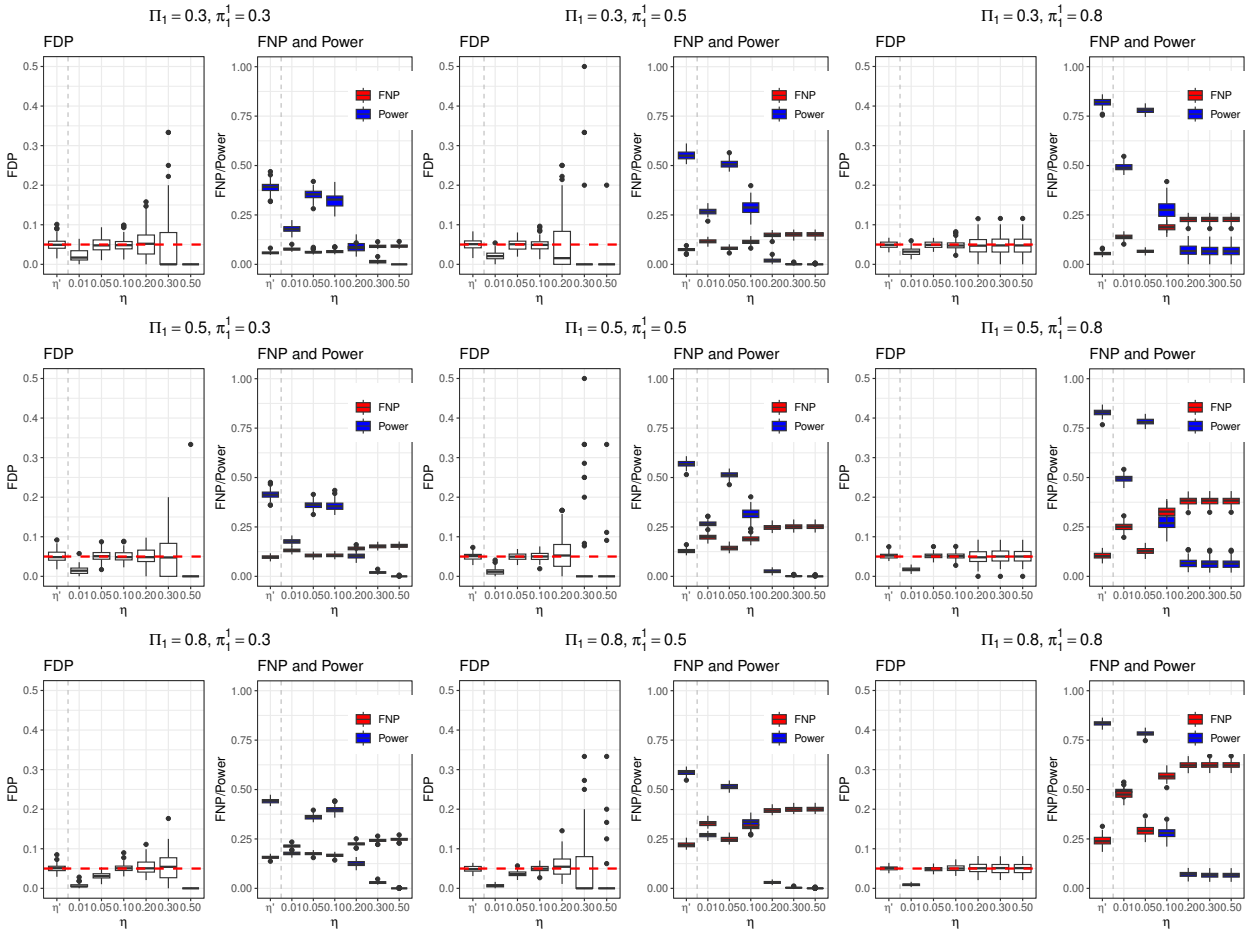


Figure B.4: Observed levels of FDP (left) and Power and FNP (right) from the proposed procure applied at the nominal FDR level of $\alpha = 0.05$ to 100 iterations of each simulation setting described in Section 3.6.1 with η selected by the data-driven criteria and for all fixed $\eta \in \{0.01, 0.05, 0.10, 0.20\}$. "Selected" represents the results for η selected by the criteria in Section 3.5.1. The horizontal, dashed line in the FDP figure represents the nominal level.

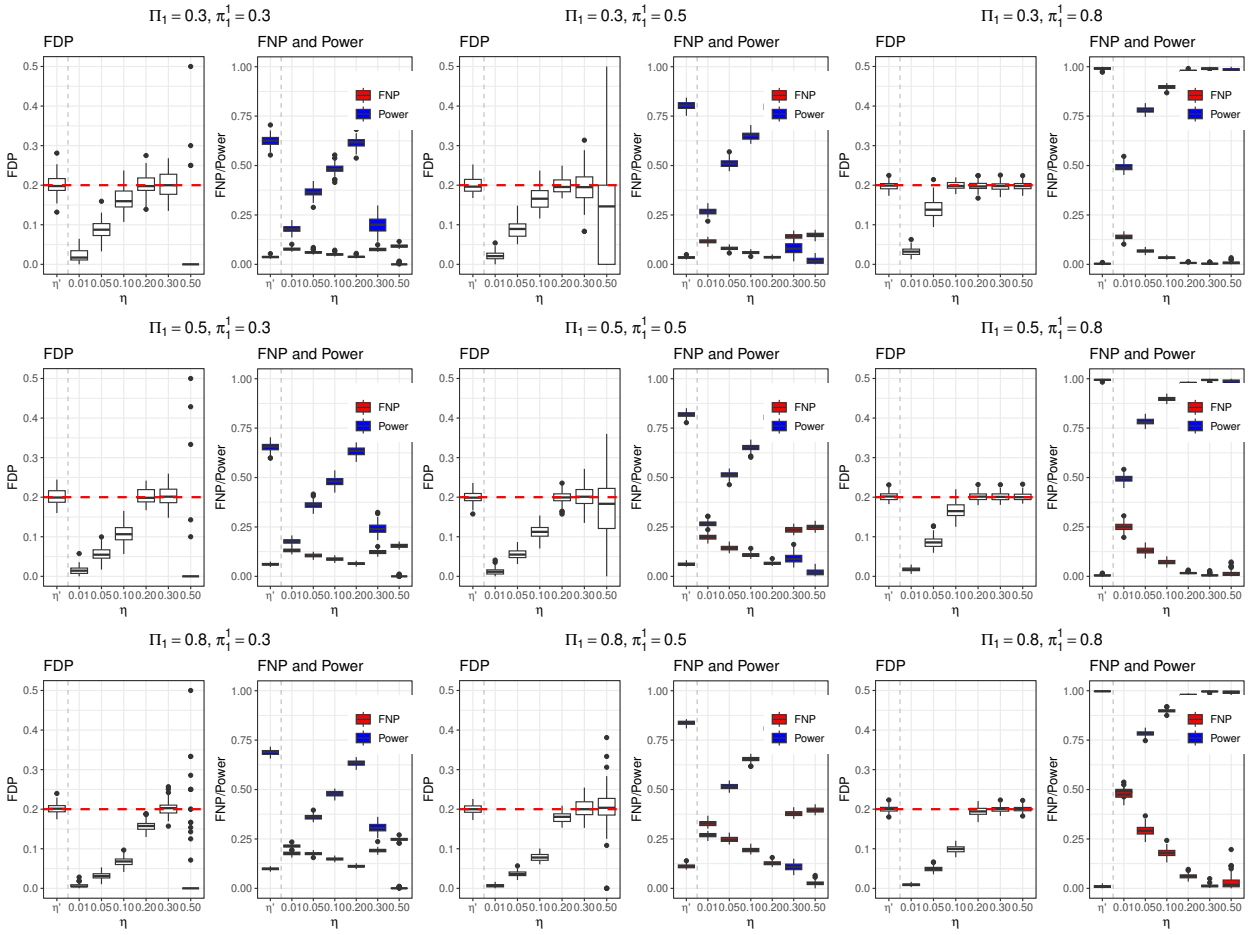


Figure B.5: Observed levels of FDP (left) and Power and FNP (right) from the proposed procure applied at the nominal FDR level of $\alpha = 0.20$ to 100 iterations of each simulation setting described in Section 3.6.1 with η selected by the data-driven criteria and for all fixed $\eta \in \{0.01, 0.05, 0.10, 0.20\}$. "Selected" represents the results for η selected by the criteria in Section 3.5.1. The horizontal, dashed line in the FDP figure represents the nominal level.

Appendix C

Supplemental materials for "*Assessing the reproducibility of results across multiple high-throughput studies using Kendall's W*"

These supplemental materials contain technical results and additional simulations for Chapter 4. Appendix C.1 details the proofs of all theoretical results from the chapter. Appendix C.2 provides the details of the simulation from Section 4.1.2. Appendix C.3 contains supplemental materials for the COVID-19 datasets.

Notations: Throughout this appendix, we assume there are n hypotheses common across m studies. The i^{th} hypothesis is denoted by $\mathbb{H}_i$. Define the sets irreproducible indices $(\mathcal{H}_0)$ and reproducible indices $(\mathcal{H}_1)$ by

$$\mathcal{H}_0 = \{i : \mathbb{H}_i \text{ is irreproduible}\} \text{ and } \mathcal{H}_1 = \{i : \mathbb{H}_i \text{ is reproduible}\}.$$

The control and test sets discussed in the chapter are denoted $\mathcal{C}_0$ and $\mathcal{D}_t$. We let $|\mathcal{C}_0| = n_0$ and $|\mathcal{D}_t| = n_1$. For each $\mathbb{H}_i$ we observe a summary statistics for each study denoted $\boldsymbol{t}_i = (t_{1i}, t_{2i}, \cdots t_{mi})$ which can be ranked within study from most (rank 1) to least (rank n) notable within experiment with ranks denoted by $(r_{1i}, r_{2i} \cdots, r_{mi})$. Throughout we use big O and big O_p notations. We use $\overset{D}{\to}$ to represent convergence in distribution.

C.1 Proofs

C.1.1 Proof of Theorem 2.1

Theorem 2.1. *Assume the global null assumption holds for a list of n hypotheses across m replicate studies. Then for $\mathbb{H}_i$,*

$$n\Delta W_{-i} \xrightarrow{D} V \equiv \frac{3(m-1)}{m} + \frac{12}{m^2}\sum_{k=1}^{m}\sum_{j\neq k} U_k(U_j - 1)$$

as $n \to \infty$ where $U_h \overset{\text{iid}}{\sim} \text{UNIF}(0,1)$ for $h \in \{1, 2, \cdots, n\}$.

Proof. Under the global null, $(r_{1i}, r_{2i}, \cdots, r_{mi})$ are marginally distributed by a independent discrete uniform distribution on $\{1, 2, \cdots, n\}$ and $(r_{j1}, \cdots, r_{jn})$ are random permutations of $\{1, 2, \cdots, n\}$ that are independent for each $j \in \{1, \cdots, m\}$. As such, it is important note that for any j, $\frac{r_{ji}}{n+1} \xrightarrow{D}$ $\text{UNIF}(0,1)$ jointly for all $j \in \{1, 2, \cdots, m\}$. Lemma C.1.1 reshapes the ΔW_{-i} statistic into the form closer to that of the asymptotic global null distribution. Using Lemma C.1.1, the desired result is proven by the joint convergence $\frac{r_{ji}}{n+1}$ independent uniform random variables, and the continuous mapping theorem. That is, we have,

$$\begin{aligned} n\Delta W_{-i} &= n\left[O_p(n^{-3/2}) + \frac{3(m-1)}{nm} + \frac{12}{nm^2}\sum_{k=1}^{m}\sum_{j\neq k}\frac{r_{ki}}{n+1}\left(\frac{r_{ji}}{n+1} - 1\right)\right] \quad \text{(Lemma C.1.1)} \\ &\xrightarrow{D} \frac{3(m-1)}{m} + \frac{12}{m^2}\sum_{k=1}^{m}\sum_{j\neq k} U_k(U_j - 1) \end{aligned}$$

where $U_h \overset{iid}{\sim} \text{UNIF}(0,1)$ for all $h \in \{1, 2, \cdots m\}$ and the convergence follows from the joint convergence of $\frac{r_{hi}}{n+1}$ to U_h for all h and the continuous mapping theorem.

□

Lemma C.1.1. *Assume the global null assumption holds for a list of n hypotheses across m replicate studies. Then for $\mathbb{H}_i$, then*

$$\Delta W_{-i} = O_p(n^{-3/2}) + \frac{3(m-1)}{nm} + \frac{12}{nm^2}\sum_{k=1}^{m}\sum_{j\neq k}\frac{r_{ki}}{n+1}\left(\frac{r_{ji}}{n+1} - 1\right)$$

Proof. The global null implies $(r_{1i}, r_{2i}, \cdots, r_{mi})$ are marginally distributed by a independent discrete uniform distribution on $\{1, 2, \cdots, n\}$ and $(r_{j1}, \cdots, r_{jn})$ are random permutations of $\{1, 2, \cdots, n\}$ that are independent for each $j \in \{1, \cdots, m\}$. Throughout the proof, we use $r_{j\ell}^{-h}$ to the ranking of hypotheses ℓ in study j when all hypotheses except $\mathbb{H}_h$ are considered. Notice

$$r_{j\ell}^{-h} = r_{j\ell} - \mathbb{I}(r_{j\ell} > r_{jh}).$$

As a reminder,

$$\Delta W_{-i} = W - W_{-i}$$

where

$$W = \frac{12\sum_{\ell=1}^{n}(R_\ell - \bar{R})^2}{m^2(n^3 - n)}$$

with $R_\ell = \sum_{j=1}^{m} r_{j\ell}$ and $\bar{R} = \frac{1}{n}\sum_{\ell=1}^{n} R_\ell$. Similarly,

$$W_{-i} = \frac{12\sum_{\ell=1}^{n}(R_\ell^{-i} - \bar{R}^{-i})^2}{m^2((n-1)^3 - (n-1))}$$

with $R_\ell^{-i} = \sum_{j=1}^{m} r_{j\ell}^{-i} = \sum$ and $\bar{R}^{-i} = \frac{1}{n-1}\sum_{\ell\neq i} R_\ell^{-i}$. Now, we establish some equalities that are critical throughout the proof.

Some Useful Equalities.

$$\bar{R} = \frac{1}{n}\sum_{i=1}^{n} R_i = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m} r_{ji} = \frac{1}{n}\sum_{j=1}^{m}\sum_{i=1}^{n} i = \frac{m}{n}\left(\frac{n(n+1)}{2}\right) = \frac{m(n+1)}{2}. \tag{C.1}$$

$$\sum_{\ell=1}^{n}\sum_{j=1}^{m} r_{j\ell}^2 = \sum_{j=1}^{m}\sum_{\ell=1}^{n} \ell^2 = m\left(\frac{n(n+1)(2n+1)}{6}\right). \tag{C.2}$$

$$\begin{aligned}
\sum_{j=1}^{m}\sum_{i\neq h} r_{ij}\mathbb{I}(r_{ji} > r_{jh}) &= \sum_{j=1}^{m}\left[(r_{jh}+1) + (r_{jh}+2) + \cdots + n\right] \\
&= \sum_{j=1}^{m}\sum_{i=r_{jh}+1}^{n} i \\
&= \sum_{j=1}^{m}\frac{(n+r_{jh}+1)(n-r_{jh})}{2} \\
&= \frac{mn(n+1)}{2} - \frac{1}{2}\sum_{j=1}^{m} r_{jh}(r_{jh}+1).
\end{aligned} \tag{C.3}$$

$$\sum_{i\neq h}\mathbb{I}(r_{ji} > r_{jh}) = \sum_{r_{ji}>r_{jh}}\mathbb{I}(r_{ji} > r_{jh}) + \sum_{r_{ji}<r_{jh}}\mathbb{I}(r_{ji} > r_{jh}) = n - r_{jh}. \tag{C.4}$$

We begin by writing ΔW_{-i} into a function of r_{ji} values. Notice

$$\begin{aligned}
W &= \frac{12\sum_{\ell=1}^{n}\left(R_\ell - \bar{R}\right)^2}{m^2(n^3-n)} \\
&= \frac{12}{m^2(n^3-n)}\left[\sum_{\ell=1}^{n}\left(R_\ell^2 - \bar{R}R_\ell + \bar{R}^2\right)\right] \\
&= \frac{12}{m^2(n^3-n)}\left[\sum_{\ell=1}^{n} R_\ell^2 - 2\bar{R}\sum_{\ell=1}^{n} R_\ell + n\bar{R}^2\right] \\
&= \frac{12}{m^2(n^3-n)}\left[\sum_{\ell=1}^{n} R_\ell^2 - 2n\bar{R}^2 + n\bar{R}^2\right] \\
&= \frac{12}{m^2(n^3-n)}\left[\sum_{\ell=1}^{n} R_\ell^2 - n\bar{R}^2\right] \\
\text{by (C.1)} \quad &= \frac{12}{m^2(n^3-n)}\left[\sum_{\ell=1}^{n} R_\ell^2 - n\left(\frac{m(n+1)}{2}\right)^2\right] \\
&= \frac{12\sum_{\ell=1}^{n} R_\ell^2}{m^2(n^3-n)} - \frac{3(n+1)^2}{(n^2-1)} \\
&= \frac{12\sum_{\ell=1}^{n}\left(\sum_{j=1}^{m} r_{j\ell}\right)^2}{m^2(n^3-n)} - \frac{3(n+1)}{n-1}
\end{aligned} \tag{C.5}$$

and following the same process,

$$\begin{aligned} W_{-i} &= \frac{12\sum_{\ell\neq i}\left(\sum_{j=1}^{m} r_{j\ell}^{-i}\right)^2}{m^2((n-1)^3-(n-1))} - \frac{3n}{n-2} \\ &= \frac{12\sum_{\ell\neq i}\left(\sum_{j=1}^{m} r_{j\ell} - \mathbb{I}(r_{j\ell} > r_{ji})\right)^2}{m^2 n(n-1)(n-2)} - \frac{3n}{n-2}. \end{aligned} \tag{C.6}$$

Using (C.5) and (C.6), we have

$$\begin{aligned} \Delta W_{-i} &= \frac{12\sum_{\ell=1}^{n}\left(\sum_{j=1}^{m} r_{j\ell}\right)^2}{m^2(n^3-n)} - \frac{3(n+1)}{n-1} - \frac{12\sum_{\ell\neq i}\left(\sum_{j=1}^{m} r_{j\ell} - \mathbb{I}(r_{j\ell} > r_{ji})\right)^2}{m^2 n(n-1)(n-2)} + \frac{3n}{n-2} \\ &= C(n) + \frac{12}{m^2 n(n-1)}\left[\frac{\sum_{\ell=1}^{n}(\sum_{j=1}^{m} r_{j\ell})^2}{n+1} - \frac{\sum_{\ell\neq i}\left(\sum_{j=1}^{m} r_{j\ell} - \mathbb{I}(r_{j\ell} > r_{ji})\right)^2}{(n-2)}\right] \end{aligned} \tag{C.7}$$

where $C(n) = \frac{3n(n-1)+3(n+1)(n-2)}{(n-1)(n-2)}$. Proceed by simplifying (C.7) by examining the term in the brackets as follows.

$$\begin{aligned} &\frac{\sum_{\ell}^{n}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{n+1} - \frac{\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell} - \mathbb{I}(r_{j\ell} > r_{ji})\right)^2}{n-2} \\ &= \frac{\sum_{\ell}^{n}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{n+1} \overset{(a)}{-} \frac{\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{n-2} \\ &\quad \overset{(b)}{+} \frac{2\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell}\right)\left(\sum_{j}^{m}\mathbb{I}(r_{j\ell} > r_{ji})\right)}{n-2} \\ &\quad \overset{(c)}{-} \frac{\sum_{\ell\neq i}\left(\sum_{j}^{m}\mathbb{I}(r_{j\ell} > r_{ji})\right)^2}{n-2}. \end{aligned} \tag{C.8}$$

We now examine (a), (b), and (c) from (C.8) on their own.

$$\begin{aligned} (a) &= \left[\frac{\left(\sum_{j}^{m} r_{ji}\right)^2}{n+1} + \frac{\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{n+1}\right] - \frac{\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{n-2} \\ &= \frac{\left(\sum_{j}^{m} r_{ji}\right)^2}{n+1} - \frac{3\sum_{\ell\neq i}\left(\sum_{j}^{m} r_{j\ell}\right)^2}{(n+1)(n-2)} \end{aligned}$$

$$(b) = \frac{2}{n-2}\sum_{\ell\neq i}\left[\sum_{j=1}^{m} r_{j\ell}\mathbb{I}(r_{j\ell} > r_{ji}) + \sum_{j=1}^{m}\sum_{k\neq j} r_{j\ell}\mathbb{I}(r_{k\ell} > r_{ki})\right]$$

$$\text{by (C.3)} \quad = \frac{2}{n-2}\left[\frac{mn(n+1)}{2} - \frac{1}{2}\sum_{j=1}^{m} r_{ji}(r_{ji}+1) + \sum_{\ell\neq i}\sum_{j=1}^{m}\sum_{k\neq j} r_{j\ell}\mathbb{I}(r_{k\ell} > r_{ki})\right]$$

$$(c) = O(1).$$

Continuing to expand and combine terms, the form from (C.8) can now be written like

$$\Delta W_{-i} = O(n^{-2}) + \frac{12}{m^2 n(n-1)}\left((a)+(b)\right) = O(n^{-2}) + \frac{12}{m^2 n(n-1)}\left((\mathrm{I})+(\mathrm{II})+(\mathrm{III})\right). \quad \text{(C.9)}$$

where

$$\begin{aligned}(\mathrm{I}) &= \left(\frac{1}{n+1} - \frac{1}{n-2}\right)\sum_{j=1}^{m} r_{ji}^2 - \frac{3}{(n+1)(n+2)}\sum_{j=1}^{m}\sum_{\ell\neq i} r_{j\ell}^2 + \frac{mn(n+1)}{n-2} \\ &= -\frac{3}{(n+1)(n-2)}\sum_{j=1}^{m}\sum_{\ell=1}^{n} r_{j\ell}^2 + \frac{mn(n+1)}{n-2} \\ \text{by (C.2)} \quad &= -\frac{mn(2n+1)}{2(n-2)} + \frac{mn(n+1)}{n-2} \\ &= \frac{2mn+1}{2(n-2)} = O(1),\end{aligned}$$

$$(\mathrm{II}) = \frac{1}{n+1}\sum_{j\neq k} r_{ji}r_{ki} - \frac{3}{(n+1)(n-2)}\sum_{j\neq k}\sum_{\ell\neq i} r_{j\ell}r_{ki}$$

and

$$\begin{aligned}(\mathrm{III}) &= -\frac{1}{n-2}\sum_{j=1}^{m} r_{ji} + \frac{2}{n-2}\sum_{\ell\neq i}\sum_{j\neq k} r_{j\ell}\mathbb{I}(r_{k\ell} > r_{ki}) \\ &= O(1) + \frac{2}{n-2}\sum_{\ell\neq i}\sum_{j\neq k} r_{j\ell}\mathbb{I}(r_{k\ell} > r_{ki}).\end{aligned} \quad \text{(C.10)}$$

Plugging (I), (II), and (III) into (C.9), we have

$$\begin{aligned}\Delta W_{-i} &= O(n^{-2}) + \left((\mathrm{I})+(\mathrm{II})+(\mathrm{III})\right) \\ &= O(n^{-2}) + \frac{12}{m^2 n(n-1)}\left((\mathrm{II})+(\mathrm{III})\right) \\ &= O(n^{-2}) + \frac{12}{m^2 n(n-1)}\sum_{j=1}^{m}\sum_{\ell=1}^{n} c_{j\ell}\frac{r_{j\ell}}{n+1}\end{aligned} \quad \text{(C.11)}$$

where

$$c_{j\ell} = \mathbb{I}(\ell = i)\sum_{k\neq j} r_{ki} + \mathbb{I}(\ell \neq i)\frac{1}{n-2}\left[2(n+1)\sum_{k\neq j}\mathbb{I}(r_{k\ell} > r_{ki}) - 3\sum_{k\neq j} r_{k\ell}\right].$$

Note that $c_{j\ell} = O(1)$ uniformly for all k and $\ell \neq i$. Notice

$$\sum_{\ell=1}^{n} c_{j\ell}\frac{r_{j\ell}}{n+1} = \frac{n\bar{c}_n}{2} + \sum_{\ell=1}^{n}(c_{j\ell}\mathbb{I}(\ell \neq i) - \bar{c}_n)\frac{r_{j\ell}}{n+1} + \frac{r_{ji}}{n+1}\sum_{k\neq j} r_{ki} \tag{C.12}$$

where

$$\begin{aligned}
\bar{c}_n &= n^{-1}\sum_{\ell=1}^{n} c_{j\ell}\mathbb{I}(\ell \neq i) \\
&= n^{-1}\left[-\frac{3}{n-2}\sum_{\ell\neq i}\sum_{k\neq j} r_{k\ell} + \frac{2(n+1)}{n-2}\sum_{\ell\neq i}\sum_{k\neq j}\mathbb{I}(r_{k\ell} > r_{ki})\right] \\
\text{by (C.4)} \quad &= n^{-1}\left[-\frac{3}{n-}\left(\frac{(m-1)n(n+1)}{2} - \sum_{k\neq j} r_{ki}\right) + \frac{2(n+1)}{n-2}(m-1)n - \sum_{k\neq j} r_{ki}\right] \\
&= n^{-1}\left(-\frac{2n-1}{n-2}\sum_{k\neq j} r_{ki} + \frac{(m-1)n(n+1)}{2(n-2)}\right) = O(1).
\end{aligned}$$

By the continuous mapping theorem and the joint convergence of $\frac{r_{ji}}{n+1}$ across j, it follows that

$$\sum_{\ell=1}^{n}(c_{j\ell}\mathbb{I}(\ell \neq i) - \bar{c}_n)\frac{r_{j\ell}}{n+1} \xrightarrow{D} T_n$$

where $T_n \sum_{\ell=1}^{n}(c_{j\ell}\mathbb{I}(\ell \neq i) - \bar{c}_n)U_\ell$ with $U_\ell \sim \text{UNIF}(0,1)$. Notice then that $\mathbb{V}\text{ar}(T_n) = O(n)$, so $\sum_{\ell=1}^{n}(c_{j\ell}\mathbb{I}(\ell \neq i) - \bar{c}_n)\frac{r_{j\ell}}{n+1} = O_p(\sqrt{n})$. Thus, it follows that the term in C.12 is simplified by

$$\begin{aligned}
\sum_{\ell=1}^{n} c_{j\ell}\frac{r_{j\ell}}{n+1} &= O_p(\sqrt{n}) + \frac{n\bar{c}_n}{2} + \frac{r_{ji}}{n+1}\sum_{k\neq j} r_{ki} \\
&= O_p(\sqrt{n}) + \frac{(m-1)n(n+1)}{4(n-2)} + \left(\frac{r_{ji}}{n+1} - \frac{2n-1}{2(n-2)}\right)\sum_{k\neq i} r_{hi}.
\end{aligned}$$

Finally, we can solve the desired result for ΔW_{-i} by starting from (C.11).

$$
\begin{aligned}
n\Delta W_{-i} &= n\left[O(n^{-2}) + \frac{12}{m^2 n(n-1)} \sum_{j=1}^{m} \sum_{\ell=1}^{n} c_{j\ell} \frac{r_{j\ell}}{n+1}\right] \\
&= O_p(n^{-3/2}) + \frac{3(m-1)}{m}\left(n^{-1} + O(n^{-2})\right) \\
&\qquad + \frac{12}{m^2} \sum_{k=1}^{m} \frac{r_{ji}}{n+1} \left(\sum_{j \neq k} \frac{r_{ji}}{n+1}\right)\left(n^{-1} + O(n^{-2})\right) \\
&\qquad - \frac{12}{m^2} \sum_{k=1}^{m} \sum_{j \neq k} \frac{r_{ji}}{n+1}(n^{-1} + O(n^{-2})) \\
&= O_p(n^{-3/2}) + \frac{3(m-1)}{nm} + \frac{12}{nm^2} \sum_{k=1}^{m} \sum_{j \neq k} \frac{r_{ki}}{n+1}\left(\frac{r_{ji}}{n+1} - 1\right).
\end{aligned}
$$

□

C.1.2 Proof of Proposition 4.2.1

Proposition 4.2.1. *Suppose $\mathcal{C}_0$ is a random sampled from $\mathcal{H}_0$. Consider $i \in \{1, 2, \cdots, d\}$ hypotheses that are in $\mathcal{D}_t$ and irreproducible. Then, the vector of conformal p-values $(p_1^{\text{con}}, p_2^{\text{con}}, \cdots, p_d^{\text{con}})$ referring to these hypotheses are positive regression dependent on a subset (PRDS).*

Proof. The proof takes the exact form of the proof of Proposition 4 from Appendix A.3 in Bates et al. (2023) with the modification because in our context, larger values for the score value, ΔW_{-i}, provide evidence of reproducibility. That is, let $Z = (S_{(1)}, S_{(2)}, \cdots, S_{(n)})$ be the order statistics of $(\Delta W_{-i})_{i \in \mathcal{C}_0^c}$, the scores calculated for hypotheses in the calibration set in the manner described in Algorithm 6 and $Y = (p_g^{\text{con}})_{g \in \mathcal{D}_t}$ be conformal p-values calculated for hypotheses in the test set. Notice that

$$
\begin{aligned}
\mathbb{P}\left(Y \in A \mid p_i^{\text{con}} = y\right) &= \int_z \mathbb{P}\left(Y \in A \mid Z = z\right) \mathbb{P}(Z = z \mid p_i^{\text{con}} = y) dz \\
&= \mathbb{E}_{Z|p_i^{\text{con}}=y} \mathbb{P}(Y \in A|Z).
\end{aligned} \tag{C.13}
$$

Now, PRDS for irreproducible hypotheses in $\mathcal{D}_t$ can be shown by proving Lemmas C.1.2 and C.1.3 (or Lemmas 4 and 5 in Appendix A.3 in Bates et al. (2023)). Effectively, Lemma C.1.2 states that as ΔW_{-i} scores for hypotheses in the calibration set increase, so too will conformal p-values for hypotheses in the test set will.

Lemma C.1.2. *Let A be a non-decreasing set, and z and z' be vectors such that $z' \preceq z$. Then*

$$\mathbb{P}(Y \in A \mid Z = z) \geq \mathbb{P}(Y \in A \mid Z = z').$$

Proof of Lemma C.1.2 This proof follows the same path as the proof of Lemma 4 in Bates et al. (2023). Since ΔW_{-i} is calculated using ranks, these scores are discrete there is a non-zero probability of ties in ΔW_{-i} statistics. Thus, in the calculation of p_i^{con}, we settle these ties randomly. That is, remember

$$p_i^{\text{con}} = \frac{\sum_{\ell \in \mathcal{C}_0^c} \mathbb{I}\left[\Delta W_{-i} < \Delta W_{-\ell}\right] + \left\lceil U_i \left(1 + \sum_{\ell \in \mathcal{C}_0^c} \mathbb{I}\left[\Delta W_{-i} = \Delta W_{-\ell}\right]\right)\right\rceil}{1 + |\mathcal{C}_0^c|} \tag{C.14}$$

where $U_i \overset{\text{iid}}{\sim} \text{UNIF}(0,1)$. Consider $U = (U_i)_{i \in \mathcal{D}_t}$. Since $U \perp Z, Y$, we have

$$\mathbb{P}(Y \in A \mid Z = z, U) = \mathbb{P}(Y \in A \mid Z = z) \text{ a.s.} \tag{C.15}$$

Now, let

$$p_i^{\text{con}}(x; z, u) = \frac{\sum_\ell \mathbb{I}[x < z_\ell] + \lceil u \left(1 + \sum_\ell \mathbb{I}[x = z_\ell]\right)\rceil}{1 + |z|} \tag{C.16}$$

be the function that takes one ΔW_{-i} statistic from the test set, the vector $(\Delta W_{-\ell})_{\ell \in \mathcal{C}_0^c}$ from the calibration set, and U and calculates p_i^{con}. We will show that $p_i^{\text{con}}(x; z, u)$ is non-decreasing in terms of z.

Take z, z' such that $z' \preceq z$. The following inequalities are immediate.

$$\begin{gathered} \sum_\ell \mathbb{I}[x < z_\ell] \geq \sum_\ell \mathbb{I}[x < z'_\ell] \\ \text{and} \\ \sum_\ell \mathbb{I}[x < z_\ell] + \sum_\ell \mathbb{I}[x = z_\ell] = \sum_\ell \mathbb{I}[x \leq z_\ell] \geq \sum_\ell \mathbb{I}[x \leq z'_\ell] = \sum_\ell \mathbb{I}[x < z'_\ell] + \sum_\ell \mathbb{I}[x = z'_\ell]. \end{gathered} \tag{C.17}$$

So, we can show $p_i^{\text{con}}(x; z, u) \geq p_i^{\text{con}}(x; z', u)$ by considering three cases:

1. Assume $\sum_\ell \mathbb{I}[x < z_\ell] = \sum_\ell \mathbb{I}[x < z'_\ell]$. Then, second inequality in (C.17) implies

$$\begin{gathered}\sum_\ell \mathbb{I}[x = z_\ell] \geq \sum_\ell \mathbb{I}[x = z'_\ell], \\ \text{and} \\ \lceil u(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \rceil \geq \lceil u(1 + \sum_\ell \mathbb{I}[x = z'_\ell]) \rceil.\end{gathered} \tag{C.18}$$

Now, it is immediate that

$$\begin{aligned} p_i^{\text{con}}(x; z, u) &= \frac{\sum_\ell \mathbb{I}[x < z_\ell] + \lceil u\left(1 + \sum_\ell \mathbb{I}[x = z_\ell]\right) \rceil}{1 + |z|} \\ &\geq \frac{\sum_\ell \mathbb{I}[x < z'_\ell] + \left\lceil u\left(1 + \sum_\ell \mathbb{I}[x = z'_\ell]\right) \right\rceil}{1 + |z'|} \\ &= p_i^{\text{con}}(x; z', u). \end{aligned} \tag{C.19}$$

2. Assume $\sum_\ell \mathbb{I}[x < z_\ell] + \sum_\ell \mathbb{I}[x = z_\ell] = \sum_\ell \mathbb{I}[x < z'_\ell] + \sum_\ell \mathbb{I}[x = z'_\ell]$. Let $a = \mathbb{I}[x < z_\ell]$, $b = 1 + \mathbb{I}[x = z_\ell]$ and $a' = \mathbb{I}[x < z'_\ell]$, $b' = 1 + \mathbb{I}[x = z'_\ell]$ and let $u \in [0, 1]$. Notice that the numerators in $p_i^{\text{con}}(x; z, u)$ and $p_i^{\text{con}}(x; z', u)$ are $a + \lceil ub \rceil$ and $a' + \lceil ub' \rceil$, respectively. Additionally, the following equality follows immediately from the assumption

   It follows that

$$\begin{aligned} a' + \lceil ub' \rceil &= a' + \lceil u(a - a') + ub \rceil \\ &\leq a' + \lceil u(a - a') \rceil + \lceil ub \rceil \\ &\leq a' + (a - a') + \lceil ub \rceil \\ &= a + \lceil ub \rceil \end{aligned} \tag{C.20}$$

   where first equality in (C.20) holds by the assumption for this case, and the second inequality in (C.20) holds because a and a' are integers, $a > a'$ by (C.17), and u is bounded by 1. Now,

the desired result is immediate.

$$\begin{aligned} p_i^{\mathrm{con}}(x; z, u) &= \frac{\sum_\ell \mathbb{I}[x < z_\ell] + \lceil u\,(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \rceil}{1 + |z|} \\ &= \frac{a + \lceil ub \rceil}{} \\ &\leq \frac{a' + \lceil ub' \rceil}{1 + |z'|} \\ &= \frac{\sum_\ell \mathbb{I}[x < z'_\ell] + \left\lceil u\left(1 + \sum_\ell \mathbb{I}[x = z'_\ell]\right) \right\rceil}{1 + |z'|} \\ &= p_i^{\mathrm{con}}(x; z', u). \end{aligned} \tag{C.21}$$

3. Assume $\sum_\ell \mathbb{I}[x < z_\ell] > \sum_\ell \mathbb{I}[x < z'_\ell]$ and $\sum_\ell \mathbb{I}[x < z_\ell] + \sum_\ell \mathbb{I}[x = z_\ell] > \sum_\ell \mathbb{I}[x < z'_\ell] + \sum_\ell \mathbb{I}[x = z'_\ell]$. Since $\sum_\ell \mathbb{I}[x < z_\ell]$, $\sum_\ell \mathbb{I}[x < z'_\ell]$, $\sum_\ell \mathbb{I}[x = z_\ell]$, and $\sum_\ell \mathbb{I}[x = z'_\ell]$ are all integers, the assumed inequalities imply

$$\begin{gathered} \sum_\ell \mathbb{I}[x < z_\ell] \geq \sum_\ell \mathbb{I}[x < z'_\ell] + 1 \\ \text{and} \\ \sum_\ell \mathbb{I}[x < z_\ell] + \sum_\ell \mathbb{I}[x = z_\ell] \geq \sum_\ell \mathbb{I}[x < z'_\ell] + \sum_\ell \mathbb{I}[x = z'_\ell] + 1. \end{gathered} \tag{C.22}$$

So it follows that

$$\begin{aligned}
\sum_\ell \mathbb{I}[x < z_\ell] + \lceil u(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \rceil &\geq \sum_\ell \mathbb{I}[x < z_\ell] + u(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \\
&= \sum_\ell \mathbb{I}[x < z_\ell] + u(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \\
&\quad + u \sum_\ell \mathbb{I}[x < z_\ell] - u \sum_\ell \mathbb{I}[x < z_\ell] \\
&= (1 - u) \sum_\ell \mathbb{I}[x < z_\ell] \\
&\quad + u(1 + \sum_\ell \mathbb{I}[x = z_\ell] + \sum_\ell \mathbb{I}[x < z_\ell]) \qquad \text{(C.23)} \\
&\geq (1 - u)(\sum_\ell \mathbb{I}[x < z'_\ell] + 1) \\
&\quad + u(2 + \sum_\ell \mathbb{I}[x = z'_\ell] + \sum_\ell \mathbb{I}[x < z'_\ell]) \\
&= \sum_\ell \mathbb{I}[x < z'_\ell] + u(1 + \sum_\ell \mathbb{I}[x < z'_\ell]) + 1 \\
&\geq \sum_\ell \mathbb{I}[x < z'_\ell] + \lceil u(1 + \sum_\ell \mathbb{I}[x < z'_\ell]) \rceil.
\end{aligned}$$

Finally, (C.23) implies the desired result

$$\begin{aligned}
p_i^{\text{con}}(x; z, u) &= \frac{\sum_\ell \mathbb{I}[x < z_\ell] + \lceil u\,(1 + \sum_\ell \mathbb{I}[x = z_\ell]) \rceil}{1 + |z|} \\
&\geq \frac{\sum_\ell \mathbb{I}[x < z'_\ell] + \left\lceil u\left(1 + \sum_\ell \mathbb{I}[x = z'_\ell]\right) \right\rceil}{1 + |z'|} \qquad \text{(C.24)} \\
&= p_i^{\text{con}}(x; z', u).
\end{aligned}$$

By the three cases, we know the function $p_i^{\text{con}}(x; z, u)$ is non-decreasing in terms of z, and thus conformal the p-values for the test set, Y, increase as the ΔW_{-i} statistics from the calibration set Z, increase. That is, by the cases presented above and the mutual independence of Z, U, and

$(\Delta W_{-i})_{i\in\mathcal{D}_t}$, we have for any $z' \preceq z$

$$\begin{aligned}\mathbb{P}(Y \in A|Z = z) &= \mathbb{P}(Y \in A|Z = z, U)\\ &\geq \mathbb{P}(Y \in A|Z = z', U)\\ &= \mathbb{P}(Y \in A|Z = z') \text{ a.s.}\end{aligned}$$

□

Lemma C.1.3. *Let hypothesis $i \in \mathcal{D}_t$ be irreproducible. Then, for $y \geq y'$, there exist $Z_1 \sim Z \mid p_i^{\text{con}} = y$ and $Z_2 \sim Z \mid p_i^{\text{con}} = y'$ such that*

$$\mathbb{P}(Z_2 \preceq Z_1) = 1.$$

Proof of Lemma C.1.3 This proof takes the same form as that for Lemma 5 in Bates et al. (2023). Suppose $i \in \mathcal{D}_t$ is irreproducible. Remember, $\left(S_{(|1|)}, S_{(2)}, \cdots, S_{(|\mathcal{C}_0^c|)}\right)$ are the order statistics for $(\Delta W_{-\ell})_{\ell\in\mathcal{C}_0^c}$ and denote the order statistics for $(\Delta W_j)_{j\in\mathcal{C}_0^c\cup\{i\}}$ as $\left(S'_{(1)}, S'_{(2)}, \cdots, S'_{(|\mathcal{C}_0^c|)}, S'_{(|\mathcal{C}_0^c|+1)}\right)$. Let $R_i = \left(1 + |\mathcal{C}_0^c|\right) p_i^{\text{con}}$. R_i is the ranking of ΔW_{-i} (with the largest ΔW_{-i} statistic rank 1) among $(\Delta W_{-\ell})_{\ell\in\mathcal{C}_0^c\cup\{i\}}$ where any ties are settled randomly. Since $R_i = k$ implies $\Delta W_{-i} = S'_{(|\mathcal{C}_0^c|+1-(k-1)|)}$, it is easy to verify

$$\left\{\left(S_{(\ell)}\right)_{\ell\in\{1,\cdots,|\mathcal{C}_0^c|\}} \mid S'_{(1)}, S'_{(2)}, \cdots, S'_{(|\mathcal{C}_0^c|)}, S'_{(|\mathcal{C}_0^c|+1)}, R_i = k\right\} = \left(S'_{(\ell)}\right)_{\ell\in\{1,\cdots,|\mathcal{C}_0^c|+1\}\setminus\{k^-\}}. \tag{C.25}$$

where $k^- = |\mathcal{C}_0^c| + 1 - (k-1)$ Now, we show that R_i is independent of $\left(S'_\ell\right)_{\ell\in\{1,\cdots,|\mathcal{C}_0^c|+1\}}$. Let σ represent a random permutation of the numbers $\{1, 2, \cdots|\mathcal{C}_0^c| + 1\}$. Note that by Definition 4.1.1, the vectors of summary statistics, $\boldsymbol{t}_j$ for all irreproducible hypotheses are independent and identically

distributed. So conditional on $\left(S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}\right) = (a_1, \cdots, a_{|\mathcal{C}_0^c|+1})$, it follows that

$$\left\{(\Delta W_{-\ell})_{\ell \in \{\mathcal{C}_0^c \cup \{i\}\}} \mid (S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}) = (a_1, \cdots, a_{|\mathcal{C}_0^c|+1})\right\} \stackrel{d}{=} \left(a_{\sigma(1)}, \cdots a_{\sigma(|\mathcal{C}_0^c|+1)}\right). \tag{C.26}$$

Let the set $I_{k^-} = \{\ell : a_\ell = a_{k^-}\}$ be the set of all indexes such that the corresponding order statistic is equal to the k^- order statistic. To show independence, we first show that, conditional on $\left(S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}\right) = (a_1, \cdots, a_{|\mathcal{C}_0^c|+1})$, $R_i = k$ if and only if

$$\Delta W_{-i} \in \{a_j : j \in I_{k^-}\} \cap U_i \in \left(\frac{\max(I_{k^-}) - k^-}{|I_{k^-}|}, \frac{\max(I_{k^-}) - k^- + 1}{|I_{k^-}|}\right]. \tag{C.27}$$

First, assume $R_i = k$, then $\Delta W_{-i} = a_{k^-}$ and thus $\Delta W_{-i} \in \{a_j : j \in I_{k^-}\}$. Additionally, we have

$$\sum_{\ell \neq i} \mathbb{I}[\Delta W_{-i} < \Delta W_{-\ell}] + \left\lceil U_i(1 + \sum_{\ell \neq i} \mathbb{I}[\Delta W_{-i} = \Delta W_{-\ell}]) \right\rceil = k. \tag{C.28}$$

Since $\Delta W_{-i} \in \{a_j : j \in I_{k^-}\}$, we know $\mathbb{I}[\Delta W_{-i} < \Delta W_{-\ell}] = (|\mathcal{C}_0^c| + 1) - \max(I_{k^-})$. Now,

$$\begin{aligned}
k - ((|\mathcal{C}_0^c| + 1) - \max(I_{k^-})) &= \max(I_{k^-}) - (k^-) + 1 \\
&= \left\lceil U_i(1 + \sum_{\ell \neq i} \mathbb{I}[\Delta W_{-i} = \Delta W_{-\ell}]) \right\rceil \\
&= \left\lceil U_i \sum_{\ell} \mathbb{I}[\Delta W_{-i} = \Delta W_{-\ell}] \right\rceil \\
&= \left\lceil U_i \sum_{\ell} \mathbb{I}[a_{k^-} = \Delta W_{-\ell}] \right\rceil \\
&= \left\lceil U_i \sum_{j} \mathbb{I}[a_{k^-} = a_j] \right\rceil \\
&= \lceil U_i |I_{k^-}| \rceil .
\end{aligned} \tag{C.29}$$

Notice, (C.29) implies $U_i \in \left(\frac{\max(I_{k^-}) - k^-}{|I_{k^-}|}, \frac{\max(I_{k^-}) - k^- + 1}{|I_{k^-}|}\right]$.

Next, assume $\Delta W_{-i} \in \{a_{(j)} : j \in I_{k^-}\}$ and $U_i \in \left(\frac{\max(I_{k^-})-k^-}{|I_{k^-}|}, \frac{\max(I_{k^-})-k^-+1}{|I_{k^-}|}\right]$, it is easy to calculate $R_i = k$. For notation sake, let $a = (a_1, \cdots, a_{|\mathcal{C}_0^c|+1})$. It follows that

$$
\begin{aligned}
&\mathbb{P}\left(R_i = k \mid (S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}) = a)\right) \\
&= \mathbb{P}\left((\text{C.27}) \mid (S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}) = (a_1, \cdots, a_{|\mathcal{C}_0^c|+1})\right) \\
&= \mathbb{P}\left(\Delta W_{-i} \in \{a_j : j \in I_{k^-}\} \mid (S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}) = a\right) \\
&\quad \times \mathbb{P}\left(U_i \in \left(\frac{\max(I_{k^-}) - k^-}{|I_{k^-}|}, \frac{\max(I_{k^-}) - k^- + 1}{|I_{k^-}|}\right] \mid (S'_{(1)}, \cdots, S'_{(|\mathcal{C}_0^c|+1)}) = a\right) \qquad \text{(C.30)} \\
&= \left(\frac{|I_{k^-}|}{|\mathcal{C}_0^c| + 1}\right)\left(\frac{1}{|I_{k^-}|}\right) \\
&= \frac{1}{|\mathcal{C}_0^c| + 1} = \mathbb{P}(R_i = k).
\end{aligned}
$$

The second equality follows from the independence of U_i. The third equality follows from (C.26) and $U_i \sim \text{UNIF}(0, 1)$.

Now, by the independence of R_i and$(S'_{(1)}, \cdots, S'_{|\mathcal{C}_0^c|})$, and (C.25),

$$
\left\{\left(S_{(\ell)}\right)_{\ell \in \{1,\cdots,|\mathcal{C}_0^c|\}} \mid p_i^{\text{con}} = \frac{k}{1 + |\mathcal{C}_0^c|}\right\} \stackrel{d}{=} \left(S'_{(\ell)}\right)_{\ell \in \{1,\cdots,|\mathcal{C}_0^c|+1\} \setminus \{k^-\}} \qquad \text{(C.31)}
$$

because $R_i = (1 + |\mathcal{C}_0^c|)p_i^{\text{con}}$. $\left(S'_{(\ell)}\right)_{\ell \in \{1,\cdots,|\mathcal{C}_0^c|+1\} \setminus \{k^-\}}$ is clearly non-increasing in k^-, so it is non-decreasing in terms of k. Thus for any $y \geq y$ we have $Z_1 \sim Z \mid p_i^{\text{con}} = y$ and $Z_2 \sim Z \mid p_i^{\text{con}} = y'$ such that

$$
\mathbb{P}(Z_2 \leq Z_1) = 1.
$$

□

Now, with Lemmas C.1.2 and C.1.3, the PRDS result is immediate. Take any y and y' such that $y \geq y'$ and the Z_1 and Z_2 that satisfy Lemma C.1.3. Now, for any irreproducible hypothesis $i \in \mathcal{D}_t$,

the following holds

$$\begin{aligned} \mathbb{P}(Y \in A \mid p_i^{\text{con}} = y) &= \mathbb{E}_{Z_1}\mathbb{P}(Y \in A \mid Z = Z_1) \\ \text{by Lemmas C.1.2 and C.1.3} \quad &\leq \mathbb{E}_{Z_2}\mathbb{P}(Y \in A \mid Z = Z_2) \\ &= \mathbb{P}(Y \in A \mid p_i^{\text{con}} = y') \end{aligned} \tag{C.32}$$

□

C.2 Additional simulations

C.2.1 Simulation setting for example in Section 4.2

Let there be n hypotheses common in $m = 2$ experiments. Additionally, denote t_{ji} is the observed summary statistic for hypothesis i in experiment j. Now, consider n_{11} hypotheses have moderate signal that is highly consistent across experiment, for example, summary statistics distributed bivariate normal distribution

$$\begin{bmatrix} t_{1i} \\ t_{2i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_{11,i} \\ \mu_{11,i} \end{bmatrix}, \begin{bmatrix} 1 & \rho_{11} \\ \rho_{11} & 1 \end{bmatrix}\right)$$

with $\rho > 0$; consider n_{10} hypotheses with incredibly strong signal in the first experiment and no signal in the second, for example, summary statistics distributed by

$$\begin{bmatrix} t_{1i} \\ t_{2i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} \mu_{10,i} \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right);$$

with $\mu_{10,i}$ large in magnitude; similarly consider n_{01} hypotheses with no signal in the first experiment and incredibly strong signal in the second, for example, summary statistics distributed by

$$\begin{bmatrix} t_{1i} \\ t_{2i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} 0 \\ \mu_{01,i} \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)$$

where $\mu_{01,i}$ is large in magnitude; finally, let the remaining $n_{00} = n - (n_{11} + n_{01} + n_{10})$ have no signal in either experiment, for example

$$\begin{bmatrix} t_{1i} \\ t_{2i} \end{bmatrix} \sim \mathbb{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right).$$

Figure 4.1 displayed a scatterplot of t_{ji} statistics and their associated rankings within experiment, r_{ji}, from this setting where $n = 1000$, $n_{11} = 300$, $n_{10} = n_{01} = 20$ with $\mu_{11,i} = \{-3.5, 3.5\}$, $\rho_{11} = 0.95$ $\mu_{10,i} = \mu_{01} = \{-20, 20\}$.

C.3 Supplemental real data applications

C.3.1 Additional real data figures

Reproducible region geometry for each pair of studies

Figures C.1-C.7 show the reproducibility/rejection regions for each of the methods for every pair of replicate studies.

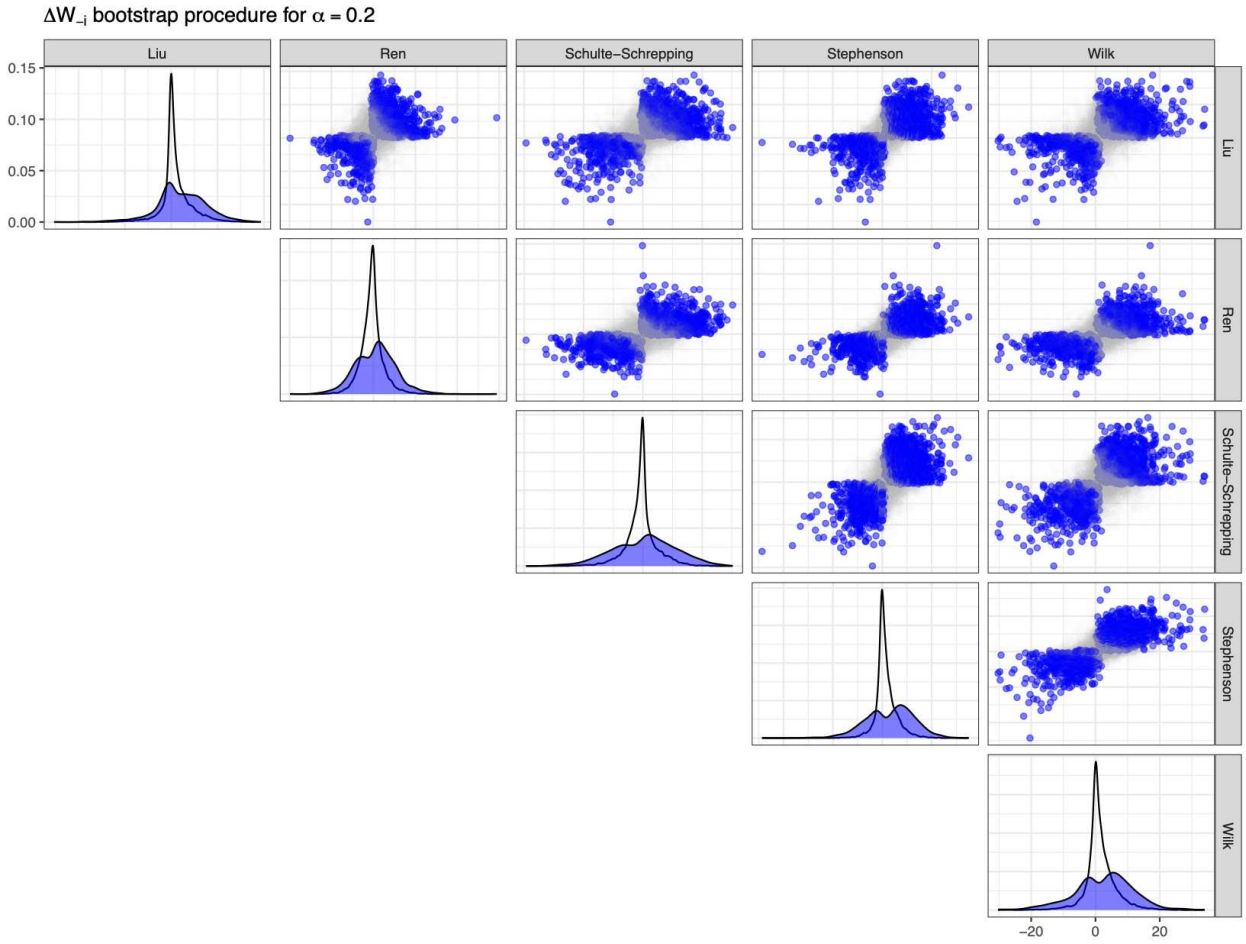


Figure C.1: Reproducibility regions for the ΔW_{-i} bootstrap procedure in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions.

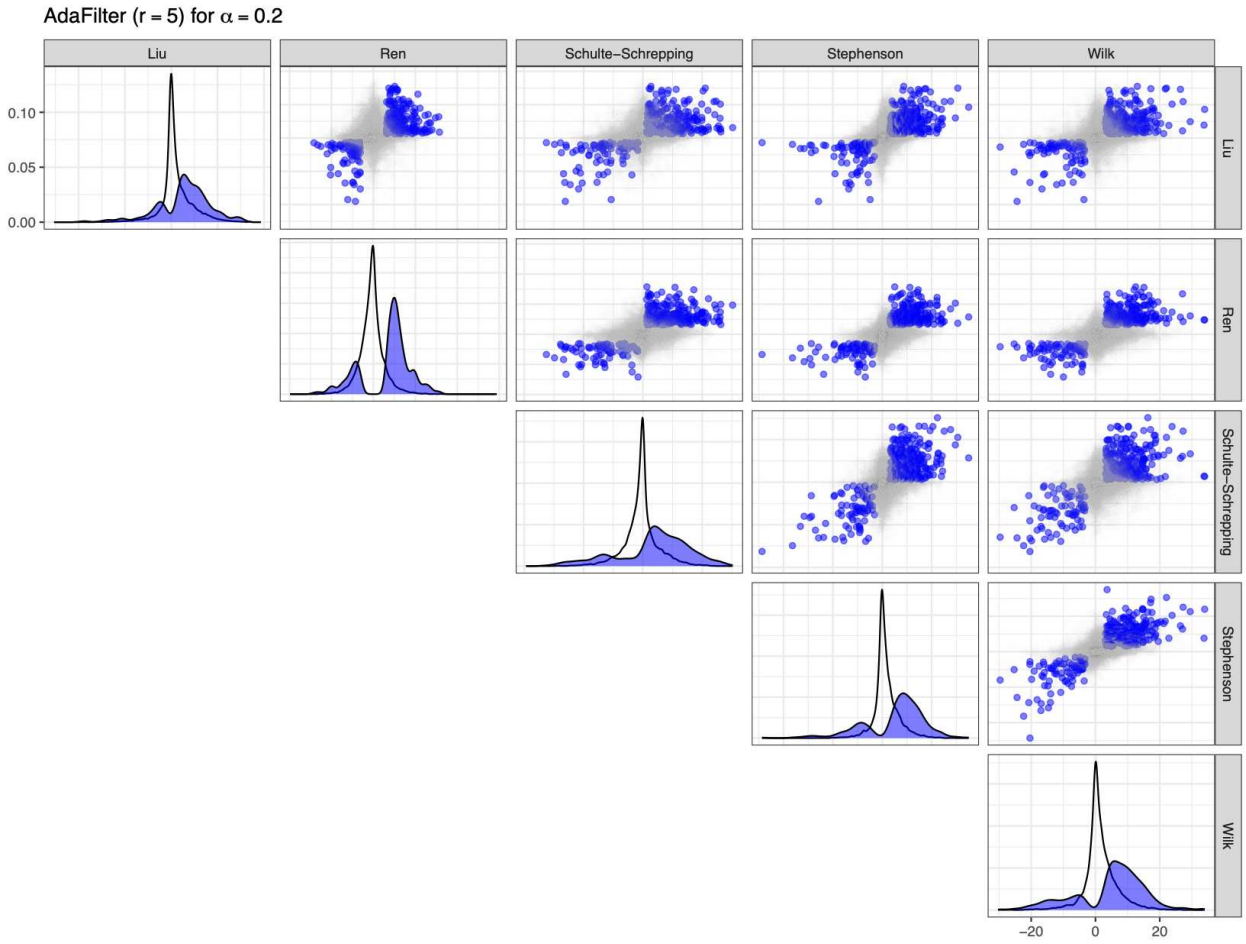


Figure C.2: Reproducibility regions for AdaFilter with $m = 5$ in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

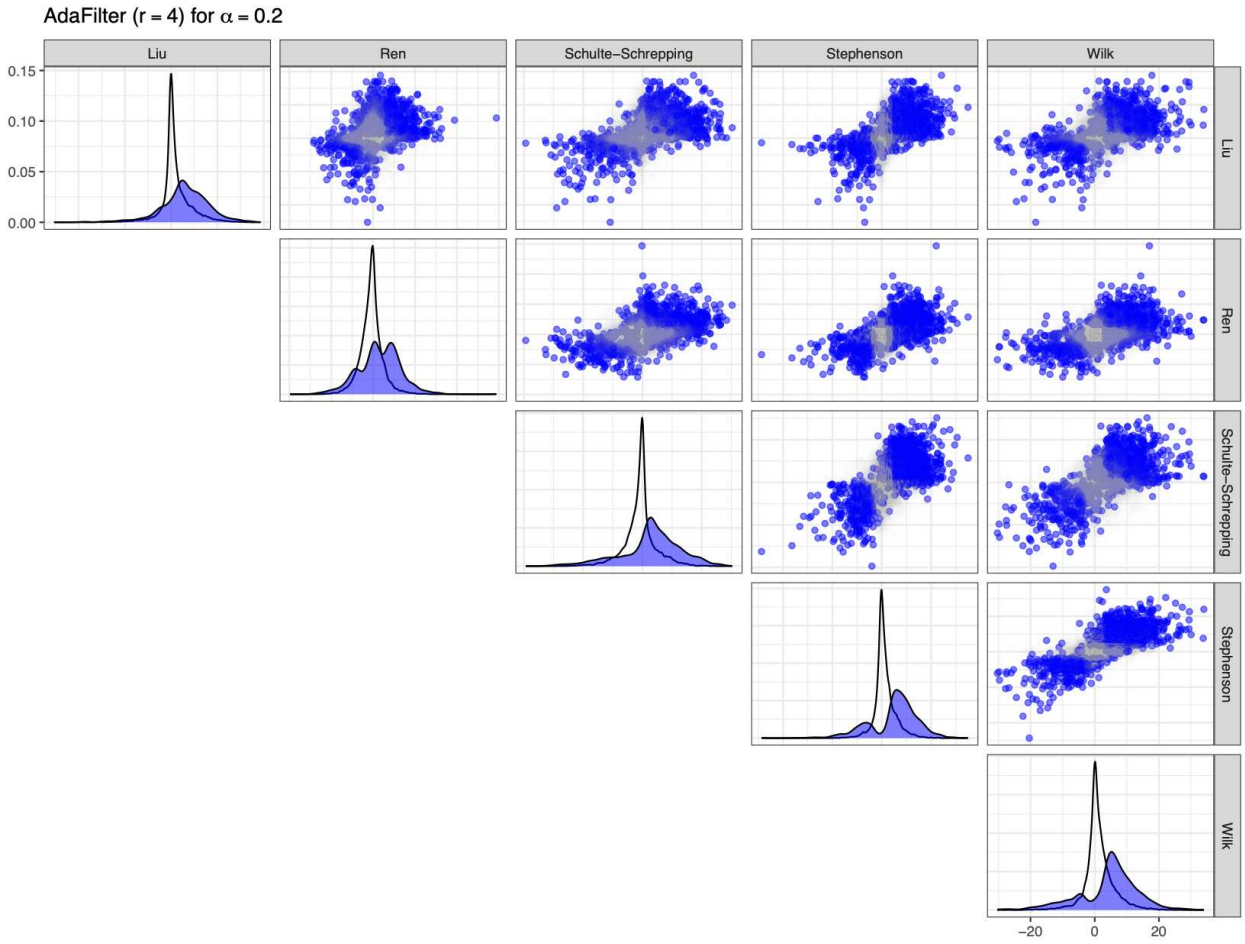


Figure C.3: Reproducibility regions for AdaFilter with $r = 4$ in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

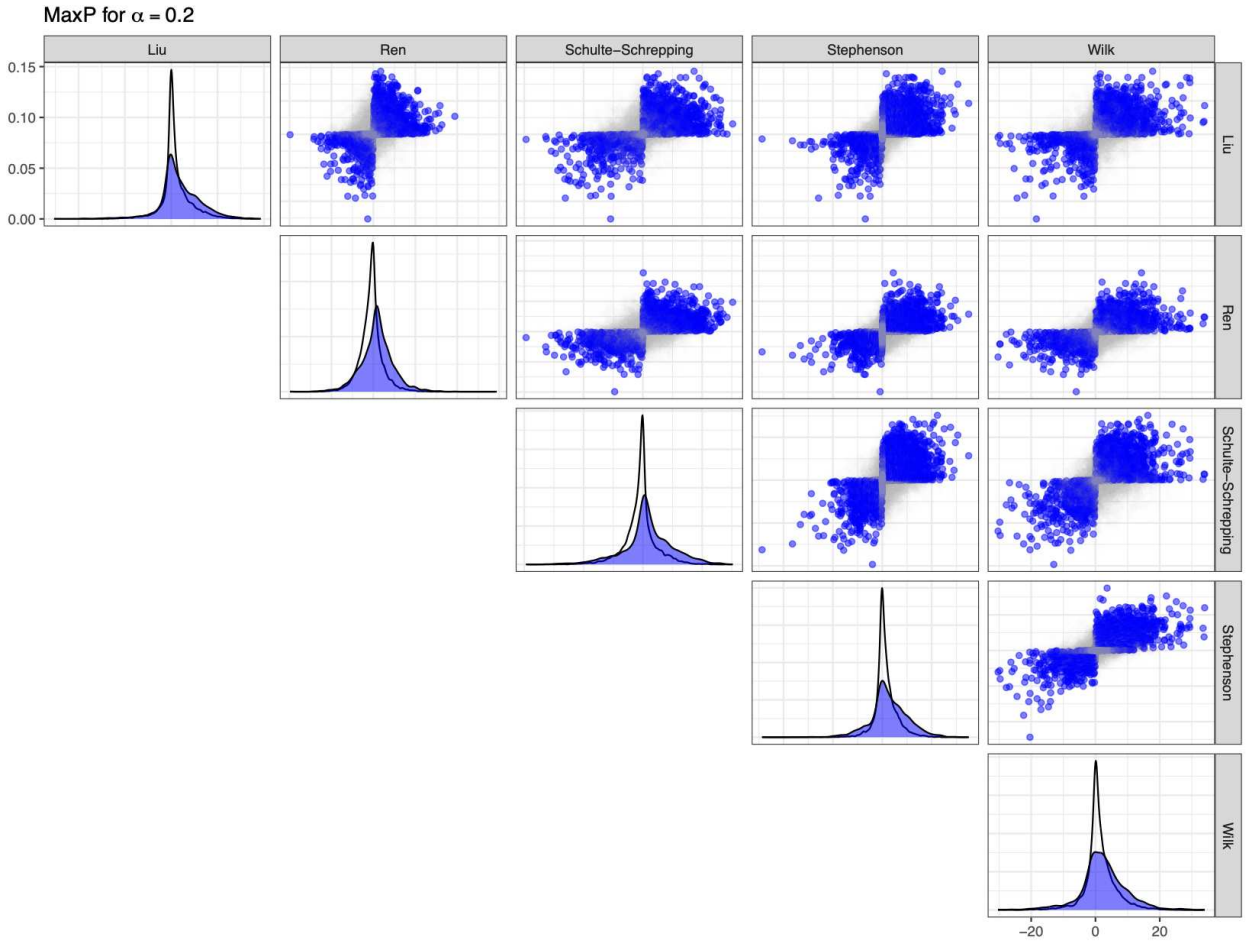


Figure C.4: Rejection regions for Max$P$ in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of α = 0.2. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

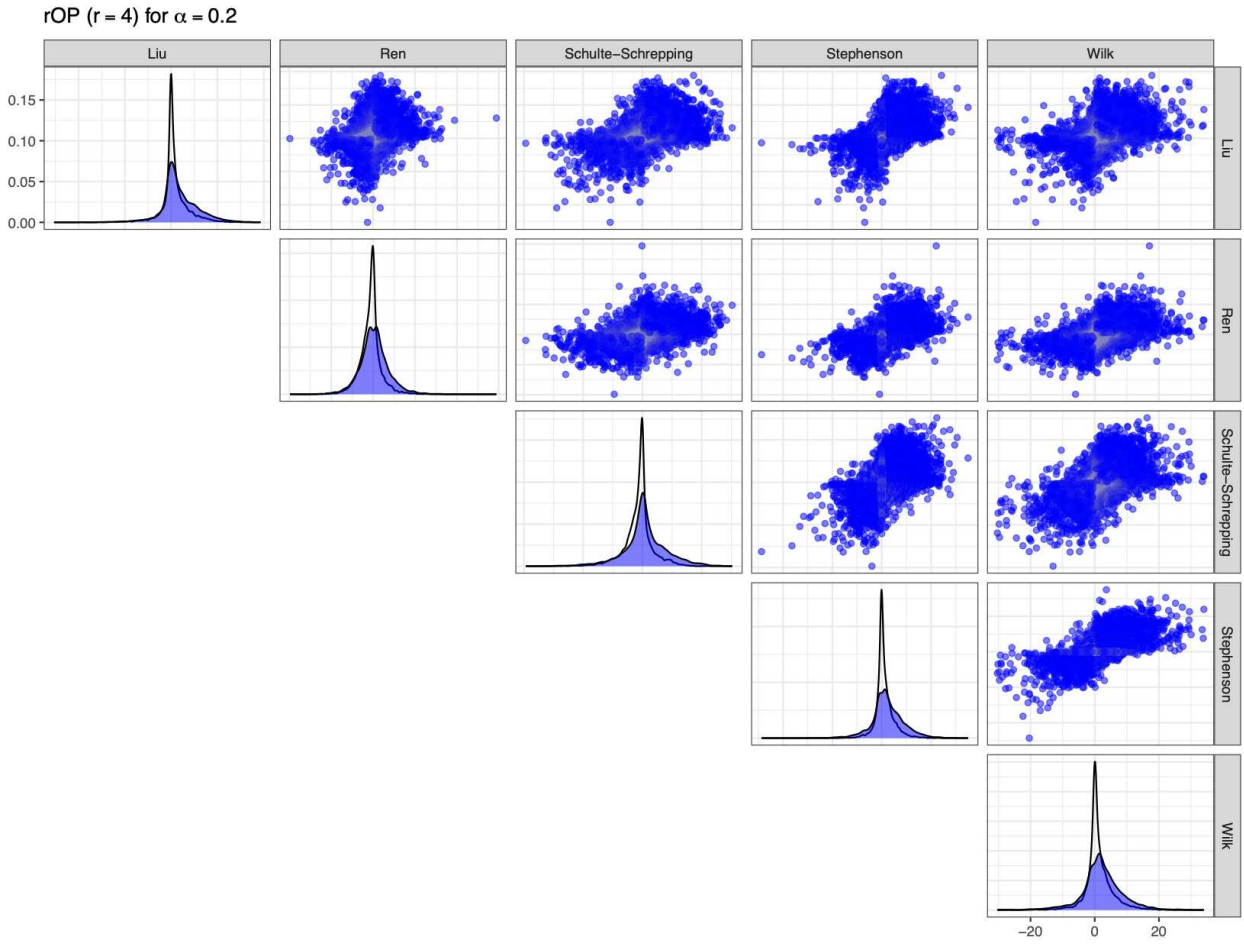


Figure C.5: Rejection regions for rOP with $r = 4$ in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

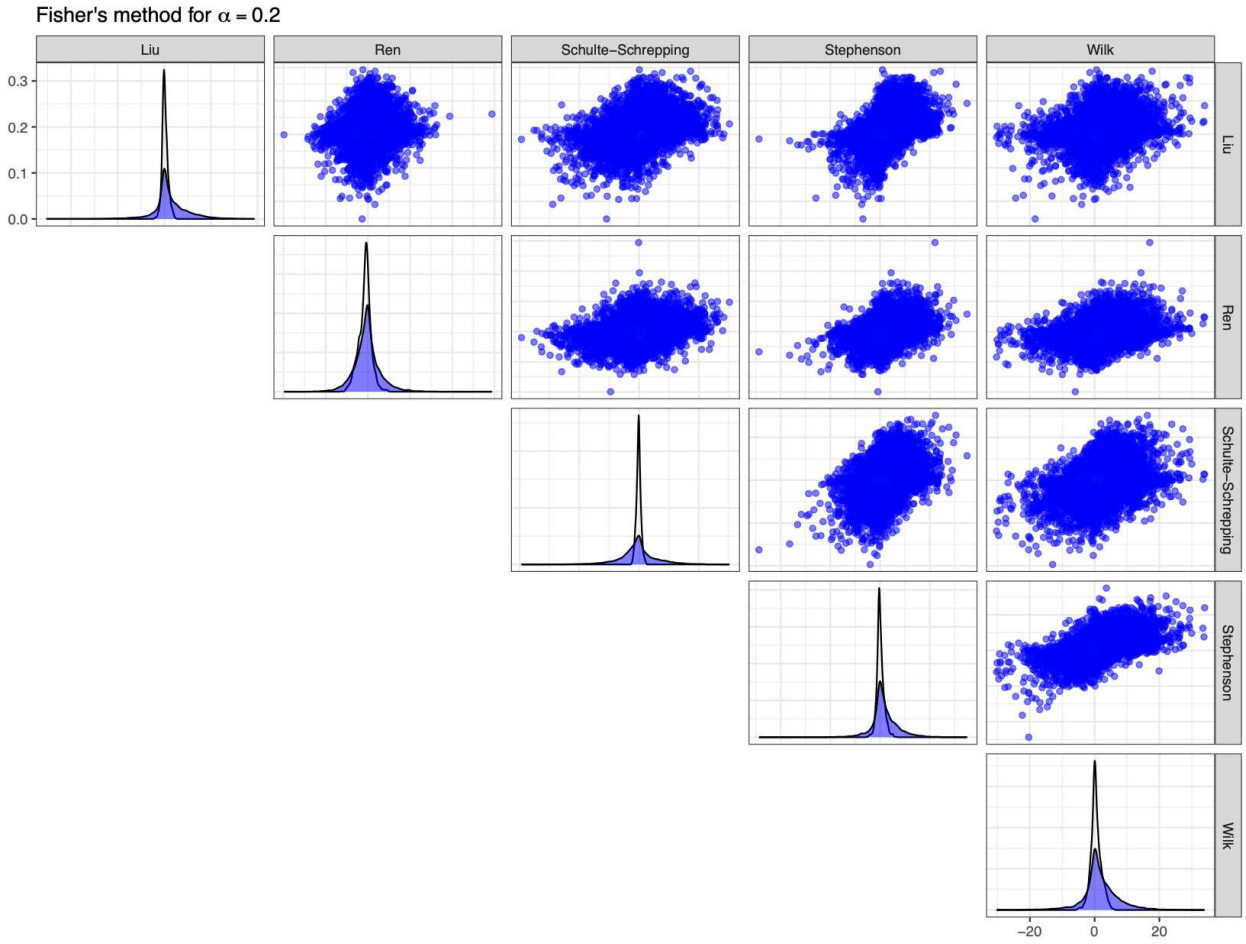


Figure C.6: Rejection regions for Fisher's method in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

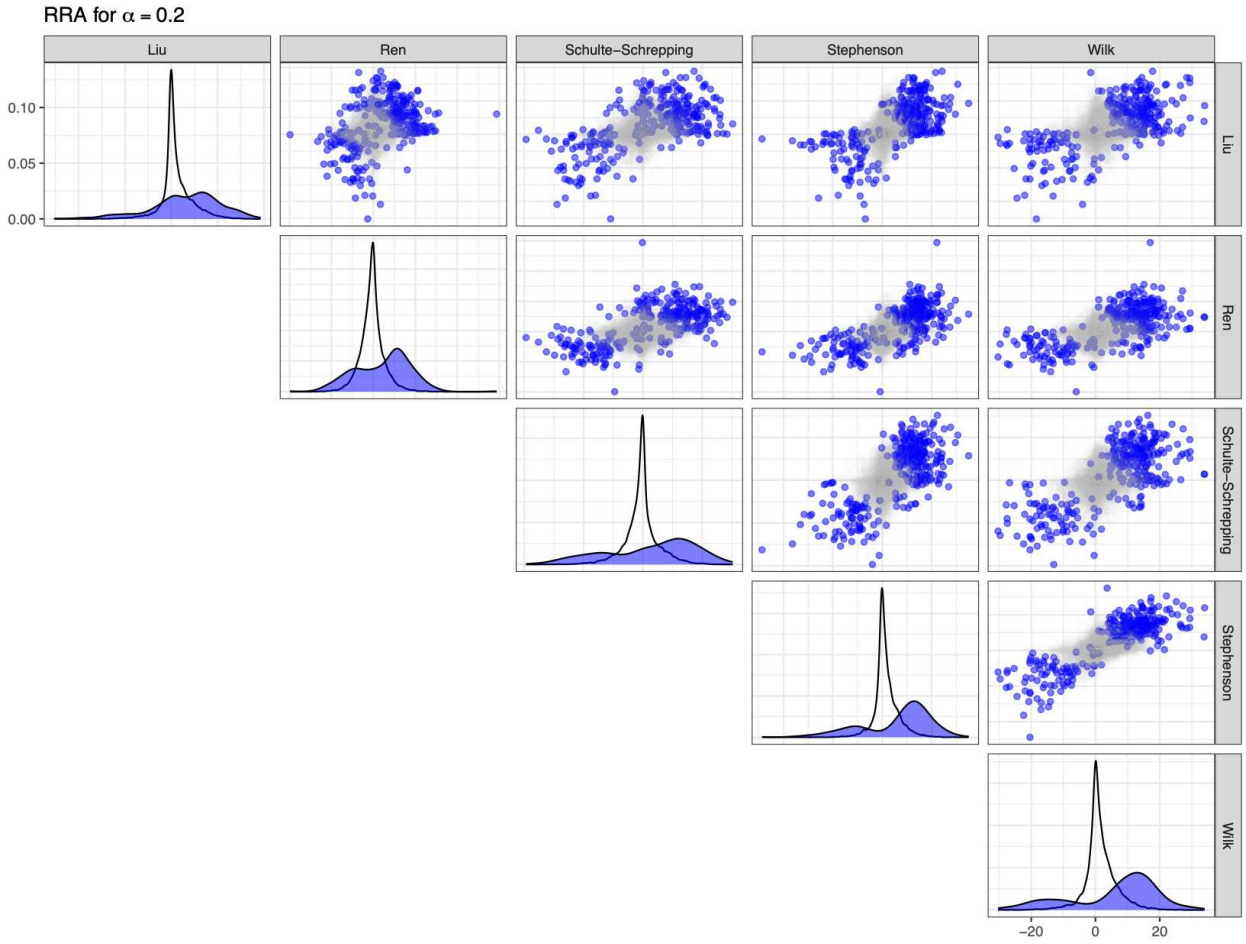


Figure C.7: Rejection regions for RRA in terms of t_{ji} statistics from each pair of studies at a nominal FDR level of $\alpha = 0.2$. Genes in the reproducible region are shown in blue. The diagonal displays the marginal density of t_{ji} statistics for genes in the reproducible (in blue) and irreproducible (in white) regions

Intersection and differences in reproducible regions for each pair of studies

Figures C.8- C.13 show the geometry of the intersections and set differences in reproducibility/rejection regions between ΔW_{-i} and each other method in terms of t_{ji} statistics from each pair of studies.

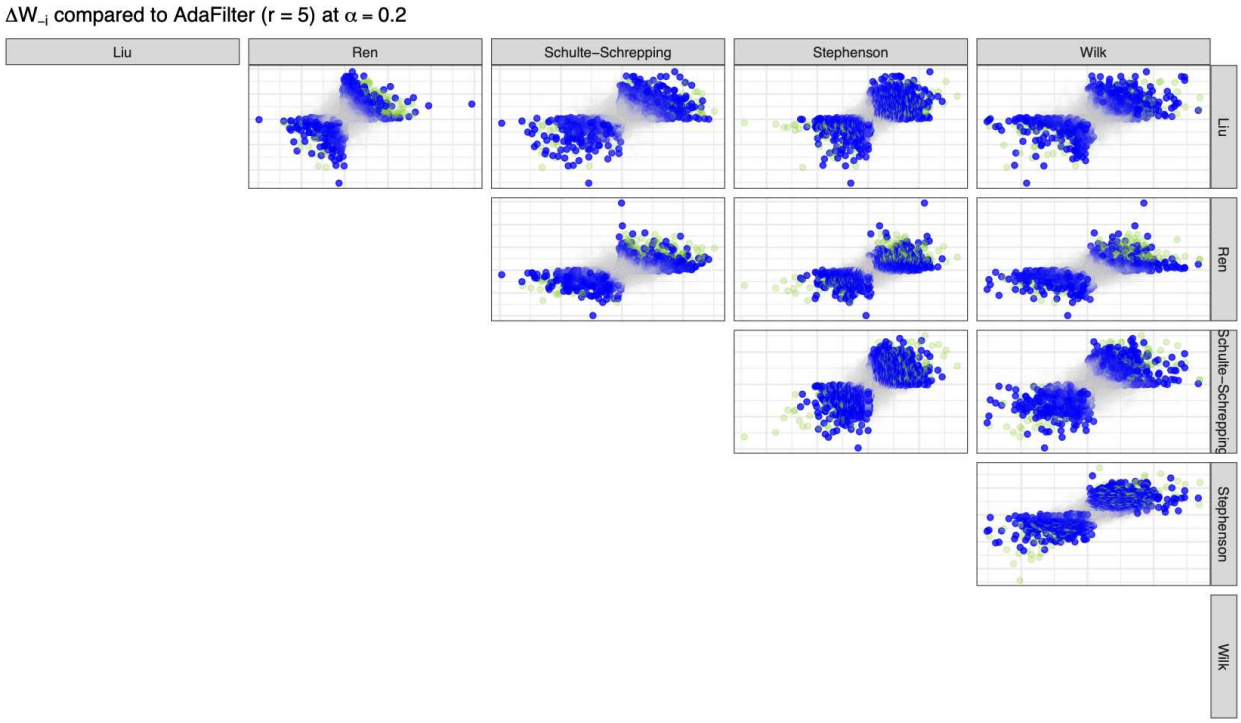


Figure C.8: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and AdaFilter with $r = 5$ in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only AdaFilter.

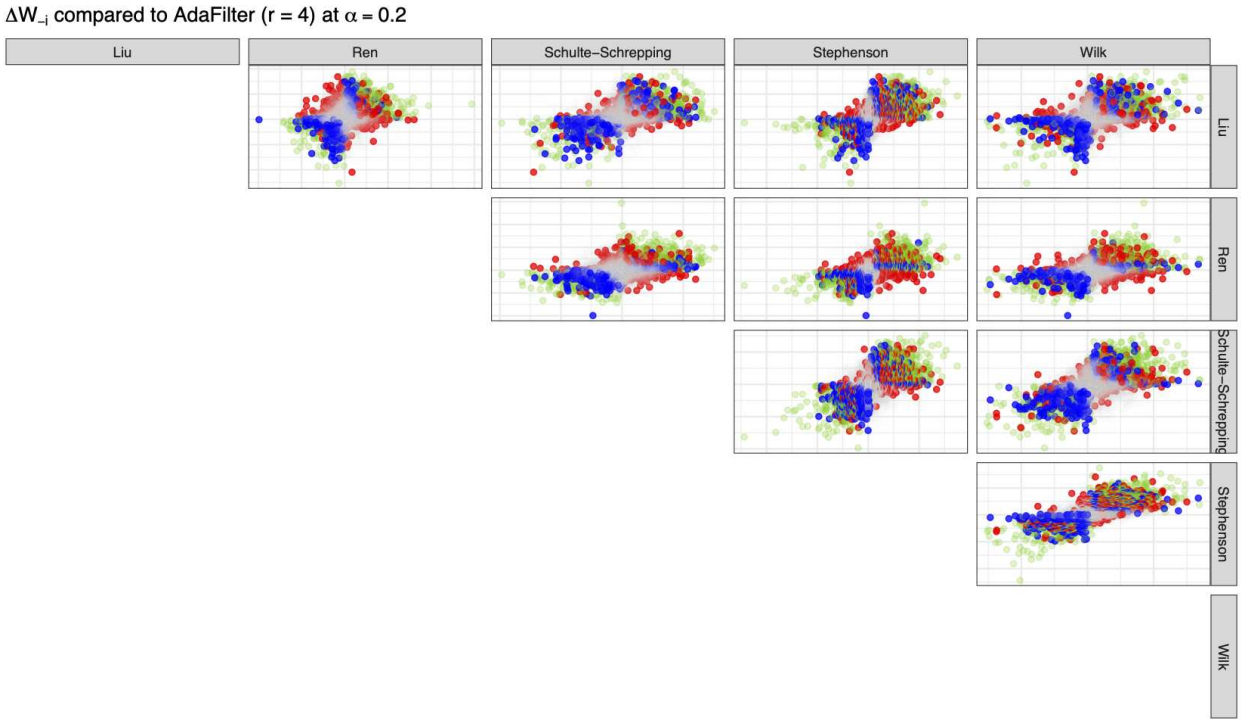


Figure C.9: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and AdaFilter with $r = 4$ in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only AdaFilter.

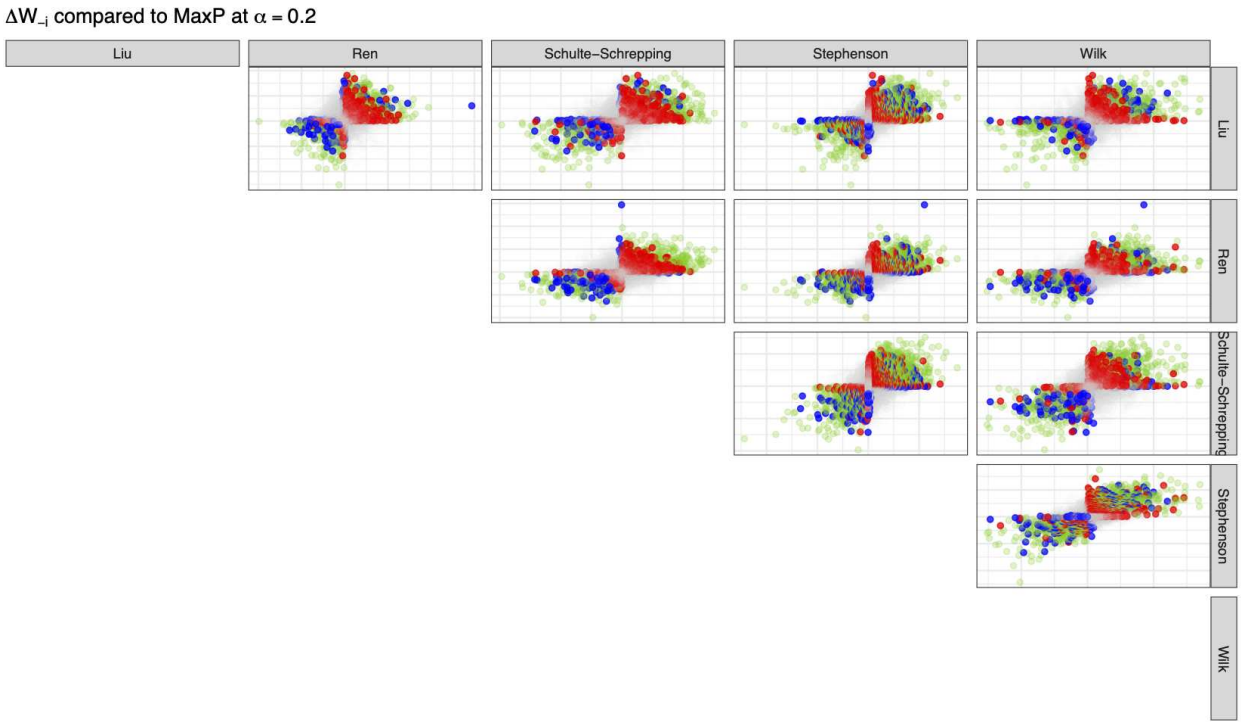


Figure C.10: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and max$P$ in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only max$P$.

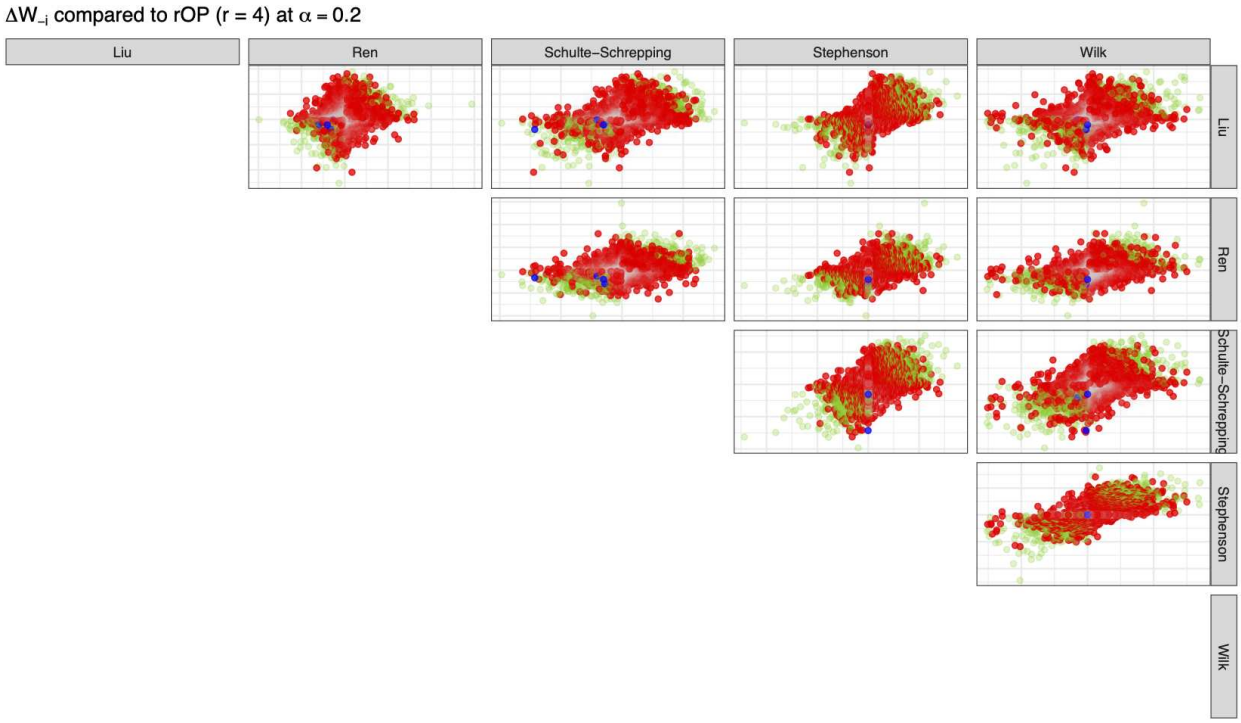


Figure C.11: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and rOP with $r = 4$ in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only rOP.

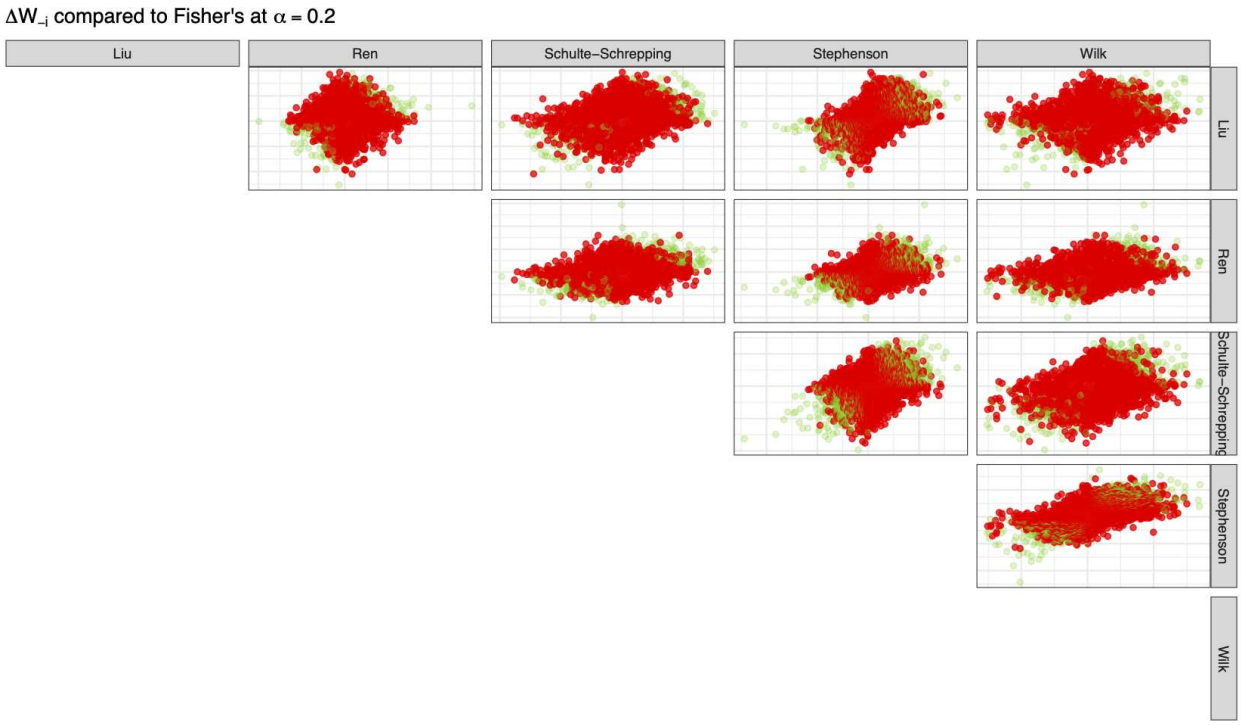


Figure C.12: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and Fisher's method in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only Fisher's.

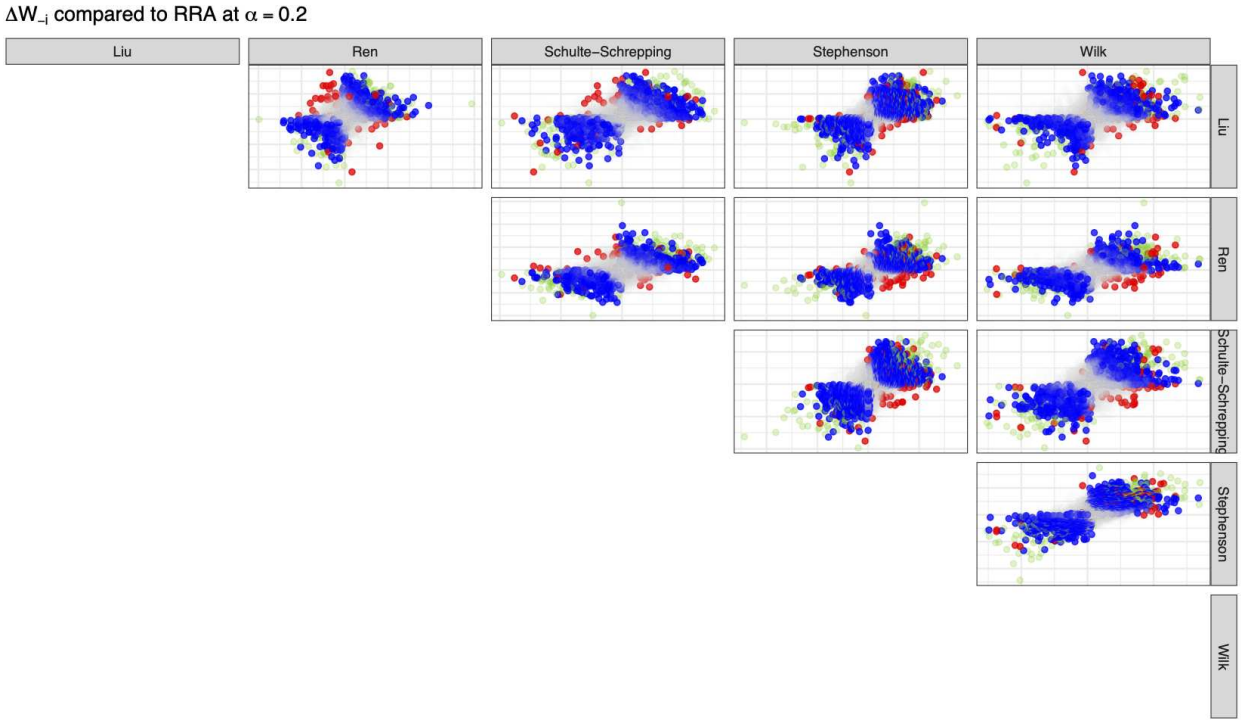


Figure C.13: Intersection and set differences in reproducibility regions at a nominal FDR level of $\alpha = 0.20$ between ΔW_{-i} and RRA in the space of t_{ji} statistics for each pair of studies. Genes in green are found to be reproducible by both methods, blue by only ΔW_{-i}, and red by only RRA.

Ranking of biological processes related to the viral process

Figures C.14 and C.15 show the rankings of biological processes that include the word “viral" among the top 1000 most represented processes and all processes in the reproducibility regions for ΔW_{-i} and AdaFilter with $(r = 4)$ with $\alpha = 0.2$.

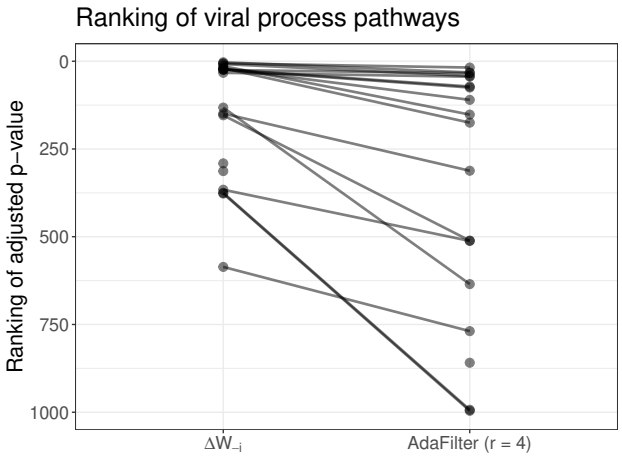


Figure C.14: Ranking of adjusted enrichment p-values for pathways related to the viral process among the top 1000 pathways identified by ΔW_{-i} and AdaFilter with $(r = 4)$.

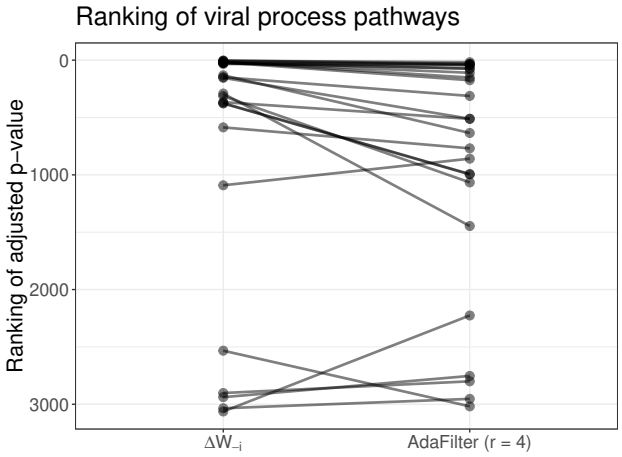


Figure C.15: Ranking of adjusted enrichment p-values for pathways related to the viral process among the top all pathways identified by ΔW_{-i} and AdaFilter with $(r = 4)$.