Methods for Combining Frequent or Sparse Signals in Omics Applications

by

Yusi Fang

B.S. in Mathematics and Applied Mathematics, Xiamen University, 2017

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2023

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Yusi Fang

It was defended on

Apirl 4, 2023

and approved by

George C. Tseng, ScD, Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Zhao Ren, PhD, Associate Professor, Department of Statistics, Kenneth P. Dietrich School of Arts & Sciences, University of Pittsburgh

Ying Ding, PhD, Associate Professor, Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh

Abdus S. Wahed, PhD, Professor, Department of Biostatistics and Computational Biology, University of Rochester

Methods for Combining Frequent or Sparse Signals in Omics Applications

Yusi Fang, PhD

University of Pittsburgh, 2023

Combining p-values to aggregate effects has been of long-standing interest. We discuss three types of p-value combination scenarios for omics studies in Chapters 2-4 of this dissertation.

Chapter 2 considers combining independent and non-sparse signals in a small group of p-values, where the number of true signals in p-values and their strengths can vary with heterogeneity. We propose the Fisher ensemble (FE) to aggregate the existing Fisher and AFp methods. The FE achieves asymptotic Bahadur optimality and integrates the strengths of Fisher and AFp. We extend FE to a variant with emphasized power for concordant effect size directions. A transcriptomic meta-analysis of the AGEMAP dataset shows the advantages of the proposed methods.

Chapter 3 proposes a simple yet truly adaptive modified Fisher's method for combining independent, weak and sparse signals in a large group of p-values. It achieves the optimal separating rate in a large-scale setup with sparse and heterogeneous signals. Our method is robust when the p-values are not exact and can maintain the optimal separating rate under mild conditions. The proposed method is applied to a neuroticism GWAS application for the pathway-based association study.

Chapter 4 considers combining dependent, weak and sparse signals in a large group of p-values. We study a family of p-value combination tests by heavy-tailed distribution transformations. We derive the conditions for a method of the family to enjoy robustness against the unknown dependency structure and to attain the optimal detection boundary for detecting weak and sparse signals. Only an equivalent class of the Cauchy test can possess robustness property. By applying our theoretical findings, we suggest a truncated Cauchy test that belongs to the class to improve the Cauchy test. A neuroticism GWAS application demonstrates the theoretical findings and advantages of the truncated Cauchy method.

Contribution to Public Health:
Omics data integration is critical for contemporary biomedical research. P-value combination ap-

proaches are widely utilized in omics studies for aggregating information from multiple sources. This dissertation establishes a robust theoretical foundation of p-value combination and offers practical, data-driven methodologies for omics data integration.

Keywords: p-value combination; global hypothesis testing; large-scale inference; meta-analysis.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation contains three research projects undertaken throughout my PhD studies. As I reach the end of my six-year journey, I am overwhelmed with emotions and reflections. A single dissertation certainly cannot fully capture the arduous yet rewarding experience, leaving me with many memories. The past six years spent with my advisors, girlfriend, lab mates, and friends have been my life's most joyful and significant time.

I would like to convey my profound gratitude to my advisors, Dr. George Tseng and Dr. Zhao Ren, for their unwavering faith in me, invaluable guidance, and steadfast support throughout the past six years. Over an extended period, I encountered challenges in finding my research interests and surmounting the obstacles and setbacks inherent in the research process. Without their patience, encouragement, and direction, this journey's end would have been unattainable.

I would also like to express my heartfelt gratitude to my girlfriend, Xiangning. Pursuing a doctoral degree is a demanding journey. Her companionship and encouragement have been invaluable during every challenging moment, such as encountering disappointing research outcomes or facing paper rejections. I am genuinely grateful for her unwavering support throughout this journey, and I commit to reciprocating by providing the same support and companionship during her PhD studies.

Moreover, I would like to thank my dissertation committee members, Dr. Ying Ding and Dr. Abdus S. Wahed, and my collaborator, Dr. Chung Chang, for their significant contributions, insightful comments, and inspiring discussions regarding my research. Additionally, I am grateful to my family, friends, and lab mates for their support over the years.

I would also like to thank my cat, Gimme, for bringing me all the joy and adorable moments.

The PhD journey has provided me not only with the doctoral degree, publications, and sound training in statistics but also with the unique experience of pursuing a long-term goal without immediate rewards. Through this process, I have learned to persevere amidst setbacks and challenges, courageously explore unknown territories, and commit to the long-term pursuit of valuable objectives. Lastly, I would like to thank myself for making the brave decision to embark on this journey six years ago. Without that courageous choice and the following persistence throughout

the years, I would never have had the opportunity to experience such a profoundly enriching and transformative chapter in my life.

1.0 Introduction

This chapter introduces background knowledge and motivations behind this dissertation. An overview of p-value combination methods is presented in Section 1.1. Section 1.2 provides a summary of omics data, along with a brief introduction to downstream analyses relevant to this dissertation. The statistical challenges associated with applying p-value combination methods for the analysis of omics data are discussed in Section 1.3, where three types of p-value combination scenarios are formulated. Finally, Section 1.4 introduces the structure of the dissertation.

1.1 Overview of P-Value Combination Methods

In statistics and applications spanning a wide range of scientific disciplines for aggregating data from multiple sources, methods for combining p-values have long been of great interest. Suppose we have n p-values $\vec{p} = (p_1, p_2, \ldots, p_n)$, where each p_i is p-value of testing $H_0^{(i)} : \theta_i \in \Theta_0^{(i)}$ versus $H_1^{(i)} : \theta_i \in \Theta^{(i)} - \Theta_0^{(i)}$. Here θ_i represents the parameter of interest for the i-th hypothesis test and $\Theta^{(i)}$ and $\Theta_0^{(i)}$ denote the corresponding total possible parameter space and null parameter space, respectively. Under this setup, a common problem is the global union-intersection test (UIT) (Roy, 1953):

$$H_0 : \cap_{1\leq i\leq n}\{\theta_i \in \Theta_0^{(i)}\} \text{ versus } H_1 : \cup_{1\leq i\leq n}\{\theta_i \in \Theta^{(i)} - \Theta_0^{(i)}\}.$$

By formulating a test statistic to combine the input p-values, the main objective of p-value combination is to perform an UIT test for detecting any signal in the n p-values. For example, suppose $\theta_i = \mu_i$ for $N(\mu_i, 1)$, $\Theta^{(i)} = \mathbb{R}$ and $\Theta_0^{(i)} = \{\mu_i = 0\}$ for a simple z-test with p-value p_i. For the UIT test $H_0 : \cap_{1\leq i\leq n}\{\mu_i = 0\}$ versus $H_1 : \cup_{1\leq i\leq n}\{\mu_i \neq 0\}$, one may consider Fisher's method $T_{\text{Fisher}} = \sum_{i=1}^{n} -2\log p_i$ (Fisher, 1992) to combine the input p-values, where T_{Fisher} follows chi-square distribution with degrees of freedom $2n$ when $p_1, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ under the null. The null hypothesis is thought to be rejected by a large value of T_{Fisher}, which implies that there exists at least one $\mu_i \neq 0$ for $1 \leq i \leq n$.

Including Fisher's method, traditional p-value combination methods (other examples include

Stouffer's method $T_{\text{Stouffer}} = \sum_{i=1}^{K} \Phi^{-1}(1-p_i)$(Stouffer et al., 1949), Edgington's method $T_{\text{Edgington}} = \sum_{i=1}^{n} p_i$ (Edgington, 1972), as well as many other methods, see Heard and Rubin-Delanchy (2018) for more examples) aim to combine relatively dense signals in a small group of independent p-values. These methods can be regarded as meta-analysis approaches to aggregate multiple small effects for improved statistical power. Besides the conventional methods, many modified Fisher's methods have been proposed in recent years to improve Fisher's method when only part of p-values contain true signals under the meta-analysis setting. Examples include rank truncated product method and its variants (Dudbridge and Koeleman, 2003; Yu et al., 2009; Li and Tseng, 2011; Song et al., 2016), and truncated product method and its variants (Zaykin et al., 2002; Zhang et al., 2020b). See Chapter 2 for a more detailed discussion of the conventional methods and the modified Fisher's methods.

Due to the increasing needs of large-scale data analysis, detecting weak and sparse signals from a huge group of independent p-values (may be more than 1000 input p-values) has attracted a lot of interest. High criticism and Berk-Jones tests (Donoho and Jin, 2004; Berk and Jones, 1979; Li and Siegmund, 2015) are the two most representative methods under this setting. The two methods can also be thought of as one-sided goodness of fit tests for determining if the input p-values are uniformly distributed across the $[0, 1]$ interval, as the input p-values $p_1, p_2, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$ under the null. Following this idea, it has been demonstrated that many other goodnesses of fit tests also enjoy comparable theoretical properties to higher criticism and Berk-Jones tests (Jager and Wellner, 2007).

All the methods described above focus on combining independent p-values. However, many modern data analyses generate the needs for combining dependent p-values. Brown's method (Brown, 1975) is an extension of Fisher's method for combining dependent p-values. As the null distribution of Fisher's method is no longer a chi-square distribution when the input p-values are dependent, Brown (1975) proposed to use a scaled chi-square distribution to approximate the null distribution of Fisher's method. Kost and McDermott (2002); Li et al. (2014); Zhang and Wu (2022); Poole et al. (2016) proposed more refined approximation methods to adapt Fisher's method for combining dependent p-values. For combining a large group of dependent p-values, Hall and Jin (2010); Barnett et al. (2017); Sun and Lin (2020) proposed to adapt the higher criticism and Berk-Jones tests by taking into account the correlation structure of p-values. On the other hand,

the harmonic mean and Cauchy combination tests were introduced by Wilson (2019a) and Liu and Xie (2020); Liu et al. (2019) to provide robustness under unknown dependency structures when inference is established under the independence assumption.

1.2 Overview of High-Throughput Omics Data

High-throughput technology may simultaneously quantify thousands to millions of molecular components for biological activity at the model organism, cellular, pathway, or molecule level using automated techniques and technologies. Numerous omics data have accumulated in public archives due to high-throughput technology's rapid development, necessitating the development of novel statistical approaches to aid in biological discoveries. The term "omics data" refers to data sets that quantify an organism's genetic materials (genomics), epigenetic alterations (epigenomics), RNA transcripts (transcriptomics), and proteins (proteomics). This section will introduce two relevant omics data types, genomics and transcriptomics data, as well as the widely used statistical techniques for the downstream analysis of the data.

1.2.1 Genomics

The study of the entire set of genetic materials of an organism, or genome, including how the genes interact with one another and the environment, is called genomics. Utilizing high throughout technology, genomics studies the genome of an organism, typically DNA (RNA for some viruses). Human DNA, distributed in 22 paired chromosomes and two sex chromosomes, contains the genetic information of humans. The human genome shares 99.9% of its components. On the contrary, the average fraction of nucleotide differences between two randomly picked individuals only ranges from 1/1500 to 1/1000 (Jorde and Wooding, 2004). The association between human phenotypes and genetic differences has been a popular topic of study for a long time.

Single nucleotide polymorphism (SNP), which refers to a variation in a single nucleotide of DNA, is the most prevalent type of genetic variation, occurring in at least 1% of the population. The identification of SNPs associated with a disease or other phenotypic traits is typically done using

genome-wide association studies (GWAS). Tens of thousands of genetic variants are examined using GWAS across numerous genomes. Generalized linear mixed models (GLMM), including mixed effects logistic regression and linear mixed models, are commonly used for the GWAS approach to model the relationship between the genotypes and phenotypes (Uffelmann et al., 2021). Common GWAS approaches test single SNPs individually (Manolio et al., 2009; Visscher et al., 2012). When the effects of individual SNPs are relatively weak, the SNP-set association test, which tests the association between the phenotypic traits and a set of SNPs in the same genetic construct such as genes and genetic pathways, can be a powerful alternative (Visscher et al., 2012). Insertion, deletion, and structural variation—which involves changing numerous base pairs—are other common types of genetic variation. When a nucleotide sequence is over-represented, it is called insertion. When a nucleotide sequence is under-represented, it is called deletion. CNV, or copy number variation, is a type of structural variation that happens when large sections of the genome are deleted or duplicated.

1.2.2 Transcriptomics

The study of the entire set of RNA molecules in an organism is referred to as "transcriptomics". Transcriptomics studies messenger RNA (mRNA) and micro RNA (miRNA) and other non-coding RNAs in a cell using high-throughput methods. mRNA carries protein-coding information, while miRNA is crucial in regulating gene expression. Gene expression can differ significantly between different types of tissues and cells, as well as in relation to developmental stages and health states, as it is regulated by a variety of transcriptional and post-transcriptional activities (e.g., alternative splicing, miRNA binding). The identification of genetic subgroups, the discovery of biomarkers, the diagnosis, and prognosis of disease have all benefited from the widespread use of transcriptomics data in biomedical research (Yang et al., 2020; Raghavachari and Garcia-Reyero, 2018).

Currently, the two main high throughput methods for measuring gene expression levels are microarray and RNAseq/scRNAseq. Microarrays are microscope slides (sometimes referred to as "chips") on which thousands of small dots, each containing a known DNA sequence or gene, are printed at particular locations (also known as a "probe"). RNA molecules are reverse transcribed to complementary DNA (cDNA) in a microarray experiment, where they are then mounted to

microscope slides and hybridized with the probes. When a laser scans a sample, a continuous fluorescence intensity score is calculated to measure the gene expression levels. RNA-seq cuts cDNA into fragments and attaches adapters to each fragment. The adapters contain functional elements that allow for sequencing. Using adapters, the fragments were sequenced and mapped to the reference genome. As an expression measurement for a gene, the number of mapped transcripts is counted.

Among all the downstream analyses of transcriptomics data, differentially expressed gene (DEG) analysis is one of the most important. Common DEG analyses perform two-sample hypothesis testing for each gene/biomarker across the genome, or other ANOVA-like tests if multiple groups comparisons are needed (Smyth, 2005; Ritchie et al., 2015; Robinson et al., 2010; Love et al., 2014). Multiple testing correction techniques, such as Bonfferoni correction and Benjamini-Hochberg procedure, are used to control the family-wise error rate (FWER) and false discovery rate (FDR) for DEG analysis in genome-wide setting (Benjamini and Hochberg, 1995; Storey and Tibshirani, 2003). After DEG analysis, pathway enrichment analysis can be a powerful option to help researchers gain more biological insight into the results of DEG analysis. Typical pathway enrichment analysis approaches test pathways in a given database for enrichment in a gene list of interest. Commonly used pathway enrichment analysis methods include gene set enrichment analysis (GSEA) (Subramanian et al., 2005) and ingenuity pathway analysis (IPA) (Krämer et al., 2014), see Maleki et al. (2020) for more details. Due to the rapid development of high-throughput techniques and the steep decline in their costs, many transcriptomic datasets are being generated in nearly all biological fields. This has led to a lot of interest in transcriptomic meta-analysis, which are ways to combine information from different transcriptomic studies. Common transcriptomic meta-analysis methods can be categoried into 4 categories: combining p-values (e.g, Li and Tseng (2011)), combining effect sizes (e.g., fixed and random effects models (FEM & REM), see Normand (1999)), direct merging (e.g., Shabalin et al. (2008)), and combining ranks (e.g., Hong et al. (2006)).

1.3 Statistical Challenges for Analyzing High Throughput Omics Data Using P-Value Combination Methods

Omics data analysis imposes new statistical challenges for the development of p-value combination methods. In this section, we outline 3 representative scenarios and corresponding challenges for applying the p-value combination methods to omics data analysis.

1.3.1 Combining Independent and Relatively Frequent Signals in A Small Group of P-Values

Common omics application examples for this scenario include transcriptomic, GWAS, CNV or methylation meta-analyses (Tseng et al., 2012; Begum et al., 2012; Guerra and Goldstein, 2016). The statistical setup considered here is that the number of combined studies n is fixed, while each study's sample size goes to infinity. Although sharing the label of "meta-analysis", there are subtle differences between the omic meta-analysis and the traditional meta-analysis scenarios when applying the p-value combination methods. Indeed, it is common that conventional meta-analysis scenarios assume all the studies share the same effect or at least all contain similar true effects, while heterogeneity between studies is frequently observed in the omic meta-analysis scenarios. More preciously, the proportions of studies that contain true signals can vary drastically, and the directions of effects can be disconcordant. There have been many efforts to modify the conventional methods for power improvement when the proportion of true signals is low. However, there is no systematic evaluation of the asymptotic properties and finite-sample performance of the proposed modified methods. In addition, there is still a lack of methods that have consistently high performance across varying scenarios.

1.3.2 Combining Independent, Weak, and Sparse Signals in A Large Group of P-Values

SNP-set association test (Arias-Castro et al., 2011, after de-correlation or SNP pruning based on high linkage disequilibrium) and multiple-sample based CNV calling (Song et al., 2016) are two common omics application for this scenario, where the number of p-values that contain true signals is assumed to be significantly smaller than n, the total number of input p-values. Due to the

theoretical advantage of the log-transformation posed on the combined p-values, it is well-known that Fisher's method has optimal Bahadur efficiency in the traditional meta-analysis setting (Littell and Folks, 1973, small n and all the p-values contain signals). However, Fisher's method is suboptimal when signals are weak and sparse (Donoho and Jin, 2004). Many modified Fisher's methods have been proposed to improve the original Fisher's method while preserving the theoretical benefits of $\log$-transformation (e.g., Li and Tseng (2011); Zaykin et al. (2002); Zhang et al. (2020b)). However, none of these methods provides a theoretical guarantee for detecting weak and sparse signals in a large-scale setup. And most of the methods consist of tuning components that heavily rely on prior or external biological knowledge. Furthermore, despite the widespread use of approximation techniques such as the central limit theorem or self-normalization for calculating p-values, the impact of such approximations in large-scale settings is rarely studied.

1.3.3 Combining Dependent, Weak, and Sparse Signals in A Large Group of P-Values

This scenario is frequently encountered in the applications of SNP-set association tests (Sun and Lin, 2020). Many efforts have been made to adapt the higher critism and Berk & Jones tests to account for the correlation structure of p-values (Sun and Lin, 2020; Hall and Jin, 2010; Barnett et al., 2017). However, such attempts are not accurate for extremely small p-values and require cumbersome and intensive computation for even moderate large n. In addition, the correlation structure between p-values can frequently become difficult to model due to the biological complexity of genetic mechanisms and limited data access to the raw GWAS data. Hence, it is of great interest to develop methods that are accurate for small p-values with fast computation, robustness to unknown dependency structure, and competitive power for detecting weak, sparse and correlated signals.

1.4 Overview of this Dissertation

This dissertation consists of 5 chapters. Chapter 1 provides an overview of p-value combination methods, omics data analysis, and challenges associated with using p-value combination methods

for omics data analysis.

In Chapter 2, we consider the scenario of combining independent and relatively frequent signals, which is summarized in Section 1.3.1. We first provide a comprehensive examination of commonly used p-value combination methods in terms of Bahadur efficiency and exact slope under independent inputs, followed by a comprehensive evaluation of the finite-sample statistical power of the methods. We conclude that the Fisher and the rank truncated product methods have top performance and complementary advantages. We consequently propose an ensemble method to combine the strengths of the two methods, with fast computation and a theoretical guarantee on asymptotic efficiency. This work is accepted by *Statistica Sinica* (Fang et al., 2023b).

Chapter 3 considers the scenario of combining independent, weak, and sparse signals, which is summarized in Section 1.3.2. In this chapter, we propose a fully adaptive modified Fisher's method based on weakly geometric system. We show the proposed method achieves the optimal separating rate in a high-dimensional setting for detecting weak, sparse, and heterogeneous signals. In terms of practical consideration, the robustness of our method when the p-values are not exact is investigated, where we show that our method still attains optimal separating rate under mild conditions.

In Chapter 4, we consider the scenario of combining dependent, weak, and sparse signals, which is summarized in Section 1.3.3. In this scenario, we investigate a family of p-value combination tests, which are formulated as the sum of transformed p-values with the transformations by a broad family of heavy-tailed distributions. We explore the conditions under which a method of the family possesses robustness to unknown dependency for type I error control and optimal detection boundary for detecting weak and sparse signals. We show that only an equivalent class of Cauchy and harmonic mean tests has sufficient practical resistance against dependency. As a consequence of the theoretical results, we propose a truncated Cauchy method, which belongs to the equivalent class of Cauchy and harmonic mean tests, to tackle the large negative penalty issue in the Cauchy method. This work is also accepted by *Statistica Sinica* (Fang et al., 2023a).

Chapter 5 contains the discussion and future work.

2.0 On P-Value Combination of Independent and Non-Sparse Signals: Asymptotic Efficiency and Fisher Ensemble

The contents of this chapter are accepted by the journal *Statistica Sinica* (Fang et al., 2023b).

2.1 Introduction

Methods for combining p-values are historically of substantial interest in statistics and in applications of many scientific fields to aggregate homogeneous or possibly heterogeneous information from multiple sources. Consider the problem of combining K p-values, $\vec{p} = (p_1, \ldots, p_K)$, where p_i is p-value of testing $H_0^{(i)} : \theta_i \in \Theta_0^{(i)}$ versus $H_1^{(i)} : \theta_i \in \Theta^{(i)} - \Theta_0^{(i)}$. Here θ_i denotes the parameter of interest and $\Theta^{(i)}$ and $\Theta_0^{(i)}$ denote the total possible parameter space and null parameter space of θ_i, respectively. For example, $\theta_i = \mu_i$ for $N(\mu_i, 1)$, $\Theta^{(i)} = \mathbb{R}$ and $\Theta_0^{(i)} = \{\mu_i = 0\}$ for a simple Z-test. The global union-intersection test for detecting any signal in the K p-values is $H_0 : \cap_{1\le i\le K}\{\theta_i \in \Theta_0^{(i)}\}$ versus $H_1 : \cup_{1\le i\le K}\{\theta_i \in \Theta^{(i)} - \Theta_0^{(i)}\}$. A general strategy is to combine the input p-values and form a test statistic for globally testing the existence of any signal. In the literature, three major categories of methods have been developed, depending on the types of input data and signal. The first category considers combination of independent p-values, where K is small and fixed (e.g., K =5-30). The sample size n_i ($1 \le i \le K$) for deriving p_i is large and can asymptotically go to infinity. This first classical scenario is closely related to meta-analysis applications to integrate multiple small effects for increasing statistical power. Traditional methods include Fisher's method $T_{\text{Fisher}} = \sum_{i=1}^{K} -2\log p_i$(Fisher, 1992) and Stouffer's method $T_{\text{Stouffer}} = \sum_{i=1}^{K} \Phi^{-1}(1-p_i)$(Stouffer et al., 1949) as well as many other transformation selections. The second category considers combining independent, sparse, and weak signals, where a large number of p-values are combined ($K \to \infty$) while only a small number ℓ of the K p-values ($\ell = K^\beta$ with $0 < \beta < \frac{1}{2}$) have weak signals and all remaining p-values have no signal. High criticism (denoted by HC test hereafter; Donoho and Jin (2004)) and Berk-Jones test (denoted by BJ test hereafter; Berk and Jones (1979); Li and Siegmund (2015)) are two representative methods

and are shown to be asymptotically optimal in terms of detection boundary across varying levels of signal sparsity ($0 < \beta < \frac{1}{2}$) as $K \to \infty$. In the third category, the integration of K p-values with unknown correlation structure and with sparse and weak signals is considered. Liu and Xie (2020) and Wilson (2019a) proposed Cauchy test (CA) and harmonic mean test (HM), respectively. These methods provide robustness under unknown dependency structure when inference is established under independence assumption and they attain optimal detection boundary for detecting highly sparse signals (with $s = K^{\beta}$, $0 < \beta < \frac{1}{4}$, but not for $\frac{1}{4} < \beta < \frac{1}{2}$) as $K \to \infty$ (Liu and Xie, 2020; Fang et al., 2023a).

In this chapter, we revisit methods of the first category, evaluate their asymptotic efficiencies, assess finite-sample numerical performance, and propose an ensemble method combining two complementary top performers for general applications. To tell the differences between the first category and the second and third categories, we emphasize that we focus on detecting independent and non-sparse signals inside a small and fixed number of p-values for scenarios of the first category, where "non-sparse" signals distinguish from the "sparse" signals in the second and third categories in the sense that the proportion of true signals varies from $1/K$ to 1 and is unknown, while the proportion in the second and third categories vanishes to zero as $K \to \infty$. Despite active needs and method development for the second and third categories, methods for the first "meta-analytic scenario with unknown heterogeneity" remain in high demand and present new challenges in many applications such as transcriptomic, GWAS, CNV or methylation meta-analyses (Li and Tseng, 2011; Tseng et al., 2012; Begum et al., 2012; Guerra and Goldstein, 2016).

Method development for the first category before the 1970-80s focuses on a class of methods aggregating transformed scores from the p-values: $T = \sum_{i=1}^{K} g(p_i) = \sum_{i=1}^{K} F_U^{-1}(p_i)$, where $F_U^{-1}(\cdot)$ is the inverse CDF of U. For example, U is chi-squared distribution for Fisher test and standard normal distribution for Stouffer test. Littell and Folks (1973) showed that Fisher's method is asymptotically optimal in terms of Bahadur relative efficiency, providing theoretical justification of the $\log$-transformation over the other types of transformations (see Section 2.2 for more details). Despite optimal asymptotic efficiency of Fisher test, its finite-sample performance in terms of statistical power is often poor if only part of the K p-values have signals. In this commonly encountered situation with unknown heterogeneous signals, many modified Fisher methods have been developed to improve the original Fisher's method. Dudbridge and Koeleman

(2003) proposed the rank truncated product (RTP) method to aggregate signals only for the top ordered (i.e., the most significant) p-values: $T_m = -2\sum_{i=1}^{m}\log p_{(i)}$, where $p_{(i)}$ is the i-th ordered p-value and $1 \leq m \leq K$ is a user-specified truncation point on ranks of input p-values. However, the choice of m is subjective and RTP can suffer substantial power loss with a mis-specified m. To address this challenge, a line of works has focused on improving RTP by adaptively determining m from an optimization criterion. For example, Song et al. (2016) developed an adaptive Fisher procedure using partial sum optimized by $\underline{z}$-standardization similar to higher criticism (denoted by AFz hereafter): $T_{\text{AFz}} = \max_{1\leq j\leq K}\frac{-\sum_{i=1}^{j}\log p_{(i)}-\sum_{i=1}^{n}w(j,i)}{\sqrt{\sum_{i=1}^{n}w^2(j,i)}}$, where $w(j,i)=\min\{1,j/i\}$. Let $\bar{F}_{\chi^2_{2j}}(t)=1-F_{\chi^2_{2j}}(t)$, where $F_{\chi^2_{2j}}(t)$ denotes the CDF of a chi-squared random variable with degrees of freedom $2j$. Li and Tseng (2011) proposed an adaptive Fisher procedure using partial sum optimized by the corresponding pseudo/surrogate "p-values"(denoted by AFs hereafter): $T_{\text{AFs}} = \max_{1\leq j\leq K} -\log(h(\vec{p},j))$. Here $h(\vec{p},j) = \bar{F}_{\chi^2_{2j}}(-2\sum_{i=1}^{j}\log p_{(i)})$ is the corresponding surrogate "p-value" of the partial sum, which is not a true and valid p-value but a surrogate for fast computation by importance sampling (Huo et al., 2020). Instead of using the surrogate p-values in AFs, Yu et al. (2009) proposed the adaptive rank truncated product (ARTP) method that is based on the exact $\underline{p}$-values of the partial sum (denoted by AFp hereafter): $T_{\text{AFp}}=\max_{1\leq j\leq K}-\log(h_j(\vec{p},j))$, where $h_j(\vec{p},j)=1-G_j(-2\sum_{i=1}^{j}\log p_{(i)})$ with $G_j(t)$ denotes the CDF function of $-2\sum_{i=1}^{j}\log p_{(i)}$ under the null. For computation, Yu et al. (2009) proposed an algorithm that requires large storage memory to achieve manageable computing.

Another related strategy in the literature is to directly filter out p-values greater than a user-specified threshold $\tau\in(0,1]$. For example, the truncated Fisher with hard-thresholding (denoted by TFhard) $T_{\text{TFhard}}(\tau)=\sum_{i=1}^{K}-\log(p_i)\text{I}_{\{p_i\leq\tau\}}$ (Zaykin et al., 2002), where $\text{I}_{\{.\}}$ denotes the indicator function. Zhang et al. (2020b) proposed truncated Fisher with soft-thresholding (TFsoft) to improve TFhard, arguing that the continuous soft-thresholding scheme can lead to more stable performance with varying input p-values. In both TFsoft and TFhard, the choice of τ is not straightforward. Zhang et al. (2020b) investigated the optimal choice of τ for TFhard under a theoretical setting of Gaussian mixture, where mixture probability and mean of the signals are known and $K\to\infty$. However, such prior information is generally unknown in applications. To this end, they replaced a single user-specified τ with a user-specified set of thresholds $\mathcal{T}$ and proposed two omnibus tests for TFhard and TFsoft, which alleviate the issue of choosing τ to some extent but

the selection of $\mathcal{T}$ is still prespecified and ad hoc.

Another line of research in p-value combination incorporates weighting in the procedure. For example, Xu et al. (2016) proposed an adaptive two-sample test for high-dimensional means, which can be regarded as a weighted test. Liptak's test (Lipták, 1958) can be considered as Stouffer's method with weights and is commonly referred to as the weighted z-test. Won et al. (2009) estimated the best weights for Liptak's method from a simple alternative hypothesis assuming expected effect size. Different choices of weights for z-test are suggested by other researchers, including Mosteller and Bush (1954) and Zaykin (2011). In addition to the weighted z-test (Liptak's test), many tests constructed by the sum of transformed p-values also have a weighted version. For example, the Cauchy and harmonic mean tests were originally proposed with weights and in this chapter we use the version of equal weights (Wilson, 2019a; Liu et al., 2019; Liu and Xie, 2020). For another example, Chen et al. (2014) proposed a test of combining p-values based on the sum of inverse gamma distribution, which can also be regarded as a weighted test in the sense that it gives larger "weights" to smaller p-values. In fact, AFs and AFp can be considered as an adaptively weighted method using binary weights and TFsoft (Zhang et al., 2020b) can be viewed as thresholding and weighting of the Fisher method.

Notwithstanding the active development of modified Fisher methods, there is a lack of comprehensive and systematic evaluation of the asymptotic properties and finite-sample numerical performance of the methods in the first category. Our chapter sets out to fill this gap. In Section 2.2, we examine asymptotic Bahadur optimality (ABO) of 7 methods in the first category: Fisher, Stouffer, AFs, AFz, AFp, TFhard, and TFsoft. The two adaptive Fisher methods, AFs and AFz, provide estimates of the subset of p-values contributing to the signal. Therefore, we also investigate whether the estimates in these two methods consistently select the subset of p-values containing true signals (signal selection consistency). For completeness, we also examine asymptotic efficiencies for methods developed for sparse signals, including Cauchy, Pareto family, minimum p-value (minP), BJ, and HC. In Section 2.3, we perform finite-sample numerical evaluations to compare statistical power of these methods under different K, signal strength, and proportions of true signals. Results of Sections 2.2 and 2.3 conclude complementary advantages of 2 top performers – Fisher and AFp –, especially in varying proportions of true signals. Consequently, we develop a Fisher ensemble (FE) method in Section 2.4 that applies a harmonic mean ensemble approach to combine Fisher

and AFp. We prove asymptotic Bahadur optimality of FE (Section 2.4.2) and demonstrate its consistently high performance in varying simulation scenarios (Section 2.4.3). Section 2.5 develops an extension of the FE method, namely FE_{CS}, for enhanced statistical power on detecting signals with concordant effect size directions. Section 2.6 applies FE and FE_{CS} as well as existing methods to a transcriptomic meta-analysis on biomarker and pathway detection for aging (Zahn et al., 2007). Section 2.7 provides final discussion and conclusion.

2.2 Asymptotic Efficiencies of Existing Methods

This section investigates the asymptotic efficiencies of existing p-value combination methods. Since our focus is on the scenarios with independent and non-sparse signals inside a finite number of p-values, we slightly generalize the setup proposed in Littell and Folks (1973) (differences are discussed in Remark 2.1), which uses the criterion of exact Bahadur relative efficiency (Bahadur, 1967b). Under this setting, Fisher's method is asymptotically Bahadur optimal (Littell and Folks, 1973) and shows theoretical advantages of $\log$-transformation. Multiple modified Fisher's methods (AFs, AFp, AFz, TFhard, and TFsoft) have been developed to improve finite-sample statistical power, but their asymptotic efficiencies have not been investigated. Section 2.2.1 introduces the problem setting and defines the exact slope of a hypothesis test, which is a natural concept derived from the exact Bahadur relative efficiency. Section 2.2.2 presents asymptotic Bahadur optimality (ABO) results of the 5 modified Fisher's methods.

2.2.1 Bahadur Relative Efficiency and Exact Slope

We first introduce the concept of exact slope of a hypothesis test (Bahadur, 1967b; Littell and Folks, 1973). Consider $(x_1, x_2, \cdots)$ an infinite sequence of independent observations of a random variable X from probability distribution P_θ with parameter $\theta \in \Theta$. Let T_n be a real-valued and continuous test statistic depending on the first n observations $(x_1, \ldots, x_n)$, where large values of T_n are considered to reject the null hypothesis. Assume that the probability distribution of T_n is the same for $\forall \theta \in \Theta_0$, which leads to $\mathbb{P}_\theta(T_n < t) = \mathbb{P}_0(T_n < t)$ for all $\theta \in \Theta_0$ and

assume $1 - \mathbb{P}_0(T_n < t)$ is uniformly distributed on $[0, 1]$ (Littell and Folks, 1973). Further denote $p^{(n)} = 1 - F_n(t_n)$ as the p-value of observed $T_n = t_n$, where $F_n(t) = \mathbb{P}_0(T_n < t)$. We then define the exact slope of T_n as follows.

Definition 2.1. For the test statistic T_n with p-value $p^{(n)}$, if there is a positive valued function $c(\theta)$, such that for any $\theta \in \Theta - \Theta_0$, $-\frac{2}{n} \log p^{(n)} \to c(\theta)$ as $n \to \infty$ with probability one. Then $c(\theta)$ is called the `exact slope` of T_n.

As a simple example, consider testing for zero mean ($\mu = 0$) with known variance under univariate Gaussian distribution and T_n is the conventional z-test. It is easily seen that $c(\mu) = \mu^2$ is the exact slope of the z-test. For more examples, see Abrahamson (1967); Bahadur (1967a). Exact slope of a test naturally connects to the exact Bahadur efficiency between test statistics. Consider two sequences of test statistics $\{T_n^{(1)}\}$ and $\{T_n^{(2)}\}$ testing the same null hypothesis with exact slopes $c_1(\theta)$ and $c_2(\theta)$ respectively. We define the ratio $\phi_{12}(\theta) = c_1(\theta)/c_2(\theta)$ as the *exact Bahadur relative efficiency* of $\{T_n^{(1)}\}$ relative to $\{T_n^{(2)}\}$, which compares the relative asymptotic efficiency between two test statistics. Indeed, considering any significance level $\alpha > 0$, for $i = 1, 2$, denote $N^{(i)}(\alpha)$ as the smallest sample size such that, for any $n \geq N^{(i)}(\alpha)$, the p-value of $T_n^{(i)}$ is smaller than α, one can show with probability one that $\lim_{\alpha \to 0} N^{(2)}(\alpha)/N^{(1)}(\alpha) = \phi_{12}(\theta)$, which asymptotically characterizes the ratio of the smallest sample sizes of two test statistics required to attain the same sufficiently small significant level α (Littell and Folks, 1973).

For $\theta \in \Theta_0$, the p-value $p^{(n)}$ follows uniform distribution Unif$(0, 1)$. Lemma 2.1 shows the analogous "exact slope" $-(2/n) \log p^{(n)}$ converges to zero with probability one.

Lemma 2.1. *For $\theta \in \Theta_0$, as n diverges, $-(2/n) \log p^{(n)} \to 0$ with probability one.*

The proof of Lemma 2.1 can be found in Supplement Section A.2.1. In this chapter, we extend the definition of exact slope to the null parameter space, where $c(\theta) = 0$ for $\theta \in \Theta_0$.

To benchmark the asymptotic efficiency of a p-value combination method, we then introduce the theoretical setup adopted from the framework in Littell and Folks (1973). Suppose we have $K < \infty$ sequences of test statistics $\{T_{n_1}^{(1)}\}, \ldots, \{T_{n_K}^{(K)}\}$ for testing $\theta_i \in \Theta_0^{(i)}$ for $1 \leq i \leq K$. Assume for all the sample sizes $n_1, \ldots, n_K$, and when $\theta_i \in \Theta_0^{(i)}$ for $1 \leq i \leq K$, $\{T_{n_1}^{(1)}\}, \ldots, \{T_{n_K}^{(K)}\}$ are independently distributed. Denote $p_i^{(n_i)}$ as the p-value of the i-th test statistic $T_{n_i}^{(i)}$. For each $1 \leq i \leq K$, assume that the sequence $\{T_{n_i}^{(i)}\}$ has exact slope $c_i(\theta_i)$ as $-(2/n_i) \log p_i^{(n_i)} \to$

$c_i(\theta_i) \geq 0$ with probability one as $n_i \to \infty$. We further assume the sample sizes $n_1, \ldots, n_K$ satisfy $n = (1/K)\sum_i^K n_i$ and $\lim_{n\to\infty}(n_i/n) = \lambda_i$, where $\lambda_i > 0$ and $\sum_i^K \lambda_i = K$. Under the above setup, the goal of any p-value combination method is to test

$$H_0 : \cap_{i=1}^K \left\{\theta_i \in \Theta_0^{(i)}\right\} \text{ versus } H_1 : \cup_{i=1}^K \left\{\theta_i \in \Theta^{(i)} - \Theta_0^{(i)}\right\}. \tag{2.1}$$

For simplicity, we assume under the null

$$\lambda_1 c_1(\theta_1) \geq \lambda_2 c_2(\theta_2) \geq \ldots \geq \lambda_K c_K(\theta_K) \geq 0,$$

where the first ℓ p-values have true signals (i.e., θ_i's belong to $\Theta^{(i)} - \Theta_0^{(i)}$ for $1 \leq i \leq \ell$) with exact slopes $c_i(\theta_i) > 0$, while $c_i(\theta_i) = 0$ for the remaining $\theta_i \in \Theta_0^{(i)}$, $\ell + 1 \leq i \leq K$.

Remark 2.1. There are 2 differences between the original setup in Littell and Folks (1973) and ours. First, Littell and Folks (1973) assumed that all the studies have strictly positive exact slopes, while we allow some studies to have zero-valued exact slopes. Second, Littell and Folks (1973) considered a general parameter space Θ while we consider a product of parameter spaces $\Theta^{(1)} \times \Theta^{(2)} \times \cdots \times \Theta^{(K)}$. Although differences exist, one can still establish the results in Littell and Folks (1973) by combining their arguments with Lemma 2.1.

Following Theorem 2 and arguments in Section 4 in Littell and Folks (1973), under the alternatives, the maximum attainable exact slope for any p-value combination method is $\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$. Hence we define the asymptotic Bahadur optimality (ABO) of a p-value combination method as follows.

Definition 2.2. Denote $\vec{\theta} = (\theta_1, \ldots, \theta_K)$. Under the above setup, a p-value combination test $H(p_1, \ldots, p_K)$ is asymptotically Bahadur optimal (ABO) if its exact slope $C_H(\vec{\theta})$ satisfies $C_H(\vec{\theta}) = \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$.

2.2.2 Asymptotic Bahadur Optimality Property of P-Value Combination Methods

Littell and Folks (1973) showed that Fisher test is ABO while Stouffer and minP tests are not. Except for these methods, there is a lack of asymptotic efficiency analysis of the other methods. This subsection focuses on discussing 5 modified Fisher methods: AFs, AFp, AFz, TFhard, and TFsoft. We additionally analyze 5 methods designed for combining sparse and weak signals: Cauchy, Pareto, BJ, and HC. As expected, the latter 4 tests do not enjoy ABO property, and the proofs are outlined in the supplement. The theoretical results of ABO, exact slope, and signal selection consistency (to be discussed in Theorems 2.4 and 2.5 and Remarks A3, A5, and A6) are summarized in Table 2.1.

Table 2.1: Results of asymptotic properties of 12 p-value combination methods: Fisher, Stouffer, 5 modified Fisher (AFs, AFp, AFz, TFhard and TFsoft) and 5 methods designed for sparse and weak signal (Cauchy, Pareto, minP, BJ and HC).

Methods	ABO	Exact slopes	Signal selection consistency	Proofs
Fisher	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem A1
Stouffer	No	$\frac{1}{K}\left[\sum_{i=1}^{\ell}(\lambda_i c_i(\theta_i))^{\frac{1}{2}}\right]^2$	–	Theorem A1
AFs	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	Yes	Theorems 2.1 & 2.4
AFp	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	Yes	Theorems 2.2 & 2.5
AFz	No	$\leq \max_{j\leq \ell} \frac{\sqrt{\sum_{i=1}^{K} \min^2\{1,1/i\}} \sum_{i=1}^{K} \lambda_i c_i(\theta_i)}{\sqrt{\sum_{i=1}^{K} \min^2\{1,j/i\}}}$	No	Theorem 2.3
TFhard	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem 2.6
TFsoft	Yes	$\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$	–	Theorem 2.6
Pareto	No	$\max_i \lambda_i c_i(\theta_i)$	–	Theorem A2
Cauchy	No	$\max_i \lambda_i c_i(\theta_i)$	–	Theorem A3
minP	No	$\max_i \lambda_i c_i(\theta_i)$	–	Littell and Folks (1973)
BJ	No	$\max_i i\lambda_i c_i(\theta_i)$	No	Theorem A4
HC	No	–	No	Proposition A1

Recall that Fisher and the 5 modified Fisher methods combine p-values using the following

test statistics:

$$T_{\text{Fisher}} = \sum_{i=1}^{K} -2\log p_{(i)};\ T_{\text{AFz}} = \max_{1\le j\le K} \frac{-\sum_{i=1}^{j}\log p_{(i)} - \sum_{i=1}^{K} w(i,j)}{\sqrt{\sum_{i=1}^{K} w^2(i,j)}};$$

$$T_{\text{AFs}} = \max_{1\le j\le K} -\log(\bar{F}_{\chi^2_{2j}}(-2\sum_{i=1}^{j}\log p_{(i)}));\ T_{\text{AFp}} = \max_{1\le j\le K} -\log(h_j(\vec{p},j));$$

$$T_{\text{TFhard}}(\tau) = \sum_{i=1}^{K}(-2\log p_i)\mathrm{I}_{\{p_i\le\tau\}};\ T_{\text{TFsoft}}(\tau) = \sum_{i=1}^{K}(-2\log p_i + 2\log\tau)_+.$$

Here $w(i,j) = \min\{1, j/i\}$. In addition, $\tau \in (0,1]$ is a user-specified constant for the two truncated Fisher methods and $(x)_+$ denotes $\max(x,0)$.

All the 6 methods can be characterized in the form of $H(-\log p_1, \ldots, -\log p_K)$ by some function $H(\cdot)$. With the $\log$-transform on p-values as a key ingredient, the above methods potentially can achieve high asymptotic efficiency. Indeed, combining with Lemma 2.1, by using almost the same arguments in Littell and Folks (1973), one can show that Fisher test attains ABO, presented in Theorem A1 for completeness.

Although achieving high asymptotic efficiency, the Fisher test has been shown to have poor performance empirically when only small part of p-values contain signals (e.g., 2 out of 10 p-values have signals); see Song et al. (2016) and Li and Tseng (2011) for more discussions. Many modified Fisher methods have been proposed to address this problem (Zaykin et al., 2002; Yu et al., 2009; Kuo and Zaykin, 2011; Zhang et al., 2020b; Li and Tseng, 2011; Song et al., 2016). The idea is to filter out large p-values that are less likely to carry signals and reduce the impact of noise, while still using the $\log$-transformation on p-values to achieve high efficiency. Particularly, AFs, AFp and AFz combine the first m smallest ordered p-values. All the three methods use some optimization criterion that adaptively selects m to achieve superior finite-sample power in varying proportions of signals. Whether AFs, AFp, and AFz retain the ABO property of Fisher is an intriguing question and is investigated below. In fact, we will surprisingly find in the following three theorems that AFs and AFp are ABO, but AFz is not.

Theorem 2.1 (AFs is ABO)**.** *Under the setup in Section 2.2.1, T_{AFs} is similar to Fisher test to be ABO with exact slope $C_{AFs}(\vec{\theta}) = \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$.*

Theorem 2.2 (AFp is ABO)**.** *Under the setup in Section 2.2.1, T_{AFp} is similar to Fisher test to be ABO with exact slope $C_{AFp}(\vec{\theta}) = \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$.*

Theorem 2.3 (AFz is not ABO)**.** *Under the setup in Section 2.2.1, consider the following test statistic $T_A = \max_{1\le j\le K} \frac{-2\sum_{i=1}^{j}\log p_{(i)} - A_j}{B_j}$, where $B_j > 0$ and A_j are some finite constants only depend on j and K. Assume there is no tie for $\frac{\sum_{i=1}^{j}\lambda_i c_i(\theta_i)}{B_j}$, j=1,....,K, and B_j is monotonic increasing. Then T_A is not ABO with exact slope*

$$C_A(\vec{\theta}) \le \max_{1\le j\le \ell} (B_1/B_j) \sum_{i=1}^{j} \lambda_i c_i(\theta_i).$$

The equality holds if and only if $\ell = 1$ (i.e., there is only one signal inside the K p-values).

By taking $A_j = 2\sum_{i=1}^{K} w(j,i)$ and $B_j = 2(\sum_{i=1}^{K} w^2(j,i))^{\frac{1}{2}}$, T_A reduces to T_{AFz}, indicating that AFz is not ABO in general (e.g., a special case that AFz is ABO is when $\ell = 1$).

The better asymptotic efficiency properties of AFp and AFs compared to AFz may be due to its attempt to estimate the subset of p-values with true signals. Consider the equivalent form of AFs for combining independent p-values:

$$T'_{\text{AFs}} = \min_{\vec{w}} \bar{F}_{\chi^2_{2(\sum_{i=1}^{K} w_i)}}(-2\sum_{i=1}^{K} w_i \log p_i),$$

where $\vec{w} = (w_1, \ldots, w_K) \in \{0,1\}^K$ is the vector of binary weights that identify the candidate subset of p-values with true signals. Note that T'_{AFs} is the original form proposed in Li and Tseng (2011). Denote by $\hat{\vec{w}} = \text{argmin}_{\vec{w}} \bar{F}_{\chi^2_{2(\sum_{i=1}^{K} w_i)}}(-2\sum_{i=1}^{K} w_i \log p_i)$ and let

$$\vec{w}^* = \{(w_1^*, \cdots, w_K^*) : w_k^* = 1 \text{ if } \theta_i \in \Theta - \Theta_0 \text{ and } w_k^* = 0 \text{ if } \theta_i \in \Theta_0\}$$

be the indicators of the true signals We can show signal selection consistency of AFs in the following theorem.

Theorem 2.4 (signal selection by AFs is consistent)**.** *Under the setup in Section 2.2.1, $\hat{\vec{w}} \to \vec{w}^*$ as $n \to \infty$ in probability for the AFs test.*

Theorem 2.5 (signal selection by AFp is consistent)**.** *Under the setup in Section 2.2.1, AFp can select true subset of p-values by selecting $p_{(1)}, \ldots, p_{(\hat{j})}$, where*

$$\hat{j} = \underset{1\le j\le K}{\text{argmax}} - \log(h_j(\vec{p}, j)).$$

The following Theorem 2.6 states that for any given value of $\tau \in (0, 1]$, TFhard and TFsoft are ABO:

Theorem 2.6 (TFhard and TFsoft are ABO)**.** *Under the setup in Section 2.2.1, TFhard and TFsoft are ABO with exact slopes* $C_{TFhard}(\vec{\theta}) = C_{TFsoft}(\vec{\theta}) = \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)$.

Although TFhard and TFsoft are ABO, the choice of τ may significantly impact their finite-sample performance (Zhang et al., 2020b). To address this issue, Zhang et al. (2020b) proposed the following omnibus tests for both methods (denoted by oTFhard and oTFsoft, respectively):

$$T_{\text{oTFhard}} = \min_{\tau \in \mathcal{T}} 1 - F_{U_{\text{TFhard}}(\tau)}(T_{\text{TFhard}}(\tau))$$

$$T_{\text{oTFsoft}} = \min_{\tau \in \mathcal{T}} 1 - F_{U_{\text{TFsoft}}(\tau)}(T_{\text{TFsoft}}(\tau)),$$

where $\mathcal{T} = \{\tau_1, \ldots, \tau_m\}$ is a user-specified set of the candidates of τ. Here $U_{\text{TFhard}}(\tau)$ and $U_{\text{TFsoft}}(\tau)$ denote the random variables that follow the null distributions of $T_{\text{TFhard}}(\tau)$ and $T_{\text{TFsoft}}(\tau)$, respectively. Although the omnibus tests alleviate the issue of sensitivity of the choice of τ for both TFhard and TFsoft to some extent, selection of $\mathcal{T}$ is still user-specified and subjective. In addition, Zhang et al. (2020b) derive the null distributions of both omnibus tests in an asymptotic sense as $K \to \infty$, which may not be accurate for small K with small p-value thresholds that are commonly used in applications, such as genomics studies, to handle multiplicity.

Proofs of theorems for Fisher and modified Fisher methods in this subsection can be found in Supplement Section A.2.1. For completeness, we also show that methods designed for combining sparse and weak signals, such as Cauchy, Pareto, BJ and HC, are not ABO (Supplement Section A.1) and the proofs are available in Supplement Section A.2.3. In conclusion, Fisher, AFs, AFp, TFhard and TFsoft are the only 5 methods with ABO property. AFs, AFp, and AFz are the 3 methods to provide signal selection (i.e., subset estimation of the true signal) and AFs and AFp are the only two methods to have consistency in the signal identification.

2.3 Power Comparison in Finite-Sample Simulations

Although Section 2.2 evaluates asymptotic efficiencies of p-value combination methods, the finite-sample statistical power of the methods under different proportions of signals has not been

assessed. In this section, we evaluate 7 methods that are designed for non-sparse signal setting described in Section 2.2: Fisher, Stouffer, AFs, AFp, AFz, TFhard, and TFsoft. Additionally, we also evaluate methods designed for combining sparse and weak signals for completeness: minimum p-value (minP), Cauchy (CA), harmonic mean (HM), Berk & Jongs (BJ), and higher criticism (HC). As TFhard and TFsoft are sensitive to the choice of tuning parameter τ, for a fair comparison, we use the corresponding omnibus tests, oTFhard and oTFsoft, instead. The tuning candidate set $\mathcal{T}$ is set to be $\{0.01, 0.05, 0.5, 1\}$, which is used in the original paper (Zhang et al., 2020b).

For better illustration, we first present the results of the 7 methods designed for combining non-sparse signals in Figures 2.1 and S1. Results comparing all 12 methods can be found in Supplement Figures A2 and A3, where modified Fisher's methods generally dominate other methods designed for sparse and weak signals unless the signals are indeed sparse and weak (e.g., cases of $\ell/K \leq 0.1$ in Figure A3). However, in such cases, methods such as AFp and AFz still have comparable power with the top-performing methods such as minP.

We simulate $X = (X_1, \ldots, X_K) \overset{D}{\sim} N(\vec{\mu}, I_K)$, where $\vec{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)$ contains ℓ non-zero signals $\mu_1 = \cdots = \mu_\ell = \mu_0$ and $K - \ell$ with no signal ($\mu_{\ell+1} = \cdots = \mu_K = 0$). We evaluate for a wide range of $K = 10, 20, 40, 80$. For each K, we vary proportions of true signals ℓ/K: $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$. We also vary $\mu_0 = 0.5, 0.65, \ldots, 6$ for a broad range of signal strength. The p-values are calculated by two-sided test $p_i = 2(1 - \Phi(|X_i|))$ for $i = 1, \ldots, K$. For each combination of parameter values, we draw 10^6 Monte Carlo samples to calculate the critical values for all the methods at a given significance level α, since the p-value calculation algorithms for some methods, such as oTFsoft and oTFhard, are not accurate for small K.

Figure 2.1 shows the empirical power of Fisher, Stouffer, and 5 modified Fisher methods with varying proportions of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ at significance level $\alpha = 0.01$. For a given K and proportion of signals ℓ/K, we choose the smallest μ_0 such that the best method has at least 0.5 statistical power, which allows optimized visualization and comparison of different methods in different signal settings. We first note that AFz is inferior to the other modified Fisher methods, consistent with our theoretical result that AFz is not ABO. We further note that AFs, AFp, oTFhard, and oTFsoft have comparable performance across varying proportions of signals. Fisher outperforms all other methods for detecting frequent signals (e.g., when the proportion of true signals is greater than 0.3).

Although the case of combining a small number of strong signals is not our primary focus, out of curiosity and for a more comprehensive evaluation of existing methods, we simulate the alternatives with fixed numbers of true signals $\ell = 1, 2, \ldots, 6$ for $K = 20, 40, 80$ following the above simulation scheme. Figure A1 shows the empirical power of Fisher, Stouffer, and 5 modified Fisher methods with varying numbers of signals $\ell = 1, 2, \ldots, 6$ at $\alpha = 0.05$. Similarly, for a given K and ℓ, we choose the smallest μ_0 such that the best method has at least 0.9 statistical power. Clearly, this simulation setting focuses more on the performance of combining less-frequent but relatively strong signals. We note that AFz, AFp, and oTFsoft have comparable statistical power across varying numbers of true signals, followed by AFs and oTFhard. While Fisher, is significantly inferior than the modified Fisher's methods when ℓ is much smaller than K (e.g., $\ell \leq 3$ for $K = 20, 40, 80$).

In many real applications (e.g., the transcriptomic meta-analysis in Section 2.6), the p-value combination test is repeated many times (i.e., for each gene). It is expected that some true biomarkers are more homogeneous with frequent true signals and some with less-frequent signals. The results in Figures 2.1 and S1 show the need to develop an ensemble method to integrate the advantages of Fisher and one of the top-performing modified Fisher methods, which is presented in the next section.

2.4 Fisher Ensemble to Combine Fisher and AFp

As shown in Sections 2.2 and 2.3, Fisher and 4 modified Fisher methods (AFs, AFp, TFhard, and TFsoft) are ABO, and have complementary strength in finite-sample evaluation of varying proportions and numbers of true signals. A natural idea is to ensemble Fisher and one of the 4 modified Fisher methods for more stable and universally competitive performance. Since oTFhard and oTFsoft methods require an ad hoc decision of user-specified set $\mathcal{T}$ and their existing computing algorithms are not accurate for small K, we choose to develop an ensemble method to combine Fisher and AFp in this section. In Section 2.4.1, we propose an ensemble approach, namely Fisher ensemble (FE), using the harmonic mean method (Wilson, 2019a; Fang et al., 2023a) to combine Fisher and AFp. In section 2.4.2, we provide theoretical support of FE and show that FE is ABO.

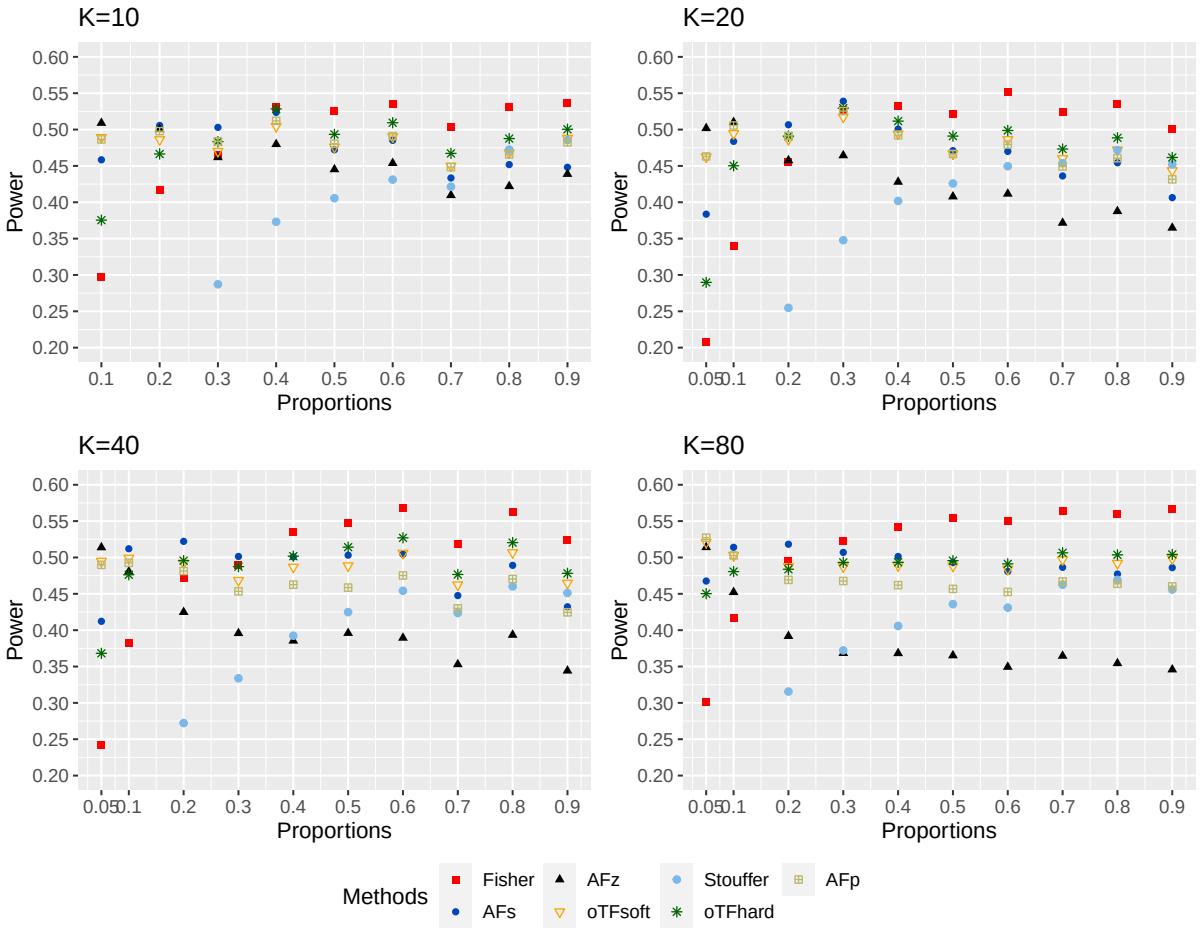


Figure 2.1: Statistical power of Fisher, Stouffer, and 5 modified Fisher's methods at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible compared to the scale of the mean power (smaller than 0.1% of the power) and hence omitted. The results of Stouffer with power smaller than 0.25 are omitted.

Section 2.4.3 presents simulation results similar to Section 2.3 to demonstrate the balanced and superior performance of FE across varying proportions of true signals.

2.4.1 Fisher Ensemble by Harmonic Mean Integration

Denote by p^{Fisher} and p^{AFp} the p-values derived from Fisher and AFp combination tests, respectively. We propose to ensemble the two methods by combining their p-values using $T_h = [h(p^{\text{Fisher}}) + h(p^{\text{AFp}})]/2$ with function h. Since p^{Fisher} and p^{AFp} can be highly dependent, one option is to use the Cauchy combination test with $h(p) = \tan(\pi(\frac{1}{2} - p))$, as theorems and simulations in Liu and Xie (2020) and Liu et al. (2019) show that the Cauchy combination test is robust to dependency of combined p-values, and further results in a fast-computing algorithm with Cauchy distribution under the null hypothesis (i.e., the null distribution is standard Cauchy). This Cauchy ensemble approach is, however, problematic when either p^{Fisher} or p^{AFp} is close to 1. In this case, the Cauchy transformation generates a $-\infty$ score and the power is greatly reduced. We propose to use the harmonic mean method (Wilson, 2019a), $h(p) = 1/p$, in our Fisher ensemble (FE) method by

$$T_{\text{FE}} = (1/2)[1/p^{\text{Fisher}} + 1/p^{\text{AFp}}], \tag{2.2}$$

where the harmonic mean method has been shown to be approximately equivalent to Cauchy in (Fang et al., 2023a). When p-value p follows Unif$(0, 1)$, the reciprocal of p follows Pareto distribution Pareto$(1, 1)$ with both the scale and shape parameters equal to 1. The purpose of using reciprocal of p-values instead of $h(p)$ is to avoid the large negative score issue of the transformation by Cauchy distribution described above; also see Fang et al. (2023a) for more details. Except for avoiding large negative score issue, ensemble by harmonic mean using the reciprocal of p-value $1/p$ performs almost identically to Cauchy $h(p)$. Supplement Section A.3.7 provides numeric results where the ensemble method using harmonic mean performs better than that using Cauchy combination test.

In the implementation, FE is fully data-driven with fast-computing algorithms. Indeed, for $p_1, \ldots, p_K \overset{\text{i.i.d.}}{\sim}$ Unif$(0, 1)$, null distribution of Fisher test follows chi-squared distribution with degrees of freedom $2K$. For p-value calculation for AFp, Yu et al. (2009) proposed an empirical

approach to avoid cumbersome two-layer permutation. Finally, Theorems 1 and 2 in Fang et al. (2023a) have shown that the harmonic approach using the reciprocal of p-values can have robust type I error control if we naively use the Pareto distribution Pareto$(1, 1)$ as the null distribution (see Supplement Section 1.2 for more details). As a result, fast p-value computation for Fisher ensemble T_{FE} is warranted. Table A1 in Section A.3.1 justifies the above fast-computing procedure, where we show the type I error control for FE is accurate for $\alpha \leq 0.05$ across a broad range of $5 \leq K \leq 100$.

2.4.2 Asymptotic Efficiency of Fisher Ensemble

In this subsection, we will show that Fisher ensemble (FE) is asymptotically Bahadur optimal (ABO). We first introduce a heavy-tailed distribution family, namely regularly-varying distribution R (Mikosch, 1999), where Cauchy and Pareto distributions are special cases of the family. Consider an ensemble method induced by a regularly-varying distribution (e.g., Pareto$(1, 1)$ for $1/p$ in our case) to combine multiple p-value combination methods (e.g., Fisher and AFp in our case). The ensemble method will be shown to be ABO if at least one of the p-value combination methods is ABO. Since both Fisher and AFp are ABO and Pareto$(1, 1)$ (corresponding to $1/p$) is a regularly varying distribution, we conclude that Fisher ensemble is also ABO. Below, we outline the definition of regularly-varying distribution and the theorem. The detailed proof is available in Supplement Section A.2.2.

Definition 2.3. A distribution F is said to belong to the regularly-varying tailed family with index γ (denoted by $F \in R_{-\gamma}$) if $\lim_{x\to\infty} \frac{\bar{F}(xy)}{\bar{F}(x)} = y^{-\gamma}$ for some $\gamma > 0$ and all $y > 0$.

We denote the whole family of regularly varying tailed distributions by R. For two positive functions $u(\cdot)$ and $v(\cdot)$, we write $u(t) \sim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} = 1$. It can be shown that every distribution F belonging to $R_{-\gamma}$ can be characterized by $\bar{F}(t) \sim L(t)t^{-\gamma}$, where $\bar{F}(t) = 1 - F(t)$ and $L(t)$ is a slowly varying function. A function L is called slowly varying if $\lim_{y\to\infty} \frac{L(ty)}{L(y)} = 1$ for any $t > 0$. Regularly varying distribution is a wide class of heavy-tailed distributions, which includes Cauchy, Pareto$(1, 1)$ (harmonic mean), and general Pareto distributions.

Consider $L < \infty$ p-value combination test statistics $T_1, \ldots, T_L$. Denote by $p_{T_1}, \ldots, p_{T_L}$ the resulting p-values of $T_1, \cdots, T_L$. In Fisher ensemble, we have $L = 2$ and (T_1, T_2) are Fisher and

AFs. Under Definition 2.3, consider the following ensemble method by a regularly varying tailed distribution:

$$T_{\text{RV}}(\gamma) = \sum_{i=1}^{L} g_\gamma(p_{T_i}) = \sum_{i=1}^{L} F^{-1}_{U(\gamma)}(1 - p_{T_i}),$$

where $F_{U(\gamma)}$ is CDF of $U(\gamma)$ and $U(\gamma) \in R_{-\gamma}$. Under the null hypothesis, the test statistic transforms all the p_{T_i}'s into regularly varying tailed random variables with index γ. The following theorem suggests that under mild conditions, the ensemble method by regularly varying tailed distribution has ABO property.

Theorem 2.7. *For each $i = 1, \ldots, L$, let $C_i(\vec{\theta})$ be the exact slope of T_i and assume $\max_{1\le i\le L} C_i(\vec{\theta}) > 0$. Let $C^{(\gamma)}_{RV}(\vec{\theta})$ be the exact slope of $T_{RV}(\gamma)$, if one of the following two conditions holds: (C-1) $F^{-1}_{U(\gamma)}(1-p)$ is bounded below: $F^{-1}_{U(\gamma)}(1-p) \ge \nu$ for some constant ν and $\forall p \in [0,1]$; (C-2) All the T_i's have non-zero exact slopes: $\min_{1\le i\le L} C_i(\vec{\theta}) > 0$. then we have $C^{(\gamma)}_{RV}(\vec{\theta}) = \max_{1\le i\le L} C_i(\vec{\theta})$.*

Remark 2.2. As $1/p$ (reciprocal of p-value) is bounded below while $h(p)$ (Cauchy) is not, using $1/p$ rather than $h(p)$ can satisfy Condition (C-1) in Theorem 2.7. In general, if Condition (C-1) is not satisfied, Condition (C-2) is a mild condition (meaning all tests T_i are at least minimally effective and have non-zero slope) but Condition (C-2) is not always easy to check or satisfied in practice. For example, when we aggregate methods combining left one-sided p-values and right one-sided p-values in Section 2.4, methods only combining left one-sided p-values will produce p-values converging to one when there exist only positive effects. See Section 2.5 and Supplement Section A.3.7 for more details.

Theorem 2.7 suggests that $T_{\text{RV}}(\gamma)$ is ABO as long as at least one of $T_1, \ldots, T_L$ methods is ABO. Consequently, Fisher ensemble is ABO since Pareto$(1, 1)$ (corresponding to $1/p$) belongs to regularly-varying tailed distribution and both Fisher and AFp are ABO.

2.4.3 Finite-Sample Power Comparison of Fisher Ensemble

In this subsection, we evaluate the finite-sample power of FE. To illustrate that FE can take advantages of integrated methods, we also include AFs and Fisher as the baseline methods. We use the same simulation scheme in Section 2.3 to generate the simulated data. Figure 2.2 shows the statistical power of FE, AFp, and Fisher with varying proportions of true signals $\ell/K =$

$0.05, 0.1, 0.2, \ldots, 0.9$ at $\alpha = 0.01$. Similar to Figure 2.1, for a given proportion of signals ℓ/K and number of combined p-values K, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 for Figure 2.2. Figure A4 shows the statistical power of FE, AFp, and Fisher for combining $K = 20, 40, 80$ p-values with varying numbers of true signals $\ell = 1, 2, \ldots, 6$ at $\alpha = 0.05$. Similar to Figure A1, for a given ℓ and K, we choose the smallest μ_0 that allows the best method to have power larger than 0.9 for Figure A4, which is supposed to focus more on combining less-frequent but strong signals. As expected, we note that FE has a stable statistical power that is comparable to the better of Fisher and AFp in different settings with either dense but weak signals or less frequent but strong signals. Specifically, when the proportion of signals is high, FE performs close to Fisher and is superior to AFp. When the number of true signals is small, FE performs close to AFp and outperforms Fisher. In Supplement Figures A5 and A6, we implement another Fisher ensemble method (FE2) combining Fisher, AFp, and minP. As expected, its power for only a small number of signals is slightly improved over FE but at the expense of a large reduction of power when signals are frequent. From the asymptotic efficiency in Section 2.4.2 and simulations above, we recommend using the Fisher ensemble method combining Fisher and AFp for general applications.

2.5 Detection of Signals with Concordant Directions

2.5.1 Fisher Ensemble Focused on Concordant Signals (FE_{CS})

For all methods we have discussed so far, the global hypothesis setting is designed for two-sided tests, regardless of the directions of the effects. Recall from Equation 2.1 that the hypothesis testing considered is $H_0 : \cap_{i=1}^{K} \{\theta_i = 0\}$ vs $H_1 : \cup_{i=1}^{K} \{\theta_i \neq 0\}$. Consider the alternative hypothesis that only the first ℓ p-values have true signals (i.e., $\theta_i \neq 0$ for $1 \leq i \leq \ell$ and $\theta_{\ell+1} = \cdots = \theta_K = 0$). The two-sided tests to obtain p_i ($1 \leq i \leq K$) cannot guarantee signals with concordant directions ($\text{sgn}(\theta_1) = \cdots = \text{sgn}(\theta_\ell)$, denoted by $\text{sgn}(\cdot)$ the sign function), which is desirable in most applications. For example, when conducting meta-analysis of K transcriptomic studies believed to be relatively homogeneous, we are interested in identifying biomarkers con-

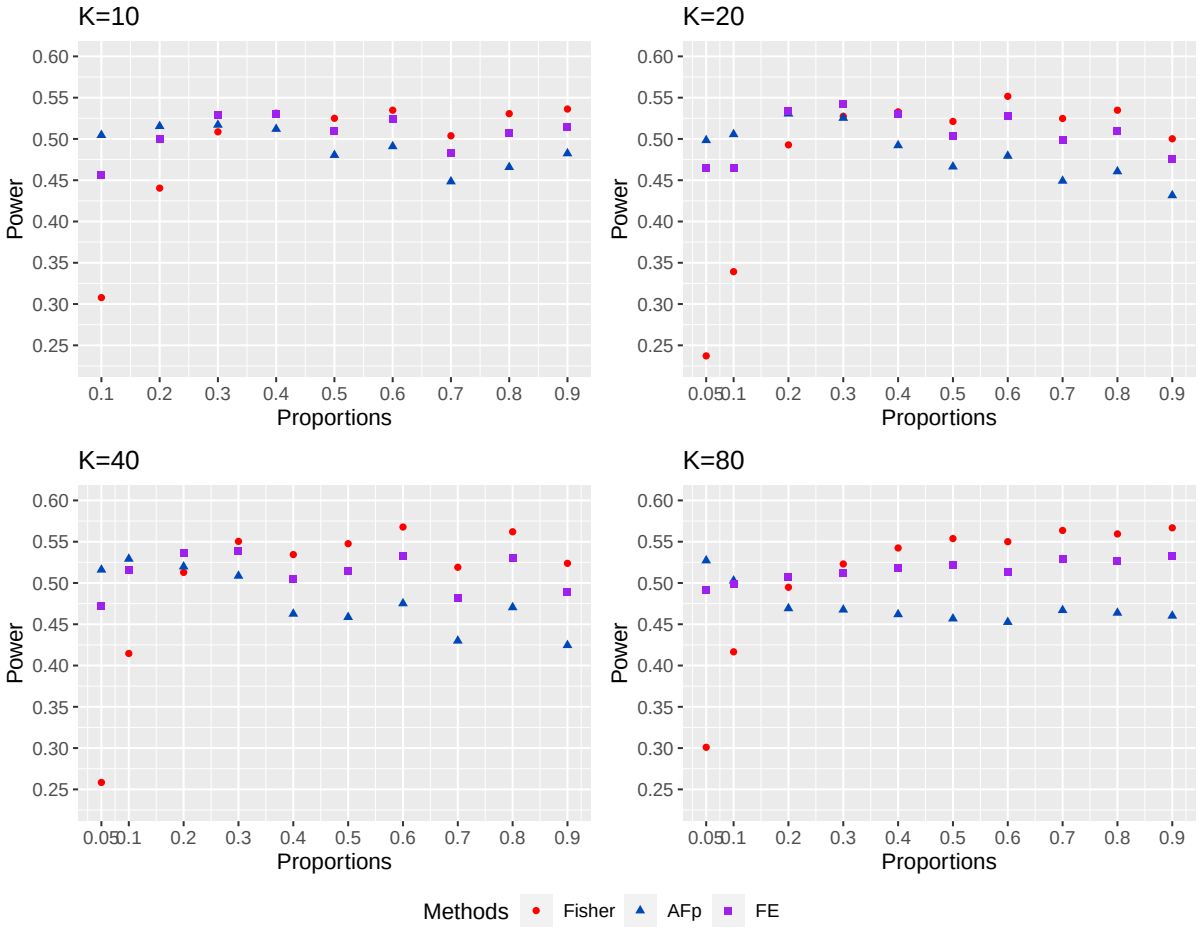


Figure 2.2: Statistical power of FE, Fisher, and AFp at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible and hence omitted.

cordantly up-regulated or down-regulated. For this problem, Owen (2009) revisited the Pearson test statistic and proposed to use $T_{\text{Pearson}} = \min\{\tilde{p}^{\text{Fisher},L}, \tilde{p}^{\text{Fisher},R}\}$, where $\tilde{p}^{\text{Fisher},L}$ and $\tilde{p}^{\text{Fisher},R}$ uses Fisher to combine the left and right one-sided p-values respectively, and the Pearson test takes the more significant one as the test statistic. In this subsection, we similarly extend the Fisher ensemble method to use the harmonic mean approach to combine the two left and right one-sided p-values of Fisher and AFs (denoted by FE_{CS}; Fisher ensemble for concordant signal):

$$T_{\text{FE}_{\text{CS}}} = (1/4)[1/\tilde{p}^{\text{Fisher},L} + 1/\tilde{p}^{\text{Fisher},R} + 1/\tilde{p}^{\text{AFp},L} + 1/\tilde{p}^{\text{AFp},R}].$$

Remark 2.3. When combining one-sided p-values, it is common to observe p-values close to 1 and it is critical to use harmonic mean, instead of Cauchy, to avoid $-\infty$ scores.

Remark 2.4. Let $C^L(\vec{\theta})$ be the maximum attainable exact slope for any p-value combination method combining left one-sided p-values, and define $C^R(\vec{\theta})$ in a similar manner for right one-sided p-values. By Theorem 2.7, the exact slope of FE_{CS} is $\max\{C^L(\vec{\theta}), C^R(\vec{\theta})\}$, indicating high asymptotic efficiency as even if one has prior knowledge of the effect size direction, it is impossible to design a p-value combination method with a larger exact slope for detecting concordant signals.

For computation, similar to FE, one can use p-value calculation by Pareto$(1, 1)$ to calculate p-value for FE_{CS}. This approximation procedure is justified by simulation results in Table A1 in Section A.3.1 for a broad range of significance levels α and numbers of input p-values K.

2.5.2 Finite-Sample Power Comparison of Fisher Ensemble for Concordant Signals

In this subsection, we evaluate the finite-sample power of FE_{CS}. To demonstrate the advantages of FE_{CS}, we also include the regular FE and Pearson as the baseline methods. We use the same simulation scheme in Section 2.3 to generate the simulated data. For FE_{CS} and Pearson, the one-sided p-values are generated by $\tilde{p}_i^{(L)} = 1 - \Phi(X_i)$ and $\tilde{p}_i^{(R)} = \Phi(X_i)$ $(i = 1, \ldots, K)$, respectively. While for the regular FE, we combine the two-sided p-values $p_i = 2(1-\Phi(|X_i|))$ for $i = 1, \ldots, K$.

Figures 2.3 and S7 show the empirical power of FE_{CS}, Pearson, and the regular FE. For Figure 2.3, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 for a given proportion of signals ℓ/K and a number of combined p-values K. Both FE_{CS} and Pearson

dominate the regular FE, indicating the former two methods perform better for the alternatives with one-sided direction consistent effects (as $\mu_1 = \ldots = \mu_s = \mu_0 > 0$ under the alternatives). In addition, FE_{CS} has a comparative performance with Pearson for $\ell/K \geq 0.2$ and outperforms Pearson when $\ell/K < 0.2$. For Figure A7, we choose the smallest μ_0 that allows the best method to have power larger than 0.9 for a given number of signals ℓ and a number of combined p-values K. This is a setting that focuses on the less frequent and strong signals. We note that FE_{CS} outperforms Pearson when the of signals is low (e.g., $\ell \leq 4$).

2.6 Real Application to AGEMAP Data

In this section, we apply different p-value combination methods to analyze the AGEMAP study (Zahn et al., 2007). The dataset contains microarray expression of 8,932 genes in sixteen tissues as well as age and sex variables of 618 mice subjects. We are interested in identifying age-associated marker genes. Following the original paper, we fit the regression model below to detect age-associated genes in each tissue:

$$Y_{ijk} = \beta_{0jk} + \beta_{\text{age},jk}\text{Age}_{ijk} + \beta_{\text{sex},jk}\text{Sex}_{ijk} + \varepsilon_{ijk} \text{ for } i = 1, \ldots, m_{jk},$$

where Y_{ijk} is the expression level of the i-th subject for the j-th gene and k-th tissue. For each gene j, we consider designs of both two-sided and one-sided tests when combining p-values across tissues. In two-sided test design, two-sided p-values $(p_{j1}, \cdots, p_{jK})$ for their corresponding $\beta_{\text{age},jk}$ coefficients are combined using Fisher, AFp, and FE methods. In this case, the association directions (positive or negative associations) are not considered. In contrast, one-sided test design combines left-tailed p-values $(\tilde{p}^L_{j1}, \cdots, \tilde{p}^L_{jK})$ or right-tailed p-values $(\tilde{p}^R_{j1}, \cdots, \tilde{p}^R_{jK})$ respectively using FE_{CS}. Figure 2.4 shows the general workflows of transcriptomic meta-analysis for the j-th gene with two-sided or one-sided designs. Compared to FE, FE_{CS} is expected to have increased power to detect age-related biomarkers with concordant signals (all positive associated or all negative associated) across tissues while have reduced power for markers with heterogeneous signals (i.e., positive associations in some tissues and negative associations in some others). In this application, both concordant and heterogeneous age-related biomarkers are of interest. Heterogeneous

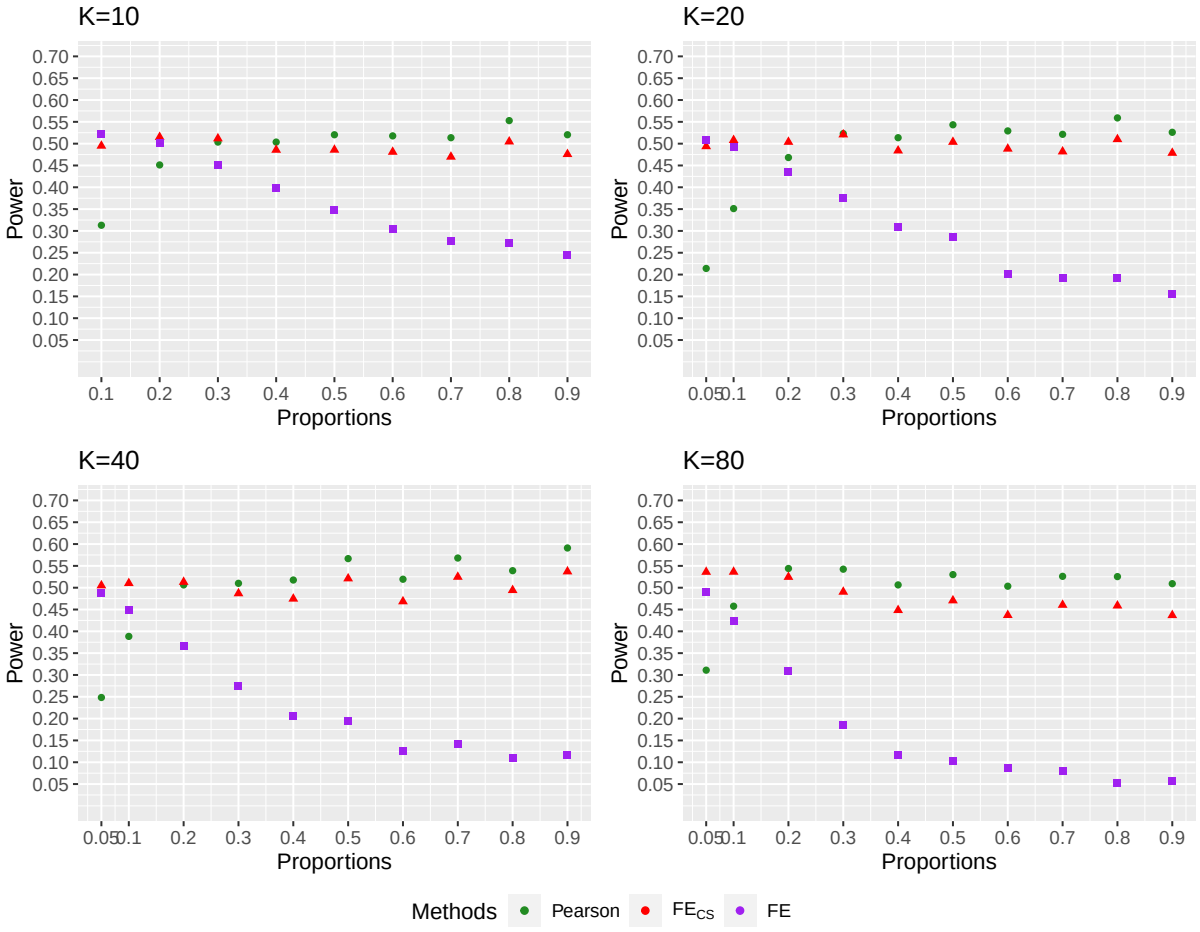


Figure 2.3: Statistical power of FE, FE_{CS}, and Pearson at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. The standard errors are negligible and hence omitted.

biomarkers detected by FE can have varying age-association (positive, negative or non-association) across tissues while concordant biomarkers detected by FE_{CS} are tissue-invariant. FE and FE_{CS} will serve as complementary tools for different biological objectives.

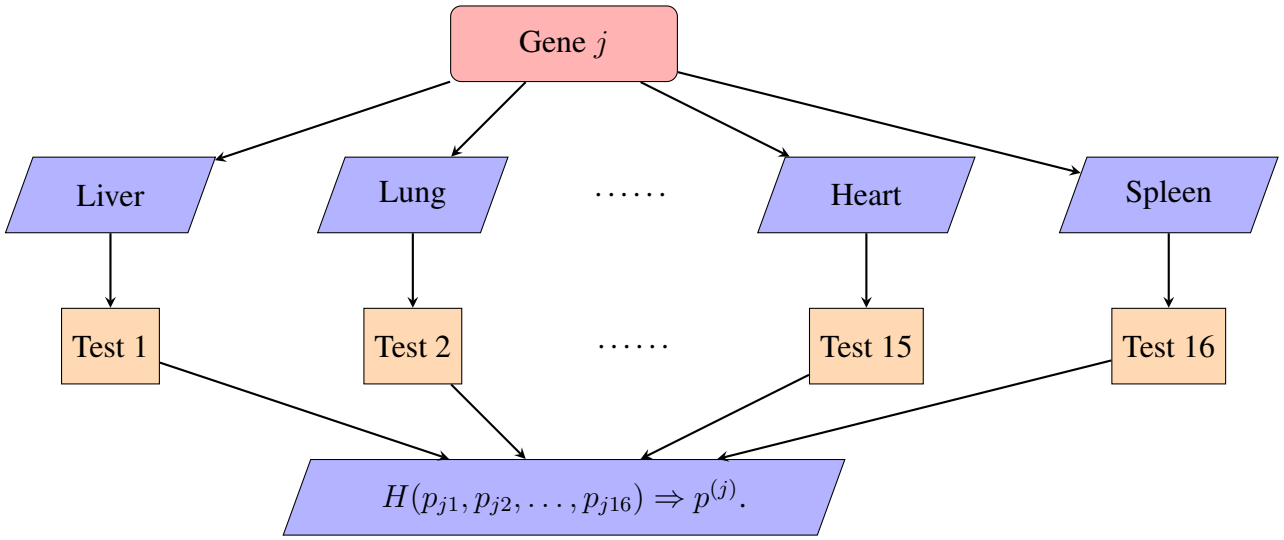
Figure 2.5(a) shows Fisher, AFp, and FE p-value combination results in the two-sided test design. Under q-value≤ 0.05, Fisher detects 576 genes (yellow color) and AFp detects 473 genes (green color), where Category II (392 genes) represents overlapped detected genes by Fisher and AFp and Categories I (184 genes) and III (81 genes) represent biomarkers uniquely detected by Fisher or by AFp. The heatmap shows age-association measure defined as: $E_{jk} = -\text{sign}(\beta_{\text{age},jk}) \log(\min\{\tilde{p}^L_{jk}, \tilde{p}^R_{jk}\})$ for gene j on the rows and tissue k on the columns; i.e., the signed $\log$-transformed (base 10) one-sided p-values. Consequently, red color of E_{jk} represents a strong positive association with age while blue means a strong negative association. As expected, FE combines the strengths of Fisher and AFs to detect 593 genes (purple color) that contain all genes in Category II and most genes in Categories I and III. By counting the number of tissues with p-values $p_{jk} \leq 0.05$, Supplement Figure A10 shows that category I genes (detected by Fisher but not by AFp) are age-associated in more tissues, while Category III (detected by AFp but not by Fisher) are age-associated in fewer tissues, which is consistent to the theoretical insight and simulation result that Fisher is more powerful for detecting frequent signals and AFp is more powerful for relatively less-frequent signals.

We next perform hierarchical clustering (using 1-correlation between tissues as dissimilarity measure and complete linkage) for the 16 tissues based on the E_{jk} values in the 593 age-related genes detected by FE, and the dendrogram is shown in Figure 2.5(a). By cutting the dendrogram, 5 clear tissue modules of similar age-association patterns are identified: (1) thymus and gonads; (2) spleen and lung; (3) eye, kidney, and heart (4) hippocampus, adrenal glands, and muscle; (5) cerebrum and spinal cord (also see Figure 2.5(b) for heatmap of pair-wise correlations). For the first module, the thymus has long been regarded as an endocrine organ that is closely related to Gonads and sexual physiology, such as sexual maturity and reproduction. (Grossman, 1985; Leposavić and Pilipović, 2018). The spleen-lung module is consistent with the finding in Zahn et al. (2007), and many reports suggest that spleen and lung share a similar aging pattern (e.g., Schumacher et al. (2008)). For the third module, literature shows that kidney and eye share structural, developmental, physiological, and pathogenic similarities and pathways. The relationships between age-related

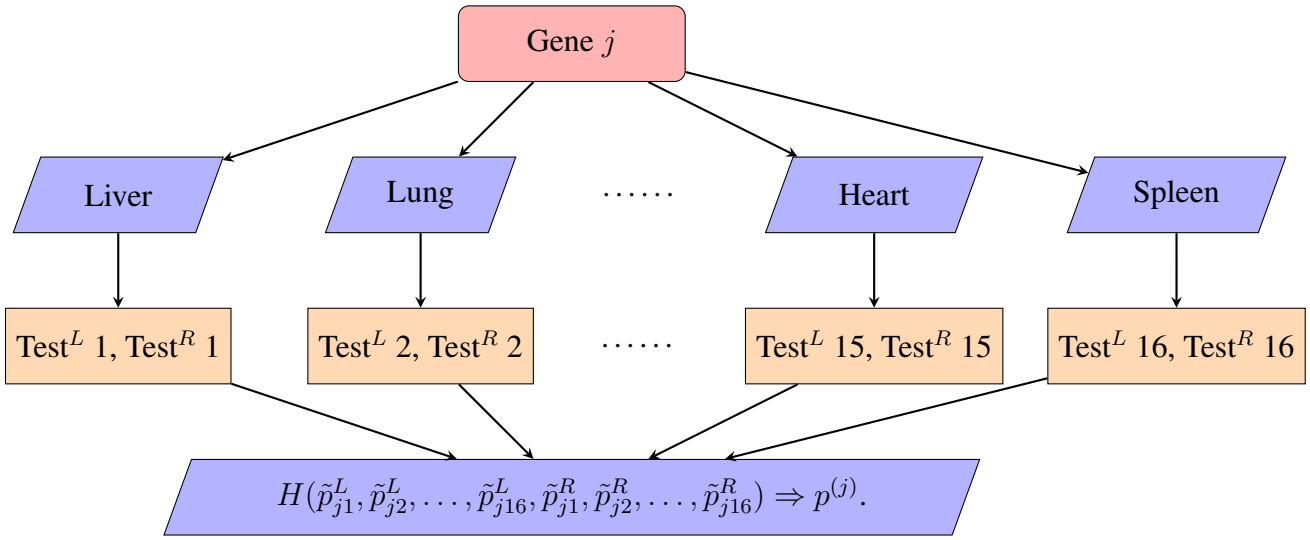
eye, kidney, and cardiovascular diseases have been widely reported (e.g., Farrah et al. (2020)). For the fourth module, extensive literature have reported the relationship between adrenal glands and hippocampal aging (e.g., Landfield et al. (1978)). For the last module, few existing studies have investigated the aging process of the spinal cord (Knight and Nigam, 2017). But it is reasonable that the cerebrum and spinal cord might share a similar aging pattern as they both belong to the central nervous system. On the other hand, the liver has intriguingly negative correlations of aging effects with muscle, adrenal glands, and several brain regions, such as the hippocampus, cerebellum, and cerebrum (also see Figure 2.5(b)).

Next, we evaluate FE_{CS} for one-sided test design and compare it with FE. We calculate $S_{\text{sign},j} = \sum_{k=1}^{16} \text{sign}(\beta_{\text{age},jk}) \mathbf{I}_{\{\min\{\tilde{p}_{jk}^L, \tilde{p}_{jk}^R\} \leq 0.05\}}$ to determine whether the detected concordant aging marker j is positively associated ($S_{\text{sign},j} > 0$) or negatively associated ($S_{\text{sign},j} \leq 0$) and use it to determine whether a detected marker is dominant with the positive association or negative association. Similar to the previous analysis, Figure 2.6 shows age-associated genes detected by FE (593 genes, Categories II(A), II(B) and III) and FE_{CS} (398 genes, Categories I(A), I(B), II(A) and II(B)), where Categories II(A) and II(B) are overlapped genes detected by FE and FE_{CS}, Category III are only detected by FE and Categories I(A) and I(B) are only detected by FE_{CS}. For genes detected by FE_{CS}, Categories I(A) and II(A) are concordant aging markers with positive association (mostly red) and Categories I(B) and II(B) are negatively associated (mostly blue), which are visually consistent with the heatmap. In contrast, genes in Category III mostly have discordant association directions (partial red and partial blue). Supplement Figure A11 shows the distributions of $S_{\text{sign},j}$ in the gene categories.

At significance level $q \leq 0.05$, FE_{CS} identifies 184 positively associated genes (Categories I(A) and II(A)) and 214 negatively associated genes (Categories I(B) and II(B)). We perform Ingenuity Pathway Analysis (IPA) to these two concordant age-associated gene lists. The result identifies 11 enriched pathways from the 184 positively associated genes and 4 enriched pathways from the 214 negatively associated genes (enrichment $p \leq 0.01$). Table A2 shows details of these enriched pathways with pathway names, enrichment p-values, and abundant supporting literature of the pathways related to aging/early development processes (see complete references in Supplement References II). The result shows the advantage of FE_{CS} to identify age-associated markers concordant across tissues and to deliver interpretable biological insights.

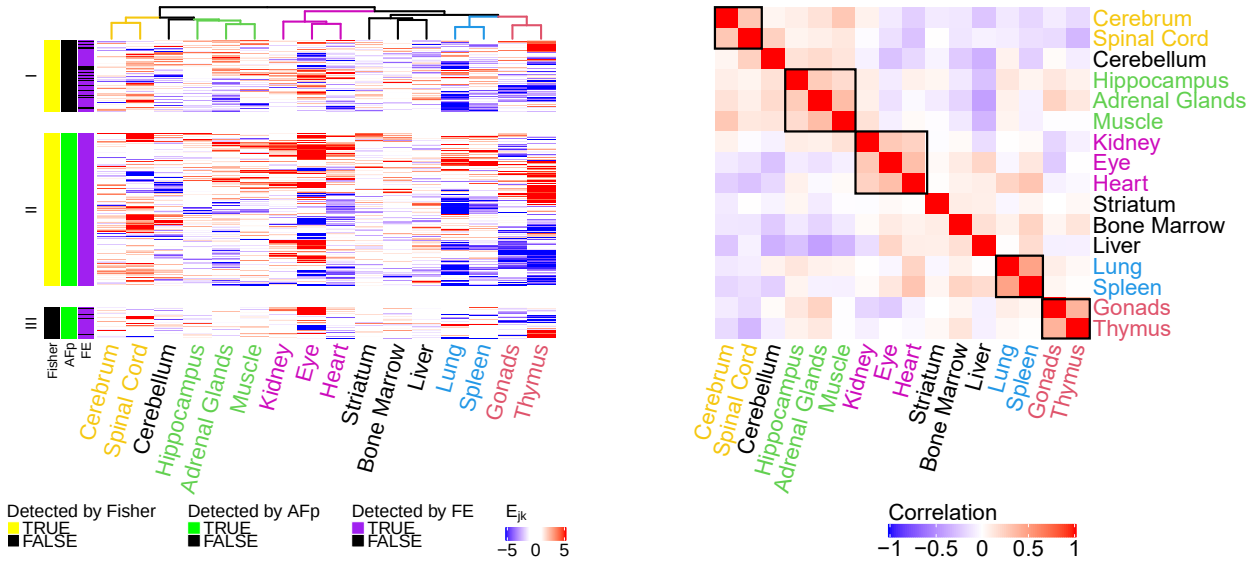


(a)



(b)

Figure 2.4: Procedures of transcriptomic meta-analysis on AGEMAP dataset (two-sided design (Figure 2.4(a)) and one-sided design (Figure 2.4(b)), where $H(\cdot)$ denotes a chosen p-value combination method and $p^{(j)}$ denotes the corresponding p-value of H with input p-values. Here p_{jk} is the two-sided p-value for j-th gene on k-th tissue, and $\tilde{p}^L_{jk}$ and $\tilde{p}^R_{jk}$ are the left-tailed and right-tailed p-values for j-th gene on k-th tissue, respectively.



(a) (b)

Figure 2.5: (a) Heatmaps of age-association measure E_{jk} of significant genes ($q \leq 0.05$) detected in the two-sided test design. Category I: genes detected by Fisher but not AFp; II: genes detected by both Fisher and AFp; III: genes detected by AFp but not Fisher. (b) Heatmap of pair-wise correlations between tissues based on the detected genes by FE ($q \leq 0.05$.) in (a).

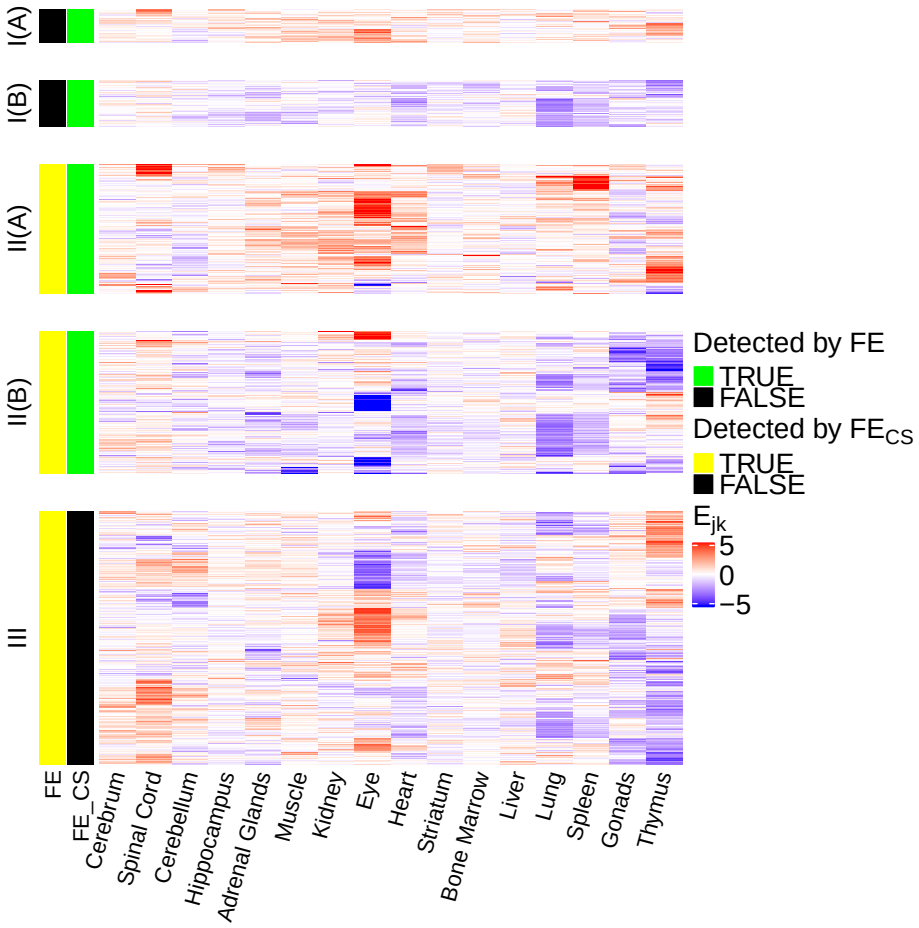


Figure 2.6: Heatmaps of age-association measure E_{jk} of genes detected by FE_{CS} or by FE ($q <= 0.05$). Heatmap I(A) represents up-regulated genes detected only by FE_{CS} (38 genes); heatmap I(B) represents down-regulated genes detected only by FE_{CS} (53 genes); heatmap II(A) represents up-regulated genes detected both by FE_{CS} and FE (146 genes); heatmap II(B) represents down-regulated genes detected both by FE_{CS} and FE (161 genes); heatmap III represents genes detected only by FE (286 genes), respectively.

2.7 Conclusion and Discussion

P-Value combination is a common and effective information synthetic tool in many scientific applications. In this chapter, we focus on a scenario of "meta-analysis with unknown heterogeneity", in which the number of combined p-values K is finite and fixed while the sample size for generating each p-value can increase to infinity (i.e., the first category described in the Introduction Section). The goal of this category is to aggregate heterogeneous independent signals, where the proportion of true signals is unknown and can range from $1/K$ to 1. We emphasize the goal of this chapter to combine independent and "non-spare" signal and distinguish it from combining "sparse" signals in the "asymptotic rare and weak (ARW)" model when $K \to \infty$, which is commonly considered in the second and third categories described in the Introduction Section.

Our contribution is three-fold. Firstly, this is the first study to comprehensively evaluate p-value combination methods for their asymptotic efficiencies in terms of asymptotic Bahadur optimality (ABO). We investigate classical methods (Fisher and Stouffer) and modified Fisher's methods (AFs, AFp, AFz, TFhard, and TFsoft). The result shows that Fisher, AFs, AFp, TFhard, and TFsoft are ABO, but Stouffer and AFz are not. We also find interesting consistency properties for the estimation of signal contributing subset in AFs and AFp (Theorems 2.4 and 2.5). Secondly, we perform an extensive finite-sample power comparison and conclude that Fisher and AFp are the 2 top performers with complementary advantages, where Fisher is more powerful with frequent signals and AFp is more powerful in relatively sparse settings. Thirdly, we propose a Fisher ensemble (FE) method to combine Fisher and AFp. A one-sided test modification, FE_{CS}, is further developed for detecting concordant signals. The advantages of FE and FE_{CS} includes: (A) Both methods have high asymptotic efficiencies (FE is ABO). (B) The harmonic mean combination avoids the $-\infty$ score in the Cauchy. (C) We numerically demonstrate their constantly high performance across varying proportions of signals. (D) Both methods have fast-computing procedures. Finally, an application to AGEMAP transcriptomic data verifies theoretical conclusions, demonstrates superior performance of FE and FE_{CS}, and discovers intriguing biological findings in age-associated biomarkers and pathways.

Modern data science faces challenges from data heterogeneity, increasingly complex data structure, and the need for effective methods for new scientific hypotheses. The ensemble methods

proposed in this chapter, FE and FE_{CS}, have solid theoretical and numerical support for their superior performance in a wide range of signal settings. We believe the methods will find impactful applications in many other scientific problems.

3.0 Adaptive Fisher's Method using Weakly Geometric Grid for Combining P-Values

3.1 Introduction

Combining p-values to aggregate information is of long-lasting interest in meta-analysis. The strategy is to construct a global test using a group of input p-values $p_1, \ldots, p_n$ for detecting the existence of signals. Classical approaches include Fisher's method (Fisher, 1992): $T_{\text{Fisher}} = \sum_{i=1}^{n} -2 \log p_i$, Edgington's method (Edgington, 1972): $T_{\text{Edgington}} = \sum_{i=1}^{n} p_i$, and Stouffer's method (Stouffer et al., 1949): $T_{\text{Stouffer}} = \sum_{i=1}^{n} \Phi^{-1}(p_i)$, where Φ is the CDF of standard normal distribution. Including the above methods, most conventional approaches can be formulated as the sum of certain transformed p-values (see, e.g., Heard and Rubin-Delanchy (2018) for further details). Among these approaches, Fisher's method is asymptotically optimal in terms of Bahadur relative efficiency for detecting frequent signals within a small number of p-values (Littell and Folks (1971, 1973)), illustrating the theoretical superiority of $\log$-transformation of p-value.

However, modern big data analysis promotes the need for detecting sparse signals within a large collection of p-values. One motivating example is to detect a small fraction of signals by combining p-values of a large number of SNPs within a SNP-set (e.g., hundreds to thousands of SNPs in a gene region or in gene regions of a pathway) in the genome-wide association studies (GWAS) (Su et al., 2016; Hoh et al., 2001). The large-scale data analysis introduces new scenarios where Fisher's method is suboptimal. Indeed, Donoho and Jin (2004) showed that under a scenario of sparse signals in the two-component Gaussian mixture model (approximately $n^{1-\beta}$ out of n p-values represent signals, for $1/2 < \beta < 1$), Fisher's method suffers from substantial power loss. By contrast, minP (Tippett et al. (1931)), an approach that simply uses the minimum p-value as the test statistic, is powerful for detecting extremely sparse signals ($3/4 < \beta < 1$), while less powerful for moderate sparse signals ($1/2 < \beta \leq 3/4$).

Intuitively, Fisher's method tends to incorporate too many noises, while only using a single minimum p-value as test statistic leads to potentially substantial information loss. One natural idea is to modify Fisher's method with a strategy to filter out noises, resulting in methods that are tailored to the scenarios of sparse signals but still enjoy the advantages of $\log$-transformation.

Along this line of research, Zaykin et al. (2002) and Zaykin et al. (2007) proposed the truncated product method (TPM), in which only the p-values below a given threshold τ is taken into account: $T_{\text{TPM}}(\tau) = \sum_{i=1}^{n} -2\log(p_i)\,\mathrm{I}_{\{p_i \le \tau\}}$, where $\mathrm{I}_{\{\cdot\}}$ is the indicator function. Replacing τ by $p_{(\ell)}$ for a given ℓ, Dudbridge and Koeleman (2003); Kuo and Zaykin (2011) proposed the rank truncated product method (RTP), where only the first ℓ-th smallest p-values are selected to form the test statistics.

The choice of truncation point for RTP and TPM is user-specified and arbitrary, while misspecification of the truncation point may lead to substantial power loss. Several adaptive procedures are proposed to address the problem. For example, Yu et al. (2009) suggested that one can separately construct multiple RTP statistics using a given collection of candidate truncation points, and then use the minimum of p-values derived from the RTP statistics. However, the choice of candidate truncation points set is still arbitrary. Li and Tseng (2011) proposed to calculate the RTP statistics for all possible truncation points $1 \le i \le n$, and choose the one leading to the minimum upper tail probability of chi-squared distributions (denoted by AFp hereafter), $T_{\text{AFp}} = \min_{1\le i\le n} \mathbb{P}(\chi^2_{2i} \ge \sum_{j=1}^{i} -2\log p_{(j)})$. However, since the tail of the null distribution of RTP for each i is much heavier than chi-squared distribution, AFp tends to choose much more than the desired number of p-values to combine. Song et al. (2016) and Heard (2021) suggested to standardize the n RTP statistics and choose the maximum z-score (denoted by AFz hereafter), $T_{AFz} = \max_{1\le i\le n} |(\sum_{j=1}^{i} - \log p_{(j)} - \sum_{j=1}^{n} w(j,i))/\sqrt{\sum_{j=1}^{n} w^2(j,i)}|$, where $w(j,i) = \min\{1, i/j\}$. This approach searches among all the possible truncation points, leading to a potential power loss, especially for the moderate sparse case. In addition, there is no theoretical justification available for AFz. Zhang et al. (2020b) investigated the theoretically optimal choice of the truncation point of TPM under a setup of two-component Gaussian mixtures with known means and variances. They further proposed an omnibus test that aggregates multiple TPM statistics for choosing the truncation points in the real practice (denoted by oTFhard hereafter). One variant of oTFhard using a soft-thresholding scheme (denoted by oTFsoft hereafter) is also proposed to improve finite-sample performance of the original omnibus test. However, their theoretical setup is formulated with the proportion of signals and signal strength (i.e., normal mean) unchanged as n diverges. Hence, a theoretical analysis using classical asymptotic efficiency theory can be applied. Such a setup is not a large-scale setting that characterizes heterogeneous and sparse signals we

consider in this chapter. Furthermore, their omnibus test relies on grid search on a prespecified truncation point set, and there is no theoretical guarantee of its performance.

The primary goal of our chapter is to develop a fully data-driven procedure adopted from Fisher's method that is tailored to the scenario of sparse signals. The development of our adaptive procedure starts from finding a suitable searching strategy of the truncation point for RTP, given that there is n candidate truncation points in total. By contrast, the choices of the truncation points of TPM can be uncountable many from the $(0, 1)$ interval. To inspire the development of our adaptive procedure, we conduct theoretical analysis on RTP and identify the optimal choice of the truncation point under the Gaussian sequence model setup. The setup is commonly used for characterizing the pattern of sparse and heterogeneous signals. Under the guidance of the theoretical results of RTP, we propose our fully adaptive procedure with a searching strategy based on the weakly geometric system. The weakly geometric system significantly reduces the number of candidate truncation points from n to a value of order $(\log n)^2$, leading to a lighter computational burden and less power loss than a "searching-all" strategy. Under the same setup, we show our adaptive procedure achieves almost the same rate-optimal theoretical performance as RTP with the oracle choice of truncation point. To the best of our knowledge, our method is the first truly adaptive procedure with theoretical guarantees under a large-scale setting of sparse signals among all modified Fisher's methods. Furthermore, we also note that in practice, people tend to assume the p-values to be exact when using the p-value combination method, although most of the p-values are derived via some approximations of distributions like the central limit theorem. Noting that there is little discussion on the impact of the gap, we investigate the robustness properties of our adaptive method using Studentization-based p-values, given the widely use of original or modified Studentized statistics in modern statistics, such as statistical genomics and genetics (Tusher et al., 2001; Smyth, 2004; Love et al., 2014; Marees et al., 2018; Svishcheva et al., 2019). Under mild moment conditions on the noises, we show that our adaptive method can still optimally distinguish between the null and alternative hypotheses in the testable region when using Studentization-based p-values. At the same time, simulations in Section 3.5.2 verify our theoretical insights.

The chapter is structured as follows. We first introduce our adaptive procedure in Section 3.2. Section 3.3 provides the theoretical justification of our method under a large-scale setup with sparse signals. Section 3.4 studies the robustness of our method for using Studentization-based p-values.

Section 3.5 contains extensive simulations to evaluate the performance of our method compared to other methods, confirming our theoretical results. A GWAS application of neuroticism is provided in Section 3.6 to compare the performance of different methods and demonstrate the advantages of our method. Section 3.7 contains final conclusion and discussions.

3.2 The Adaptive Testing Procedure

In this section, we introduce our adaptive testing procedure. Consider n p-values $p_1, \dots p_n$ derived from n independent hypothesis tests. Let s be the largest possible number of p-values that represent true signals such that $s \ll n$. As discussed in Section 3.1, Fisher's method fails in this setting with a large number of noises since it combines all the p-values and the noises impose substantial impact on the test statistic. Hence a truncation strategy on the p-values is preferred to eliminate the impact of noises. One way is to use the RTP:

$$T(\ell) = \sum_{i=1}^{\ell} -2 \log p_{(i)},$$

where ℓ is a prespecified truncation point on the rank of p-values. Intuitively, a choice of ℓ that is much larger than s can introduce too many noises into $T(\ell)$. At the same time, a choice of ℓ that is much smaller than s (e.g., 1) can lead to a potential loss of information. With this in mind, a natural choice of ℓ would be s, which is justified in Section 3.3.

However, it is rare to have prior knowledge of s in real practice. To address this problem, we propose our adaptive procedure, namely adaptive Fisher based on weakly geometric system (AFg). We divide our adaptive procedure in the following 3 steps:

Step 1. Denote by $\lceil \cdot \rceil$ the ceiling operator. Generate the candidate set $\mathcal{S}$ of s: $s_0 = 1$, $s_1 = \lceil \log n \rceil$, $s_2 = \lceil \log n \, (1 + 1/\log n) \rceil$, $s_3 = \lceil \log n \, (1 + 1/\log n)^2 \rceil, \dots, s_M = \lceil \log n \, (1 + 1/\log n)^{M-1} \rceil$, where M is the smallest integer such that $\log n \, (1 + 1/\log n)^{M-1} \geq \sqrt{n}/\log n$.

The novel weakly geometric system was rooted and mostly applied in nonparametric statistics to achieve sharp adaptation, see, e.g., Goldenshluger and Tsybakov (2001); Cavalier and Tsybakov

(2002); Tsybakov (2008). In this chapter, we assume $s = O(n^{1-\beta})$ for $1/2 < \beta < 1$ with the fraction of true signals small but not vanishingly small, a subtle scenario considered in Donoho and Jin (2004). To this end, it suffices to only consider candidate truncation points that are smaller than $\sqrt{n}/\log n$. To enhance the finite-sample performance of our method for detecting extremely sparse signals, we incorporate $s_0 = 1$ into $\mathcal{S}$ as minP is powerful in this case. The intuition of the weakly geometric system is that for a sufficiently large n, the ratio $(1 + 1/\log n)$ between any two s_i and s_{i+1} is very close to 1. Hence s is not too far away from at least one of the candidate truncation points as there always exists some i such that $s_i \leq s \leq s_{i+1}$. By our analysis in Section 3.3, the proposed design leads to almost the same performance as $T(s)$ under the same conditions imposed on the alternative, indicating this design introduces almost no extra cost into our adaptive procedure. Compared to the strategy by searching all the n possible points $1 \leq \ell \leq n$, the weakly geometric design significantly reduces the number of candidate truncation points to a value of order $O((\log n)^2/2)$. Hence the new design can lead to a lighter computational burden and a potentially more powerful test. We also try the geometric design in which all the ratios s_{i+1}/s_i equal to some constant $\kappa > 1$, and find its finite-sample power is generally worse than the weakly geometric design.

Step 2. Calculate the $M + 1$ RTP test statistics based on $\mathcal{S}$:

$$T(s_i) = \sum_{j=1}^{s_i} -2 \log p_{(j)}, \quad i = 0, 1, ..., M$$

Let $U(s_i)$ be the random variable following the same distribution of $T(s_i)$ under the null, that is, $p_1, \ldots, p_n \overset{\text{i.i.d.}}{\sim} \text{Unif}(0, 1)$, where $\text{Unif}(0, 1)$ denotes the uniform distribution. For $s_0 = 1$, $U(s_0)$ is a monotonic transformation of $p_{(1)}$, where $p_{(1)}$ follows a $\text{Beta}(1, n)$ distribution with shape parameters 1 and n. By Proposition B1 in Section B.2, for each $k = 1, \ldots, M$, $U(s_i)$ can be written as the following sum of independent random variables:

$$U(s_i) = \chi^2_{2s_i} + \sum_{i=1}^{n-s_i} U_i,$$

where $\chi^2_{2s_i}$ denotes the chi-squared random variable with degrees of freedom $2s_i$, and $U_i \sim \text{GAM}(2s_i/(n - i + 1), 1)$. Here $\text{GAM}(a, b)$ denotes the gamma distribution with shape and rate

parameters a and b, respectively.

Step 3. For each $k = 0, \ldots, M$, derive the p-value of $T(s_i)$, i.e., $\mathbb{P}(U(s_i) > T(s_i))$. We consider the minimum of the derived p-values as our new test statistic T_{AFg}:

$$T_{\text{AFg}} = \min_{s_i \in \mathcal{S}} \mathbb{P}\left(U(s_i) > T(s_i)\right). \tag{3.1}$$

Calculate the critical value C_α of T_{AFg} at a given significance level α using a selected sampling-based method. Reject the null if $T_{\text{AFg}} < C_\alpha$.

Remark 3.1. For a commonly used significance level α (e.g., $\alpha = 0.05$ or 0.01), one can derive the critical value C_α by Monte-Carlo samples of T_{AFg} under the null. However, in some real applications like GWAS, p-value of T_{AFg} is often more desired. Under such scenarios, generating p-values of T_{AFg} by Monte-Carlo samples is often computationally infeasible as the desired p-values can be extremely small. For the scenarios where extremely small p-values are desired, we develop an efficient sampling-based method to calculate the p-value of T_{AFg}. See Section B.3 for the details of the computational method.

Remark 3.2. T_{AFg} is essentially an omnibus test. It is common to use omnibus tests for constructing adaptive testing procedure (e.g., Shah and Bühlmann (2018); Janková et al. (2020); Zhang et al. (2020b)).

We summarize the three steps of our adaptive procedure:

Algorithm 3.1. Pseudo algorithm for AFg.

Input n, $\vec{p} = (p_1, \ldots, p_n)$, α

Set M the smallest integer such that $\log n \left(1 + 1/\log n\right)^{M-1} \geq \sqrt{n}/\log n$

Set $\mathcal{S} = \{s_0, \ldots, s_M\}$, where $s_0 = 1$, $s_1 = \lceil \log n \rceil$, $s_2 = \lceil \log n \left(1 + 1/\log n\right) \rceil$,
$s_3 = \lceil \log n \left(1 + 1/\log n\right)^2 \rceil, \ldots, s_M = \lceil \log n \left(1 + 1/\log n\right)^{M-1} \rceil$

For $i = 0$ to $i = M$

$T(s_i) = \sum_{j=1}^{s_i} -2 \log p_{(j)}$

$p_{T(s_i)} = \mathbb{P}\left(U(s_i) > T(s_i)\right)$

Set $T_{\text{AFg}} = \min_{s_i \in \mathcal{S}} p_{T(s_i)}$

Calculate critical value C_α for T_{AFg} by a selected sampling-based method

Output $\mathrm{I}_{\{T_{\mathrm{AFg}}<C_\alpha\}}$

Remark 3.3. Compared to the adaptive procedure proposed by Yu et al. (2009) and Zhang et al. (2020b), Algorithm 3.1 is a tuning-free adaptive procedure.

As Littell and Folks (1973) have shown that Fisher's method is optimal for combining frequent signals in terms of Bahadur efficiency, we slightly modify Algorithm 3.1 to incorporate Fisher for better finite-sample performance when signals are frequent or moderate sparse (e.g., when β closes to 0.5 under Setup 3.1 in Section 3.3):

Algorithm 3.2. Pseudo algorithm for AFg.

Input n, $\vec{p} = (p_1, \ldots, p_n)$, α

Set M the smallest integer such that $\log n(1 + 1/\log n)^{M-1} \geq \sqrt{n}/\log n$

Set $\mathcal{S} = \{s_0, \ldots, s_{M+1}\}$, where $s_0 = 1$, $s_1 = \lceil \log n \rceil$, $s_2 = \lceil \log n\,(1 + 1/\log n) \rceil$,

$s_3 = \lceil \log n\,(1 + 1/\log n)^2 \rceil, \ldots, s_M = \lceil \log n(1 + 1/\log n)^{M-1} \rceil, s_{M+1} = n$

For $i = 0$ to $i = M + 1$

$T(s_i) = \sum_{j=1}^{s_i} -2 \log p_{(j)}$

$p_{T(s_i)} = \mathbb{P}\left(U(s_i) > T(s_i)\right)$

Set $T_{\mathrm{AFg}} = \min_{s_i \in \mathcal{S}} p_{T(s_i)}$

Calculate critical value C_α for T_{AFg} by some sampling-based method

Output $\mathrm{I}_{\{T_{\mathrm{AFg}}<C_\alpha\}}$

Compared to Algorithm 3.1, we only add one more point $s_{M+1} = n$ into the candidate truncation point set $\mathcal{S}$ in Algorithm 3.2. In Sections 3.5 and 3.6, AFg stands for Algorithm 3.2 unless further notice.

3.3 Theoretical Justification of T(s) and AFg

In this section, we provide theoretical guarantees on the performance of AFg from a minimax point of view. Our goal is to show that AFg is able to optimally distinguish between the null and alternative hypotheses in the testable region. Step 3 in Algorithm 3.1 is for the real practice where a given significance level α is used to reject or not reject the null hypothesis. We note that the

critical value C_α for each α in step 3 corresponds to a set of $M+1$ critical values $C_{i,\alpha}$ so that we reject the null once any test rejects the null, i.e., $T(s_i) > C_{i,\alpha}$. To illustrate the good theoretical properties of our proposed method, we consider the following step $3'$, which is a modified version of step 3. With a larger choice of critical value $C_i = 2s_i(1+2\delta_n)\log(n/s_i)$ compared to $C_{i,\alpha}$ for each i and any fixed α, we show that the proposed test is still powerful in the testable region with type I error tending to 0.

Step $3'$. Conduct $M+1$ tests using all the $T(s_i)$'s with critical values $C_i = 2s_i(1+2\delta_n)\log(n/s_i)$ $(i = 0, 1, .., M)$ with $\delta_n = 1/\sqrt{\log\log n}$. Reject the null once any test rejects the null, i.e., $T(s_i) > C_i$.

For the whole following Section 3.3, AFg stands for the procedure based on Algorithm 3.1 using step $3'$ unless further notice. We firstly present Theorems 3.1 and 3.2, which respectively show the type I and type II errors of $T(s)$ and AFg simultaneously tend to zero under the same mild conditions imposed on the alternative. To further justify the performance of AFg, we summarize the conditions in Theorem 3.1 and 3.2, and formulate an alternative parameter space that characterizes sparse and heterogeneous signals. By adopting the arguments in Collier et al. (2017), we derive the lower bound of separating rate of discriminating the alternative parameter space from the null and show both $T(s)$ and AFg are rate-optimal testing procedures (all the concepts are defined latter in this section).

Consider the following Gaussian sequence model where each Y_i's is a z-score test statistic for the i-th hypothesis test:

Model 3.1.

$$Y_i = \mu_i + \xi_i, \quad i = 1, \ldots, n, \tag{3.2}$$

where $\theta = (\mu_1, \ldots, \mu_n)' \in \mathbb{R}^n$ is an unknown mean vector. As one can always rescale the test statistics, we assume $\xi_1, \xi_2, \ldots, \xi_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, denoted by $\mathrm{N}(0,1)$ the standard normal distribution. The above model is commonly used to provide theoretical justifications for p-value combination methods (e.g., Li and Tseng (2011); Owen (2009)). Our goal is to globally test if there is any $\mu_i \neq 0$ using the input p-values derived by $p_i = 2(1-\Phi(|Y_i|))$ for each $i = 1, \ldots, n$.

To characterize the signal sparsity, we assume $\|\theta\|_0 \leq n^{1-\beta}$, where $\|x\|_0$ denotes the number of non-zero entries of the vector x. A larger β leads to a higher level signal sparsity. We further use $\|\theta\|_2$, the ℓ_2 norm of θ to quantify the signal strength. The following Theorem 3.1 indicates the type I and type II errors of $T(s)$ simultaneously tend to zero as long as $\|\theta\|_2^2 \geq C^{(0)} n^{1-\beta} \log n$ for some constant $C^{(0)}$. This is a weak signal strength condition since the lowest required signal strength matches the lower bound of the separating rate to discriminate the null and alternatives under our setup (see Corollary 3.1 for further details).

Theorem 3.1. *Assume $\|\theta\|_0 \leq n^{1-\beta}$ with $1/2 < \beta < 1$ and $\|\theta\|_2^2 \geq C^{(0)} \cdot n^{1-\beta} \log n$, where $C^{(0)}$ denotes any constant that is strictly greater than 2β. Consider the RTP test statistic $T(s)$ truncated on $s = \lceil n^{1-\beta} \rceil$ with critical value $C^{(n)} = 2\beta n^{1-\beta}(1 + 2/\sqrt{\log \log n}) \log n$:*

$$T(s) = \sum_{i=1}^{s} -2 \log p_{(i)}.$$

Denote the sum of type I and type II errors of $T(s)$ by

$$\mathcal{R}_{T(s)} = \mathbb{P}_0(\varphi_{T(s)} = 1) + \mathbb{P}_\theta(\varphi_{T(s)} = 0),$$

where $\varphi_{T(s)} = I_{\{T(s) > C^{(n)}\}}$. Then we have $\lim_{n\to\infty} \mathcal{R}_{T(s)} = 0$.

Remark 3.4. Besides from the perspective of separating rate demonstrated in Corollary 3.1, one can show the signal strength requirement is mild from another aspect. Indeed, Proposition B1 in the Supplement Section B.2 shows that under the null

$$\mathbb{E}(T(s)) = \sum_{i=1}^{n} 2 \min\{1, s/i\} = O(s \log n),$$

which is the same order of the lower bound of $\|\theta\|_2^2$ in Theorem 3.1.

Under the same setup and conditions imposed on $\|\theta\|_2$, the following Theorem 3.2 shows the type I and type II errors of AFg also tend to zero simultaneously, indicating there is almost no extra cost for introducing the adaptive procedure.

Theorem 3.2. *Under Model 3.1 and the same conditions on θ in Theorem 3.1, denote the sum of type I and type II errors of AFg by*

$$\mathcal{R}_{T_{AFg}} = \mathbb{P}_0(\varphi_{AFg} = 1) + \mathbb{P}_\theta(\varphi_{AFg} = 0),$$

where $\varphi_{AFg} = I_{\left\{\cup_{i=0}^{M}\{T(s_i)>C_i\}\right\}}$. Then we have $\lim_{n\to\infty} \mathcal{R}_{T_{AFg}} = 0$.

To justify the performance of $T(s)$ and T_{AFg}, we further include the lower bound of separating rate for separating the null parameter space ($\theta = \vec{0}$) from the following alternative parameter space, formulated by the conditions on θ in Model 3.1.

Setup 3.1. Consider the following alternative parameter space $\Theta_{0,\beta}(L_n)$ and the null parameter space with $\theta = \vec{0}$,

$$\Theta_{0,\beta}(L_n) = \{\theta \in B_0(\beta) : \|\theta\|_2^2 \geq L_n > 0\},$$
$$\text{where } B_0(\beta) = \{\theta : \|\theta\|_0 \leq n^{1-\beta}\}, 1/2 < \beta < 1.$$

Then the hypothesis testing problem considered in Theorems 3.1 and 3.2 is reformulated as the following hypothesis test:

$$H_0 : \theta = \vec{0}$$
$$H_1 : \theta \in \Theta_{0,\beta}(L_n).$$

We consider the following minimax risk for separating the above alternative and null parameter spaces:

$$\mathcal{R}_{0,\beta}(L_n) = \inf_{\varphi}\{\mathbb{P}_0(\varphi = 1) + \sup_{\theta\in\Theta_{0,\beta}(L_n)} \mathbb{P}_\theta(\varphi = 0)\},$$

where $\varphi = \mathrm{I}_{\{T\in R(T)\}}$ and $R(T)$ denotes the rejection region of any test statistic T. $\mathcal{R}_{0,\beta}(L_n)$ is the minimum possible sum of type I and type II errors that can be achieved by a testing procedure under Setup 3.1. $\mathcal{R}_{0,\beta}(L_n)$ can also be regarded as a measure of difficulty level for the testing problem given L_n. A higher $\mathcal{R}_{0,\beta}(L_n)$ corresponds to a more difficult testing problem. Different choices of the order of L_n can lead to different difficulty levels for the testing problem, i.e., different levels of

$\mathcal{R}_{0,\beta}(L_n)$. And there is certainly a subtle choice of the order of L_n, denoted by λ_n, such that,
(i) For any $\varepsilon \in (0,1)$, there exists a constant A_ε such that, for all $L_n \geq A_\varepsilon \lambda_n$,

$$\mathcal{R}_{0,\beta}(L_n) \leq \varepsilon. \tag{3.3}$$

(ii) For any $\varepsilon \in (0,1)$, there exists a constant a_ε such that, for all $L_n \leq a_\varepsilon \lambda_n$,

$$\mathcal{R}_{0,\beta}(L_n) \geq 1 - \varepsilon. \tag{3.4}$$

We define λ_n as the separating rate for distinguishing the alternative and the null parameter spaces. Besides, a test statistic that satisfies (i) is called a rate-optimal test.

The following Theorem 3.3 derives the lower bound of separating rate λ_n for the distinguishing $\Theta_{0,\beta}(L_n)$ against the null parameter space.

Theorem 3.3. *Under Setup 3.1, the separating rate of testing on $\Theta_{0,\beta}(L_n)$ against the null is $\lambda_n = n^{1-\beta}\log n$.*

As demonstrated in Theorems 3.1 and 3.2, the order of the minimum signal strength ($\|\theta\|_2^2 \geq C^{(0)} \cdot n^{1-\beta}\log n$) that allows the sum of type I and II errors of $T(s)$ or AFg to tend to zero is also λ_n. Hence both testing procedures are rate-optimal:

Corollary 3.1. *Under Setup 3.1 with $1/2 < \beta < 1$, $T(s)$ and AFg are rate-optimal testing procedures.*

3.4 Robustness Properties of T(s) and AFg using Studentization-Based P-Values

This section conducts theoretical analyses on the robustness of $T(s)$ and AFg using p-values derived by the normal approximation to the Student's t-statistics. In real practice, people tend to assume the derivation of the p-values is based on the exact distribution of the test statistics, despite the common applications of approximation techniques. More precisely, certain approximations of distributions (e.g., the central limit theorem) are widely used for constructing the test statistics, resulting in p-values valid only in an asymptotic sense. There is little discussion on the validity of the consequent global test due to this approximation in the context of Fisher's methods, while

its impact may not be negligible under a large-scale setting where the total number of p-values diverges and may be larger than the sample size in each test. Among all the approximation techniques, self-normalization or Studentization (Shao et al., 2013) is probably the simplest but most common one. For example, Studentization and its modified variants are widely used in genetics and genomics data analysis (e.g., Smyth (2004); Love et al. (2014); Svishcheva et al. (2019)). To better understand the behavior of our methods in a more practical sense, we investigate the impact on $T(s)$ and AFg for combining Studentization-based p-values. Consider the following model:

Model 3.2.

$$Y_{ij} = \mu_i + \xi_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, m, \tag{3.5}$$

where $\theta = (\mu_1, ..., \mu_n)' \in \mathbb{R}^n$ is the unknown mean vector, and ξ_{ij}'s are centered independent and identically distributed random variables with variance σ^2 and CDF function F such that $F(0) < 1$. Model 3.1 and Model 3.2 are equivalent if ξ_{ij}'s are standard normal. Here we consider Model 3.2 since it is a good example where people consider t-test, as there are m samples for each study and σ^2 is unknown. With some mild moment conditions imposed on ξ_{ij}'s, we adopt arguments in Delaigle et al. (2011) to show that AFg and $T(s)$ using Studentization-based p-values are robust even when ξ_{ij}'s are heavy-tailed distributed. More precisely, we show that the signal strength requirement is not changed for the sum of type I and type II errors to go to zero.

Setup 3.2. Consider the following alternative parameter space $\Theta'_{0,\beta}(L_n)$ associated with Model 3.2 versus the null parameter space with $\theta = \vec{0}$,

$$\Theta'_{0,\beta}(L_n) = \{\theta \in B_0(\beta) : \|\theta\|_2^2 \geq \sigma^2 L_n/m > 0\},$$
$$\text{where } B_0(\beta) = \{\theta : \|\theta\|_0 \leq n^{1-\beta}\},\ 1/2 < \beta < 1.$$

One major difference between $\Theta'_{0,\beta}(L_n)$ and $\Theta_{0,\beta}(L_n)$ is that the lower bound of $\|\theta\|_2^2$ is scaled by m as there are m samples in each hypothesis test. The goal is to test

$$H_0 : \theta = \vec{0}$$
$$H_1 : \theta \in \Theta'_{0,\beta}(L_n),$$

with p-values $p_i = 2(1 - \Phi(|T_i|))$ derived from the normal approximation of the Studentized statistics $T_i = \sqrt{m}\bar{Y}_i/L_i$, where $\bar{Y}_i = (1/m)\sum_{j=1}^m Y_{ij}$ and $L_i = \sqrt{(1/m)\sum_{j=1}^m (Y_{ij} - \bar{Y}_i)^2}$ ($i =$

$1, \ldots, n$). Here we assume $m^\eta = n$ with some constant $\eta > 0$. When $\eta > 1$, m goes to infinity as n diverges, but with a slower rate. Such a relationship between m and n is quite common in genomics and genetics data. Theorems 3.4 and 3.5 below investigate the performance of $T(s)$ and AFg under some mild moment conditions, respectively.

Theorem 3.4. *Under Setup 3.2 and assume* $\mathbb{E}(|\xi_{ij}|^D) \leq B_0\sigma^D$ *for* $i = 1, \ldots, n; j = 1, \ldots, m$, *where* B_0 *is some finite constant and* $D = \max\{6\eta + \varepsilon, 4\}$ *with some* $\varepsilon > 0$, *then for* $T(s)$ *we have*

$$\lim_{n\to\infty} \mathcal{R}_{T(s)} = 0,$$

as long as $\|\theta\|_2^2 = \sum \mu_i^2 \geq \sigma^2 C^{(0)} n^{1-\beta} \log n/m$, *where* $C^{(0)}$ *denotes any constant strictly greater than* 2β.

Theorem 3.5. *Under Setup 3.2 and the same conditions in Theorem 3.4, then we have* $\lim_{n\to\infty} \mathcal{R}_{T_{AFg}} = 0$, *as long as* $\|\theta\|_2^2 = \sum \mu_i^2 \geq \sigma^2 n^{1-\beta} C^{(0)} \log n/m$, *where* $C^{(0)}$ *denotes any constant strictly greater than* 2β.

Remark 3.5. The moment condition on ξ_{ij} in Theorems 3.4 and 3.5 is slightly more stringent than the condition in Theorem 3 of Delaigle et al. (2011) for the higher criticism test. This is due to the fact that one has to bound the the sum of deviations of several p-values for Fisher-type test statistics such as $T(s)$ and AFg, while for higher criticism, only the deviation of each individual p-value needs to be controlled at a certain level.

The above two theorems show that under some mild conditions both $T(s)$ and AFg maintain the same theoretical performance for separating the alternative and the null parameter space. In addition, these mild conditions can be carefully characterized by the relationship between moment conditions of ξ_{ij}'s and the sample size requirement in each hypothesis test. This surprising result is partially due to the good theoretical property of Studentization with a sharp moderate deviation bound. Indeed, as Shao et al. (2013) and Delaigle et al. (2011) suggested, studentized statistics can approximate the normal distributed random variables well under some rather mild moments assumptions. On the other hand, the weakly geometric system in AFg only introduces a small candidate set of truncation points, leading to little extra cost for the adaptive procedure.

3.5 Simulations

In this section, we perform simulations to evaluate the finite-sample performance of AFg. Section 3.5.1 studies the statistical power of AFg and other six methods across varying sparsity and signal strength levels. Section 3.5.2 evaluates robustness of AFg under the violation of normality assumption and verifies the theoretical results in Section 3.4.

3.5.1 Power Comparison

In this subsection, we perform power comparisons across varying sparsity levels and signal strengths. We include methods discussed in Sections 3.1 and 3.2, including our method (AFg), minP, AFz, AFp, oTFhard, and oTFsoft. In addition, we include the higher criticism test (denoted by HC hereafter) proposed by Donoho and Jin (2004), which is designed for a slightly different setting. The major difference is that we allow arbitrary alternatives, while the setup of HC imposes a mixture of the null and a single distribution representing the alternative. To make a fair and comprehensive comparison, we consider an even more subtle scenario from Donoho and Jin (2004), where the signal strength $\|\theta\|_2^2$ may be below the requirement in Theorems 3.1 and 3.2. This is a setting that supposedly favors HC.

We simulate $n = 1000$, $X = (X_1, ..., X_n)' \sim \mathrm{N}_n(\mu, I_{n\times n})$, where $\mu = (\mu_1, \dots \mu_n)'$, $\mu_1, \dots \mu_s \overset{\text{i.i.d.}}{\sim} \mathrm{N}(\sqrt{2(r+\Delta)\log n}, \sigma^2)$ with $s = \lceil n^{1-\beta} \rceil$ ($\beta = 0.55, \dots, 0.9$), and $\mu_{s+1}, \dots, \mu_n$ equal 0. We further set r to be:

$$r = \rho^*(\beta) = \begin{cases} \beta - \frac{1}{2}, & \frac{1}{2} < \beta \leq \frac{3}{4} \\ (1 - \sqrt{1-\beta})^2, & \frac{3}{4} < \beta < 1. \end{cases} \tag{3.6}$$

For different levels of signal strength, we set $\Delta = 0.05$, 0.1, 0.2. p-values are calculated through two-sided z-score test $p_i = 2(1 - \Phi(|X_i|))$ for $i = 1, \dots, n$.

To calculate the empirical power, we first draw 10^5 Monte-Carlo samples to calculate the critical values for the seven methods at significance level 0.05. We then perform 10^4 simulations to calculate the empirical power of each method for each combination of simulation parameters. Each simulation setting is repeated 30 times to calculate the mean empirical power and corresponding standard error for each method. We note that potentially one can get even better performance of

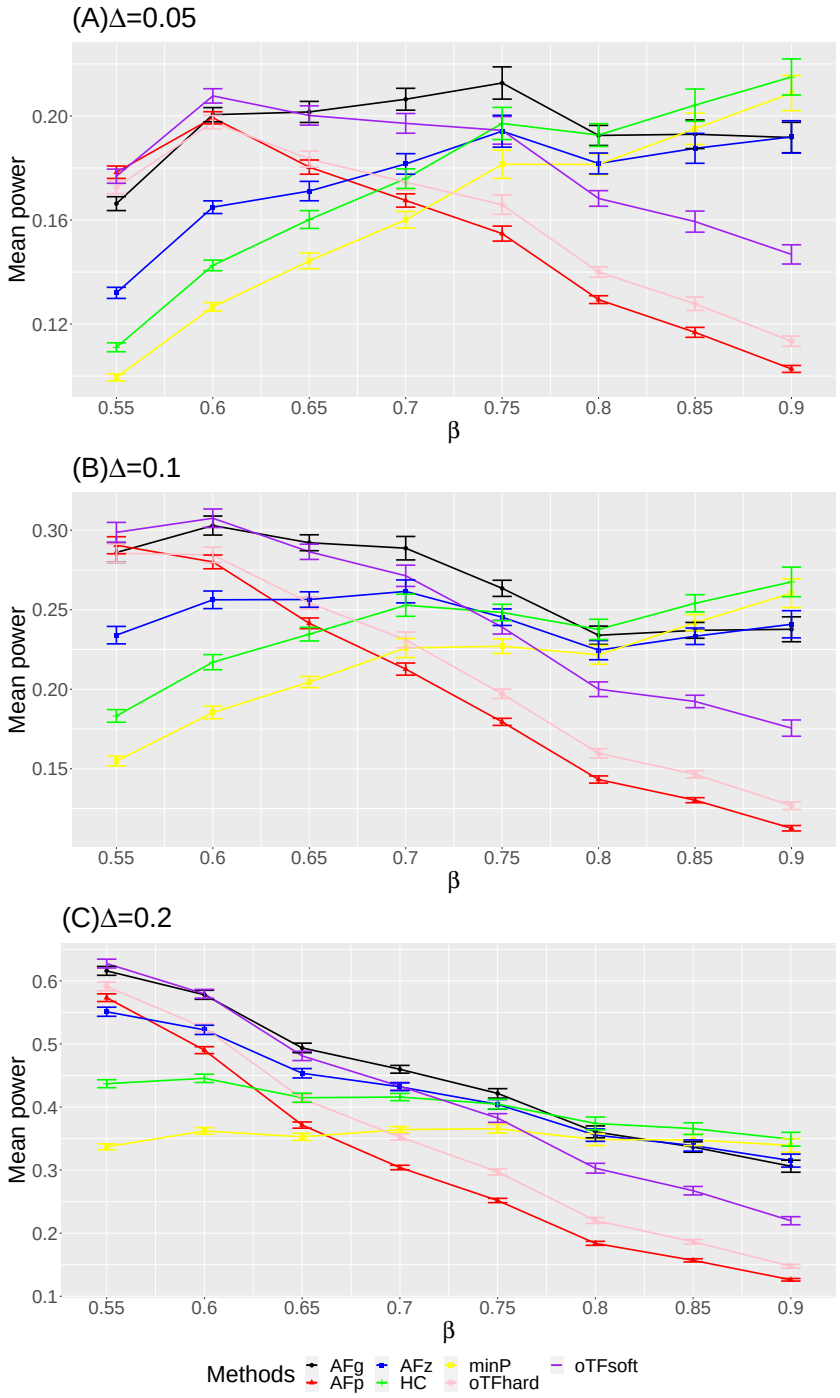


Figure 3.1: Simulations with $\sigma = 0.2$. (A)-(C) represent mean power (significance threshold $p < 0.05$) of seven p-value combination methods AFp, AFg, AFz, Higher Criticism (HC), minP, oTFhard and oTFsoft under different levels of signal strength $\Delta = 0.05$, 0.1 and 0.2, across different levels of sparsity β =0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85 and 0.9. The number of true signals $s = \lceil n^{1-\beta} \rceil$. A larger value of β leads to more sparse signals.

AFg by calculating M via finding the smallest integer such that $\log n(1+1/\log n)^{M-1} \geq n/\log n$ since $\log n$ may not be negligible and $\sqrt{n}/\log n$ can be relatively small in the finite-sample case. All the aforementioned theoretical results can be shown in a similar manner for our method after this modification. We use this version of AFg hereafter unless further notice.

Figure 3.1 shows statistical power of the seven methods with $\sigma = 0.2$, other choices of σ lead to similar results. The results show that most of the considered methods are only powerful either in the cases of relatively sparse signals or cases of relatively dense signals. For example, in Figure 3.1(A), AFz, HC and minP are powerful when $\beta \geq 0.8$ (≤ 4 true signals out of 1000 p-values) while powerless when $\beta \leq 0.7$. By contrast, AFp, oTFhard, and oTFsoft are only powerful when $\beta \leq 0.6$ and their power decreases significantly as the sparsity level increases ($\beta > 0.6$). On the contrary, AFg (black line) is always among the best across the whole range of β. AFg's performance is slightly worse than minP and HC when $\beta = 0.9$, but one can note that in this case, there is only $\lceil 1000^{0.1} \rceil = 2$ true signals out of 1000 p-values, which is the favorable case for minP and HC.

One possible reason that AFg has better performance than HC and minP for the moderate sparse signals is the use of $\log$-transformation. Intuitively, rather than focusing on the effects of a few extreme p-values, $\log$-transformation favors incorporating p-values with relatively moderate effects. With the help of a suitable selection strategy, this property of $\log$-transformation can lead to more balanced performance across a wide range of sparsity levels.

3.5.2 Robustness of AFg in the Finite-Sample Cases

In this subsection, we investigate the robustness of our method in the finite-sample case. We simulate the data by generating the following random samples:

$$Y_{ij} = \mu_i + X_{ij}, \quad i = 1, \ldots, n; j = 1, \ldots, m,$$

where $X_{ij} = (U_{ij} - E(U_{ij}))/(\text{Var } U_{ij})^{1/2}$. Here U_{ij}'s are independent and identically distributed. The choices of the distribution for U_{ij} are standard normal distribution, Student's t distribution with degrees of freedom of 5, chi-squared distribution with degrees of freedom of 10, and log-normal

distribution with zero mean and standard deviation $\sigma = 0.1$. The 4 choices of distribution correspond to the cases of baseline (normal), the case of moderate deviation from normality assumption (Student's t distribution and chi-squared distribution), and the case of asymmetry and heavy-tailed distributions (log-normal distribution), respectively. We set $n = 1000$, $m = 500, 1000$, and $\mu_1 = \ldots = \mu_s = \sqrt{2(r + \Delta) \log n}$ with $s = \lceil n^{1-\beta} \rceil$ and $\mu_i = 0$ for $i = s + 1, \ldots, n$, where $\beta = 0.55$, 0.6, ..., 0.9 for varying levels of sparsity of signals and r is determined by equation (3.6). Δ are set to be $-r$ for the robustness of the type I error control, 0.1 and 0.2 for the power with different levels of signal strength under the alternatives. The p-values to be combined are calculated by two-sided z-score tests as $p_i = 2(1 - \Phi(|T_i|))$ for $i = 1, \ldots, n$.

We perform 10^4 simulations using the above sampling procedure, the critical value at significance level 0.05 is calculated from a 10^5 Monte-Carlo sample that is sampled from the above procedure when U_{ij}'s follow the standard normal distribution and $\Delta = -r$. We repeat the whole simulation scheme 30 times.

Figure 3.2 shows the results of the robustness of AFg under various combinations of simulation parameters. AFg controls type I errors well for all the distributions at significance level 0.05. For statistical power, under all the circumstances, AFg has quite similar empirical power under the Student's t distribution case compared to the standard normal case, slightly losing power under the log-normal distribution and chi-squared distribution cases.

3.6 Application

In this subsection, we apply p-value combination methods to analyze the GWAS of neuroticism (Okbay et al., 2016), a personality trait characterized by easily experiencing negative emotions. The study investigates $6,524,432$ genetic variants (SNPs) across $179,811$ individuals, where p-values are calculated for all SNPs to evaluate the association between the variant and neuroticism. We consider four pathways from the KEGG database (Kanehisa and Goto, 2000), hsa05012, hsa05010, hsa05014, and hsa04730, which relate to four neurological diseases, Parkinson disease, Alzheimer disease, amyotrophic lateral sclerosis, and long-term depression, respectively. For each pathway, we collect SNPs within the genic or intergenic regions that belong to the pathway and

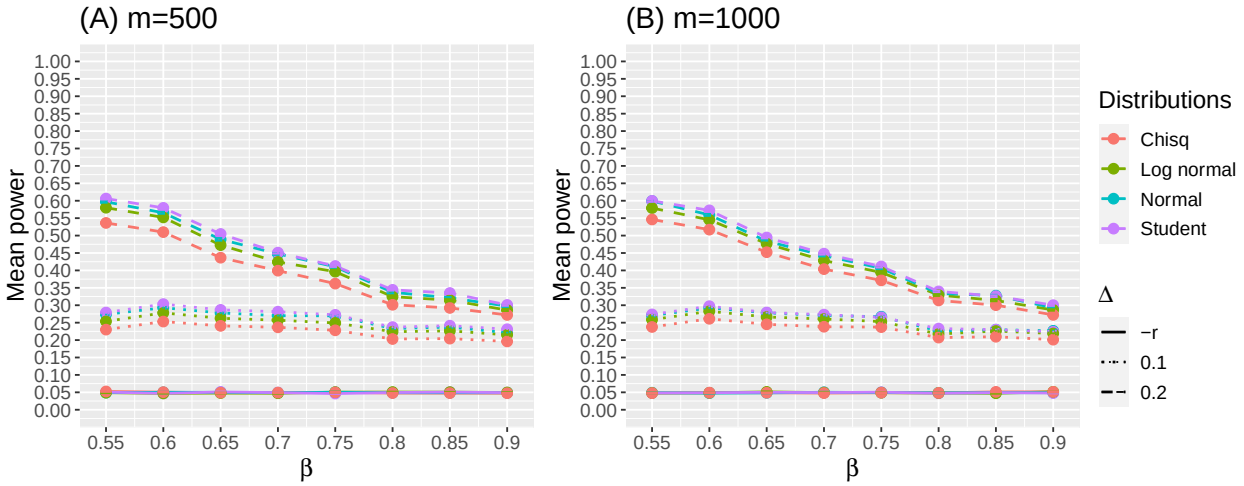


Figure 3.2: Robustness of AFg under different distributions: standard normal distribution (reference), log-normal distribution with $\mu = 0$ and $\sigma = 0.1$, chi-squared distribution with degrees of freedom of 10, and Student's t distribution with degrees of freedom 5. We evaluate the empirical power of AFg under different distributions, various levels of signal strength $\Delta = 0.1$ (dotted lines) and 0.2 (dashed lines), and different levels of sparsity of signals $\beta = 0.55$, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85 and 0.9. We also evaluate the performance of type I error control of AFg under different distributions ($\Delta = -r$, solid lines). The significance threshold in this figure is $p < 0.05$.

form a SNP set with the corresponding p-value set. To de-correlate the p-values corresponding to the SNPs in each SNP set, we filter out SNPs by a linkage disequilibrium threshold $R^2 > 0.1$, which leads to a SNP set of 1603 SNPs for hsa05012, a SNP set of 2939 SNPs for hsa05010, a SNP set of 2673 SNPs for hsa05014, and a SNP set of 1205 SNPs for hsa04730. To reduce the signal strength for higher difficulty levels of SNP-set test with varying sparsity levels, we apply a random filtering strategy to the SNPs with p-values smaller than 0.1 within each SNP set. More precisely, let H and R be the numbers of SNPs with p-values ≤ 0.1 within a SNP set before and after the filtering. For each SNP set, we randomly filter out 90%, 85%, 80% and 75% of SNPs with p-values ≤ 0.1 (corresponding to $R/H = 10\%, 15\%, 20\%, 25\%$), respectively. We repeat the above random filtering scheme 500 times. p-value combination methods are applied to the set of p-values corresponding to each SNP set to evaluate their empirical power. Here we only apply oTFsoft, AFg, minP, and HC, as AFp and oTFhard have generally worse performance than oTFsoft shown in Section 3.5.1, and AFz also has generally worse performance than AFg (also see Section 3.5.1) but without fast-computing algorithm. We apply the efficient sampling-based algorithm in Supplement Section B.3 for the fast computation of AFg. The fast-computing algorithm proposed in Zhang et al. (2020a) is applied for HC.

Table 3.1 shows empirical power results of the four methods at significance level $p < 10^{-3}$. Similar to the simulation results in Section 3.5.1, we observe that the methods can be divided into 3 categories based on their empirical power across different sparsity levels. Indeed, for example, for pathway hsa05012 (Parkinson's disease), minP and HC are the most powerful with sparse signals, with power greater than 0.38 under $R/H = 10\%$, compared to the power of 0.192 for oTFsoft. On the contrary, though powerless under the sparse case, oTFsoft is the most powerful for detecting relatively dense signals, with power greater 0.99 when $R/H \geq 20\%$, while minP and HC only have power below 0.75 under the same case. However, compared to the first two categories of methods, AFg always has comparable power to the top-performer under all the sparsity levels R/H, which is consistent with the simulation results.

Table 3.1: Empirical power with significance threshold $p < 10^{-3}$ for AFg, oTFsoft, minP and HC across different levels of sparsity ($R/H = 10\%, 15\%, 20\%, 25\%$ for the 4 pathways hsa05012 (Parkinson disease), hsa05010 (Alzheimer disease), and hsa05014 (amyotrophic lateral sclerosis), hsa04730 (long-term depression) from KEGG.

Pathways	Methods	$R/H = 10\%$	$R/H = 15\%$	$R/H = 20\%$	$R/H = 25\%$
hsa05012	AFg	0.378	0.588	0.936	1.000
	oTFsoft	0.192	0.466	0.996	1.000
	minP	0.382	0.534	0.686	0.730
	HC	0.392	0.552	0.698	0.748
hsa05010	AFg	0.278	0.452	1.000	1.000
	oTFsoft	0.052	0.492	1.000	1.000
	minP	0.262	0.352	0.516	0.550
	HC	0.270	0.364	0.520	0.552
hsa05014	AFg	0.086	0.842	1.000	1.000
	oTFsoft	0.020	0.992	1.000	1.000
	minP	0.118	0.142	0.180	0.304
	HC	0.130	0.160	0.210	0.340
hsa04730	AFg	0.164	0.280	0.962	1.000
	oTFsoft	0.040	0.296	1.000	1.000
	minP	0.164	0.270	0.372	0.445
	HC	0.164	0.270	0.372	0.445

3.7 Discussion

In this chapter, we revisit modified Fisher's methods and develop a new adaptive p-value combination method tailored to the scenario of detecting sparse and heterogeneous signals. Our contributions are threefold. First, we propose an adaptive p-value combination procedure that improves

Fisher's method under the large-scale setting. Taking the advantages of the weakly geometric system and the $\log$-transformation on p-values, our proposed method is powerful across a variety of sparsity levels of signals inside combined p-values. The weakly geometric system also alleviates the computational burden of our procedure. Second, we show that our approach is rate-optimal for separating the null parameter space and the sparse alternative parameter spaces in a classical Gaussian sequence model. To the best of our knowledge, this is the first result for a fully data-driven p-value combination procedure among all modified Fisher's methods with an optimal separating rate under the large-scale setup. Third, realizing the gap between the assumption that p-values are exact and the widely use of approximation techniques (e.g., CLT) to calculate p-values, we investigate the robustness property of our method. Under mild moment conditions, we show that our testing procedure is still rate-optimal and hence robust to p-values calculated using Studentized test statistics. This result suggests that it is worthwhile to investigate similar robustness properties for other existing p-value combination methods.

Modern data science faces challenges from larger data dimension, high-level data variability, and the need for inference tailored to the subject domain. Motivated by the need to detect sparse signals within a large number of p-values in real data analysis, in this chapter, starting from solid theoretical analysis, we modify the classical Fisher's method to a novel adaptive procedure tailored to the high-dimensional scenario with sparse signals. We conclude that our method is powerful and robust for detecting heterogeneous and sparse signals. There are several future directions to be investigated. Firstly, it is of interest to extend our method to the scenarios of combining dependent p-values, as complex dependency structures widely exist in modern large-scale data, such as genomics data. Secondly, it is desired to conduct a refined theoretical analysis to precisely quantify the minimal signal strength, including a sharp constant factor in front of the separating rate, for our method to be powerful, as it would provide a more accurate prediction of the finite-sample performance. Thirdly, it is interesting to assign weights to each combined p-values to incorporate prior information (e.g., functional annotations for GWAS) or upweight some p-values to increase power (e.g., p-values correspond to rare variants in GWAS). These topics are currently under exploration and will be reported in the future work.

4.0 Heavy-tailed Distribution for Combining Dependent P-Values with Asymptotic Robustness

The contents of this chapter are accepted by the journal *Statistica Sinica* (Fang et al., 2023a).

4.1 Introduction

Combining p-values to aggregate information from multiple sources is popular in the social sciences and biomedical research. Classical methods focus on combining multiple independent and frequent signals to increase the statistical power, which can be viewed as a type of meta-analysis. Consider the combination of n independent p-values, $\boldsymbol{p} = (p_1, ..., p_n)$. Early methods used $T(\boldsymbol{p}) = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} F_U^{-1}(1 - p_i)$ to sum transformed p-values, where the transformation $g(p)$ is the inverse cumulative distribution function (CDF) of a random variable U. Conventional methods in this category include Fisher's method (Fisher, 1992) where $T = \sum_{i=1}^{n} -2\log(p_i)$ and U is a chi-squared distribution, and Stouffer's method (Stouffer et al., 1949) where $T = \sum_{i=1}^{n} -\Phi^{-1}(p_i)$ and U is a standard normal distribution, among others (Edgington, 1972; Pearson, 1933; Mudholkar and George, 1979). These methods use a classical meta-analysis to combine independent and relatively frequent signals, and apply a light-tailed distribution (i.e., tails thinner than an exponential function) for U. The efficiency of such methods is mostly considered under the asymptotic framework that the number of p-values n is fixed and sample size m used to derive each p-value goes to infinity, where $p = O(e^{-m})$ in most cases. Under this setting, it has been shown that only the equivalent class of Fisher's method is asymptotically Bahadur optimal (ABO), meaning that the efficiency of the combined p-value statistics is asymptotically optimal under fixed n and $m \to \infty$ (Littell and Folks, 1971).

With the advent of big data, many studies now combine p-values with large n. The seminal paper by Donoho and Jin (2004) established a framework for combining p-values with weak and sparse signals, and proposed the higher-criticism test with the asymptotically optimal property. This second category of methods considers $n \to \infty$, and only a small number s of the n p-values

($s = n^{\beta}$ where $0 < \beta < \frac{1}{2}$) have weak signals ($p = O(n^{-r}/(\log n)^{\frac{1}{2}})$ with $0 < r < 1$), while all remaining p-values have no signal (i.e., $p \overset{D}{\sim} Unif(0,1)$). Under this setting, the classical minimum p-value method (minP) $T = min_{1 \leq i \leq n} p_i$ is asymptotically optimal in terms of the detection boundary only for $0 < \beta < 1/4$, whereas higher criticism attains an optimal detection boundary for all possible $0 < \beta < 1/2$. Several methods, including the Berk-Jones test (Berk and Jones, 1979; Li and Siegmund, 2015), have subsequently been proposed to improve the finite-sample power of higher criticism, while maintaining an optimal detection boundary.

All of the aforementioned methods were developed to combine independent p-values. However, many modern large-scale data analyses need to combine a large number of dependent p-values that have sparse and weak signals, which we categorize as methods of the third category. A notable application is to combine p-values of multiple correlated SNPs (there may be tens to hundreds or thousands) in an SNP set (e.g., all SNPs in a gene region or in gene regions of a pathway) in a genome-wide association study (GWAS). In this case, neighboring SNPs often have unknown dependency structures, prompting efforts to extend existing tests to account for dependency using permutations or other numerical simulation approaches (Liu and Xie, 2019, e.g.,). However, permutation or simulation-based methods are not practical when n is large, and a precise p-value is needed for multiple comparisons. The null hypothesis may also be difficult to simulate using a permutation. Barnett et al. (2017) developed an analytic approximation for higher criticism that incorporated a dependency structure. However, the method is still computationally intensive and not sufficiently accurate for the small p-values needed for multiple comparisons. Motivated by these needs, Liu and Xie (2020) and Wilson (2019a) independently proposed the Cauchy combination test ($T = \sum_{i=1}^{n} \tan\{(0.5 - p_i)\pi\}$) and the harmonic mean combination test ($T = \sum_{i=1}^{n} \frac{1}{p_i}$), respectively, to combine p-values under an unspecified dependency structure. Wilson (2019a) also provided a convenient R package called harmonicmeanp (function p.mamml) to implement the harmonic test. A remarkable property of both methods is that the null distributions and testing procedures derived from the independence assumption are robust under a dependency structure in an asymptotic, but practical sense; see Section 4.3.1. Motivated by this observation, we consider a rich family of test statistics that includes the Cauchy and harmonic mean tests. More precisely, we

consider the test statistic

$$T = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} F_U^{-1}(1 - p_i),$$

where the transformation $g(p)$ corresponds to U from a regularly varying distribution family, which is a broad family of heavy-tailed distributions. We investigate the conditions required to achieve the practical robustness to dependency of the Cauchy and harmonic mean methods. Note that selections of U in classical meta-analysis settings (fixed n and $m \to \infty$) are all from thin-tailed distributions (e.g., chi-squared distribution for Fisher's methods and the Gaussian distribution for Stouffer's method). This is reasonable, because a thin-tailed distribution produces contributions that are more even from marginally significant p-values in meta-analyses of frequent signals. In contrast, the Cauchy and harmonic mean methods correspond to heavy-tailed distributions of U, which focus on small p-values and down-weigh marginally significant p-values. Figure 4.1 shows the transformation function of $g(p)$ in $\log$-scale. For Fisher's method, the contributions of the p-values 10^{-2} and 10^{-6} to the test statistics are 4.6 and 13.8, respectively. For heavy-tailed transformation methods, the contributions become 100 and 10^6 for the harmonic mean, and 31.82052 and 3.18×10^5 for the Cauchy method. With an increased focus on small p-values, the methods are more powerful for detecting sparse signals. Note that Vovk and Wang (2020) also considered the sum of transformed p-values to combine p-values, and showed an upper bound of the significance level inflation under an arbitrary dependence structure. The comparison of our results with theirs is provided in the remark following Theorem 4.2. Wilson (2019b), Wilson (2020), and Vovk et al. (2021) also did related work involving combining dependent p-values.

Throughout this paper, when we refer to a thin-tailed, heavy-tailed, or regularly varying method, we mean that its corresponding U is a thin-tailed, heavy-tailed, or regularly varying distribution. The remainder of the paper is structured as follows. We first investigate the Box-Cox transformation for $g(p)$ in Section 4.2, which is equivalent to a Pareto distribution for U. In Section 4.2.1, we discuss existing methods, including the minP, harmonic mean, Cauchy, and Fisher methods in this framework. In particular, we show that the Cauchy method is approximately equivalent to the harmonic mean method, which is a special case of the Box-Cox transformation. In Section 4.2.1, we observe that the Cauchy method may suffer from a large negative penalty for p-values close to one. To avoid this problem, we improve the Cauchy method by introducing a new test, called the truncated Cauchy method, and develop a fast computing algorithm for it. In Section 4.3, we

introduce a family of heavy-tailed distributions, namely, regularly varying distributions, and investigate the conditions in the family that provide robustness for the dependency structure, as in the Cauchy and harmonic mean methods (Sections 4.3.1 and 4.3.2). Section 4.3.3 studies the power of the family of methods in terms of the detection boundary under the sparse and weak alternatives considered in Donoho and Jin (2004). Section 4.4 contains extensive simulations that demonstrate the type-I error control and power of various methods. Here, we also verify the theoretical results numerically. In Section 4.5, we apply the proposed method to data from a GWAS application of neuroticism to compare the performance of the methods and demonstrate the improvement of the truncated Cauchy method over the Cauchy method. Section 4.6 concludes the paper.

4.2 Connection between MinP, Harmonic Mean, Cauchy, and Fisher

4.2.1 Using A Pareto Distribution to Connect Four Existing Methods

As mentioned in Section 4.1, many methods of the first category combine independent and relatively frequent signals from thin-tailed distributions for U, and many methods of the second and third categories for combining sparse and weak signals, respectively, use heavy-tailed distributions. In this subsection, we consider a Pareto distribution for U, which is equivalent to a Box-Cox transformation for $g(p)$. Based on this transformation family, we connect four existing methods: minP, harmonic mean, Cauchy, and Fisher. The insights gained from the Pareto distribution also help when we introduce the regularly varying distribution as an extended richer family in the next section. Finally, we prove the approximate equivalency of the harmonic mean and Cauchy combination methods. Consider the following family of p-value combination methods: $T = \sum_{i=1}^{n} g(p_i)$, where $g(p) = \frac{1}{p^\eta}$, for some $\eta > 0$. We can show that $g(p) = F_U^{-1}(1-p)$, such that $U \overset{D}{\sim} Pareto(\frac{1}{\eta}, 1)$. In other words, $P(U > t) = t^{-\frac{1}{\eta}}$ for $t > 1$, which means U is a heavy-tailed distribution. A larger η corresponds to a heavier tail. In particular, the harmonic mean method corresponds to $\eta = 1$ in the Pareto distribution. Note that by denoting $\lambda = -\eta$, we can rewrite $h(p;\lambda) = \frac{g(p;\eta)-1}{\lambda} = \frac{p^\lambda - 1}{\lambda}$, which is the Box-Cox transformation. Proposition 4.1 shows that minP and Fisher are limiting cases in the Pareto distribution when $\eta \to +\infty$ and when

$\eta \to 0$, respectively. Proposition 4.2 shows that the Cauchy combination method is approximately identical to the harmonic mean for relatively small p-values.

Proposition 4.1. *For fixed n, minP is a limiting case in the Pareto distribution when $\eta \to \infty$. Similarly, Fisher's method is a limiting case of Pareto when $\eta \to 0$.*

Proof. Denote $T_{\gamma_m} = \sum_{i=1}^n \frac{1}{p_i^{\gamma_m}} = \sum_{i=1}^n \frac{1}{p_{(i)}^{\gamma_m}}$, where $p_{(i)}$ are ordered p-values. Note that T_{γ_m} is equivalent to $T^*_{\gamma_m} = \left(\sum_{i=1}^n \frac{1}{p_i^{\gamma_m}}\right)^{\frac{1}{\gamma_m}} = \frac{1}{p_{(1)}}\left(\sum_{i=1}^n \left(\frac{p_{(1)}}{p_{(i)}}\right)^{\gamma_m}\right)^{\frac{1}{\gamma_m}}$. As $\gamma_m \to \infty$, $T^*_{\gamma_m} \to \frac{1}{p_{(1)}}$, which is equivalent to minP.

To prove the result for Fisher's method, note that T_{γ_m} is equivalent to $T^{**}_{\gamma_m} = \sum_{i=1}^n \frac{p_i^{-\gamma_m}-1}{-\gamma_m}$. By L'Hospital's rule , we have $\lim_{\gamma_m \to 0} \frac{p^{-\gamma_m}-1}{-\gamma_m} = \log(p)$. Hence, $T^{**}_{\gamma_m} \to \sum_{i=1}^n \log(p_i)$ almost surely, and is equivalent to Fisher's method. □

Proposition 4.2. *The Cauchy combination test is approximately identical to the harmonic mean for relatively small p-values, in the sense that $\frac{\pi \cdot g^{(CA)}(p) - g^{(HM)}(p)}{g^{(HM)}(p)} = O(p^2)$.*

Proof. By Taylor's expansion, $g^{(CA)}(p) = \tan\{(0.5-p)\pi\} \approx \frac{1}{\pi p} - \frac{\pi p}{3} - \frac{(\pi p)^3}{45} + \cdots$. The result follows immediately. Chen et al. (2021) also showed a similar result. □

It is somewhat surprising that even though the forms of the Cauchy and harmonic mean transformations are different, they are approximately equivalent when p is small. Furthermore, the behavior of both when p is small is characterized by the index $\eta = 1$ of the Box-Cox transformation (note that these two transformations behave differently when p is close to one). It is natural to ask whether other methods exist for combining p-values in an extended rich heavy-tailed distribution family that enjoy a similar finite-sample robustness property to that of the Cauchy and harmonic mean methods. To answer this question, we introduce a family of regularly varying distributions, and investigate its properties in Section 4.3.

Figure 4.1 shows a minus log-scaled p transformation $g(p)$ versus a minus log-scaled transformation $g(p)$ for $\text{BC}_{0.5}$ (i.e., Box-Cox transformation with $\eta = 0.5$), HM (the harmonic mean method, equivalent to BC_1), CA (the Cauchy method), $\text{BC}_{1.5}$, Fisher's method and Stouffer's method. We find that as η increases, smaller p-values become more dominant and the effect of marginally significant p-values rapidly diminishes, yielding stronger power for sparse signal appli-

cations. CA and HM are approximately proportional when p is sufficiently small (roughly when $p < 10^{-2}$).

Although HM and CA are approximately equivalent when combining relatively small p-values, when a p-value is very close to one, the contribution in the Cauchy method is close to negative infinity, which can cause numerical issues and substantial power loss; we refer to this as the "large negative penalty issue" in relation to the Cauchy method. A p-value close to one happens often in tests of discrete data, in which case, the p-values under the null hypothesis may not necessarily be $Unif(0, 1)$. The p-values may also be close to one when n is large or when the model used to derive the p-values is misspecified. As a simple remedy, we propose a truncated Cauchy test (CA$^{\text{tr}}$) that truncates any of the n p-values greater than $1 - \delta$ to be $1 - \delta$. For example, when $\delta = 0.01$, we have $p^{tr} = p$ if $p < 0.99$, and $p^{tr} = 0.99$ if $p \geq 0.99$. We recommend using $\delta = 0.01$. Conceptually, δ should be sufficiently large so that it avoids the large negative penalty issue in Cauchy. However, for computational purposes, it cannot be too large, or the approximation by our fast-computing procedures may not be accurate. A detailed justification for choosing $\delta = 0.01$, with support from simulation results, is given in the Supplement Section C.2.4. The proposed method can also be viewed as a sum of transformed p-values. Indeed, the CA$^{\text{tr}}$ statistic can be written as

$$T_{CA^{tr}} = \sum_{i=1}^{n} \tan(\pi(\frac{1}{2} - p_i))1(p_i < 1 - \delta) + \tan(\pi(\delta - \frac{1}{2}))1(p_i \geq 1 - \delta).$$

For more details on CA$^{\text{tr}}$, see the Supplement Section C.2.

4.3 Asymptotic Properties of Regularly Varying Methods for P-Value Combination

4.3.1 Disbributions with Regularly Varying Tails

Before introducing the regularly varying distributions, we first define some notations. Throughout this paper, denote by $\bar{F}$ the survival function of the distribution F (i.e., $\bar{F}(t) = 1 - F(t)$, for any t). The limits and asymptotic properties are assumed to be for $t \to \infty$, unless stated otherwise. For two positive functions $u(\cdot)$ and $v(\cdot)$, we write $u(t) \sim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} = 1$. In addition,

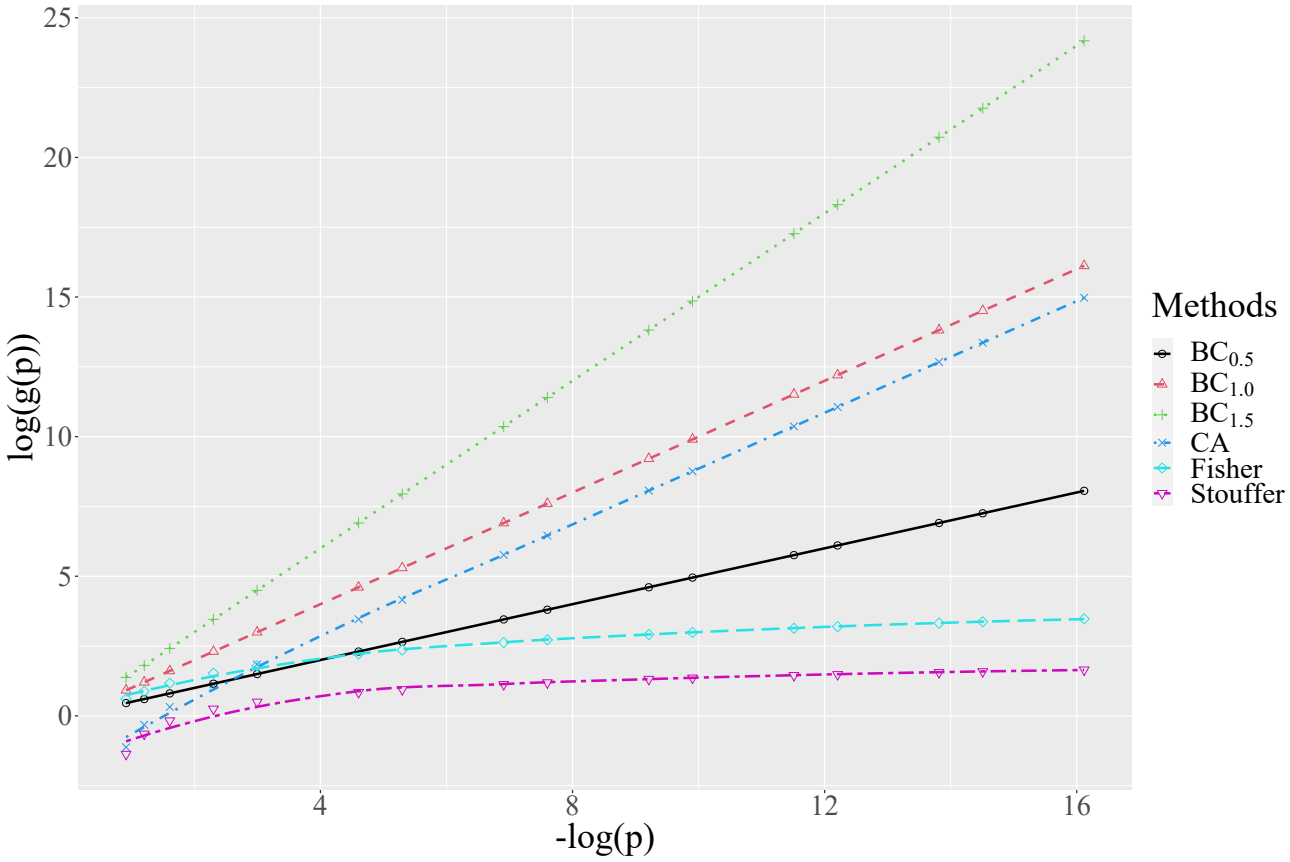


Figure 4.1: Comparison of transformations. We show six transformations of p-values, $g(p)$, i.e., $\text{BC}_{0.5}$, BC_1 (HM), $\text{BC}_{1.5}$, CA, Fisher, and Stouffer. The x-axis is $-\log(p)$, and the y-axis shows $\log(g(p))$.

if $\lim_{t\to\infty} \frac{u(t)}{v(t)} > 1$, we write $u(t) \gtrsim v(t)$, and if $\lim_{t\to\infty} \frac{u(t)}{v(t)} < 1$, we write $u(t) \lesssim v(t)$. A distribution with a regularly varying tail is defined as follows:

Definition 4.1. A distribution F is said to belong to the family of distributions with regularly varying tails with index γ (denoted by $F \in R_{-\gamma}$) if

$$\lim_{x\to\infty} \frac{\bar{F}(xy)}{\bar{F}(x)} = y^{-\gamma},$$

for some $\gamma > 0$ and all $y > 0$.

We denote the family of distributions with regularly varying tails as R. Then, we can show that every distribution F belonging to $R_{-\gamma}$ can be characterized by

$$\bar{F}(t) \sim L(t)t^{-\gamma},$$

where $L(t)$ is a slowly varying function (Karamata, 1933). A function L is called slowly varying if $\lim_{y\to\infty} \frac{L(ty)}{L(y)} = 1$, for any $t > 0$. Some examples of slowly varying functions $L(t)$ are $1, \ln(t)^{\nu}, and \ln(\ln(t))$. Given the property of a slowly varying function $L(t)$, the tail of a regularly varying distribution converges to zero at a relatively slow rate, which leads to the heavy-tailed property.

The family of distributions with regularly varying tails includes the Pareto distribution, Cauchy distribution, log-gamma distribution, and inverse gamma distribution. Indeed, the survival function of Pareto(a,b) is $\bar{F}(t) = \frac{b}{t^a},\ t > b$, and hence $U \in R_{-a}$. In addition, the survival function of the Cauchy distribution is $\bar{F}(t) \sim \frac{1}{t\pi}$, and therefore $U \in R_{-1}$.

An important property of distributions with regularly varying tails is as follows: Assume $U_1, \ldots, U_n$ are independent and identically distributed (i.i.d.) random variables with distribution function $F \in R_{-\gamma}$. Then,

$$P(U_1 + \ldots + U_n > t) \sim nP(U_1 > t). \tag{4.1}$$

4.3.2 Asymptotic Tail Probability Approximation and Robustness to Dependence

The first theorem approximates the null distribution of the test statistic. Assume that the p-values are obtained from z-scores; that is, the test statistics all follow normal distributions. Specifically, let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be the random vector (z-scores) for the n test statistics. The mean of $\boldsymbol{X}$ is $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and the correlation matrix is $\boldsymbol{\Sigma}$. Because we can always rescale test statistics, we assume each X_i has variance one. Under the null hypothesis, $H_0 : \mu_i = 0, \forall i = 1, \ldots, n$; hence, the p-value for the ith study is $p_i = 2(1 - \Phi(|X_i|))$, for $i = 1, \ldots, n$. We consider the test statistic $T(\boldsymbol{X}) = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} g(2(1-\Phi(|X_i|)))$, which is the sum of transformed p-values. When $p_i \overset{D}{\sim} Unif(0, 1)$ under the null hypothesis, $g(p_i)$ is a random variable, where we denote $g(p_i) \overset{D}{\sim} U$, which is consistent with the previously introduced relationship $g(p_i) = F_U^{-1}(1 - p_i)$ when U is a continuous random variable. We further assume the following conditions for $T(\boldsymbol{X})$:

(A1) $\forall 1 \leq i < j \leq n$, X_i and X_j are bivariate normally distributed.

(A2) Let $U_i = g(p_i)$, for $i = 1, \ldots, n$, with $U_i \overset{D}{\sim} U \in R_{-\gamma}$ under H_0. Assume that the function $g(p)$ is continuous and satisfies one of two situations: (A2.1) $g(p)$ is strictly decreasing in $(0, 1)$; (A2.2) $g(p)$ is bounded below (i.e., $g(p) > c'$, for a certain constant c') and is strictly decreasing in $(0, c)$, for some constant $0 < c < 1$.

(A3) (*balance condition*) Under H_0, let F be the CDF of U and $G(t) = P(|U| > t) = t^{-\gamma}L(t)$, where $L(t)$ is a slowly varying function. Assume $\frac{\bar{F}(t)}{G(t)} \to p$ and $\frac{F(-t)}{G(t)} \to q$ as $t \to \infty$, where $0 < p \leq 1$ and $p + q = 1$.

Condition (A1) is mild and is also assumed in Liu and Xie (2020) when investigating the robustness of the Cauchy method under an unspecified dependence structure. Throughout this paper, the term "unspecified dependence structure" indicates an unspecified Gaussian correlation structure. This condition guarantees that the tail distributions of each pair of U_i and U_j are asymptotically tailed independent; see the precise definition of asymptotically tailed independence for a pair of random variables in the Supplement Section C.1.

Condition (A2) includes the Box-Cox transformation (satisfying A2.1), Cauchy transformation (satisfying A2.1), and truncated Cauchy transformation (satisfying A2.2) introduced in Section 4.2.1. Condition (A3) is called the "balance condition", and is a common condition for random variables with regularly varying tails (Goldie and Klüppelberg, 1998). For example, for the har-

monic mean method, $p = 1$ and $q = 0$, for the Cauchy method, $p = q = 1/2$, and for the truncated Cauchy method, $p = 1$ and $q = 0$.

Theorem 4.1. *Under conditions (A1), (A2), and (A3) and assuming* $\rho_{ij}, for 1 \leq i < j \leq n$, *the* (i, j)*th element of* $\boldsymbol{\Sigma}$ *satisfies* $-1 < \rho_{ij} < 1$. *Then, under* $H_0 : \boldsymbol{\mu} = \mathbf{0}$ *and for any correlation matrix* $\boldsymbol{\Sigma}$, *we have*

$$P(T(\boldsymbol{X}) > t) \sim nP(U > t).$$

Here, $T(\boldsymbol{X}) = \sum_{i=1}^{n} U_i$ is the sum of correlated random variables with regularly varying tails. The theorem is somewhat surprising and a general result, because it applies to any regularly varying method and any correlation structure $\boldsymbol{\Sigma}$ with $-1 < \rho_{ij} < 1$, as long as no perfect correlation exists. This theorem is related to Theorem 3.1 in Chen and Yuen (2009), i.e., Lemma S2 in the Supplementary Material. Roughly speaking, because of the heaviness of the tail of each U_i and the asymptotic, tailed independence between each pair of U_i and U_j, asymptotically, the correlation structure has limited influence on the tail of $T(\boldsymbol{X})$. Because the approximated tail probability is independent of $\boldsymbol{\Sigma}$, an immediate application is to derive the p-value of a regularly varying method under the independence assumption (i.e., $P(U_1 + \cdots + U_n > t)$, with i.i.d. $U_1, \cdots, U_n$; see Equation (4.1)). The theorem is asymptotically robust to an unspecified dependence structure, as shown for the harmonic mean and Cauchy methods (Wilson, 2019a; Liu and Xie, 2020). Alternatively, one may approximate the tail probability by $nP(U > t)$. However, note that the robustness to an unspecified dependence structure is in an asymptotic sense, meaning that we may require an extremely large t (corresponding to an extremely small test size α) for different tail heaviness in U and correlation structures in order to guarantee a good approximation. Throughout this paper, we approximate $P(U_1 + \cdots + U_n > t)$ under a dependence structure by calculating $P(U_1 + \cdots + U_n > t)$ under the independence assumption using a Monte Carlo simulation.

Below, we perform a simple simulation to demonstrate and investigate Theorem 4.1. Assume $n = 3$, and $\boldsymbol{X} = (X_1, X_2, X_3)$ is multivariate normal with unit variance and common pairwise correlation $\rho_{ij} = \rho$ $(1 \leq i < j \leq 3)$. In this simulation, we set $\rho = 0$, 0.3, 0.6, 0.9, and 0.99. Here, we consider seven Box-Cox tests, $\text{BC}_{0.75}$, $\text{BC}_{0.8}$, $\text{BC}_{0.9}$, BC_{1}, $\text{BC}_{1.1}$, $\text{BC}_{1.25}$, and $\text{BC}_{1.5}$. From Theorem 4.1, we calculate $y(\alpha) = \frac{nP(U > t_\alpha)}{P(T(\boldsymbol{X}) > t_\alpha)}$ from simulations, where t_α is chosen so that $P(T(\boldsymbol{X}) > t_\alpha) = \alpha$ and $\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$. We expect $\lim_{t_\alpha \to \infty} \log(y(\alpha)) = 0$ when

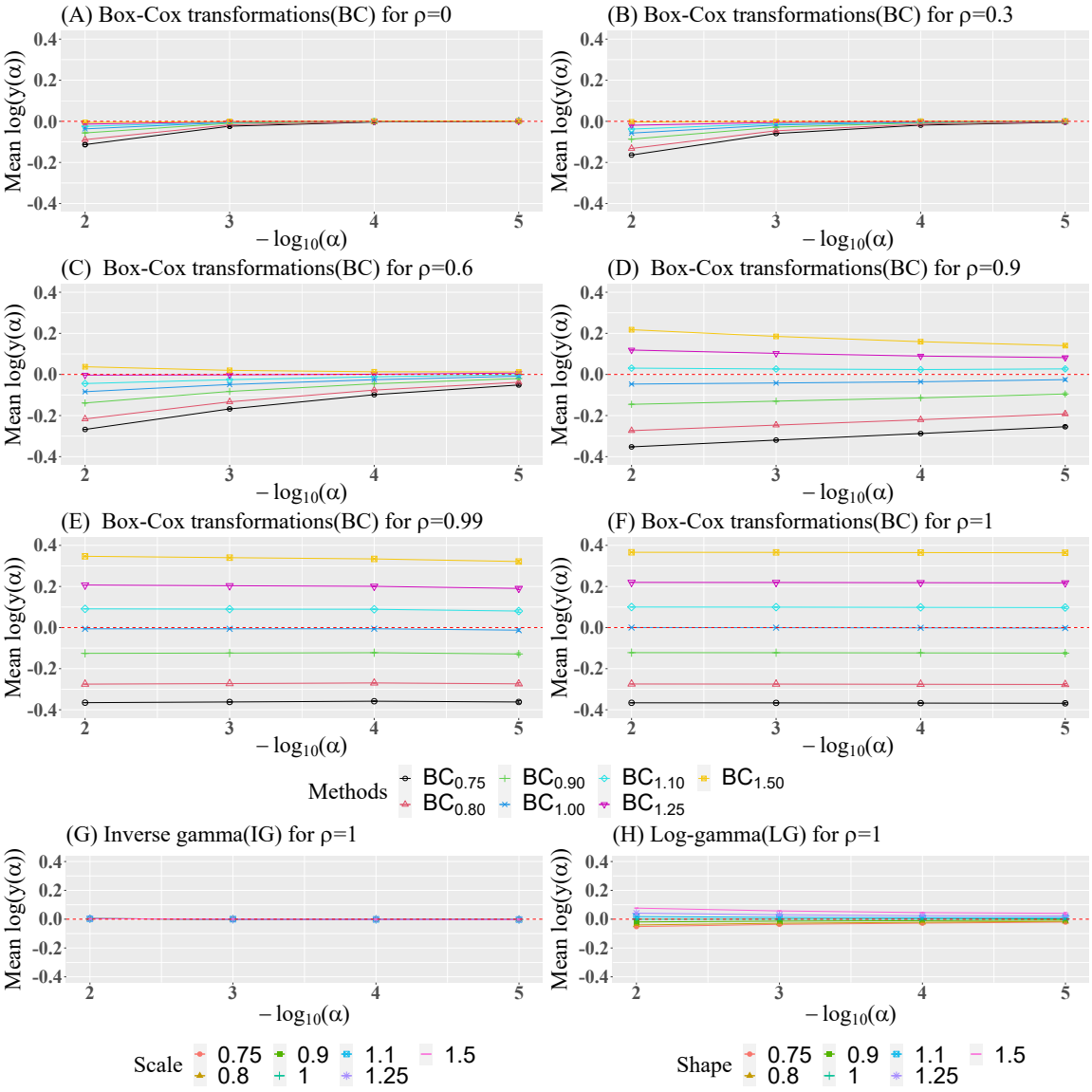


Figure 4.2: The mean log-scaled $y(\alpha)$ for Box-Cox transformations, inverse gamma and log-gamma across different significance levels α. (A)-(F) represent the results of Box-Cox transformations with values of $\eta = 0.75, 0.8, 0.9, 1, 1.1, 1.25, 1.5$ for correlation level ρ =0, 0.3, 0.6, 0.9, 0.99, and 1, respectively. (G) represents the results of the inverse gamma with shape parameter one and scale parameter values 0.75, 0.8, 0.9, 1, 1.1, 1.25, 1.5, for correlation level $\rho = 1$. (H) represents the results of the log-gamma with rate parameter one and scale parameter values 0.75, 0.8, 0.9, 1, 1.1, 1.25, 1.5, for correlation level $\rho = 1$. The x-axis is the negative logarithm of significance level α to base 10, where α is set to $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$, and the red dash line is the reference line $\log(y(\alpha)) = 0$ in all sub-figures.

$-1 < \rho < 1$. Figures 4.2A-4.2E show $\log_{10}$-scale α on the x-axis and the mean $\log(y(\alpha))$ on the y-axis for different ρ =(0, 0.3, 0.6, 0.9, 0.99). Note that as ρ increases, a smaller α will be required for a good approximation. Theorem 4.2 further characterizes what would happen if some of the p-values have perfect correlations $\rho_{ij} = 1$ or -1.

Theorem 4.2. *Suppose conditions (A1), (A2), and (A3) in Theorem 4.1 hold. Define an arbitrary weight vector* $\boldsymbol{w} = (w_1, \cdots, w_n) \in R_+^n$, $T_{n,\boldsymbol{w}}(\boldsymbol{X}) = \sum_{i=1}^n w_i g(p_i)$. *Furthermore, assume* $\rho_{ij} = 1$ *or* -1 *for* $1 \leq i < j \leq m$, *and* $|\rho_{ij}| < 1$ *for* $i > m$ *or* $j > m$. *Then, under the null hypothesis* $H_0 : \boldsymbol{\mu} = \mathbf{0}$, *we have*

$$P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) \sim \left\{ \left(\sum_{i=1}^{m} w_i \right)^{\gamma} + \sum_{i=m+1}^{n} w_i^{\gamma} \right\} P(U > t).$$

Note that Theorem 4.2 is a more general result, of which Theorem 4.1 is a special case. Consider a special scenario $\boldsymbol{w} = (1, \cdots, 1)$. An immediate consequence of Theorem 4.2 is that only when $\gamma = 1$ (e.g., the HM, CA, or CA$^{\text{tr}}$ method) can satisfy $\{(\sum_{i=1}^m w_i)^{\gamma} + \sum_{i=m+1}^n w_i^{\gamma}\} = m^{\gamma} + (n-m) = n$, which produces the asymptotic robustness of Theorem 4.1. In other words, Figures 4.2 (A)-4.2(E) already show a hint that the convergence of Theorem 4.1 becomes increasingly difficult when ρ increases to almost one. When some of the p-values have perfect correlation, only index $\gamma = 1$ of the regularly varying distribution is asymptotically robust to an unspecified dependence structure. Figure 4.2 (F) shows a simulation with $\rho = 1$, which satisfies the condition of Theorem 4.2. By assuming $w_1 = w_2 = w_3 = 1$ and $\rho = 1$, we have $P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) \sim 3^{\gamma} P(U > t)$. Figure 4.2 (F) verifies Theorem 4.2 that only BC_1 can reach the convergence $\lim_{t_\alpha \to \infty} \log(y(\alpha)) = 0$, showing robustness to perfect correlation. Although Figure 4.2 (E) ($\rho = 0.99$) and Figure 4.2 (F) ($\rho = 1$) are visually similar, all BC methods in Figure 4.2 (E) eventually converge to zero as $\alpha \to 0$, by Theorem 4.1, although very slowly. On the other hand, in Figure 4.2 (F), only BC_1 converges to zero, by Theorem 4.2.

Corollary 4.1. *Suppose the conditions in Theorem 4.2 hold and assume* $\sum_{i=1}^n w_i = n$, *then we have*

$$\begin{cases} P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) & \sim nP(U > t) \quad \textit{if } \gamma = 1, \\ P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) & \gtrsim nP(U > t) \quad \textit{if } \gamma > 1, \\ P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) & \lesssim nP(U > t) \quad \textit{if } \gamma < 1. \end{cases}$$

From Corollary 4.1, note that when $w_1 = \cdots = w_n = 1$ and the transformation $g(p) = 1/p^{1/\gamma}$, the test statistic $T_{n,\boldsymbol{w}}$ corresponds to the statistic $\text{BC}_\eta, \eta = 1/\gamma$. Hence, the BC tests with $\eta < 1$ (i.e., $\gamma > 1$) are anti-conservative in this situation; the higher the value of γ, the more anti-conservative the test is. This is verified by Figure 4.2F for $\text{BC}_{0.9}$, $\text{BC}_{0.8}$, and $\text{BC}_{0.75}$ when $\rho = 1$. As $\eta \to 0$ (i.e., $\gamma \to \infty$), BC_η is asymptotically equivalent to Fisher's method, and is the most anti-conservative under dependence. On the other hand, for $\eta > 1$ (i.e., $\gamma < 1$), all the corresponding tests BC_η ($\eta > 1$) are conservative under this dependence structure, which is confirmed by Figure 4.2F for $\text{BC}_{1.1}$, $\text{BC}_{1.25}$, and $\text{BC}_{1.5}$. In particular, when $\eta \to \infty$ ($\gamma \to 0$), BC_η becomes minP, which hence is expected to be very conservative. Figures 4.2G and 4.2H verify that because the inverse gamma and log-gamma are also regularly varying distributions with index $\gamma = 1$, they enjoy an asymptotic robustness to the correlation structure, similar to that of HM (BC_1) and Cauchy, even when perfect correlation exists. Another important implication of this corollary is that among all the tests that use transformations of regularly varying distributions, only the type-I errors of those corresponding to $\gamma \leq 1$ are well preserved asymptotically (i.e., tests that are at least not anti-conservative asymptotically) under any correlation structure.

Corollary 4.2. *If we further assume $-1 < \rho_{i,j} < 1, \forall 1 \leq i < j \leq n$ (i.e., $m = 0$), then we have*

$$P(T_{n,\boldsymbol{w}} > t) \sim \sum_{i=1}^{n} w_i^\gamma P(U > t).$$

Corollary 4.2 shows that the tail probability of the weighted test statistic $T_{n,\boldsymbol{w}}$ can be approximated by $\sum_{i=1}^{n} w_i^\gamma P(U > t)$. Similarly to the unweighted version in Theorem 4.1, because the approximation in Corollary 4.2 is independent of the correlation structure, $P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t)$ under the dependence structure can be approximated by calculating $P(w_1 U_1 + \ldots + w_n U_n > t)$ under the independence assumption using a Monte Carlo simulation, as long as there are no perfect correlations between U_i. Furthermore, note that this formula can be considered an extension of Corollary 1.3.8 in (Mikosch, 1999), in which $U_1, \ldots, U_n$ are assumed to be independent regularly varying distributed random variables.

Remark 4.1. Note that the robustness property of Theorems 4.1 and 4.2 is similar to (Liu and Xie, 2020; Wilson, 2019a) and describes only the asymptotic behavior of the tail probability of our proposed family. Indeed, the results of Theorems 4.1 and 4.2 guarantee only that the type-I

errors of the corresponding tests ($\gamma = 1$, equivalent to the harmonic mean and Cauchy) can be well controlled for a small size α, given fixed n and $\mathbf{\Sigma}$. Intuitively, as n increases, a more stringent cutoff corresponding to a small α is needed to ensure the robustness of type I error control. An ideal robustness property for the type-I error should achieve a uniform upper tail bound in the sense of $P(T(\boldsymbol{X}) > t_\alpha) \leq c \cdot \alpha$ under any dependence structure $\mathbf{\Sigma}$, where t_α is the tail threshold when a nominal α is controlled under the independence assumption, c is independent of n, and $\mathbf{\Sigma}$ is in a reasonable magnitude (e.g., $c = 1.5$, meaning the inflation of the type-I error is at most 50%, in the worst scenario). However, this uniform bound is not achievable, in general. Vovk and Wang (2020) recently provided a remarkable uniform bound for arbitrary dependency structure (note that ours is an unspecified dependency structure), but dependent on n for the HM method:

$$P(HM > t) \leq n a_n^{HM} P\left(U > t\right) = \frac{n a_n^{HM}}{t}, \text{where } U \overset{D}{\sim} Pareto(1, 1),$$

where the adjusted factor α_n^{HM} is between $\log(n)$ and $e \cdot \log(n)$ (see Proposition 6 in Vovk and Wang (2020)). However, this bound is not practical in general applications because, considering $n = 100$ or 1000, the inflation bound $\alpha_n^{HM} \geq \log(n)$ is at least 4.6- or 6.9-fold greater Furthermore, the factor α_n^{HM} is comparable with the type-I error in the case of a perfect correlation (i.e., $\rho = 1$), instead of the nominal size α under independence. On this issue, Goeman et al. (2019) pointed out an extreme case that when $n = 10^5$ and $\mathbf{\Sigma}$ has exchangeable correlation $\rho = 0.2$, HM has a more than threefold type-I error inflation (true type-I error = 0.164 under nominal $\alpha = 0.05$). In Section 4.4.1, we perform extensive simulations for a wide range of n and size α to investigate the limitation and develop practical guidance for applying the HM method.

The discussion above (Remark 4.1) indicates that with a mild normality assumption, the upper bound for the inflation of type-I errors is much smaller than that under an arbitrary dependence structure. This is especially useful, because using a smaller upper bound of the inflation of type I errors to adjust the significance level increases the power of the test. Furthermore, based on Theorem 4.2 and simulations, we can develop a practical guideline to adjust the significance level for the HM test ($\eta = 1$), and for any test that is a sum of transformations by a distribution with a regularly varying tail, including any BC_η test.

4.3.3 Detection Boundary of Regularly Varying Methods

In this subsection, we investigate the power of regularly varying methods by deriving the detection boundary of the test $T(\boldsymbol{X})$ under sparse alternatives as $n \to \infty$ (Theorem 4.3), which is a popular measurement of power performance when detecting weak and sparse signals. Below, we introduce the standard setup of weak and sparse signals by Donoho and Jin (2004), which we refer to in Theorem 4.3.

Consider testing the null hypothesis $H_0 : \boldsymbol{\mu} = (\mu_1, \cdots, \mu_n) = \vec{0}$ for the bivariate normal $\boldsymbol{X}$. For the alternative, we consider the conventional "weak" and "sparse" signals setting in Donoho and Jin (2004) by assuming a small number of the n signals are nonzero with $|\mu_i| = \sqrt{2\tau \log(n)}$, for $i \in S = \{1 \leq i \leq n : \mu_i \neq 0\}$ with $|S| = s$ and $0 < \tau < 1$, and the rest $\mu_i = 0$, for $i \in S^c$. In addition, the sparsity of the signals is of order $s = n^\beta$, with $0 < \beta < \frac{1}{2}$.

Under the above setup, for any fixed value of β, a larger value of τ makes it easier for a method to detect the existence of signals. Indeed, for any given $\beta \in (0, \frac{1}{2})$, Donoho and Jin (2004) reported a threshold effect of τ; the sum of the type-I and type-II errors of a method tends to be zero or one depending on whether τ exceeds the detection boundary $\rho(\beta)$ or not.

For Theorem 4.3, in addition to the setup of Donoho and Jin (2004) and conditions (A2) and (A3), we need two additional conditions:

(C1): We assume $\boldsymbol{X} \overset{D}{\sim} N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and that $\boldsymbol{\Sigma}$ is a banded correlation matrix; i.e., its $(i,j)th$ element $\rho_{ij} = 0$, for any $|i - j| > d_0$, for some positive constant $d_0 > 0$.

(C2): There exist $h \geq 0$ and $t_1 > 0$ such that

$$\frac{1}{t^\gamma (\ln(t))^h} \leq \bar{F}(t) \leq \frac{(\ln(t))^h}{t^\gamma},$$

for all $t > t_1$.

Condition (C2) is for the tail probability of U_i and is a mild condition because $\bar{F}(t) = P(U_i > t) = \frac{L(t)}{t^\gamma}$ ($L(t)$ is a slowly varying function). This condition holds for all commonly used distributions with regularly varying tails with index γ. In the Supplement Section C.1 (Remark C4), we show that the BC, Cauchy, and truncated Cauchy methods all satisfy Condition (C2).

Theorem 4.3. *Under conditions (A2), (A3), (C1), and (C2), for any* $0 < \gamma \leq 1$*, any significance level* $0 < \alpha < 1$*, and* τ *satisfying* $\sqrt{\tau} + \sqrt{\beta} > 1$*, then under the alternative hypothesis, we have*

$$\lim_{n\to\infty} P(T(\boldsymbol{X}) > t_\alpha) = 1,$$

where t_α *is the* p*-value cutoff. That is, the detection boundary for* $T(\boldsymbol{X})$ *is* $\rho(\beta) = (1 - \sqrt{\beta})^2$*.*

Remark 4.2. Under the same conditions of Theorem 4.3, one can show that for $T_{n,\boldsymbol{w}} = \sum_{i=1}^{n} w_i g(p_i)$ with $\boldsymbol{w} \in R_+^n$ and $\sum_{i=1}^{n} w_i = n$, if $\max_i w_i \leq (\log n)^{\eta_1}$ and $\min_i w_i \geq 1/(\log n)^{\eta_2}$ for some fixed constants $\eta_1, \eta_2 > 0$, the result of Theorem 4.3 still holds. See the Supplementary Material, Remark S6 for more details.

Theorem 4.3 states that the power of this test $T(\boldsymbol{X})$ converges to one for any significance level $\alpha > 0$ and $0 < \gamma \leq 1$, or equivalently, that the sum of the Type-I and Type-II errors goes to zero, given the setup. Moreover, Theorem 4.3 implies that the methods with $0 < \gamma \leq 1$ attain the optimal detection boundary defined in Donoho and Jin (2004) in the strong sparsity situation $0 < \beta < 1/4$. Liu and Xie (2020) showed a similar result for their proposed Cauchy test. As discussed in Section 4.2, the Cauchy distribution has a regularly varying tail with index $\gamma = 1$. This theorem is valid for methods of distributions with regularly varying tails with index $0 < \gamma \leq 1$. Therefore, this theorem can be considered a generalization of Theorem 4.3 in Liu and Xie (2020).

4.4 Simulations

In this section, we perform simulations to compare the robustness of different p-value combination methods under varying correlation levels between p-values in order to verify the theoretical results presented in Sections 4.2 and 4.3. We include the seven methods discussed in Section 4.2, minP, $\text{BC}_{1.25}$, CA, CA^{tr}, $HM(BC_1)$, $\text{BC}_{0.75}$, and Fisher's method, as well as HC (Higher criticism) and BJ (Berk-Jones test). Section 4.4.1 first evaluates the type-I error control of the methods under independence and varying levels of correlation to verify the robustness of the HM and Cauchy methods. Furthermore, because the robustness in Theorem 4.2 for HM and Cauchy is an asymptotic result, we further investigate the type-I error control for HM under a wide range of

n, ρ, and γ to ensure that the robustness of HM and Cauchy is preserved and useful in a practical sense. Section 4.4.2 assesses the statistical power under different dependency structures and sparsity of signals in the alternative hypothesis. In Section 4.4.3, we evaluate the improvement of the truncated Cauchy method over the Cauchy method in a discrete data simulation.

4.4.1 Type-I Error Control

In this subsection, we first simulate $n = 100$, $\boldsymbol{X} = (X_1, \cdots, X_n) \overset{D}{\sim} N(0, \boldsymbol{\Sigma})$, $p_i = 2(1 - \Phi(|X_i|))$, and $T = \sum_{i=1}^{n} g(p_i)$ for the various methods. We further assume that $\boldsymbol{\Sigma}$ has unit variance on the diagonal line, and is exchangeable with the common correlation $\rho = \text{cor}(X_i, X_j)$, for $1 \leq i \neq j \leq n$, where ρ is evaluated at 0 (independence), 0.3, 0.6, 0.9, and 0.99. Table 4.1 shows the type-I errors of the nine methods with different levels of correlations at $\alpha = 0.001$ using 10^6 simulations under the null hypothesis. As expected, all methods control the type-I error perfectly under the independence assumption (i.e., $\rho = 0$). When correlations exist between p-values, we find that minP is the most conservative in terms of the type-I error control, followed by $\text{BC}_{1.25}$, as expected from the theoretical result in Corollary 4.1. CA, CA^{tr}, and HM exhibit perfect type-I error control in all correlation settings, showing robustness to the dependency structure. Fisher and BJ are the most anti-conservative methods in the presence of correlation, followed by slight anti-conservativeness for HC and $\text{BC}_{0.75}$.

Note that according to Theorems 4.1 and 4.2 for regularly varying distribution transformation, the tail probability $P(T(\boldsymbol{X}) > t)$ under dependence can be asymptotically approximated by that under independence. However, the asymptotic result guarantees only the dependence robustness for very large t (or equivalently very small α). We also expect that larger n will require a larger t (smaller α) to ensure a good approximation. Specifically, Goeman et al. (2019) noted that with $\rho = 0.2$ and $n = 10^5$, the much inflated type-I error of 0.164 is obtained for size $\alpha = 0.05$. Therefore, it is of interest to explore the robustness property of $T(\boldsymbol{X})$ for dependence in HM for varying n, α, and ρ in order to provide practical guidance in real applications. In Table 4.2, we extend the simulation for HM with $n = (25, 50, 100, 500, 1000, 2000, 10000)$, $\alpha = (0.05, 0.01, 0.001, 0.0001)$, and $\rho = (0, 0.3, 0.6, 0.9, 0.99)$. Given each combination of α and n, we calculate the maximum percent of inflation (PI) across different ρ, which is defined as

Table 4.1: Type-I errors for nine tests: Fisher, CA, CA^{tr} (truncated Cauchy), $BC_{0.75}$, BC_1 (HM), $BC_{1.25}$, minP, HC, and BJ, across correlation level ρ =0, 0.3, 0.6, 0.9, 0.99.

Method/Correlation	ρ=0	ρ=0.3	ρ=0.6	ρ=0.9	$\rho = 0.99$
Fisher	0.0010	0.1160	0.1960	0.2483	0.2610
$BC_{0.75}$	0.0010	0.0016	0.0031	0.0041	0.0043
CA	0.0010	0.0011	0.0013	0.0011	0.0010
CA^{tr}	0.0010	0.0011	0.0013	0.0011	0.0010
BC_1 (HM)	0.0010	0.0011	0.0013	0.0011	0.0010
$BC_{1.25}$	0.0010	0.0010	0.0009	0.0005	0.0004
minP	0.0010	0.0010	0.0007	0.0002	0.00003
HC	0.0010	0.0012	0.0047	0.0173	0.0227
BJ	0.0010	0.0850	0.1744	0.2506	0.2712

PI $= (\max_\rho$ type-I error $- \alpha)/(\alpha)$. The result confirms the theoretical result that a larger n generates greater type-I error inflation under dependence for a fixed α, and requires a much smaller α to improve the type-I error inflation. For example, when $\alpha = 0.01$, we have $PI = 30\%$ for $n = 25$, compared with $PI = 80\%$ for $n = 10,000$. On the other hand, when $n = 10,000$, PI decreases from 80% to 49% when α decreases from 0.01 to 0.0001. In general, this result shows robust type-I error control under varying correlation levels, in a practical sense, when $n \leq 1,000$ and $\alpha \leq 0.05$ with the maximum $PI = 50\%$, which inflates type I error from $\alpha = 0.01$ to 0.015 at $n = 1000$ and $\rho = 0.3$. Even when n increases to 10,000, PI only minimally increases to 80%. When multiple comparisons are needed, such as in GWAS applications, a small α is targeted, and HM achieves robust type-I error control, in general, in a practical sense. However, if a single test is performed with a very large n, we need to be careful with the type-I error inflation (e.g., type-I error is 0.072 for $\alpha = 0.05$ when $n = 10,000$ and $\rho = 0.3$).

Table 4.2: Type-I error control of HM evaluated for the total number of p-values $n = 25$, 50, 100, 500, 1000, 2000, 10000 and $\rho = 0$, 0.3, 0.6, 0.99 for different sizes of test $\alpha = 0.05$, 0.01, 10^{-3}, and 10^{-4}. We also calculate the percent of inflation (PI) to reflect the extent of inflation of the type-I error under various cases, given n and α. PI is defined as PI $= (\max_\rho$ type I error $- \alpha)/\alpha$.

n	ρ	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 10^{-3}$	$\alpha = 10^{-4}$
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.058	0.012	1.06×10^{-3}	1.01×10^{-4}
25	$\rho = 0.6$	0.061	0.013	1.19×10^{-3}	1.11×10^{-4}
	$\rho = 0.9$	0.052	0.011	1.09×10^{-3}	1.08×10^{-4}
	$\rho = 0.99$	0.048	0.010	1.00×10^{-3}	9.93×10^{-5}
	PI	22%	30%	20%	11%
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.057	0.012	1.08×10^{-3}	1.02×10^{-4}
50	$\rho = 0.6$	0.053	0.012	1.23×10^{-3}	1.15×10^{-4}
	$\rho = 0.9$	0.041	0.010	1.08×10^{-3}	1.09×10^{-4}
	$\rho = 0.99$	0.038	0.010	9.99×10^{-4}	1.01×10^{-4}
	PI	14%	20%	23%	15%
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.06	0.012	1.12×10^{-3}	1.04×10^{-4}
100	$\rho = 0.60$	0.053	0.013	1.29×10^{-3}	1.22×10^{-4}
	$\rho = 0.9$	0.040	0.010	1.09×10^{-3}	1.10×10^{-4}
	$\rho = 0.99$	0.037	0.010	1.00×10^{-3}	1.01×10^{-4}
	PI	20%	30%	29%	22%
	$\rho = 0$	0.05	0.010	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.065	0.014	1.20×10^{-3}	1.07×10^{-4}
500	$\rho = 0.6$	0.052	0.013	1.39×10^{-3}	1.32×10^{-4}
	$\rho = 0.9$	0.038	0.010	1.10×10^{-3}	1.11×10^{-4}
	$\rho = 0.99$	0.035	0.010	9.94×10^{-4}	1.01×10^{-4}
	PI	30%	40%	39%	32%
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.068	0.015	1.26×10^{-3}	1.08×10^{-4}
1000	$\rho = 0.6$	0.052	0.014	1.42×10^{-3}	1.35×10^{-4}
	$\rho = 0.9$	0.037	0.010	1.08×10^{-3}	1.09×10^{-4}
	$\rho = 0.99$	0.034	0.010	9.94×10^{-4}	1.00×10^{-4}
	PI	36%	50%	42%	35%
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.069	0.016	1.31×10^{-3}	1.12×10^{-4}
2000	$\rho = 0.6$	0.051	0.018	1.46×10^{-3}	1.40×10^{-4}
	$\rho = 0.9$	0.036	0.010	1.09×10^{-3}	1.11×10^{-4}
	$\rho = 0.99$	0.033	0.009	9.93×10^{-4}	1.01×10^{-4}
	PI	38%	80%	46%	40%
	$\rho = 0$	0.05	0.01	1×10^{-3}	1×10^{-4}
	$\rho = 0.3$	0.072	0.018	1.48×10^{-3}	1.25×10^{-4}
10000	$\rho = 0.6$	0.049	0.014	1.50×10^{-3}	1.49×10^{-4}
	$\rho = 0.9$	0.034	0.010	1.07×10^{-3}	1.12×10^{-4}
	$\rho = 0.99$	0.031	0.009	9.79×10^{-4}	1.01×10^{-4}
	PI	44%	80%	50%	49%

4.4.2 Statistical Power

In this subsection, we follow the simulation setting in Section 4.4.1 to evaluate the statistical power of the methods under different values of correlation ρ and strengths of the signal. Following the sparse and weak signal setting in Donoho and Jin (2004), we design the n signals $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ to contain $n - s$ with no signal ($\mu_{s+1} = \cdots = \mu_n = 0$), and the first s to have nonzero signals $\mu_1 = \cdots = \mu_s = \mu_0 = \frac{\sqrt{4\log(n)}}{s^{0.1}}$, where $s/n = (5\%, 10\%, 20\%)$. We first compare the power of the methods under varying correlations, where the rejection threshold is obtained from the independence assumption, and is uncorrected for dependence. Furthermore, we compare the power of the methods, where the rejection threshold is corrected with precise type-I error control under dependency. Note that the correction applies only in simulations, and is not accessible, in general, without applying extensive permutation tests or simulation-based methods.

4.4.2.1 Power Comparison with an Uncorrected Rejection Threshold from the Independence Assumption

In Section 4.4.1, BJ, HC, $\text{BC}_{0.75}$, and Fisher's method are anti-conservative when using the rejection threshold from the independence assumption. In other words, the methods lose control of the type-I error when a dependence structure exists. As a result, we compare only HM, CA, CA^{tr}, $\text{BC}_{1.25}$, and minP here to evaluate the power of the methods in varying levels of correlation ρ. Table 4.3 shows the power of the five methods. As expected, the statistical power decreases as ρ increases. HM, CA, and CA^{tr} have almost identical power and are superior to $\text{BC}_{1.25}$. minP is the least powerful method among the five. Different proportions of signals give similar patterns and conclusions.

Table 4.3: Mean uncorrected power for tests CA, CA$^{\text{tr}}$ (truncated Cauchy), HM, BC$_{1.25}$, and minP across correlation $\rho = 0, 0.3, 0.6, 0.9, 0.99$ and proportion of signals $s/n = 5\%, 10\%, 20\%$. The standard error is far less than the mean power, and hence is not shown here.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
	CA	0.749	0.629	0.518	0.392	0.347
	CA$^{\text{tr}}$	0.749	0.629	0.518	0.393	0.347
5%	BC$_1$(HM)	0.749	0.629	0.518	0.393	0.347
	BC$_{1.25}$	0.735	0.617	0.505	0.374	0.321
	minP	0.712	0.596	0.482	0.339	0.256
	CA	0.870	0.690	0.533	0.371	0.319
	CA$^{\text{tr}}$	0.870	0.690	0.533	0.371	0.318
10%	BC$_1$(HM)	0.870	0.690	0.533	0.371	0.318
	BC$_{1.25}$	0.850	0.670	0.512	0.342	0.282
	minP	0.814	0.639	0.479	0.292	0.194
	CA	0.955	0.738	0.542	0.353	0.299
	CA$^{\text{tr}}$	0.955	0.738	0.542	0.353	0.299
20%	BC$_1$(HM)	0.954	0.737	0.542	0.353	0.298
	BC$_{1.25}$	0.936	0.712	0.513	0.314	0.250
	minP	0.895	0.670	0.469	0.249	0.145

4.4.2.2 Power Comparison with a Corrected Rejection Threshold Considering the Dependence Structure

Because methods other than CA, CA^{tr}, and HM are either conservative or anti-conservative in terms of type-I error control in the presence of correlation, the power comparison in the previous subsection is not completely fair. Here, we evaluate the power of each method using the rejection threshold corresponding to the accurate type-I error control in each case under each correlation setting. Thus, we obtain the corrected rejection thresholds, considering the dependence structure, for each ρ, and simulate 10^6 Monte Carlo samples for each method using the same sampling procedure as in Section 4.4.1, with assumed correlation. Then, we calculate the empirical rejection threshold from the Monte Carlo samples under the null hypothesis as the critical value for each method.

Note that although this comparison is theoretically a fairer comparison, with accurate type-I error control, it is less practical, unless we know the dependency structure or perform computationally intensive approaches to precisely control the type-I error.

Table 4.4 shows the results for all nine methods. We order the methods by the index η of the Box-Cox transformation, as introduced in Section 4.2: minP, $BC_{1.25}$, HM, CA, CA^{tr}, $BC_{0.75}$, Fisher, and then add HC and BJ for comparison. We first observe almost identical results for CA, CA^{tr}, and HM, and decreasing power when ρ increases, as expected. We next compare the five methods minP, CA/CA^{tr}/HM, and Fisher with varying proportions of signals and ρ. When $\rho = 0$, Fisher is the least powerful when $s/n = 5\%$ (power $= 0.640$), but becomes more powerful than CA/CA^{tr}/HM and minP when $s/n = 10\%$ and 20%, showing its superior performance in frequent signals. CA/CA^{tr}/HM consistently have good power between that of minP and Fisher. When ρ increases, Fisher quickly drops to almost zero power, even with accurate type-I error control. For each given s/n, minP is slightly less powerful than CA/CA^{tr}/HM at small ρ, but becomes much more powerful than CA/CA^{tr}/HM when ρ is large. This is reasonable because at a very high correlation (e.g., $\rho = 0.99$), all signals can be viewed as coming from one source, so taking the smallest p-value gives sufficiently complete information. For $BC_{0.75}$ and $BC_{1.25}$, we observe that, in general, the performance of $BC_{1.25}$ lies between that of minP and CA/CA^{tr}/HM, and that of $BC_{0.75}$ lies between that of CA/CA^{tr}/HM and Fisher. We next compare HC and BJ with the

other methods. Although these two methods lose control of the type-I error under a dependency structure, and are not the focus of this study, we are curious about their power performance if the correlation structure is correctly considered with type-I error control. As shown in Table 4.4, BJ is surprisingly powerful for all three proportions of signals when $\rho = 0$ (e.g., power$= 0.91$ compared with power $= 0.640 - 0.778$ for the other seven methods when $s/n = 5\%$). However, similarly to Fisher's method, the power of BJ drops quickly to almost zero with the existence of dependency. The power of HC is, in general, similar to that of CA/CA$^{\text{tr}}$/HM, but becomes weaker than CA/CA$^{\text{tr}}$/HM for larger ρ. Both HC and BJ lose much power when ρ increases. One possible explanation is that both tests compare the ordered p-values $p_{(i)}$ with the reference value i/n, which is not the correct reference under the null with a dependence structure (Liu and Xie, 2020).

4.4.3 Simulation for the Large Negative Penalty Issue in the Cauchy Method

As discussed in Section 4.2.1, p-values close to one lead to large negative penalties in the Cauchy method, which can cause significant power loss. Below, we design a Fisher's exact (hypergeometric) test for a 2×2 contingency table to illustrate the issue and evaluate the improvement offered by the truncated Cauchy method.

We first evaluate the type-I error, similarly to Section 4.1. We randomly generate $n = 20$ 2×2 contingency tables with fixed row and column margins equal to 200. The table has only one degree of freedom, assuming the upper-left cell of each table is undetermined. Under the null hypothesis, the rows and columns are independent, and we generate the value of the upper-left cell from $Hypergeometric(400, 200, 200)$. We then apply Fisher's exact test to the simulated data of each table, and combine the $n = 20$ p-values using the HM and CA methods. We repeat the simulation 10^5 times, set the significance level at $\alpha = 0.05, 0.01, 0.005, 0.001, 0.0005$, and 0.0001, and calculate the proportions of rejections at each α. As shown in Table 4.5 (effect size $p_{11} = 0$), the type-I errors for HM are slightly smaller than the desired significance level under the null hypothesis (e.g., 0.00077 versus 0.001), whereas those for CA are much lower (e.g., 0.00016 versus 0.001). The main reason for the conservativeness in both tests is that the null distribution under the simulation setting is skewed towards one, instead of $Unif(0, 1)$, in which case CA is more sensitive because it imposes a greater penalty for p-values close to one. As shown in Table

4.5, the type-I error control of CA$^{\text{tr}}$ under $\delta = 0.01$ is largely improved for all α; for example, the type I error is now 0.00077, identical to that of HM, when $\alpha = 0.001$.

We next evaluate the power for HM and CA. Similarly to Section 4.4.2, we simulate 10^5 Monte Carlo samples. All settings are identical to the last paragraph in terms of the type-I error control except that we now generate 2×2 tables with row-column correlations. We first simulate Y from $Hypergeometric(400, 200, 200)$ under the independence assumption. We then simulate $Z \overset{D}{\sim} Bin(200 - Y, p_{11})$, and take $Y + Z$ as the value for the upper-left cell. Note that $p_{11} = 0$ corresponds to the original null hypothesis, and a larger effect size p_{11} means a stronger signal. We set $p_{11} = 0.2$, and 0.3 and the power values under different α are shown in Table 4.5. As expected, a larger p_{11} generates higher power for both HM and CA. CA produces much smaller power than HM, mainly because the p-values are skewed toward one. CA$^{\text{tr}}$ largely alleviates the issue and performs almost identically to HM.

4.5 Application

We apply the HM, CA, CA$^{\text{tr}}$, and minP tests to analyze a GWAS of neuroticism (Okbay et al., 2016), a personality trait characterized by easily experiencing negative emotions. The data set contains $6,524,432$ genetic variants (SNPs) across $179,811$ individuals, and the p-values are calculated for all SNPs to represent the association between the variant and neuroticism. We use genome annotations to locate the genic or intergenic region for each variant. The total number of intergenic and genic regions is $78,895$. Within each genic or intergenic region, we combine the p-values of the variants using the HM, CA, CA$^{\text{tr}}$, and minP methods. Figure 4.3 shows three Manhattan plots for the combined p-values using the HM, CA, and minP methods, respectively. As shown in Figure 4.3, the combined p-values using CA and HM are almost identical, and are slightly more significant than those obtained from minP. The bottom-right plot in Figure 4.3 shows the numbers of significant genic or intergenic regions, with the significance thresholds determined using the Bonferroni procedure (controlling the family-wise error rate at 0.05) and the Benjamini–Hochberg FDR procedure (controlling the false discovery rate at 0.05; the significant threshold is $p_{(k)}$, where k is the largest integer such that $p_{(k)} \leq \frac{0.05k}{n}$), or p-value threshold at 10^{-4}, 10^{-5}, or 10^{-6}. For all

Table 4.4: Mean corrected power for tests Fisher, $BC_{0.75}$, CA, CA^{tr}(truncated Cauchy), HM, $BC_{1.25}$, minP, HC, and BJ across correlations $\rho = 0, 0.3, 0.6, 0.9, 0.99$ and proportions of signals $s/n = 5\%, 10\%, 20\%$. The standard errors are far less than the mean power, and hence are omitted.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
	Fisher	0.640	0.0039	0.0021	0.0017	0.0016
	$BC_{0.75}$	0.778	0.615	0.437	0.308	0.269
	CA	0.749	0.620	0.490	0.387	0.348
5%	CA^{tr}	0.749	0.621	0.490	0.388	0.348
	BC_1(HM)	0.749	0.621	0.491	0.389	0.348
	$BC_{1.25}$	0.735	0.618	0.509	0.438	0.402
	minP	0.712	0.603	0.522	0.532	0.600
	HC	0.760	0.623	0.415	0.216	0.195
	BJ	0.912	0.0015	0.0001	0.001	0.001
	Fisher	0.992	0.013	0.0044	0.003	0.003
	$BC_{0.75}$	0.908	0.689	0.461	0.301	0.258
	CA	0.870	0.680	0.503	0.365	0.320
10%	CA^{tr}	0.870	0.681	0.503	0.366	0.319
	BC_1(HM)	0.869	0.681	0.504	0.366	0.319
	$BC_{1.25}$	0.850	0.672	0.517	0.407	0.361
	minP	0.814	0.646	0.520	0.480	0.514
	HC	0.887	0.691	0.432	0.213	0.206
	BJ	0.998	0.017	0.001	0.001	0.001
	Fisher	1.000	0.0745	0.017	0.009	0.008
	$BC_{0.75}$	0.982	0.752	0.484	0.300	0.255
	CA	0.955	0.728	0.511	0.347	0.299
20%	CA^{tr}	0.955	0.729	0.512	0.348	0.299
	BC_1(HM)	0.955	0.729	0.512	0.349	0.299
	$BC_{1.25}$	0.936	0.713	0.518	0.378	0.329
	minP	0.895	0.678	0.511	0.429	0.436
	HC	0.973	0.749	0.451	0.227	0.231
	BJ	1.000	0.202	0.016	0.008	0.013

Table 4.5: Mean proportion of rejection of CA, HM and CA^{tr} (truncated CA) across $\rho_{11} = 0$ (type I error), 0.2 (power), 0.3 (power). The standard errors are far less than the mean proportion and hence are omitted.

ρ_{11}	Methods/Cutoffs	0.05	0.01	0.005	0.001	5×10^{-4}	10^{-4}
$\rho_{11} = 0$	CA	0.00825	0.00182	0.000862	0.00016	0.0000687	0.0000100
	BC$_1$(HM)	0.0386	0.00894	0.00417	0.00077	0.000334	0.0000487
	CA^{tr}	0.0285	0.00729	0.00417	0.00077	0.0000334	0.0000487
$\rho_{11} = 0.2$	CA	0.333	0.202	0.146	0.0582	0.0408	0.0135
	BC$_1$(HM)	0.863	0.525	0.379	0.154	0.108	0.0357
	CA^{tr}	0.848	0.522	0.377	0.154	0.108	0.0361
$\rho_{11} = 0.3$	CA	0.431	0.428	0.420	0.355	0.310	0.190
	BC$_1$(HM)	1.000	0.992	0.972	0.822	0.717	0.440
	CA^{tr}	1.000	0.991	0.971	0.822	0.716	0.440

significance thresholds, the numbers of statistically significant genes for HM and CA are almost identical, and are larger, in general, than those from minP. In particular, HM and CA both identify 750 regions under FDR$= 5\%$, whereas minP finds only 476 regions.

We input the 750 regions identified by HM/CA under FDR $= 5\%$ into the Ingenuity Pathway Analysis package for pathway enrichment analysis. The top enriched pathways include NEUROD1 and NEUROG2, which are transcription factors with important functions in neurogenesis. The top diseases and causal networks identify "neurological disease", which is related to neuroticism. In contrast, by applying the pathway analysis to the top 456 regions identified using minP, we do not find enriched pathways potentially related to neuroticism. The top causal network is MKNK1, which has not been found to play a role in neurological functions.

We next investigate two regions, SLC2A9 and PCSK6, with small combined p-values, using HM $p = 9.534 \times 10^{-4}$ for SLC2A9 and $p = 1.527 \times 10^{-3}$ for PCSK6; q = 0.0759 for SLC2A9 and q = 0.0939 for PCSK6), but not using CA ($p = 0.9999$ and 0.9999 and q-values both equal one). The SLC29A9 gene has been found to be related to Alzheimer's disease, and PCSK6 is related to structural asymmetry of the brain and handedness. We suspect the difference between the results of HM and CA is because the p-values are close to one as described in Section 4.4.3. Figure C2 shows two jitter plots of the p-values for the SNPs in genes SLC2A9 (right) and PCSK6

(left). Both genes contain multiple SNPs with very small p-values (e.g., 17 SNPs with $p < 10^{-4}$ in SLC2A9, and eight SNPs for PCSK6), thus, the gene regions could be significant. However, both genes also contain many SNPs with p-values close to one (five SNPs with $p > 0.99$ for SLC2A9, and nine SNPs for PCSK6), CA is affected and produces larger combined p-values than those of HM, a situation similar to that described in Section 4.4.3. Because there are more than 500 p-values to combine for both genes, by applying CA^{tr} at $\delta = 0.99$ with an approximation by GCLT (Proposition S1), the p-values improve to 9.531×10^{-4} for SLC2A9 and 1.532×10^{-3} for PCSK6, which are almost identical to the p-values calculated by HM.

4.6 Discussion

We have investigated methods for combining dependent p-values using transformations corresponding to regularly varying distributions, which is a rich family of heavy-tailed distributions, and includes the Pareto distribution (Box-Cox transformation) as a special case. We first present the aggregating of multiple p-values in three major historical scenarios: (1) a classical meta-analysis of combining independent and frequent signals (e.g., Fisher), (2) methods for aggregating independent weak and sparse signals (e.g., minP, higher criticism, and Berk-Jones), and (3) recent methods for combining p-values with sparse signals and an unknown dependency structure (i.e., Cauchy and harmonic mean). We then examine popular methods designed for these three settings under the Pareto and regularly varying distributions to provide theoretical insight. Lastly, we present the condition that heavy-tailed transformation methods be robust to the dependency structure.

Our results contribute to the literature in four ways. First, in Section 4.2, we use the family of Box-Cox transformations, or equivalently, transformations by the CDF of Pareto distributions, to connect the Fisher, CA, HM, and minP methods, which are designed to specialize in the three scenarios. We also show that two recent methods, CA and HM, are approximately identical. Second, in Section 4.3, we focus on the dependent p-value scenario, and investigate the condition that p-value combination methods with regularly varying distributions be robust to the dependency structure, where CA and HM are special cases. We show that only methods of the equivalent class of CA and HM (i.e., index $\gamma = 1$) in the regularly varying distributions have the robustness prop-

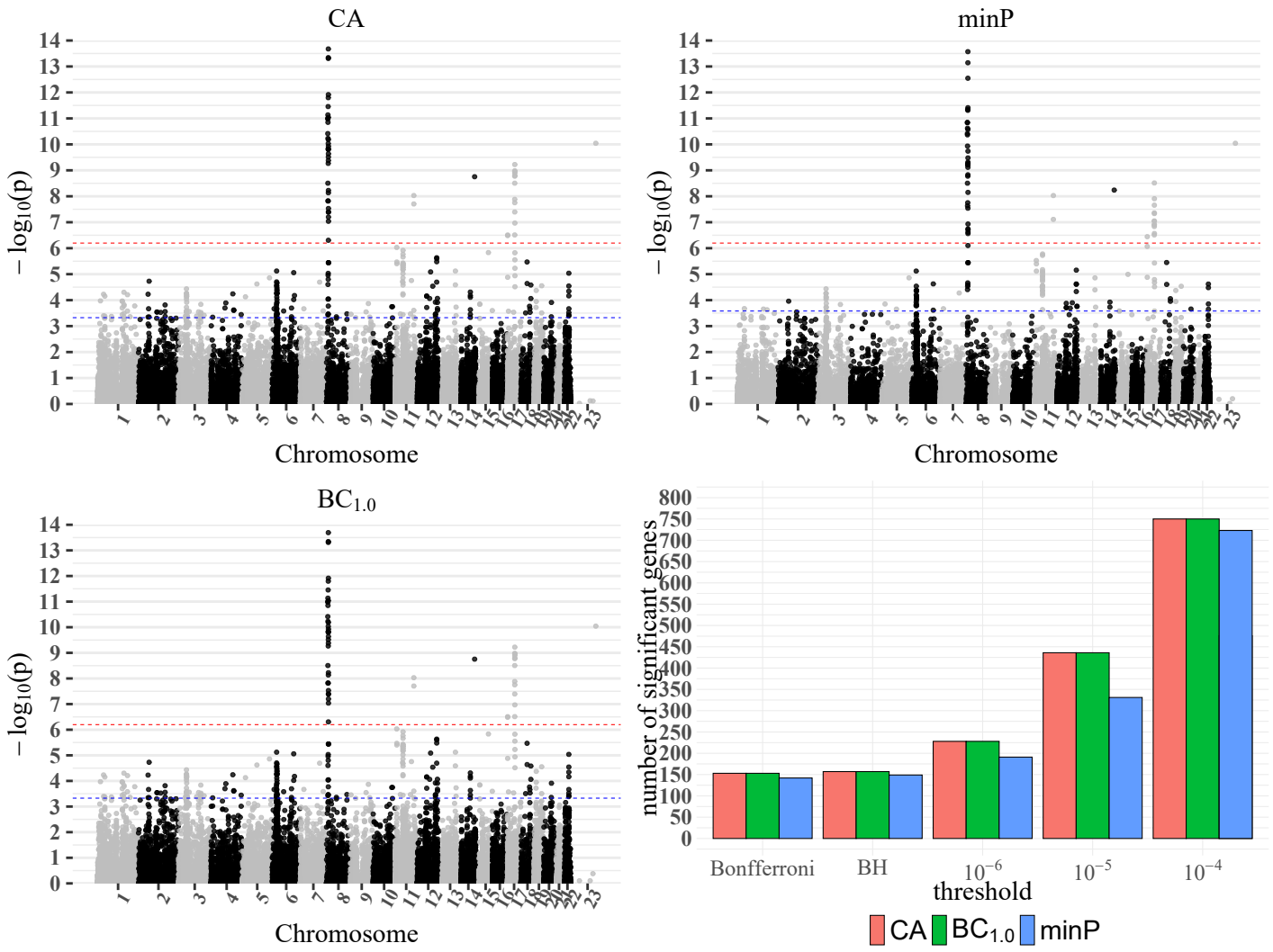


Figure 4.3: Mahattan plots and number of significant p-values for CA, BC_1(HM), and minP. The red dash lines are the cutoffs of the Bonferroni correction for $\alpha = 5\%$, and the blue dash lines are the cutoffs of the Benjamini-Hochberg correction for FDR = 5%. The significant regions (FDR = 5%) detected by HM and CA are the same, except for two regions, DDX58 (q = 0.0499 by CA and q = 0.0501 by HM) and POU2F3 (q = 0.0509 by CA and q = 0.0492 by HM).

erty. Third, we demonstrate an occasional drawback of the Cauchy method when some p-values are close to one, which contributes to the large negative penalty and causes a loss of power. We propose a simple, yet practical solution using a truncated Cauchy method with fast and accurate computation. Finally, the simulations and a real GWAS application confirm our theoretical insights, and provide a practical guideline for using the harmonic mean and Cauchy methods. Specifically, Table 4.2 in Section 4.4.1 shows the degree of possible type-I error inflation of the harmonic mean method under varying n (number of combined p-values), ρ (correlation level between p-value), and α (test size).

Modern data science faces challenges from larger data dimensions, increased structural complexity, and the need for models and inference tailored to subject domains. The three categories of p-value combination methods have motivated the development of numerous methods, and is a good example of how statistical theories can provide insight into method development and a guide toward real applications. We conclude that the condition that regularly varying distributions must be robust to the dependency structure when combining p values is satisfied by those distributions with index $\gamma = 1$, which includes the Cauchy and harmonic mean methods. In future research, we would like to determine whether other methods (e.g., the inverse gamma or log-gamma families) that satisfy this condition may enjoy robustness and obtain better statistical power in some applications of interest.

5.0 Future Directions

There are several directions for the three projects in Chapters 2-4. In Chapter 2, we consider the input p-values to be independent with each other. However, dependence structure between a small group of p-values are common in real data practice (Brown, 1975). It is of great interest to consider dependency structure under the scenario. Similarly, it is of great interest to investigate the performance of AFg under dependency in Chapter 3. For regularly varying methods in Chapter 4, it is of great interest to develop a fully data-driven procedure to determine the choice of index γ. In addition, the non-asymptotic rate for approximating the tail probability of regularly varying methods under certain dependency structures between p-values is also of great interest.

Appendix A Supplementary Materials for Chapter 2

A.1 Supplementary Theoretical Results

A.1.1 Asymptotic Efficiencies of P-Value Combination Methods

In this subsection, we outline the asymptotic efficiencies of multiple p-value combination methods mentioned in Section 2: Fisher, Stouffer, Pareto, Cauchy (CA), Berk-Jones (BJ) and higher criticism (HC). All technical proofs are left to Section A.2. For Fisher test, combined with Lemma 2.1 and by almost the same argument in Littell and Folks (1973), one can show that Fisher test attains ABO in the modified partial signal setting. Similarly, by similar argument as above, the exact slope of Stouffer is

$$C_{\text{Stouffer}}(\vec{\theta}) = \frac{1}{K}\Big[\sum_{i=1}^{\ell}(\lambda_i c_i(\theta_i))^{\frac{1}{2}}\Big]^2.$$

Although generally $C_{\text{Stouffer}} \leq \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$ and Stouffer is not ABO, Stouffer becomes ABO when all the p-values contain true signals with equal effects $\lambda_1 c_1(\theta_1) = \ldots = \lambda_K c_K(\theta_K) > 0$. Theorem A1 below describes the asymptotic efficiency property of Fisher and Stouffer.

Theorem A1 (extended from Littell and Folks (1973); Fisher is ABO and Stouffer is generally not ABO)**.** *Under the setup in Section 2.1, Fisher is ABO with exact slope* $C_{\textit{Fisher}}(\vec{\theta}) = \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$. *Stouffer is generally not ABO with exact slope* $C_{\textit{Stouffer}}(\vec{\theta}) = \frac{1}{K}\Big[\sum_{i=1}^{\ell}(\lambda_i c_i(\theta_i))^{\frac{1}{2}}\Big]^2$. *Stouffer is ABO when all signals combined have equal sample-size adjusted exact slope:* $\lambda_1 c_1(\theta_1) = \ldots = \lambda_K c_K(\theta_K) > 0$.

We next study two methods by heavy-tailed distribution transformation, Pareto and CA, as follows

$$T_{\text{Pareto}}(\eta) = \sum_{i=1}^{K}\frac{1}{p_i^{\eta}} \text{ with some } \eta > 0, \quad T_{\text{CA}} = \frac{1}{K}\sum_{i=1}^{K}\tan(\pi(\frac{1}{2} - p_i)).$$

Methods of this category are in the form of statistics $T(\vec{p}) = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} F_U^{-1}(1-p_i)$ to sum up transformed p-values, where the transformation $g(p)$ is the inverse CDF of U. Indeed, for Cauchy, $U \overset{D}{\sim} \text{CAU}(0,1)$ (standard Cauchy), and $U \overset{D}{\sim} \text{Pareto}(\frac{1}{\eta}, 1)$ for Pareto. Intuitively, methods

by light-tailed distribution transformations (e.g., Stouffer and Fisher) achieve better asymptotic efficiency as a thin-tailed distribution generates more comparable contributions from marginally significant p-values with frequent signals, while methods by heavy-tailed distribution focus more on the extreme effects and downweigh the frequent small effects. For example, Stouffer test transforms p-values 10^{-2} and 10^{-6} to 2.32 and 4.75 while $\tan(\pi(1/2-p))$ for Cauchy test transforms the same p-values to 31.82 and 3.82×10^5, which makes the contribution from very small p-value (10^{-6}) dominate that from the moderate one (10^{-2}). The following two theorems show that CA and Pareto are generally not ABO and they are ABO if and only if there is only one true signal among p-values.

Theorem A2. *Under the setup in Section 2.1,* $T_{Pareto}(\eta)$ *is generally not ABO with exact slope* $C^{(\eta)}_{Pareto}(\vec{\theta}) = \max_{1\le i\le K} \lambda_i c_i(\theta_i)$.

Theorem A3. *Under the setup in Section 2.1,* T_{CA} *is generally not ABO with exact slope* $C_{CA}(\vec{\theta}) = \max_{1\le i\le K} \lambda_i c_i(\theta_i)$.

Both exact slopes of CA and Pareto are $\max_{1\le i\le K} \lambda_i c_i(\theta_i)$, which is also the exact slope of minP, shown in Littell and Folks (1973). This suggests that CA and Pareto are more powerful for detecting sparse signals as minP.

We continue investigating the asymptotic efficiencies of BJ and HC, which can be viewed as goodness-of-fit tests:

$$T_{\text{HC}} = \max_{1\le i\le K} \sqrt{K}\frac{i/n - p_{(i)}}{\sqrt{p_{(i)}\left(1-p_{(i)}\right)}}$$
$$T_{\text{BJ}} = \max_{1\le i\le K} \mathrm{I}_{\{p_{(i)}<\frac{i}{K}\}}\left[\frac{i}{K}\log\left(\frac{i/K}{p_{(i)}}\right) + \left(1-\frac{i}{K}\right)\log\left(\frac{1-i/K}{1-p_{(i)}}\right)\right].$$

As goodness-of-fit tests, both test statistics can test whether the underlying distribution is $\text{Unif}(0,1)$ given K independent observed p-values $p_1,\ldots,p_K$. Both BJ and HC are mainly applied to the scenarios of detecting weak and sparse signals (Donoho and Jin, 2004; Berk and Jones, 1979; Li and Siegmund, 2015). The following theorem shows that BJ is generally not ABO.

Theorem A4. *Under setup in Section 2.1, T_{BJ} is not ABO with exact slope*

$$C_{BJ}(\vec{\theta}) = \max_{1\le i\le K} i\lambda_i c_i(\theta_i).$$

The following proposition shows that HC generally is not ABO even for combining two p-values with equal effects.

Proposition A1. *Suppose p_1 and p_2 are two independent p-values such that*

$$-\frac{2}{n}\log(p_i) \to c_i(\theta_i) \text{ as } n \to \infty \text{ for } i = 1, 2,$$

with probability one. Then for $c_1(\theta_1) = c_2(\theta_2) = c_0 > 0$, T_{HC} is not ABO with exact slope $C_{HC}(\vec{\theta}) = c_0$.

A.1.2 Type I Error Control of FE and FE_{CS}

In this subsection, we provide more details on the type I error control of FE and FE_{CS} computation using the Pareto$(1, 1)$ distribution. Assume X follows Pareto$(1, 1)$. As suggested by Theorems 1 and 2 in Fang et al. (2023a), under the null, the upper tail of distribution of the average of $1/p_1, \ldots, 1/p_L$ with unknown dependence structure can be approximated by that of Pareto$(1, 1)$, in a sense that for sufficiently large $t > 0$ (corresponding to sufficiently small significant level α),

$$\frac{\mathbb{P}\big(\frac{1}{L}\sum_{i=1}^{L} 1/p_i > t\big)}{\mathbb{P}\big(X > t\big)} \approx 1. \tag{A1}$$

Hence for FE and FE_{CS} respectively, one can show that for sufficiently large $t > 0$ (corresponding to sufficiently small α),

$$\frac{1 - F_{T_{\text{FE}}(t)}}{1 - F_{\text{Pareto}(1,1)}(t)} \approx 1$$
$$\frac{1 - F_{T_{\text{FE}_{\text{CS}}}(t)}}{1 - F_{\text{Pareto}(1,1)}(t)} \approx 1,$$

which justifies the type I error control procedures for FE and FE_{CS} using Pareto$(1, 1)$, respectively. Table A1 in Section A.3.1 numerically justifies accuracy of the above fast-computing procedure, where we show that type I error control procedures for FE and FE_{CS} are accurate for $\alpha = 0.0001 \sim$

0.05 across a broad range of K (5 to 100). Note $1 - F_{\text{Pareto}(1,1)}(t) = 1/t$, combined with equation (A1), one can show that the above procedures are equivalent to directly using

$$\frac{L}{\sum_{i=1}^{L} 1/p_i}$$

as p-value for statistic $\frac{1}{L}\sum_{i=1}^{L} 1/p_i$, which is suggested by Wilson (2019a).

A.2 Technical Arguments

In this section, we present the technical arguments for proving the theoretical results. For any random variable X with CDF F, the corresponding survival function is denoted by $\bar{F} = 1 - F(t)$. For two positive functions $u(\cdot)$ and $v(\cdot)$, we denote by $u(t) \sim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} = 1$. Also, $u(t) \gtrsim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} > 1$ and $u(t) \lesssim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} < 1$.

A.2.1 Proofs of Results of Modified Fisher's Methods: Lemma 2.1 and Theorems 2.1-2.6

In this subsection, we prove Theorems 2.1-2.6. Before proceeding to the proofs, we first prove Lemma 2.1 and introduce Lemmas A1- A3.

Proof of Lemma 2.1. For $\theta \in \Theta_0$, note that $-\log p^{(n)}$ follows exponential distribution with rate parameter 1 (denoted by EXP(1)) since the p-value p_n is distributed uniformly in $(0, 1)$. Consider the sequence of random variables $Y_1, Y_2, \ldots$, where $Y_1, Y_2 \ldots$, identically follow EXP(1). Define event $A_n = \{\frac{Y_n}{n} > \frac{\log(n(n+1))}{n}\}$. Then since $\sum_{i=1}^{+\infty} \mathbb{P}(A_n) < \infty$, by the Borel–Cantelli lemma, we have $\mathbb{P}\left(\limsup_{n\to+\infty} A_n\right)$=0. Hence $\frac{Y_n}{n}$ converges to zero with probability one. □

Lemma A1 (Bahadur et al. (1960))**.** *Let $F_{\chi_k}(x) = \mathbb{P}(\chi_k \le x)$, where $\chi_k = \sqrt{\chi_k^2}$ and χ_k^2 follows chi-squared distribution with degrees of freedom k. Then $\log(\bar{F}_{\chi_k}(x)) \to -\frac{1}{2}x^2(1 + o(1))$ as $x \to \infty$.*

Lemma A2 (Savage (1969))**.** *Suppose $\left\{T^{(n)}\right\}$ is a sequence of test statistics which satisfies the following two properties:*

1. *There exists a function $b(\theta), 0 < b(\theta) < \infty$, such that $T^{(n)}/\sqrt{n} \to b(\theta)$ with probability one.*

2. *There exists a function $f(t), 0 < f(t) < \infty$, which is continuous in some open set containing the range of $b(\theta)$ such that for each t in the open set:*

$$-\frac{1}{n}\log\left[1 - F_n(\sqrt{n}t)\right] \to f(t),$$

where F_n is the continuous CDF function of some random variable X_n.

Then

$$-\frac{2}{n}\log\left[1 - F_n(T^{(n)})\right] \to 2f(b(\theta))$$

with probability one.

Remark A1. The condition $0 < f(t) < \infty$ implicitly puts restrictions on the choice of X_n (corresponding to F_n). For example, the rate of the upper tail of X_n should not be too fast. Indeed, $X_n \overset{D}{\sim} \text{Unif}(0,1)$ leads to a too fast rate ($F_{\text{Unif}(0,1)}(\sqrt{n}t) = 0$ for any $\sqrt{n}t > 1$), resulting in $f(t) = +\infty$ that clearly does not satisfy the conditions of Lemma A2.

Remark A2. When F_n is the CDF of $T^{(n)}$, Lemma A2 becomes Theorem 1 in Littell and Folks (1973), which will be used in the proof of Theorem A2; We will use Lemma A2 in the proofs of Theorems 2.1 and 2.3 to 2.6, where $F_n = F$ for some F and all n.

Lemma A3. *Under the setup in Section 2.1, define the following two index sets:*

$$\mathcal{D}^* = \{i : c_i(\theta_i) > 0\};\ \hat{\mathcal{D}} = \left\{i : p_i \le p_{(\ell)}\right\}.$$

Then we have, as $n \to \infty$, $\hat{\mathcal{D}} \to \mathcal{D}^*$ with probability one. And if $\lambda_i c_i(\theta) > \lambda_{i'} c_{i'}(\theta) > 0$, $\frac{p_{i'}}{p_i} \to +\infty$ with probability one.

Proof. To prove the first claim, first denote $\mathcal{D}^{*c} = \{i : i \notin \mathcal{D}^*\}$. For any $i' \in \mathcal{D}^{*c}$, and any $i \in \mathcal{D}^*$, by Lemma 2.1, we have $\log\left(p_{i'}/p_i\right)/n \to \lambda_i c_i(\theta_i) - 0 > 0$ with probability one. Hence p_i is smaller order of $p_{i'}$ as $n \to \infty$, which completes the proof. For the second claim, simply note for any $\lambda_i c_i(\theta) > \lambda_{i'} c_{i'}(\theta) > 0$, $\log\left(p_{i'}/p_i\right)/n \to \lambda_i c_i(\theta) - \lambda_{i'} c_{i'}(\theta) > 0$ with probability one. Then the result follows. □

Corollary A1. *Under the alternative in Section 2.1, with probability one, we have*

$$-\frac{2}{n}\sum_{i=1}^{j}\log p_{(i)} \to \begin{cases} \sum_{i=1}^{j}\lambda_i c_i(\theta_i) & 1 \le j \le \ell \\ \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i) & \ell < j \le K. \end{cases}$$

Proof. Combine the results of Lemmas 2.1 and A3, the results follow. □

The proof of Theorem 2.1 below will use the second equivalent form of AFs to derive the exact slope:

$$T'_{\text{AFs}} = \min_{\vec{w}} \bar{F}_{\chi^2_{2(\sum_{i=1}^K w_i)}}(-2\sum_{i=1}^K w_i \log p_i),$$

where $\vec{w} = (w_1, \ldots, w_K) \in \{0,1\}^K$ is the vector of binary weights that identify the candidate subset of p-values with true signals. In addition, denote

$$\hat{\vec{w}} = \operatorname*{argmin}_{\vec{w}} \bar{F}_{\chi^2_{2(\sum_{i=1}^K w_i)}}(-2\sum_{i=1}^K w_i \log p_i),$$

and $\vec{w}^* = \{\vec{w} : w_k = 1 \text{ if } \theta_i \in \Theta_0 \text{ or } 0 \text{ if } \theta_i \in \Theta_0\}$ as the weight vector identifying the true signals. Also denote $\hat{\vec{w}} = (\hat{w}_1, \ldots, \hat{w}_K)$. For the original form

$$T_{\text{AFs}} = \max_{1\le j\le K} -\log(\bar{F}_{\chi^2_{2j}}(-2\sum_{i=1}^j \log p_{(i)})),$$

we denote correspondingly $\hat{j} = \text{argmax}_j -\log \bar{F}_{\chi^2_{2j}}\big(-2\sum_{i=1}^j \log p_{(i)}\big)$. Since $p_1, \ldots, p_K$ are independent with each other, we have

$$-2\sum_{i=1}^{\hat{j}} \log p_{(i)} = -2\sum_{i=1}^K \hat{w}_i \log p_i.$$

Proof of Theorem 2.1. Denote

$$T(\vec{w};\vec{p}) = -2\sum_{i=1}^K w_i \log p_i$$
$$L(T(\vec{w};\vec{p})) = \bar{F}_{\chi^2_{2d(\vec{w})}}(T(\vec{w};\vec{p})),$$

where $d(\vec{w}) = \sum_{i=1}^K w_i$ and $\vec{p} = (p_1, \ldots, p_K)$. Essentially, $T'_{\text{AFs}} = \min_{\vec{w}} L(T(\vec{w};\vec{p}))$. Further denote by $L_{\text{obs}} = \min_{\vec{w}} L(T(\vec{w};\vec{p}_{\text{obs}}))$ the observed value of T'_{AFp}.

Let $\mathbb{P}_0$ be the probability measure of $\vec{p} = (p_1, \ldots, p_K)$ under the null and U_{AFs} be the random variable that follows the same distribution of T'_{AFs} under the null. For any fixed $\vec{w}$, denote

by $U(\vec{w},\vec{p})$ the random variable follows the same distribution of $\bar{F}_{\chi^2_{2d(\vec{w})}}(T(\vec{w};\vec{p}))$ under the null. Further denote $\Omega_j = \{\vec{w} : d(\vec{w}) = j\}$ for $j = 1,\ldots,K$. Then we have:

$$\begin{aligned} p_{\text{AFs}} = F_{U_{\text{AFs}}}(L_{\text{obs}}) &= 1 - \bar{F}_{U_{\text{AFs}}}(L_{\text{obs}}) \\ &= 1 - \mathbb{P}_0\Big(\bigcap_{j=1}^{K}\bigcap_{\vec{w}\in\Omega_j} U(\vec{w},\vec{p}) \geq L_{\text{obs}}\Big). \end{aligned} \tag{A2}$$

By Bonferroni's inequality,

$$\begin{aligned} (A2) &\leq 1 - \Big[1 - \sum_{j=1}^{K}\mathbb{P}_0\Big(\bigcup_{\vec{w}\in\Omega_j} U(\vec{w},\vec{p}) \leq L_{\text{obs}}\Big)\Big] \\ &\leq \sum_{j=1}^{K}\sum_{\vec{w}\in\Omega_j} \bar{F}_{\chi^2_{2j}}\big(\bar{F}^{-1}_{\chi^2_{2j}}\big(L_{\text{obs}}\big)\big) = \big(2^K - 1\big)\, L_{\text{obs}}, \end{aligned} \tag{A3}$$

where $\bar{F}^{-1}_{\chi^2_{2j}}(\alpha)$ denotes the $1-\alpha$ quantile of χ^2_{2j}. Further note

$$\begin{aligned} -\frac{2}{n}\log\big(L_{\text{obs}}\big) &= -\frac{2}{n}\log\big(1 - F_{\chi^2_{2d(\hat{\vec{w}})}}\big(-2\sum_{i=1}^{K}\hat{w}_i\log p_i\big)\big) \\ &= -\frac{2}{n}\log\bar{F}_{\chi_{2d(\hat{\vec{w}})}}\big(\big(-2\sum_{i=1}^{K}\hat{w}_i\log p_i\big)^{\frac{1}{2}}\big). \end{aligned}$$

Since $1 \leq d(\hat{\vec{w}}) \leq K$,

$$\begin{aligned} -\frac{2}{n}\log\bar{F}_{\chi_{2K}}\big(\big(-2\sum_{i=1}^{K}\hat{w}_i\log p_i\big)^{\frac{1}{2}}\big) &\leq -\frac{2}{n}\log\big(L_{\text{obs}}\big) \\ &\leq -\frac{2}{n}\log\bar{F}_{\chi_2}\big(\big(-2\sum_{i=1}^{K}\hat{w}_i\log p_i\big)^{\frac{1}{2}}\big). \end{aligned} \tag{A4}$$

Denote $\hat{j} = \text{argmax}_j - \log\bar{F}_{\chi^2_{2j}}\big(-2\sum_{i=1}^{j}\log p_{(i)}\big)$, as $p_1,\ldots,p_K$ are independent with each other, we have

$$-2\sum_{i=1}^{K}\hat{w}_i\log p_i = -2\sum_{i=1}^{\hat{j}}\log p_{(i)}.$$

In the proof of Theorem 2.4 latter, we will show that $\hat{j} \geq \ell$ with probability one (equation (A11) in the proof of Theorem 2.4). Hence by Corollary A1, under the alternative, we have

$\frac{\sqrt{-2\sum_{i=1}^{K}\hat{w}_i \log p_i}}{\sqrt{n}} = \frac{\sqrt{-2\sum_{i=1}^{\hat{j}}\log p_{(i)}}}{\sqrt{n}} \to (\sum_{i=1}^{\ell}\lambda_i c_i(\theta_i))^{\frac{1}{2}}$ with probability one. Further combined with and Lemmas A1 and A2 and equation (A4), we have

$$-\frac{2}{n}\log\left(L_{\text{obs}}\right) \to \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i).$$

Combined with (A3), we have under the alternative

$$-\frac{2}{n}\log p_{\text{AFs}} \geq \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$$

with probability one. Then the result follows.

Proof of Theorem 2.2. Note that by Theorem A1, we have that Fisher is ABO with exact slope $C_{\text{Fisher}}(\vec{\theta}) = \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)$, then combine with Theorem 2.6 in Berk and Jones (1978), the result follows. □

□

Proof of Theorem 2.3. Case when $\ell \geq 2$:
Assume $j^* = \text{argmax}_j \frac{\sum_{i=1}^{j}\lambda_i c_i(\theta_i)}{B_j}$. We first prove that under alternative,

$$\underset{j}{\text{argmax}}\, T_A \to j^* \tag{A5}$$

with probability one as $n \to \infty$. Indeed, for $\forall j' \neq j^*$, suppose the following event holds:

$$\begin{aligned}
&\frac{-2\sum_{i=1}^{j'}\log p_{(i)} - A_{j'}}{B_{j'}} > \frac{-2\sum_{i=1}^{j^*}\log p_{(i)} - A_{j^*}}{B_{j^*}} \\
&\Leftrightarrow -2B_{j^*}\sum_{i=1}^{j'}\log p_{(i)} + A_{j^*}B_{j'} > -2B_{j'}\sum_{i=1}^{j^*}\log p_{(i)} + A_{j'}B_{j^*} \\
&\Leftrightarrow \frac{-2B_{j^*}\sum_{i=1}^{j'}\log p_{(i)}/n + A_{j^*}B_{j'}/n}{-2B_{j'}\sum_{i=1}^{j^*}\log p_{(i)}/n + A_{j'}B_{j^*}/n} > 1.
\end{aligned} \tag{A6}$$

Here without loss of generality we assume both $A_{j^*}B_{j'}$ and $A_{j'}B_{j^*}$ in the second inequality are positive. Otherwise one can always move the smaller term to the other side of the inequality and

still use almost the same arguments as follows. However, under the setup in Section 2.1, note that by Lemmas 2.1 and A3 and $j^* = \text{argmax}_j \frac{\sum_{i=1}^{j} \lambda_i c_i(\theta_i)}{B_j}$, under the alternative we have

$$\frac{-2B_{j^*} \sum_{i=1}^{j'} \log p_{(i)}/n + A_{j^*} B_{j'}/n}{-2B_{j'} \sum_{i=1}^{j^*} \log p_{(i)}/n + A_{j'} B_{j^*}/n} \to \frac{B_{j^*}}{B_{j'}} \cdot \frac{\sum_{i=1}^{j'} \lambda_i c_i(\theta_i)}{\sum_{i=1}^{j^*} \lambda_i c_i(\theta_i)} < 1$$

as $n \to \infty$ with probability one, which contradicts to equation (A6). Hence (A5) holds. Let U_{A} be the random variable that follows the same distribution of T_A under the null. Denote by F_{U_A} the CDF of the U_{A} and $\bar{F}_{U_A} = 1 - F_{U_A}$ as the corresponding survival function, respectively. Similarly, for the following test statistic

$$T_{A_j} = \frac{-2 \sum_{i=1}^{j} \log p_{(i)} - A_j}{B_j},$$

let U_{A_j} be the random variable that follows the same distribution of T_{A_j} under the null. And define $F_{U_{A_j}}$ and $\bar{F}_{U_{A_j}}$ as the CDF and survival function of U_{A_j}, respectively. Furthermore, define the test statistic

$$T_j = -2 \sum_{i=1}^{j} \log p_{(i)}$$

and U_j as the random variable that follows the same distribution of T_j under the null and let F_{U_j} and $\bar{F}_{U_j}$ be the CDF and survival function of U_j, respectively. Pick $j = 1$, then we have:

$$\bar{F}_{U_A}(T_A) \geq \bar{F}_{U_{A_1}}(T_A) = \bar{F}_{U_1}(B_1 T_A + A_1),$$

Denote $T^{(n)} = \sqrt{B_1 T_A + A_1}$, with (A5) holds, by Lemmas 2.1 and A3, under the alternative, we have,

$$\frac{T^{(n)}}{\sqrt{n}} = n^{-\frac{1}{2}} (B_1 T_A + A_1)^{\frac{1}{2}} \to \left[(B_1/B_{j^*}) \sum_{i=1}^{j^*} \lambda_i c_i(\theta_i) \right]^{\frac{1}{2}}$$

with probability one. Note for $t > 0$ we have

$$\bar{F}_{\chi_2}(\sqrt{n}t) = \bar{F}_{\chi_2^2}(nt^2) \leq \bar{F}_{U_1}(nt^2) \leq \bar{F}_{\chi_{2K}^2}(nt^2) = \bar{F}_{\chi_{2K}}(\sqrt{n}t).$$

Hence by Lemma A1,

$$-\frac{1}{n} \log \bar{F}_{U_1}(nt^2) = -\frac{1}{n} \log \bar{F}_{\sqrt{U}_1}(\sqrt{n}t) \to \frac{t^2}{2}.$$

Hence by Lemma A2, under the alternative, we have

$$-\frac{2}{n}\log\left(\bar{F}_{U_A}(T_A)\right) \leq -\frac{2}{n}\log\bar{F}_{U_1}\left((T^{(n)})^2\right) = -\frac{2}{n}\log\bar{F}_{\sqrt{U}_1}(T^{(n)})$$
$$\to \frac{B_1}{B_{j^*}}\sum_{i=1}^{j^*}\lambda_i c_i(\theta_i) < \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i) \qquad \text{(A7)}$$

with probability one. Here the last inequality is due to $\ell \geq 2$ and B_j is a strictly increasing function. Hence T_A is still not ABO.

Case when $\ell = 1$:

First we prove that $j^* \to 1$ with probability one under the alternative. Note here we assume $\ell = 1$ and B_j increases as j increases. Note that $c_i(\theta_i) = 0$ with probability one for all $i > 1$, hence

$$\max_j \frac{\sum_{i=1}^{j}\lambda_i c_i(\theta_i)}{B_j} \to \frac{\lambda_1 c_1(\theta)}{B_1} \qquad \text{(A8)}$$

with probability one. Hence $j^* \to 1$ with probability one. Then we have:

$$\bar{F}_{U_A}(T_A) \leq \sum_{j=1}^{K}\bar{F}_{U_{A_j}}(T_A) = \sum_{j=1}^{K}\bar{F}_{T_j}\left(B_j T_A + A_j\right) \leq K\cdot\bar{F}_{\chi^2_{2K}}\left(B_1 T_A + \min_j A_j\right).$$

By combining (A8) and Lemmas 2.1 and A3, under the alternative, we have

$$\frac{\sqrt{B_1 T_A + \min_j A_j}}{\sqrt{n}} \to \sqrt{\lambda_1 c_1(\theta)}$$

with probability one. And by Lemma A1

$$-\frac{1}{n}\log\left(1 - F_{\chi_{2K}}(\sqrt{n}t)\right) \to \frac{1}{2}t^2.$$

In addition,

$$-\frac{2}{n}\log\bar{F}_{U_A}(T_A) \geq -\frac{2}{n}\Big[\log\bar{F}_{\chi^2_{2K}}\left(B_1 T_A + \min_j A_j\right) + \log K\Big]. \qquad \text{(A9)}$$

Hence by Lemma A2, under alternative, $(A9) \to \lambda_1 c_1(\theta)$ with probability one. Then we conclude that when $\ell = 1$, T_A is ABO. □

Remark A3. It can be shown that T_A generally does not has signal selection consistency. Recall that T_A picks $j^* = \text{argmax}_j \frac{\sum_{i=1}^{j} \lambda_i c_i(\theta_i)}{B_j}$ with probability one as shown in the proof. To give a counter example, we consider $B_j = \sqrt{\sum_{i=1}^{K} w(i,j)}$ (corresponding to T_{AFz}), where $K = 2$. We assume there is only two signals, with $\lambda_1 c_1(\theta_1) = 9$ and $\lambda_2 c_2(\theta_2) = 1$. Then one can show $j^* = 1$ here, i.e., T_A picks the wrong subset of p-values with probability one. Since B_j is a strictly increasing function, we can easily show that $j^* \leq \ell$ always holds and $j^* < \ell$ in general.

The proof of Theorem 2.4 will use the first equivalent form of AFs,

$$T_{\text{AFs}} = \max_{1 \leq j \leq K} -\log \bar{F}_{\chi^2_{2j}}(-2 \sum_{i=1}^{j} \log p_{(i)}).$$

Proof of Theorem 2.4. The goal is to prove $\hat{\vec{w}} \to \vec{w}^*$ in probability as $n \to \infty$ under the alternative. Recall by Corollary A1, we have, under the alternative,

$$-\frac{2}{n} \sum_{i=1}^{j} \log p_{(i)} \to \begin{cases} \sum_{i=1}^{j} \lambda_i c_i(\theta_i) & 1 \leq j \leq \ell \\ \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i) & \ell < j \leq K \end{cases} \tag{A10}$$

with probability one as $n \to +\infty$. Define index sets

$$\mathcal{S}_1 = \{i : \ w_i^* = 1 \text{ and } \hat{w}_i = 0\}; \ \mathcal{S}_2 = \{i : \ w_i^* = 0 \text{ and } \hat{w}_i = 1\}.$$

Recall that we assume the first $\ell \leq K$ studies are with exact slopes $c_i(\theta) > 0$. The following arguments are based on the first equivalent form of AFs, denoted by T_{AFs}.

We first prove $\mathcal{S}_1 \to \varnothing$ in probability. Indeed, we claim a stronger result that $\mathcal{S}_1 \to \varnothing$ with probability one under the alternative. By Lemmas 2.1 and A3, as $n \to +\infty$, the first smallest ℓ p-values converge to the first ℓ p-values with exact slopes strictly greater than 0. Hence it suffices to prove that for

$$\hat{j} = \underset{j}{\text{argmax}} -\log \bar{F}_{\chi^2_{2j}}\big(-2 \sum_{i=1}^{j} \log p_{(i)}\big),$$

as $n \to +\infty$, we have $\hat{j} \geq \ell$ with probability one. Indeed, for any $j' < \ell$, by Lemmas A2 and A3 and equation (A10),

$$\begin{aligned}\frac{-\log \bar{F}_{\chi^2_{2j'}}\big(-2\sum_{i=1}^{j'}\log p_{(i)}\big)}{-\log \bar{F}_{\chi^2_{2\ell}}\big(-2\sum_{i=1}^{\ell}\log p_{(i)}\big)} &= \frac{-(1/n)\log \bar{F}_{\chi^2_{2j'}}\big(-2\sum_{i=1}^{j'}\log p_{(i)}\big)}{-(1/n)\log \bar{F}_{\chi^2_{2\ell}}\big(-2\sum_{i=1}^{\ell}\log p_{(i)}\big)} \\ &= \frac{-(1/n)\log \bar{F}_{\chi_{2j'}}\big((-2\sum_{i=1}^{j'}\log p_{(i)})^{\frac{1}{2}}\big)}{-(1/n)\log \bar{F}_{\chi_{2\ell}}\big((-2\sum_{i=1}^{\ell}\log p_{(i)})^{\frac{1}{2}}\big)} \\ &\to \frac{\sum_{i=1}^{j'}\lambda_i c_i(\theta)}{\sum_{i=1}^{\ell}\lambda_i c_i(\theta)} < 1\end{aligned}$$

with probability one. Hence as $n \to +\infty$,

$$\hat{j} \geq \ell \tag{A11}$$

with probability one, i.e., $\mathcal{S}_1 \to \varnothing$ with probability one.

We then prove $\mathcal{S}_2 \to \varnothing$ in probability under the alternative, which is essentially to prove $\hat{j} \leq \ell$ in probability. To prove this, pick arbitrary $j > \ell$, and note event $\hat{j} = j$ is equivalent to event

$$\frac{\bar{F}_{\chi^2_{2j}}\big(-2\sum_{i=1}^{j}\log p_{(i)}\big)}{\bar{F}_{\chi^2_{2\ell}}\big(-2\sum_{i=1}^{\ell}\log p_{(i)}\big)} \leq 1. \tag{A12}$$

Then we have

$$\begin{aligned}(A12) &\Leftrightarrow \sum_{i=0}^{j-1}\frac{1}{i!}\big(-\sum_{k=1}^{j}\log p_{(k)}\big)^i \exp\big(\sum_{k=1}^{j}\log p_{(k)}\big) \\ &\leq \sum_{i=0}^{\ell-1}\frac{1}{i!}\big(-\sum_{k=1}^{\ell}\log p_{(k)}\big)^i \exp\big(\sum_{k=1}^{\ell}\log p_{(k)}\big) \qquad (A13) \\ &\Leftrightarrow \exp\Big\{\sum_{k=\ell+1}^{j}\log p_{(k)}\Big\} \leq \frac{\sum_{i=0}^{\ell-1}\frac{1}{i!}\big(-\sum_{k=1}^{\ell}\log p_{(k)}\big)^i}{\sum_{i=0}^{j-1}\frac{1}{i!}\big(-\sum_{k=1}^{j}\log p_{(k)}\big)^i} \\ &\Leftrightarrow \underbrace{\prod_{k=\ell+1}^{j} p_{(k)}}_{I} \leq \underbrace{\frac{\sum_{i=0}^{\ell-1}\frac{1}{i!}\big(-\sum_{k=1}^{\ell}\log p_{(k)}\big)^i}{\sum_{i=0}^{j-1}\frac{1}{i!}\big(-\sum_{k=1}^{j}\log p_{(k)}\big)^i}}_{II}. \qquad (A14)\end{aligned}$$

(A13) is due to relationship between Poisson distribution and chi-squared distribution. Note

$$II = \underbrace{\frac{\sum_{i=0}^{\ell-1}\frac{1}{i!}\big(-\sum_{k=1}^{\ell}\log p_{(k)}\big)^i/\big(\frac{n}{2}\big)^{\ell-1}}{\sum_{i=0}^{j-1}\frac{1}{i!}\big(-\sum_{k=1}^{j}\log p_{(k)}\big)^i/\big(\frac{n}{2}\big)^{j-1}}}_{III} \cdot \frac{1}{\big(\frac{n}{2}\big)^{j-\ell}},$$

and

$$III \to \frac{(j-1)!}{(\ell-1)!} \cdot \frac{1}{\big(\sum_{k=1}^{\ell} \lambda_i c_i(\theta_i)\big)^{j-\ell}}$$

with probability one. Hence $II = O\big(\frac{1}{n^{j-\ell}}\big)$ with probability one. While for I, with probability one, it is the product of the first $(j-\ell)$-th smallest p-values of $K-\ell$ i.i.d. p-values following Unif$(0,1)$ as $n \to +\infty$. Hence $I = O_p(1)$ under the alternative. Hence the probability that event (A14) holds converges to zero as $n \to +\infty$. Then the result follows. □

Let $R_j = -\sum_{i=1}^{j} \log p_{(i)} = \frac{T_j}{2}$, to prove Theorem 2.5, we need the following Lemma to carefully quantify the upper tails of R_j when $1 < j < K$ and under the null:

Lemma A4 (Nagaraja (2006))**.** *Let $F_{R_j}(t)$ and $\bar{F}_{R_j}(t)$ be the CDF and survival function of R_j under the null, separately. For $1 < j < K$, we have:*

$$\bar{F}_{R_j}(t) = \sum_{i=1}^{K-j} w_i \exp\{-c_i t / c_{K-j+1}\} \frac{1}{(j-1)!} \int_0^t \exp(d_i y)\, y^{j-1} dy + \sum_{k=0}^{j-1} e^{-t} \frac{t^k}{k!},$$

where $c_i = K - i + 1$, $d_i = \frac{c_i}{c_{K-j+1}} - 1$. And

$$w_i = \prod_{k=1; k\neq i}^{K-j} \frac{K-k+1}{i-k}.$$

Further calculation leads to

$$\bar{F}_{R_j}(t) = \sum_{i=1}^{K-j} w_i \exp\{-t\} \frac{1}{(j-1)!} \left\{ \sum_{m=0}^{j-1} (-1)^m t^{j-1-m} \frac{1}{d_i^{m+1}} \frac{(j-1)!}{(j-1-m)!} \right\} + \sum_{k=0}^{j-1} e^{-t} \frac{t^k}{k!}. \tag{A15}$$

Proof of Theorem 2.5. We consider the $T_{\text{AFp}} = \max_{j \in \mathcal{S}} - \log \bar{G}_j(-2\sum_{i=1}^{j} \log p_{(i)})$, where $\bar{G}_j = 1 - G_j(t)$ and $G_j(t)$ denotes the CDF function of $T_j = -2\sum_{i=1}^{j} \log p_{(i)}$ under the null. Let

$$\hat{j} = \operatorname*{argmax}_j - \log \bar{G}_j\big(-2\sum_{i=1}^{j} \log p_{(i)}\big).$$

By Lemma A3, it suffices to show $\hat{j} \to \ell$ in probability. We First show that the choice of $\hat{j} \geq \ell$ with probability one as n diverges under the alternative. Indeed, by the following inequality

$$\mathbb{P}(\chi^2_{2j} > t) \leq \bar{G}_j(t) \leq \mathbb{P}(\chi^2_{2K} > t), \tag{A16}$$

we have

$$-2\log\bar{G}_j\big(-2\sum_{i=1}^{j}\log p_{(i)}\big) \le -2\log\bar{F}_{\chi^2_{2j}}\big(-2\sum_{i=1}^{j}\log p_{(i)}\big)$$
$$-2\log\bar{G}_j\big(-2\sum_{i=1}^{j}\log p_{(i)}\big) \ge -2\log\bar{F}_{\chi^2_{2K}}\big(-2\sum_{i=1}^{j}\log p_{(i)}\big).$$

Then by Lemma 2.1, Lemmas A1- A3, and Corollary A1, we have:

$$-\frac{2}{n}\log\bar{G}_j\big(-2\sum_{i=1}^{j}\log p_{(i)}\big) \to \begin{cases} \sum_{i=1}^{j}\lambda_i c_i(\theta_i) & j<\ell \\ \sum_{i=1}^{\ell}\lambda_i c_i(\theta_i) & j\ge\ell \end{cases}$$

with probability one. Hence $\hat{j} \ge \ell$ with probability one as n goes to infinity under the alternative. Now we show $\hat{j} \le \ell$ in probability as n goes to infinity. Indeed, for any $j > \ell$, consider the following event:

$$\frac{A}{B} = \frac{\bar{G}_j\left(\sum_{i=1}^{j} -2\log p_{(i)}\right)}{\bar{G}_\ell\left(\sum_{i=1}^{\ell} -2\log p_{(i)}\right)} \le 1. \tag{A17}$$

It suffices to show probability of the above event goes to zero under the alternative.

For the case $1 < \ell < j < K$, by Lemma A4,

$$\begin{aligned}
&\frac{A}{B} \le 1 \\
&\Leftrightarrow \frac{\bar{F}_{R_j}\left(\sum_{i=1}^{j} -\log p_{(i)}\right)}{\bar{F}_{R_\ell}\left(\sum_{i=1}^{\ell} -\log p_{(i)}\right)} \le 1 \\
&\Leftrightarrow \underbrace{\prod_{i=\ell+1}^{j} p_{(i)}}_{I} \le \underbrace{\frac{\sum_{i=1}^{K-\ell} w_i \frac{1}{(\ell-1)!}\left\{\sum_{m=0}^{\ell-1}(-1)^m R_\ell^{\ell-1-m}\frac{1}{d_i^{m+1}}\frac{(\ell-1)!}{(\ell-1-m)!}\right\} + \sum_{k=0}^{\ell-1}\frac{R_\ell^k}{k!}}{\sum_{i=1}^{K-j} w_i \frac{1}{(j-1)!}\left\{\sum_{m=0}^{j-1}(-1)^m R_j^{j-1-m}\frac{1}{d_i^{m+1}}\frac{(j-1)!}{(j-1-m)!}\right\} + \sum_{k=0}^{j-1}\frac{R_j^k}{k!}}}_{II}.
\end{aligned}$$

Note that

$$II \cdot \frac{\left(\frac{n}{2}\right)^{j-1}}{\left(\frac{n}{2}\right)^{\ell-1}} = II \cdot \left(\frac{n}{2}\right)^{j-\ell} \to C_{K,j,\ell}\big(\sum_{i=1}^{\ell}\lambda_i c_i(\theta_i)\big)^{j-\ell}$$

with probability one, where $C_{K,i,\ell}$ is some constant that depends on K, i and ℓ. Hence $II \to 0$ with probability one as n diverges. By Lemma A3, we note I is the product of the first $(j-\ell)$-th smallest p-values of $K-\ell$ i.i.d. p-values following Unif$(0,1)$ as $n \to +\infty$. Hence $I = O_p(1)$.

And the probability of event $A/B \leq 1$ goes to 0 in probability. For the case $1 < \ell < i = K$, we note that

$$\begin{aligned}
&\frac{A}{B} \leq 1 \\
&\Leftrightarrow \frac{\bar{F}_{\chi^2_{2K}}\left(\sum_{i=1}^{K} - \log p_{(i)}\right)}{\bar{F}_{R_\ell}\left(\sum_{i=1}^{\ell} - \log p_{(i)}\right)} \leq 1 \\
&\Leftrightarrow \underbrace{\prod_{i=\ell+1}^{K} p_{(i)}}_{III} \leq \underbrace{\frac{\sum_{i=1}^{K-\ell} w_i \frac{1}{(\ell-1)!}\left\{\sum_{m=0}^{\ell-1}(-1)^m R_\ell^{\ell-1-m} \frac{1}{d_i^{m+1}} \frac{(\ell-1)!}{(\ell-1-m)!}\right\} + \sum_{k=0}^{\ell-1} \frac{R_\ell^k}{k!}}{\sum_{i=0}^{K-1} \frac{1}{i!}\left(R_K\right)^i}}_{IV}.
\end{aligned}$$

Note that

$$IV \cdot \frac{\left(\frac{n}{2}\right)^{K-1}}{\left(\frac{n}{2}\right)^{\ell-1}} = IV \cdot \left(\frac{n}{2}\right)^{K-\ell} \to C_{K,\ell}\Big(\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)\Big)^{K-\ell}$$

with probability one, where $C_{K,\ell}$ is some constant that depends on K and ℓ. Hence $IV \to 0$ with probability one as n diverges. By Lemma A3, we note III is the product of $K - \ell$ i.i.d. p-values following Unif$(0, 1)$ as $n \to +\infty$. Hence $III = O_p(1)$. And the probability of event $A/B \leq 1$ goes to 0 in probability.

Now we consider the case $1 = \ell < j < K$. By inequality $(1 + x)^K \geq 1 + Kx$ for $x > -1$, we have

$$\begin{aligned}
\mathbb{P}(A/B \leq 1) &= \mathbb{P}\Big(\frac{\bar{F}_{R_j}(\sum_{i=1}^{j} - \log p_{(i)})}{\bar{F}_{R_1}(-\log p_{(1)})} \leq 1\Big) \\
&= \mathbb{P}\Big(\frac{\bar{F}_{R_j}(\sum_{i=1}^{j} - \log p_{(i)})}{1 - (1 - \exp(\log p_{(1)}))^K} \leq 1\Big) \\
&\leq \mathbb{P}\Big(\frac{\bar{F}_{R_j}(\sum_{i=1}^{j} - \log p_{(i)})}{K \exp(-R_1)} \leq 1\Big).
\end{aligned}$$

Hence it suffices to show $\mathbb{P}\Big(\frac{\bar{F}_{R_j}(\sum_{i=1}^{j} - \log p_{(i)})}{K \exp(-R_1)} \leq 1\Big) \to 0$ as n diverges. Note that by Lemma A4, we have

$$\begin{aligned}
&\frac{\bar{F}_{R_j}(\sum_{i=1}^{j} - \log p_{(i)})}{K \exp(-R_1)} \leq 1 \\
&\Leftrightarrow \underbrace{\prod_{i=2}^{j} p_{(i)}}_{V} \leq \underbrace{\frac{K}{\sum_{i=1}^{K-j} w_i \frac{1}{(j-1)!}\left\{\sum_{m=0}^{j-1}(-1)^m R_j^{j-1-m} \frac{1}{d_i^{m+1}} \frac{(j-1)!}{(j-1-m)!}\right\} + \sum_{k=0}^{j-1} \frac{R_j^k}{k!}}}_{VI}
\end{aligned}$$

Note that

$$VI \cdot \big(\frac{n}{2}\big)^{K-1} \to C_K\big(\lambda_1 c_1(\theta_1)\big)^{K-1}$$

with probability one, where C_K is some constant that depends on K. Hence $VI \to 0$ with probability one as n diverges. By Lemma A3, we note V is the product of $j-1$ smallest p-values of $K-1$ i.i.d. p-values following Unif$(0,1)$ as $n \to +\infty$. Hence $V = O_p(1)$. And the probability of event $A/B \leq 1$ goes to 0 in probability. The arguments for the case $1 = \ell < j = K$ is quite similar, hence we omit the details. Combine the above results, we have $\hat{j} \leq \ell$ in probability as n diverges. Then the conclusion follows. □

We prove Theorem 2.6 by proving the following test statistic in a more general form is ABO:

$$T(\tau_1, \tau_2) = \sum_{i=1}^{K} \left(-2\log(p_i) + 2\log(\tau_2)\right) \mathrm{I}_{\{p_i \leq \tau_1\}} \text{ with } 0 \leq \tau_1, \tau_2 \leq 1.$$

When $\tau_1 = \tau$ and $\tau_2 = 1$, $T(\tau_1, \tau_2) = T_{\text{TFhard}}(\tau)$; and when $\tau_1 = \tau_2 = \tau$, $T(\tau_1, \tau_2) = T_{\text{TFsoft}}(\tau)$. The proof of the Theorem 2.6 requires the following additional lemma:

Lemma A5 (Zhang et al. (2020b))**.** *Assume $p_1, \ldots, p_K \sim$ Unif$(0,1)$ independently and identically. Denote by $U(\tau_1, \tau_2)$ the random variable that follows the same distribution of $T(\tau_1, \tau_2)$ under the null. Then*

$$\bar{F}_{U(\tau_1,\tau_2)}(t) = \big(1-\tau_1\big)^K I_{\{t \leq 0\}} + \sum_{i=1}^{K} \binom{K}{i} \tau_1^i \big(1-\tau_1\big)^{K-i} \bar{F}_{\chi^2_{2i}}\big(t + 2i\log\big(\tau_1/\tau_2\big)\big) \tag{A18}$$

Proof of Theorem 2.6. We only prove the case of $\tau_2 \leq \tau_1$ as the case of $\tau_2 > \tau_1$ can be proved by similar arguments. Let $F_{U(\tau_1,\tau_2)}(t)$ and $\bar{F}_{U(\tau_1,\tau_2)}(t)$ be the CDF and survival function of $U(\tau_1, \tau_2)$. Consider test statistic $\sqrt{T(\tau_1, \tau_2)}$. Under the setup in Section 2.1 and the alternative, by Lemmas 2.1 and A3, we have

$$\frac{\sqrt{T(\tau_1, \tau_2)}}{\sqrt{n}} = \frac{\sqrt{\sum_{i=1}^{K}\big(-2\log p_i + 2\log\tau_2\big)\mathrm{I}_{\{p_i \leq \tau_1\}}}}{\sqrt{n}} \to \big(\sum_{i=1}^{\ell} \lambda_i c_i(\theta_i)\big)^{\frac{1}{2}} \tag{A19}$$

with probability one as $n \to \infty$. In addition, by Lemma A1, for each $i = 1, \ldots, K$,

$$-\frac{1}{n}\log\bar{F}_{\chi^2_{2i}}\left(nt^2 + 2i\log(\tau_1/\tau_2)\right) = -\frac{1}{n}\log\bar{F}_{\chi_{2i}}\left(\sqrt{nt^2 + 2i\log(\tau_1/\tau_2)}\right) \to \frac{t^2}{2}$$

as $n \to \infty$. Note by Lemma A5, for $t > 0$ we have

$$\begin{aligned}
\bar{F}_{\sqrt{U(\tau_1,\tau_2)}}(\sqrt{n}t) &= \bar{F}_{U(\tau_1,\tau_2)}(nt^2) \\
&\geq \bar{F}_{\chi^2_2}\left(nt^2 + 2K\log(\tau_1/\tau_2)\right) \sum_{i=1}^{K} \binom{K}{i} \tau_1^i (1-\tau_1)^{K-i} \\
\bar{F}_{\sqrt{U(\tau_1,\tau_2)}}(\sqrt{n}t) &= \bar{F}_{U(\tau_1,\tau_2)}(nt^2) \\
&\leq \bar{F}_{\chi^2_{2K}}\left(nt^2 + 2\log\left(\tau_1/\tau_2\right)\right) \sum_{i=1}^{K} \binom{K}{i} \tau_1^i (1-\tau_1)^{K-i}.
\end{aligned}$$

Hence

$$-\frac{1}{n}\log \bar{F}_{\sqrt{U(\tau_1,\tau_2)}}\left(\sqrt{n}t\right) \to \frac{t^2}{2} \tag{A20}$$

with probability one as $n \to \infty$. By combining (A19) and (A20) and applying Lemma A2, we have for the exact slope of $T(\tau_1, \tau_2)$,

$$\begin{aligned}
C_{T(\tau_1,\tau_2)} &= -\frac{2}{n}\log \bar{F}_{U(\tau_1,\tau_2)}\big(T(\tau_1,\tau_2)\big) \\
&= -\frac{2}{n}\log \bar{F}_{\sqrt{U(\tau_1,\tau_2)}}\big(\sqrt{T(\tau_1,\tau_2)}\big) \to \sum_{i=1}^{\ell} \lambda_i c_i(\theta_i).
\end{aligned}$$

Hence $T(\tau_1, \tau_2)$ is ABO. □

A.2.2 Proof of Theorem 2.7

Proof of Theorem 2.7. Let $U_{\text{RV}}(\gamma)$ be the random variable that follows the same distribution of $T_{\text{RV}}(\gamma)$ under the null. Denote by $F_{U_{\text{RV}}(\gamma)}(t)$ and $\bar{F}_{U_{\text{RV}}(\gamma)}(t)$ the CDF and survival function of $T_{\text{RV}}(\gamma)$ under the null. Furthermore, under the null, let $U(\gamma)$ be the random variable such that $U(\gamma) \in R_{-\gamma}$. Hence $g_\gamma(p_{T_i}) = F^{-1}_{U(\gamma)}(1-p_{T_i})$ follows the same distribution of $U(\gamma)$ under the null. Let $t_i = F^{-1}_{U(\gamma)}(1-p_{T_i})$. Consequently, under the alternative, for i such that $C_i(\vec{\theta}) > 0$, $p_{T_i} = \bar{F}_{U(\gamma)}(t_i)$ and $\bar{F}_{U(\gamma)}(t_i)/(L(t_i)t_i^{-\gamma}) \to 1$ with probability one. We have, under the alternative, as $n \to +\infty$,

$$-\frac{2}{n}\log\left(\bar{F}_{U(\gamma)}(t_i)/(L(t_i)t_i^{-\gamma})\right) - \frac{2}{n}\log\left(L(t_i)t_i^{-\gamma}\right) = -\frac{2}{n}\log(p_{T_i}) \to C_i(\vec{\theta})$$

with probability one. Hence $-\frac{2}{n}\log\left(L(t_i)t_i^{-\gamma}\right) \to C_i(\vec{\theta})$ with probability one. By the basic property of slowly varying function, we have $L(t_i) = o(t_i^\gamma)$ with probability one for any γ. Hence for i such that $C_i(\vec{\theta}) > 0$,

$$-\frac{2}{n}\log(t_i^{-\gamma}) \to C_i(\vec{\theta}) \tag{A21}$$

with probability one. Let $t_0 = \sum_{i=1}^L F^{-1}_{U(\gamma)}(1-p_{T_i}) = \sum_{i=1}^L t_i$, then by Bonferroni's inequality, we have $\bar{F}_{U_{\text{RV}}(\gamma)}(t_0) \le L \cdot \bar{F}_{U(\gamma)}(\frac{t_0}{L})$ with probability one. then we have

$$\begin{aligned}
&-\frac{2}{n}\log\bar{F}_{U_{\text{RV}}(\gamma)}(T_{\text{RV}}(\gamma)) \\
&\ge -\frac{2}{n}\log\left(L\bar{F}_{U(\gamma)}(\frac{t_0}{L})/(L(t_0)L^{\gamma+1}t_0^{-\gamma})\right) - \frac{2}{n}\log(L(t_0)L^{\gamma+1}t_0^{-\gamma}) \\
&= \underbrace{-\frac{2}{n}\log\left(\bar{F}_{U(\gamma)}(\frac{t_0}{L})/(L(t_0)L^{\gamma}t_0^{-\gamma})\right)}_{(A)} + \underbrace{\frac{2\gamma\log t_0 - 2\log L^{\gamma+1}L(t_0)}{n}}_{(B)}.
\end{aligned}$$

Under the alternative, for (A), with $\max_{1\le i\le L} C_i(\vec{\theta}) > 0$ and either Conditions (C1) or (C2) holds, we have $t_0 \to +\infty$ with probability one. Then we have

$$\bar{F}_{U(\gamma)}(\frac{t_0}{L})/(L(t_0)L^{\gamma}t_0^{-\gamma}) = \left[\bar{F}_{U(\gamma)}(\frac{t_0}{L})/[L(\frac{t_0}{L})(\frac{t_0}{L})^{-\gamma}]\right] \cdot \left[L(\frac{t_0}{L})/L(t_0)\right] \to 1$$

with probability one, where the first term converges to 1 by the regularly varying tailed distribution definition and the second term converges to 1 by the definition of slow-varying distribution. Hence

we have $(A) \to 0$ with probability one. For (B), we first assume Condition (C-2) holds. Let $C_{i^*}(\vec{\theta}) = \max_{1\le i\le L} C_i(\vec{\theta})$, then under the alternative, by (A21) we have

$$\frac{2\gamma}{n}\log t_0 = \frac{2\gamma}{n}\log\big(\sum_{i=1}^{L} t_i\big) \ge \frac{2\gamma}{n}\max_{1\le i\le L}\{\log(t_i)\} \to C_{i^*}(\vec{\theta})$$
$$\frac{2\gamma}{n}\log t_0 = \frac{2\gamma}{n}\log\big(\sum_{i=1}^{L} t_i\big) \le \frac{2\gamma}{n}\max_{1\le i\le L}\{\log(t_i)\} + \frac{2\gamma\log L}{n} \to C_{i^*}(\vec{\theta})$$

with probability one. Suppose Condition (C-1) holds and Condition (C-2) does not hold, it suffices to consider the worst case that $F^{-1}_{U(\gamma)}(1-p) \ge \nu$ for some $\nu < 0$ and $\forall p \in (0,1]$. Denote by index set $\mathcal{B} = \{i : C_i(\vec{\theta}) > 0\}$. Then under the alternative, with probability one we have

$$\begin{aligned}\frac{2\gamma}{n}\log t_0 = \frac{2\gamma}{n}\log\big(\sum_{i\in\mathcal{B}} t_i + \sum_{i\in\mathcal{B}^c} t_i\big) &= \frac{2\gamma}{n}\log\big(\sum_{i\in\mathcal{B}} t_i\big) + \frac{2\gamma}{n}\log\big(1+\frac{\sum_{i\in\mathcal{B}^c} t_i}{\sum_{i\in\mathcal{B}} t_i}\big)\\ &\ge \underbrace{\frac{2\gamma}{n}\log\big(\sum_{i\in\mathcal{B}} t_i\big)}_{(C)} + \underbrace{\frac{2\gamma}{n}\log\big(1+\frac{|\mathcal{B}^c|\nu}{\sum_{i\in\mathcal{B}} t_i}\big)}_{(D)},\end{aligned}$$

where $|\mathcal{B}^c|$ denotes the cardinality of index set $\mathcal{B}^c$. For term (C), by (A21), under the alternative we have

$$\frac{2\gamma}{n}\log\big(\sum_{i\in\mathcal{B}} t_i\big) \ge \frac{2\gamma\max_{i\in\mathcal{B}}\{\log(t_i)\}}{n} = \frac{2\gamma\max_{1\le i\le L}\{\log(t_i)\}}{n} \to C_{i^*}(\vec{\theta})$$
$$\begin{aligned}\frac{2\gamma}{n}\log\big(\sum_{i\in\mathcal{B}} t_i\big) &\le \frac{2\gamma\max_{i\in\mathcal{B}}\{\log(t_i)\}}{n} + \frac{2\gamma\log|\mathcal{B}|}{n}\\ &= \frac{2\gamma\max_{1\le i\le L}\{\log(t_i)\}}{n} + \frac{2\gamma\log|\mathcal{B}|}{n} \to C_{i^*}(\vec{\theta})\end{aligned}$$

with probability one. Here we can also show that term (D) converges to zero with probability one as $n \to +\infty$. Hence $\frac{2\gamma}{n}\log t_0 = C_{i^*}(\vec{\theta})$ with probability one under the alternative. Further note $L(t_0) = o(t_0^{\gamma})$ with probability one, then we have $(B) = C_{i^*}(\vec{\theta})$ with probability one. Hence under the alternative

$$-\frac{2}{n}\log\bar{F}_{U_{\text{RV}}(\gamma)}(T_{\text{RV}}(\gamma)) = C_{i^*}(\vec{\theta})$$

as $n \to +\infty$ with probability one. □

Remark A4. The result of Theorem 2.7 also holds for the weighted version of $T_{\text{RV}}(\gamma)$ by the similar arguments in the above proof:

$$T^{\epsilon}_{\text{RV}}(\gamma) = \sum_{i=1}^{L} \epsilon_i g_\gamma(p_{T_i}) = \sum_{i=1}^{L} \epsilon_i F^{-1}_{U(\gamma)}(1 - p_{T_i})$$

with $\sum_{i=1}^{L} \epsilon_i = 1$ and $\epsilon_i > 0$ for each $i = 1, \ldots, L$.

A.2.3 Proofs of Theorems A2- A4 and Proposition A1

Lemma A6 (Mikosch (1999))**.** *Assume $U_1(\gamma), \ldots, U_K(\gamma)$ are i.i.d. random variables with distribution function $F \in R_{-\gamma}$. Then as $t \to \infty$, we have*

$$\mathbb{P}\left(U_1(\gamma) + \ldots + U_K(\gamma) > t\right) / \left(K\mathbb{P}\left(U_1(\gamma) > t\right)\right) \to 1. \tag{A22}$$

proof of Theorem A2. Denote $T_\eta = \sqrt{(1/\eta)\log\left(\sum_{i=1}^{K} 1/p_i^\eta\right)}$. Let $U(\eta)$ be the random variable that follows the same distribution of T_η under the null. Denote by $F_{U(\eta)}(t)$ and $\bar{F}_{U(\eta)}(t)$ the CDF and the survival function of T_η under the null. Further denote by $\mathbb{P}_0$ the probability measure of $\vec{p} = (p_1, \ldots, p_K)$ under the null. First note that $\bar{F}_{U(\eta)}(\sqrt{n}t) = \mathbb{P}_0\big(\sum_{i=1}^{K} 1/p_i^\eta > \exp(\eta n t^2)\big)$. Further note that $\frac{1}{p_i^\eta} \overset{D}{\sim} \text{Pareto}(\frac{1}{\eta}, 1) \in R_{-\frac{1}{\eta}}$ under the null, where the explicit form of survival function of Pareto distribution is $\bar{F}_{\text{Pareto}(\frac{1}{\eta},1)}(t) = t^{-\frac{1}{\eta}}$. Hence by Lemma A6 we have

$$\begin{aligned}
&\bar{F}_{U(\eta)}(\sqrt{n}t) / (K\bar{F}_{\text{Pareto}(\frac{1}{\eta},1)}(\exp(\eta n t^2)) \\
&= \mathbb{P}_0\big(\sum_{i=1}^{K} 1/p_i^\eta > \exp(\eta n t^2)\big) / (K(\exp(\eta n t^2))^{-\frac{1}{\eta}}) \to 1,
\end{aligned}$$

as $n \to +\infty$. Then we have,

$$-\frac{1}{n}\log\left(1 - F_{U(\eta)}(\sqrt{n}t)\right) \to t^2, \tag{A23}$$

as $n \to \infty$. We further claim under the alternative,

$$\frac{T_\eta}{\sqrt{n}} \to \sqrt{\max_{1 \le i \le K}\left\{\lambda_i c_i(\theta_i)\right\}/2} \tag{A24}$$

with probability one. Indeed, note

$$\max_{1\le i\le K}\{\log(1/p_i^\eta)\} \le \log\big(\sum_{i=1}^K 1/p_i^\eta\big) \le \log K + \max_{1\le i\le K}\{\log(1/p_i^\eta)\}.$$

Hence under the alternative, we have

$$\frac{1}{n}\log\big(\sum_{i=1}^K 1/p_i^\eta\big) \to \eta \max_{1\le i\le K}\lambda_i c_i(\theta)/2 \tag{A25}$$

with probability one. Then we have

$$\frac{T_\eta}{\sqrt{n}} = \frac{\sqrt{(1/\eta)\log\big(\sum_{i=1}^K 1/p_i^\eta\big)}}{\sqrt{n}} \to \sqrt{\max_{1\le i\le K}\lambda_i c_i(\theta)/2}$$

with probability one. Hence (A24) holds. Combining (A23) and (A24) and by Lemma A2, the result follows. □

Proof of Theorem A3. Note $T_{\text{CA}} = \frac{1}{K}\sum_{i=1}^K \cot(\pi p_i)$. Under the alternative, recall without loss of generality we assume that the first ℓ p-values correspond to non-zero exact slopes $c_i(\theta_i) > 0$ ($1 \le i \le \ell$), while the remaining p-values correspond to the zero exact slopes ($p_i \sim \text{Unif}(0,1)$ for $\ell+1 \le i \le K$). For the p-values with non-zero exact slopes, by the Taylor's expansion $x\cot x - 1 = -\frac{x^2}{3} + o(x^2)$, under the alternative we have,

$$\frac{1}{K}\sum_{i=1}^\ell\big[\frac{1}{\pi p_i} - \frac{2\pi p_i}{3}\big] \le \frac{1}{K}\sum_{i=1}^\ell \cot(\pi p_i) \le \frac{1}{K}\sum_{i=1}^\ell \frac{1}{\pi p_i}$$

with probability one. Note $\frac{1}{K}\sum_{i=1}^\ell\big[\frac{1}{\pi p_i} - \frac{2\pi p_i}{3}\big] = \frac{1}{K}\Big(1 - \frac{\sum_{i=1}^K 2\pi p_i/3}{\sum_{i=1}^K 1/\pi p_i}\Big)\sum_{i=1}^\ell \frac{1}{\pi p_i}$ and under the alternative, with probability one,

$$\begin{aligned}
&\Big(1 - \frac{\sum_{i=1}^\ell 2\pi p_i/3}{\sum_{i=1}^\ell 1/\pi p_i}\Big) \to 1\\
&\frac{1}{n}\log\big(\frac{1}{K}\sum_{i=1}^\ell 1/\pi p_i\big) \to \frac{1}{2}\max_{1\le i\le \ell}\lambda_i c_i(\theta),
\end{aligned} \tag{A26}$$

where (A26) is due to similar arguments for (A25) in the proof of Theorem A2 for $\eta = 1$. Hence we have

$$\frac{1}{n}\log\big(\frac{1}{K}\sum_{i=1}^\ell \cot(\pi p_i)\big) \to \frac{1}{2}\max_{1\le i\le \ell}\lambda_i c_i(\theta) \tag{A27}$$

with probability one.

Note that $\cot(\pi p_{\ell+1}), \ldots, \cot(\pi p_K) \overset{\text{i.i.d.}}{\sim} \text{CAU}(0,1)$. Hence we have

$$\frac{1}{K}\sum_{i=\ell+1}^{K} \cot\left(\pi p_i\right) \overset{D}{\sim} \frac{K-\ell}{K} U_{\text{CAU}(0,1)},$$

where $U_{\text{CAU}(0,1)}$ denotes standard Cauchy random variable. Note that under the null, $T_{\text{CA}} \overset{D}{\sim} \text{CAU}(0,1)$. Hence $F_{\text{CAU}(0,1)}(t)$ and $\bar{F}_{\text{CAU}(0,1)}(t)$ are the CDF and survival function of T_{CA} under the null. Hence under the alternative, we have

$$\begin{aligned}
\bar{F}_{\text{CAU}(0,1)}(T_{\text{CA}}) &= \mathbb{P}(U_{\text{CAU}(0,1)} > \frac{1}{K}\sum_{i=1}^{\ell} \cot(\pi p_i) + \frac{1}{K}\sum_{i=\ell+1}^{K} \cot(\pi p_i)) \\
&= \mathbb{P}\Big(\Big(1 + \frac{K-\ell}{K}\Big) U_{\text{CAU}(0,1)} > \frac{1}{K}\sum_{i=1}^{\ell} \cot(\pi p_i)\Big) \\
&= \mathbb{P}\Big(U_{\text{CAU}(0,1)} > \frac{K}{2K-\ell} \cdot \frac{1}{K}\sum_{i=1}^{\ell} \cot(\pi p_i)\Big). \qquad \text{(A28)}
\end{aligned}$$

In addition, for $t > 1$,

$$\begin{aligned}
\bar{F}_{\text{CAU}(0,1)}\left(t\right) &= \frac{1}{2} - \frac{1}{\pi}\arctan t = \frac{1}{\pi} \cdot \arctan\left(1/t\right) \leq \frac{1}{\pi t} \\
\bar{F}_{\text{CAU}(0,1)}\left(t\right) &= \frac{1}{\pi}\arctan\frac{1}{t} \geq \frac{1}{\pi t} \cdot \frac{t^2}{1+t^2}.
\end{aligned}$$

By combining the above two inequalities with (A27) and (A28), under the alternative we have

$$-\frac{2}{n}\log\left(\bar{F}_{\text{CAU}(0,1)}(T_{\text{CA}})\right) \to \max_{1\leq i\leq \ell} \lambda_i c_i(\theta) = \max_{1\leq i\leq K} \lambda_i c_i(\theta)$$

with probability one. □

To prove Theorem A4, we introduce the following notations adopted from Zhang et al. (2020a). Define

$$f_1^{\phi}(x,y) = x\log(\frac{x}{y}) + (1-x)\log(\frac{1-x}{1-y}).$$

Further define

$$\begin{aligned} f(x,y) &= \sqrt{2Kf_1^{\phi}(x,y)} \qquad \text{if } y \le x \\ &= -\sqrt{2Kf_1^{\phi}(x,y)} \quad \text{if } y > x. \end{aligned}$$

Note that $f(x,y)$ is strictly decreasing in y. When $T_{\text{BJ}} > 0$, we have

$$\sqrt{2KT_{\text{BJ}}} = \max_{1\le i\le K} f(\frac{i}{K}, p_{(i)}).$$

For each fixed x, define the inverse function of $f(x,\cdot)$ as $g(x,\cdot)$, i.e.,

$$g(x,\cdot) = f^{-1}(x,\cdot).$$

Proof of Theorem A4. Let $i^* = \text{argmax}_i\, i\lambda_i c_i(\theta)$ and note that by Lemma 2.1, $i^* \le \ell$. We first show that under the alternative,

$$2KT_{\text{BJ}}/n \to i^*\lambda_{i^*}c_{i^*}(\theta) \tag{A29}$$

with probability one. Denote

$$\hat{i} = \underset{i}{\text{argmax}} \left\{ \frac{i}{K}\log\left(\frac{i/K}{p_{(i)}}\right) + \left(1-\frac{i}{K}\right)\log\left(\frac{1-i/K}{1-p_{(i)}}\right)\right\}\mathrm{I}_{\{p_{(i)}<\frac{i}{K}\}}.$$

We show that under the alternative $\hat{i} \to i^*$ with probability one. Indeed, for any $i \ne i^*$ and $i \le \ell$, by Lemma A3, we have

$$\frac{(1/n)\left[\frac{i}{K}\log\left(\frac{i/K}{p_{(i)}}\right) + \left(1-\frac{i}{K}\right)\log\left(\frac{1-i/K}{1-p_{(i)}}\right)\right]\mathrm{I}_{\{p_{(i)}<\frac{i}{K}\}}}{(1/n)\left[\frac{i^*}{K}\log\left(\frac{i^*/K}{p_{(i^*)}}\right) + \left(1-\frac{i^*}{K}\right)\log\left(\frac{1-i^*/K}{1-p_{(i^*)}}\right)\right]\mathrm{I}_{\{p_{(i^*)}<\frac{i^*}{K}\}}} \to \frac{i\lambda_i c_i(\theta)}{i^*\lambda_{i^*}c_{i^*}(\theta)} < 1$$

with probability one. For any $i > \ell$, note that $1 - p_i$ still follows Unif$(0, 1)$. Hence for any $i' > \ell$, by Lemmas 2.1 and A3, we have $-\frac{1}{n}\log p_{(i')} \to 0$ and $-\frac{1}{n}\log(1-p_{(i')}) \to 0$ with probability one. Hence

$$\begin{aligned}
&\frac{(1/n)\left[\frac{i'}{K}\log\left(\frac{i'/K}{p_{(i')}}\right) + \left(1-\frac{i'}{K}\right)\log\left(\frac{1-i'/K}{1-p_{(i')}}\right)\right]\mathbf{I}_{\{p_{(i')}<\frac{i'}{K}\}}}{(1/n)\left[\frac{i^*}{K}\log\left(\frac{i^*/K}{p_{(i^*)}}\right) + \left(1-\frac{i^*}{K}\right)\log\left(\frac{1-i^*/K}{1-p_{(i^*)}}\right)\right]\mathbf{I}_{\{p_{(i^*)}<\frac{i^*}{K}\}}} \\
&\leq \frac{(1/n)\left[\frac{i'}{K}\log\left(\frac{i'/K}{p_{(i')}}\right) + \left(1-\frac{i'}{K}\right)\log\left(\frac{1-i'/K}{1-p_{(i')}}\right)\right]}{(1/n)\left[\frac{i^*}{K}\log\left(\frac{i^*/K}{p_{(i^*)}}\right) + \left(1-\frac{i^*}{K}\right)\log\left(\frac{1-i^*/K}{1-p_{(i^*)}}\right)\right]\mathbf{I}_{\{p_{(i^*)}<\frac{i^*}{K}\}}} \to 0
\end{aligned}$$

with probability one. Hence under the alternative $2KT_{\text{BJ}}/n \to i^*\lambda_{i^*}c_{i^*}(\theta_{i^*})$ with probability one.

Denote by U_{BJ} the random variable follows the same distribution of $\sqrt{2KT_{\text{BJ}}}$ under the null, and let

$$\mu_i = g(\frac{i}{K}, b) = f^{-1}(\frac{i}{K}, b),\ i = 1, 2, \ldots, K.$$

Let $F_{U_{\text{BJ}}}$, $\bar{F}_{U_{\text{BJ}}}$ be the CDF and survival function of U_{BJ}, respectively. Also let $F_{\text{Beta}(\alpha,\beta)}$ and $\bar{F}_{\text{Beta}(\alpha,\beta)}$ be the CDF and survival function of Beta(α, β), respectively. By Theorem 5.1 in Zhang et al. (2020a), we have,

$$F_{U_{\text{BJ}}}(b) = \bar{F}_{\text{Beta}(K,1)}(\mu_K) - \sum_{i=1}^{K-1}\frac{\mu_i^i}{i!}a_{i+1}, \tag{A30}$$

where

$$\begin{aligned}
&a_K = K!\bar{F}_{\text{Beta}(1,1)}(\mu_K) \\
&a_i = \frac{K!}{(K-i+1)!}\bar{F}_{\text{Beta}(K-i+1,1)}(\mu_K) - \sum_{j=1}^{K-i}\frac{\mu_{i+j-1}^j}{j!}a_{i+j} \\
&\text{for } i = K-1, K-2, \ldots, 1.
\end{aligned}$$

Since $\mu_i = g(\frac{i}{K}, b) = f^{-1}(\frac{i}{K}, b)$, for sufficiently large b, we have

$$b = \sqrt{2Kf_1^{\phi}(\frac{i}{K}, \mu_i)}.$$

Hence

$$\frac{b^2}{2K} = f_1^{\phi}(\frac{i}{K}, \mu_i) = \frac{i}{K}\log\frac{\frac{i}{K}}{\mu_i} + (1-\frac{i}{K})\log\frac{1-\frac{i}{K}}{1-\mu_i}.$$

Then

$$e^{-[\frac{b^2}{2K}-\frac{i}{K}\log\frac{i}{K}-(1-\frac{i}{K}\log(1-\frac{i}{K}))]}=\mu_i^{\frac{i}{K}}(1-\mu_i)^{1-\frac{i}{K}}. \tag{A31}$$

Note $f(x,y)$ is strictly decreasing in y, for $b\to\infty$, $\mu_i\to 0$. Denote

$$\mu_i = C_{i,b}e^{-\frac{b^2}{2i}},$$

where $C_{i,b}$ depends on i and b. We show that there exist $C_i>0$ only depending on i, such that

$$\lim_{b\to\infty} C_{i,b} = C_i. \tag{A32}$$

Indeed, from equation (A31), we have

$$\lim_{b\to\infty}\frac{e^{-[\frac{b^2}{2K}-\frac{i}{K}\log\frac{i}{K}-(1-\frac{i}{K}\log(1-\frac{i}{K}))]}}{(C_{i,b}e^{-\frac{b^2}{2i}})^{\frac{i}{K}}(1-C_{i,b}e^{-\frac{b^2}{2i}})^{1-\frac{i}{K}}}=1.$$

Hence

$$\lim_{b\to\infty}\frac{e^{\frac{i}{K}\log\frac{i}{K}+(1-\frac{i}{K}\log(1-\frac{i}{K}))}}{C_{i,b}^{\frac{i}{K}}}=1.$$

Hence we have $\lim_{b\to\infty}C_{i,b}=C_i>0$ for $i=1,\ldots,K$. Hence for sufficiently large b, we have

$$\mu_i=(C_i+o(1))e^{-\frac{b^2}{2i}}.$$

As $\lim_{b\to\infty}\mu_i=0$, for sufficiently large b, we have

$$a_k = K!\bar{F}_{\text{Beta}(1,1)}(\mu_K)=K!+o(1).$$

Similarly, for $i=1,\ldots,K-1$ and sufficiently large b,

$$a_i=\frac{K!}{(K-i+1)!}+o(1).$$

For $\bar{F}_{U_{\text{BJ}}}=1-F_{U_{\text{BJ}}}$, we have

$$\bar{F}_{U_{\text{BJ}}}(b)=\underbrace{F_{\text{Beta}(K,1)}(\mu_K)}_{I}+\underbrace{\sum_{i=1}^{K-1}\frac{\mu_i^i}{i!}a_{i+1}}_{II}. \tag{A33}$$

As $\mu_i^i = C_{i,b}^i e^{-\frac{b^2}{2}} = (C_i^i + o(1))e^{-\frac{b^2}{2}}$ for sufficiently large b, we have

$$I = F_{\text{Beta}(K,1)}(\mu_K) = \int_0^{\mu_K} Kx^{K-1}dx = \mu_K^K = (C_K^K + o(1))e^{-\frac{b^2}{2}}.$$

Similarly,

$$II = \sum_{i=1}^{K-1} \frac{\mu_i^i}{i!} a_{i+1} = \sum_{i=1}^{K-1} \frac{(C_i^i + o(1))e^{-\frac{b^2}{2}}}{i!} \Big[\frac{K!}{(K-i+1)!} + o(1)\Big].$$

Hence for sufficiently large b,

$$\begin{aligned}(A33) = I + II &= \Big[C_K^K + \sum_{i=1}^{K-1} \frac{C_i^i}{i!} \frac{K!}{(K-i+1)!} + o(1)\Big] e^{-\frac{b^2}{2}} \\ &= (C(K) + o(1))e^{-\frac{b^2}{2}}, \end{aligned} \tag{A34}$$

where $C(K)$ only depends on K. Let $b = \sqrt{2KT_{\text{BJ}}}$, combine equations (A29) and (A34), under the alternative, we have

$$-\frac{2\log \bar{F}_{U_{\text{BJ}}}\big(\sqrt{2KT_{\text{BJ}}}\big)}{n} \to i^* \lambda_{i^*} c_{i^*}(\theta_{i^*})$$

with probability one. □

Remark A5. It can be shown that T_{BJ} generally does not has signal selection consistency. Recall that T_{BJ} picks $i^* = \text{argmax}_i\, i\lambda_i c_i(\theta_i)$ with probability one as shown in the proof. Consider $K = 2$ and there is only two signals, with $\lambda_1 c_1(\theta_1) = 9$ and $\lambda_2 c_2(\theta_2) = 1$. Then one can show $i^* = 1$ here, i.e., T_{BJ} picks the wrong subset of p-values with probability one.

Below we use a counter example to show that higher criticism is generally not ABO. Let U_{HC} be the random variable that follows the same distribution of T_{HC} under the null. Denoted by $F_{U_{\text{HC}}}$ and $\bar{F}_{U_{\text{HC}}}$ the CDF and survival function of U_{HC}, respectively. To prove Proposition A1, we need the following Lemma to derive the survival function $\bar{F}_{U_{\text{HC}}}$ under the finite-sample case.

Lemma A7 (Barnett and Lin (2014)). *For each $k = 1, \ldots, K$, let*

$$t_k = \Phi^{-1}\left[1 - \frac{2(K-k+1) + h^2 - h\left\{h^2 + 4(K-k+1) - 4(K-k+1)^2/K\right\}^{1/2}}{4\left(h^2+K\right)}\right].$$

Denote $q_{1,a} = \mathbb{P}(S(t_1) = a)$ for $a = 0, 1, \ldots, K-1$. Here $S(t) = \sum_{j=1}^{K} I_{\{(|Z_j| \geq t)\}}$ is the binomial random variable with $Z_1, \ldots, Z_K \overset{i.i.d.}{\sim} N(0,1)$. Let

$$q_{k,a} = \sum_{m=0}^{K-k+1} I_{\{a \leq m\}} \binom{m}{a} \left\{\bar{\Phi}\left(t_k\right) / \bar{\Phi}\left(t_{k-1}\right)\right\}^a \times \left\{1 - \bar{\Phi}\left(t_k\right) / \bar{\Phi}\left(t_{k-1}\right)\right\}^{m-a} \frac{q_{k-1,m}}{\sum_{\ell=0}^{K-k+1} q_{k-1,\ell}}$$

for $k = 2, \ldots, K$ and $a = 0, 1, \ldots, K-k$. Then we have

$$\bar{F}_{U_{HC}}(h) = 1 - \prod_{k=1}^{K} \sum_{a=0}^{K-k} q_{k,a}.$$

Proof of Proposition A1. We first derive the exact form of $\bar{F}_{U_{\text{HC}}}(h)$ for $K = 2$. By Lemma A7,

$$\bar{F}_{U_{\text{HC}}}(h) = 1 - \prod_{k=1}^{2} \sum_{a=0}^{2-k} q_{k,a} = 1 - (q_{1,0} + q_{1,1}) q_{2,0}. \tag{A35}$$

We note

$$t_1 = \Phi^{-1}\left[1 - \frac{2(2-1+1) + h^2 - h\{h^2 + 8 - 4 \cdot 4/2\}^{\frac{1}{2}}}{4(h^2+2)}\right] = \Phi^{-1}\left[1 - \frac{4}{4(h^2+2)}\right].$$

And

$$q_{1,0} = \mathbb{P}(S(t_1) = 0) = \left[1 - 2(1 - \Phi(t_1))\right]^2 = \left[1 - \frac{2}{h^2+2}\right]^2.$$

Also

$$q_{1,1} = \mathbb{P}(S(t_1) = 1) = \frac{4}{h^2+2}\left(1 - \frac{2}{h^2+2}\right).$$

Hence $q_{1,1} + q_{1,0} = \left(1 - \frac{2}{h^2+2}\right)\left(1 + \frac{2}{h^2+2}\right)$. Further more,

$$t_2 = \Phi^{-1}\left[1 - \frac{2 + h^2 - h\{h^2 + 4 - 4/2\}^{\frac{1}{2}}}{4(h^2+2)}\right] = \Phi^{-1}\left[\frac{3}{4} + \frac{h}{4\sqrt{h^2+2}}\right].$$

Then we have $\bar{\Phi}(t_2) = \frac{1}{4} - \frac{h}{4\sqrt{h^2+2}}$, also $\bar{\Phi}(t_1) = \frac{1}{h^2+2}$. Hence

$$\begin{aligned} q_{2,0} &= \sum_{m=0}^{1} \mathbf{I}_{\{0 \le m\}} \binom{m}{0} \left[\frac{\bar{\Phi}(t_2)}{\bar{\Phi}(t_1)}\right]^0 \left[1 - \frac{\bar{\Phi}(t_2)}{\bar{\Phi}(t_1)}\right]^m \cdot \frac{q_{1,m}}{q_{1,0}+q_{1,1}} \\ &= I_1 + I_2, \end{aligned}$$

where

$$\begin{aligned} I_1 &= \frac{q_{1,0}}{q_{1,0}+q_{1,1}} = \frac{1 - \frac{2}{h^2+2}}{1 + \frac{2}{h^2+2}} \\ I_2 &= \left[1 - \frac{h^2+2-h\sqrt{h^2+2}}{4}\right] \frac{q_{1,1}}{q_{1,0}+q_{1,1}} = \left[\frac{1}{2} - \frac{h^2}{4} + \frac{h\sqrt{h^2+2}}{4}\right] \frac{\frac{4}{h^2+2}}{1+\frac{2}{h^2+2}}. \end{aligned}$$

Hence

$$I_1 + I_2 = \frac{1 + \frac{2}{(h^2+2)(\sqrt{h^2+2}+h)}}{1+\frac{2}{h^2+2}}.$$

By plugging in all the quantities into (A35), we have,

$$\begin{aligned} \bar{F}_{U_{\text{HC}}}(h) &= 1 - \left(1 - \frac{2}{h^2+2}\right)\left(1 + \frac{2}{h^2+2}\right) \frac{1 + \frac{2}{(h^2+2)(\sqrt{h^2+2}+h)}}{1+\frac{2}{h^2+2}} \\ &= \frac{2}{h^2+2} - \frac{2}{(h^2+2)(\sqrt{h^2+2}+h)} + \frac{4}{(h^2+2)^2(\sqrt{h^2+2}+h))}. \end{aligned} \tag{A36}$$

Recall

$$T_{\text{HC}} = \max_{1 \le i \le 2} \sqrt{2} \frac{\frac{i}{2} - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}} = \max_{1 \le i \le 2} \frac{i}{\sqrt{2p_{(i)}(1-p_{(i)})}} - \sqrt{\frac{p_{(i)}}{1-p_{(i)}}}.$$

Under the alternative, note $T_{\text{HC}}/(\sqrt{2}\exp(nc_0/4)) \to 1$ with probability one given $c_1(\theta_1) = c_2(\theta_2) = c_0 > 0$. Plugging into (A36), we have under the alternative in Proposition A1,

$$-\frac{2}{n} \log \bar{F}_{U_{\text{HC}}}(T_{\text{HC}}) \to c_0$$

with probability one as $n \to \infty$. □

Remark A6. One can note that under the alternative of combining two p-values with $c_1(\theta_1) = 2c_2(\theta_2) = 2c_0 > 0$, $\hat{i} = \text{argmax}_i \sqrt{2} \frac{\frac{i}{2} - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}} \to 1$ with probability one. Hence, HC is not consistent for selecting the subset of p-values with true signals.

A.3 Supplementary Simulation Results

A.3.1 Type I Error Control of FE and FE_{CS}

In this subsection, we numerically evaluate accuracy of type I error control using fast algorithm of independent Cauchy for the two methods proposed in Sections 4 and 5, FE and FE_{CS}. We simulate K p-values $p_1, \ldots, p_K \overset{D}{\sim} \text{Unif}(0,1)$, and calculate the test statistics for the two methods respectively, where $1-p_1, \ldots, 1-p_K$ with the previously generated p-values are used as one-sided p-values for FE_{CS}. We vary $K = 5, 10, 20, 40, 60, 80, 100$ for a wide range of numbers of combined p-values. Table A1 shows type I error control for the two methods under different significance levels α =0.05, 0.01, 0.001, 0.005, 0.001 using 10^5 times of simulations. Across wide ranges of K and $\alpha \leq 0.01$, type I error by the fast computing has less than 10% inflation, with improved accuracy for smaller α. As the worst case, the type I error control of FE_{CS} when $\alpha = 0.05$ is slightly anti-conservative but acceptable (in the range of 0.0539$\sim$0.0578 for different K).

A.3.2 Statistical Power Comparison for Modified Fisher Methods in the Case of Combining A Small Group of Strong Signals

In this subsection, we demonstrate the statistical power of Stouffer, Fisher, and 5 modified Fisher methods for combining a small group of strong signals. We simulate the alternatives with fixed numbers of true signals $\ell = 1, 2, \ldots, 6$ for $K = 20, 40, 80$ following the same simulation scheme in Section 3.3. For a given K and ℓ, we choose the smallest μ_0 such that the best method has at least 0.9 statistical power at $\alpha = 0.05$. The results are shown in Figure A1.

A.3.3 Statistical Power Comparison for 12 Existing P-Value Combination Methods

In this subsection, we demonstrate the statistical power of 12 p-value combination methods: Fisher, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer. For Figure A2, the signal strength μ_0 is chosen the same as Figure 1 in Section 3.3 for a given proportion of signals ℓ/K and number of combined p-values K. As expected, 4 added methods (HC, minP, HM, CA) that are designed for sparse signals and have very weak power for frequent signals. Although

Table A1: Accuracy of type I error control for FE and FE_{CS}

Methods	K	0.05	0.01	5×10^{-3}	1×10^{-3}
FE	5	5.02×10^{-2}	1.02×10^{-2}	5.23×10^{-3}	1.07×10^{-3}
	10	5.12×10^{-2}	9.96×10^{-3}	5.05×10^{-3}	1.17×10^{-3}
	20	5.12×10^{-2}	1.02×10^{-2}	4.90×10^{-3}	9.40×10^{-4}
	40	5.11×10^{-2}	9.80×10^{-3}	5.15×10^{-3}	1.02×10^{-3}
	60	5.15×10^{-2}	1.01×10^{-2}	5.13×10^{-3}	1.17×10^{-3}
	80	5.31×10^{-2}	1.10×10^{-2}	5.72×10^{-3}	1.05×10^{-3}
	100	5.36×10^{-2}	1.06×10^{-2}	5.37×10^{-3}	1.04×10^{-3}
FE_{CS}	5	5.39×10^{-2}	1.03×10^{-2}	5.16×10^{-3}	1.02×10^{-3}
	10	5.51×10^{-2}	1.02×10^{-2}	5.23×10^{-3}	1.15×10^{-3}
	20	5.50×10^{-2}	1.01×10^{-2}	4.95×10^{-3}	9.30×10^{-4}
	40	5.52×10^{-2}	1.06×10^{-2}	5.02×10^{-3}	9.80×10^{-4}
	60	5.35×10^{-2}	1.04×10^{-2}	5.55×10^{-3}	1.13×10^{-3}
	80	5.78×10^{-2}	1.13×10^{-2}	5.25×10^{-3}	9.90×10^{-4}
	100	5.70×10^{-2}	1.15×10^{-2}	5.77×10^{-3}	1.17×10^{-3}

BJ is also designed for sparse signal scenarios, it has relatively higher power, comparable to AFz but much lower than Fisher and AFp. For Figure A3, the signal strength μ_0 is chosen the same as Figure A1 for a given number of signals ℓ and number of combined p-values K. 4 added methods (HC, minP, HM, CA) that are designed for sparse signals outperform Fisher and Stouffer, but still are comparative with modified Fisher's methods such as AFp and AFz.

A.3.4 Statistical Power Comparison for FE in the Case of Combining A Small Group of Strong Signals

In this subsection, we demonstrate the statistical power of Fisher, AFp, and FE for combining a small group of strong signals. We simulate the alternatives with fixed numbers of true signals $\ell = 1, 2, \ldots, 6$ for $K = 20, 40, 80$ following the same simulation scheme in Section 3.3. For a given K and ℓ, we choose the smallest μ_0 such that the best method has at least 0.9 statistical power at $\alpha = 0.05$. The results are shown in Figure A4.

A.3.5 Statistical Power Comparison for FE and FE2

In this subsection, we evaluate the statistical power of Fisher, AFp, FE, and the following FE2 that integrates Fisher, AFp and minP:

$$T_{\text{FE2}} = [1/p^{\text{Fisher}} + 1/p^{\text{AFp}} + 1/p^{\text{minP}}]/3.$$

The following Figures A5 and A6 present the results in settings similar to that of Figures 2 and A4, respectively. For Figure A5, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 for a given proportion of signals ℓ/K and a number of combined p-values K. For Figure A6, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 for a given proportion of signals ℓ and a number of combined p-values K. Although FE2 improves power over FE when the signal is very sparse, its power is much reduced when the signal is frequent, which is an important scenario in most applications. As a result, FE combining Fisher and AFp but not minP is recommended for general applications.

A.3.6 Statistical Power Comparison for FE_{CS} in the Case of Combining A Small Group of Strong Signals

In this subsection, we demonstrate the statistical power of Pearson, FE, and FE_{CS} for combining a small group of strong signals. We simulate the alternatives with fixed numbers of true signals $\ell = 1, 2, \ldots, 6$ for $K = 20, 40, 80$ following the same simulation scheme in Section 3.3. For a given K and ℓ, we choose the smallest μ_0 such that the best method has at least 0.9 statistical power at $\alpha = 0.05$. The results are shown in Figure A7.

A.3.7 Numeric Examples where Harmonic Mean Outperforms Cauchy for Fisher Ensemble

This subsection provides numeric examples that using harmonic mean is better than Cauchy in the FE and FE_{CS} construction (Equation (2) in the manuscript). Below, we follow the simulation scheme in Section 5.2 to generate data and the combined p-values, where we evaluate the power of FE_{CS} (using the harmonic mean), Pearson, and $\text{FE}_{\text{CS}}^{\text{Cauchy}}$ (using Cauchy). Figures A8 and A9 show the empirical power of the three methods. For figure A8, we choose the smallest μ_0 that allows the best method to have power larger than 0.5 at significance level $\alpha = 0.01$ for a given proportion of signals ℓ/K and a number of combined p-values K. For figure A9, we choose the smallest μ_0 that allows the best method to have power larger than 0.9 at significance level $\alpha = 0.05$ for a given proportion of signals ℓ and a number of combined p-values K. The results show that FE_{CS} largely outperforms the latter for $\ell/K \geq 0.4$ in Figure A8, as a consequence of the "$-\infty$ score" issue when using Cauchy.

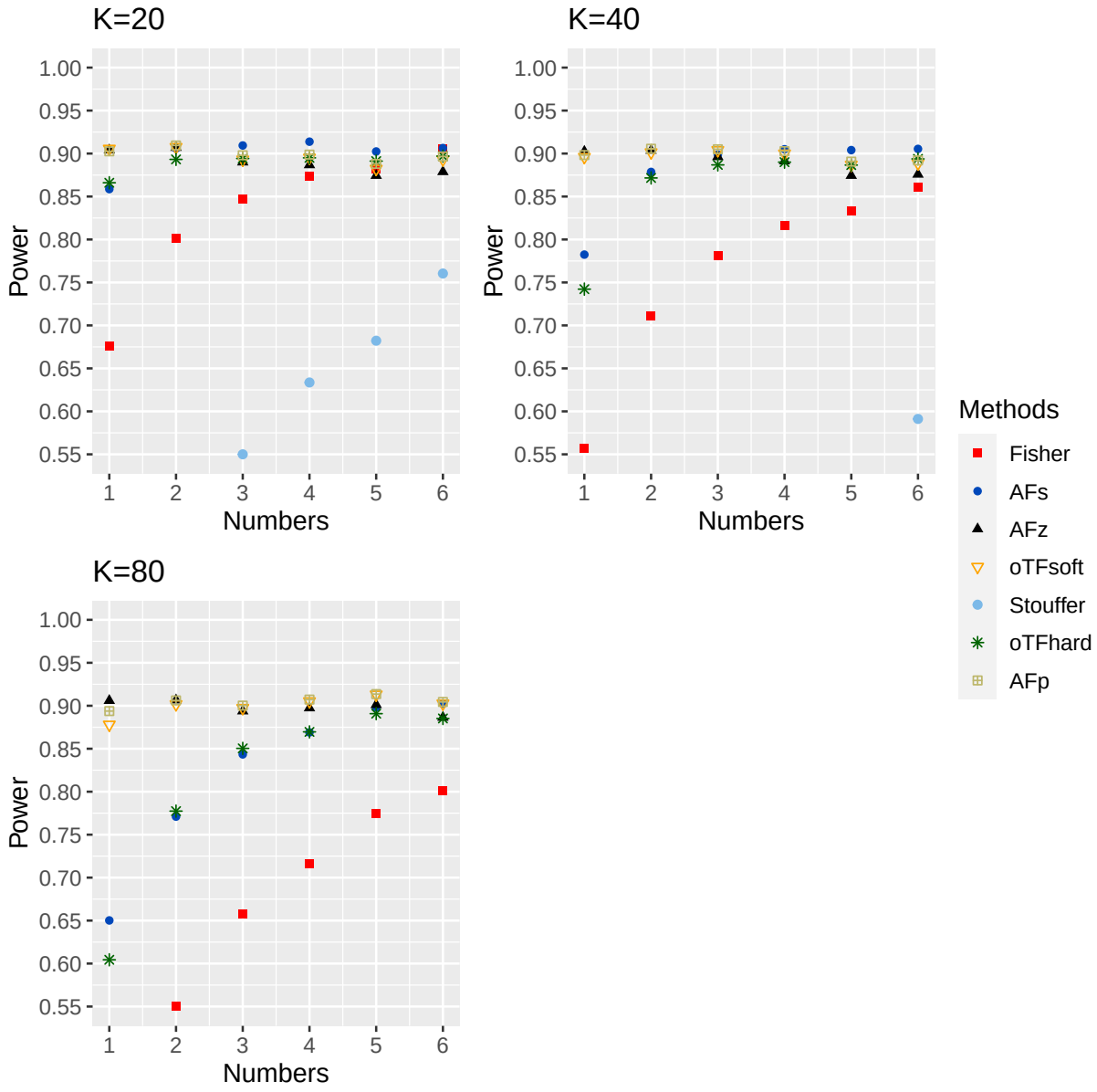


Figure A1: Statistical power of Fisher, Stouffer, and 5 modified Fisher's methods at significance level $\alpha = 0.05$ across varying numbers of true signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible compared to the scale of the mean power (smaller than 0.1% of the power) and hence omitted. The results of Stouffer and Fisher with a power smaller than 0.55 are omitted.

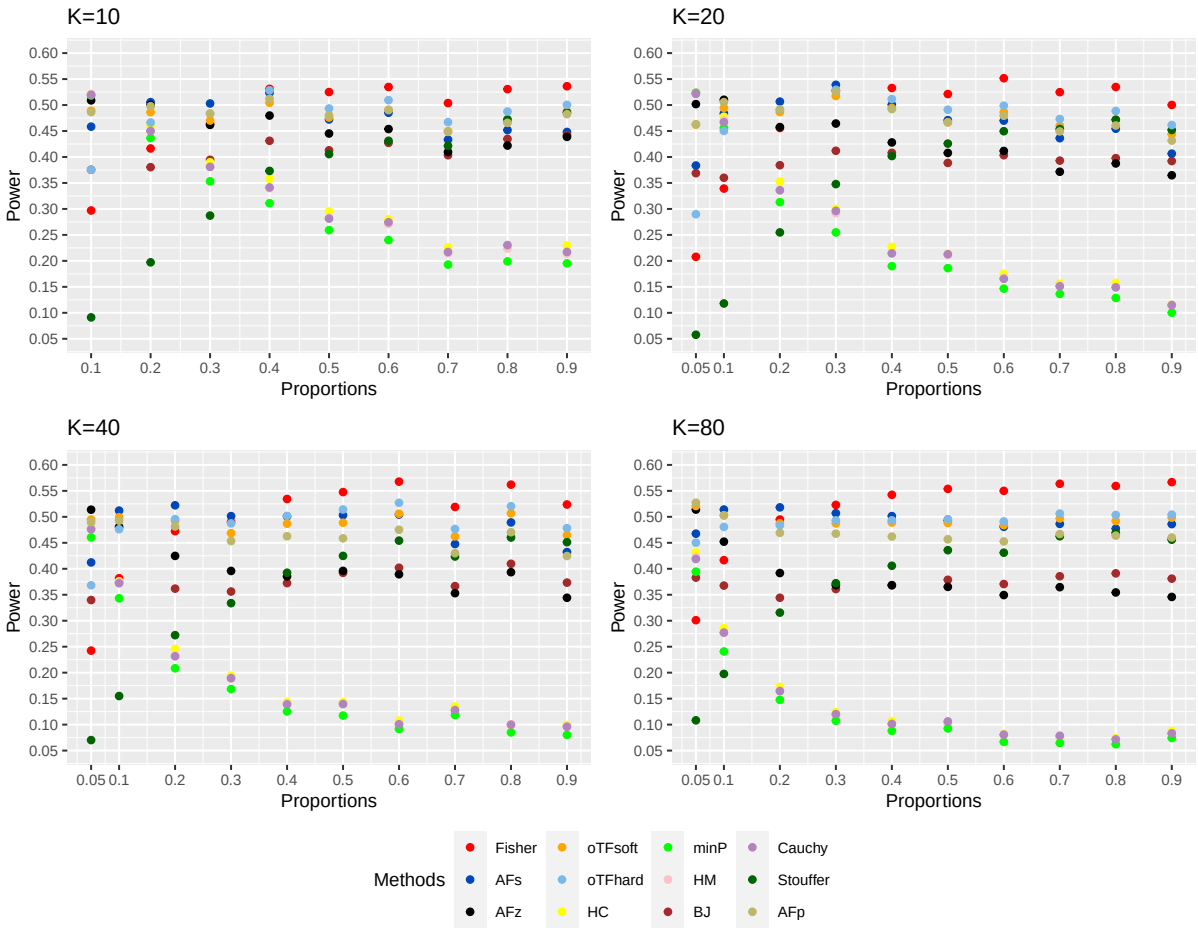


Figure A2: Statistical power of Fisher, AFs, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer at significance level $\alpha = 0.01$ across varying proportions of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

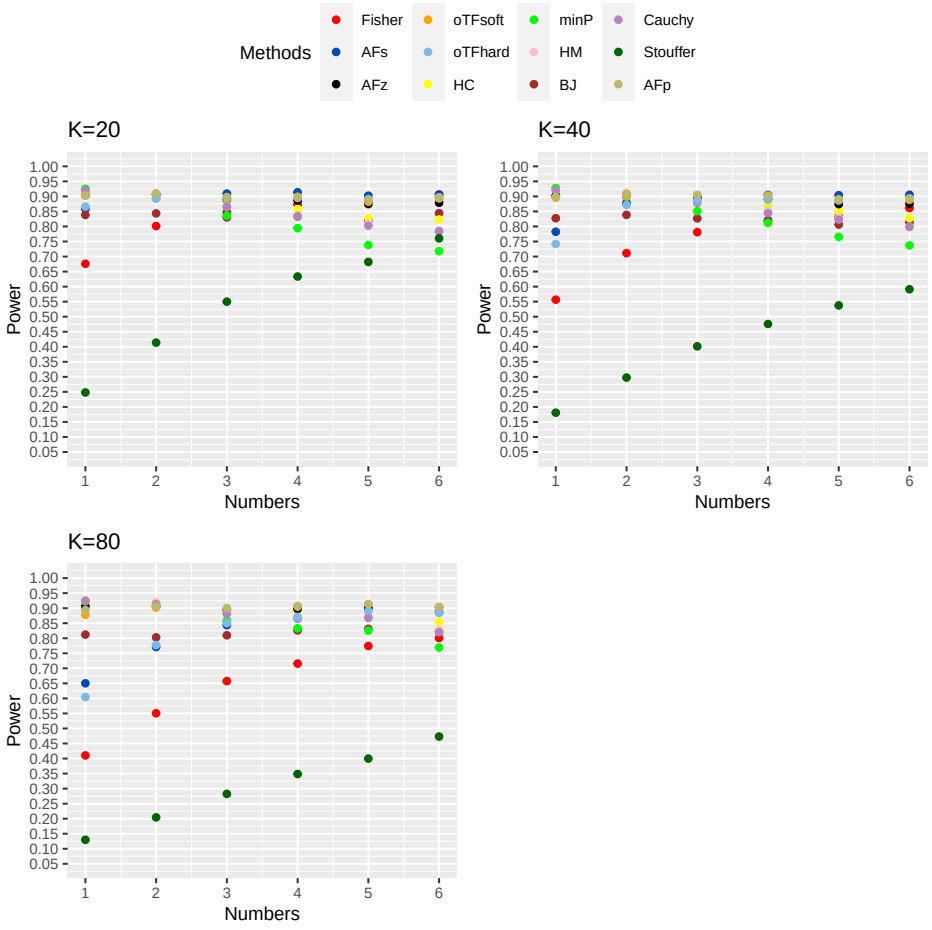


Figure A3: Statistical power of Fisher, AFs, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, 3, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

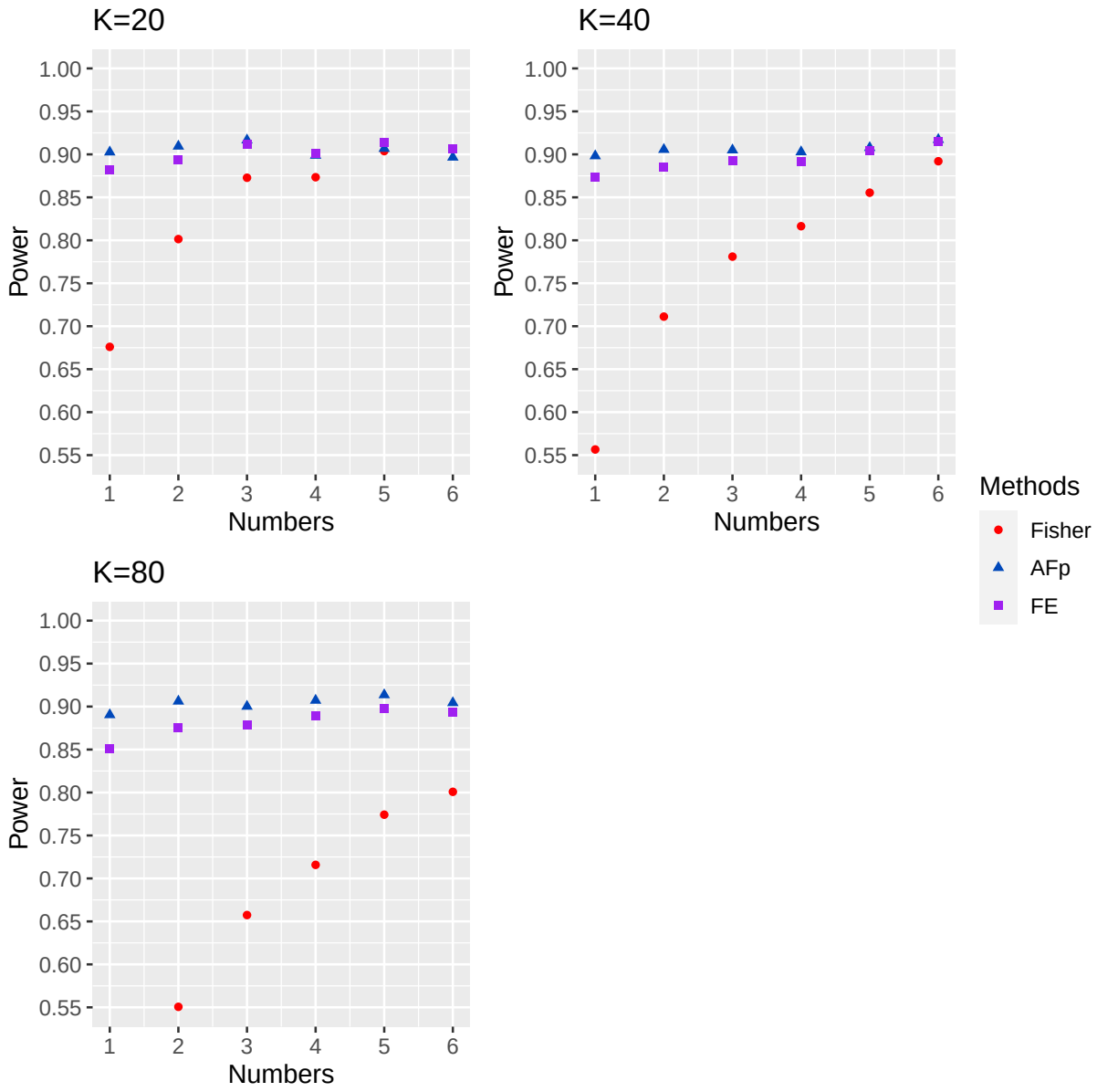


Figure A4: Statistical power of FE, Fisher, and AFp at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted. Dots smaller than 0.55 are also omitted.

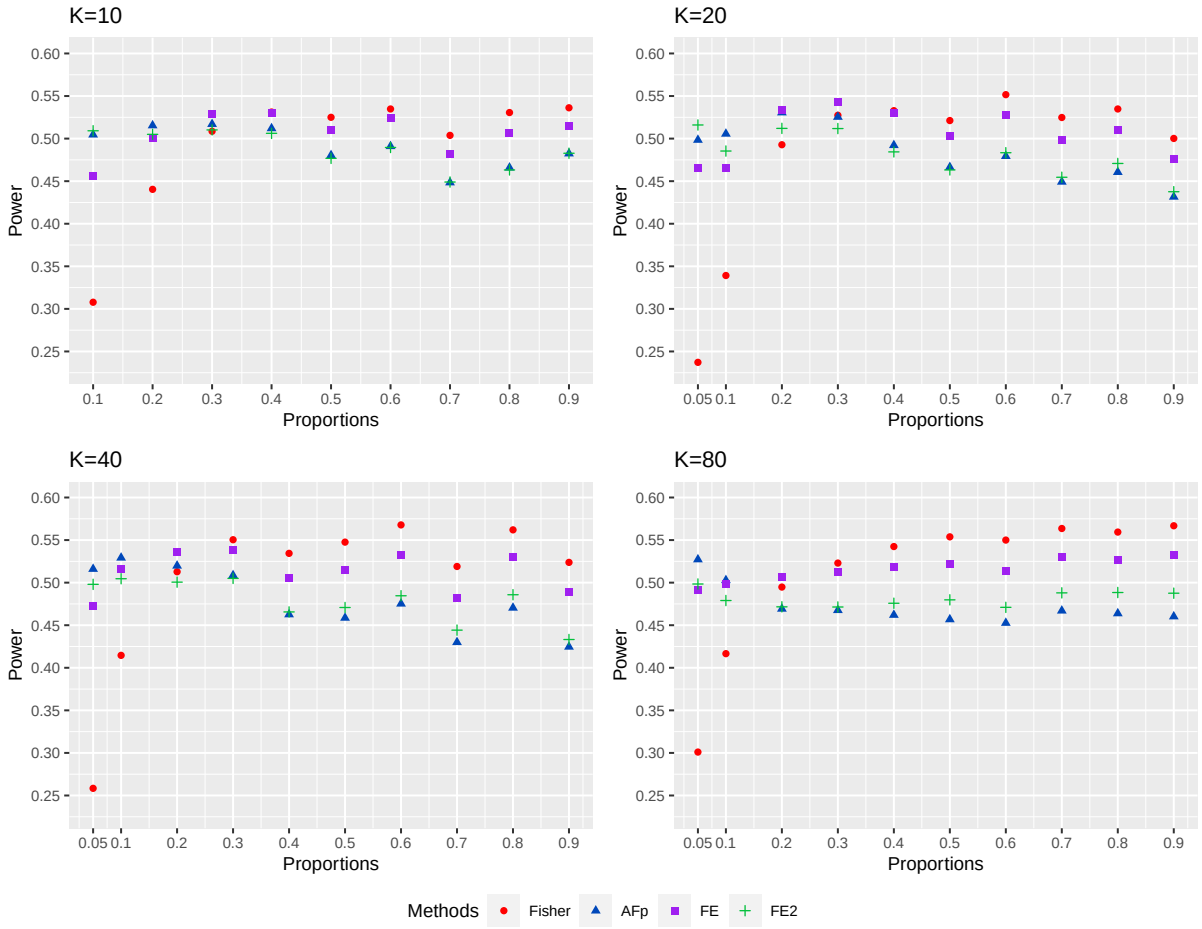


Figure A5: Statistical power of Fisher, AFp, FE, and FE2 at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

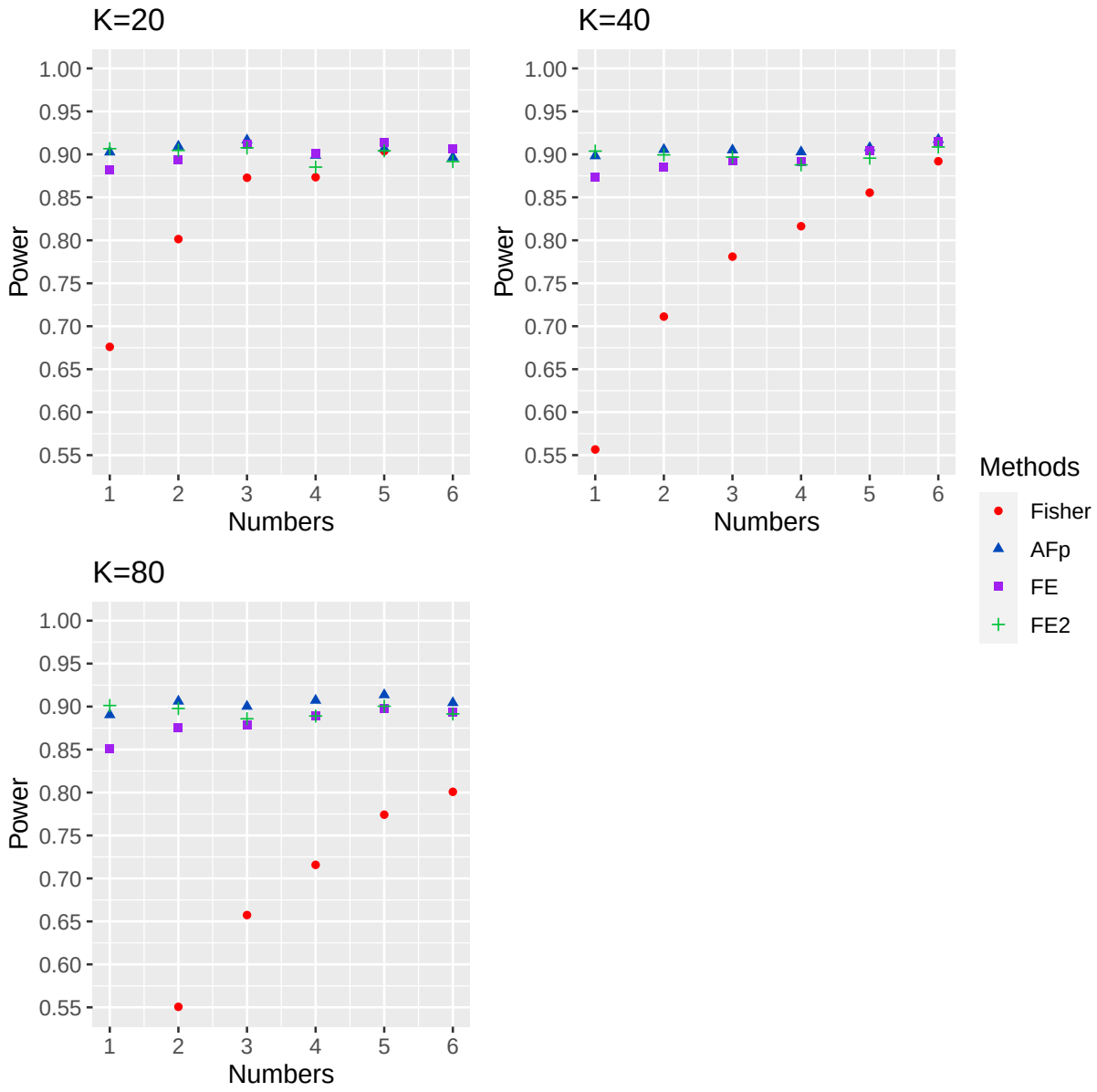


Figure A6: Statistical power of Fisher, AFp, FE, and FE2 at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted. results of Fisher smaller than 0.55 are omitted.

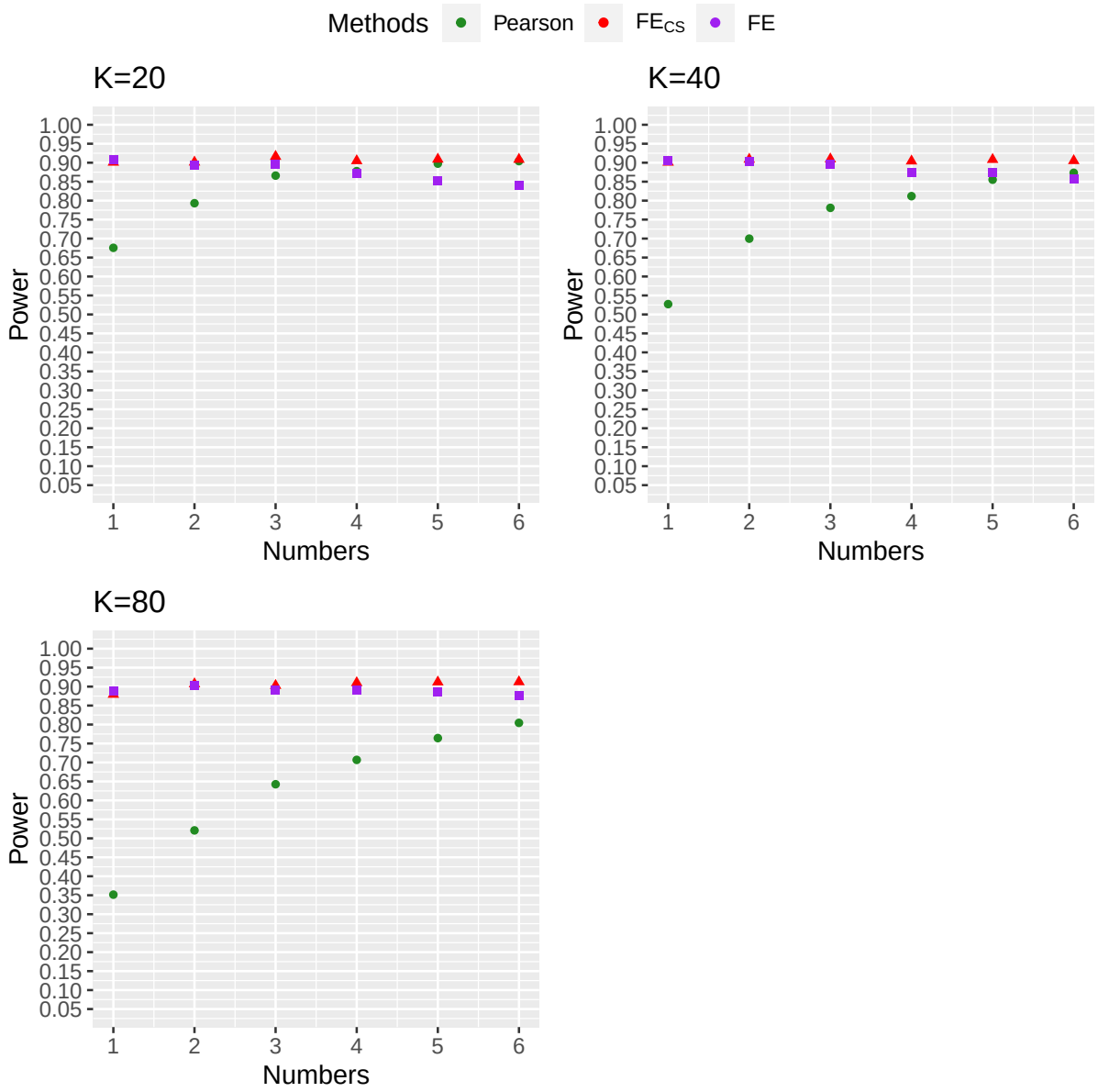


Figure A7: Statistical power of FE, FE_{CS}, and Pearson at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. The standard errors are negligible and hence omitted.

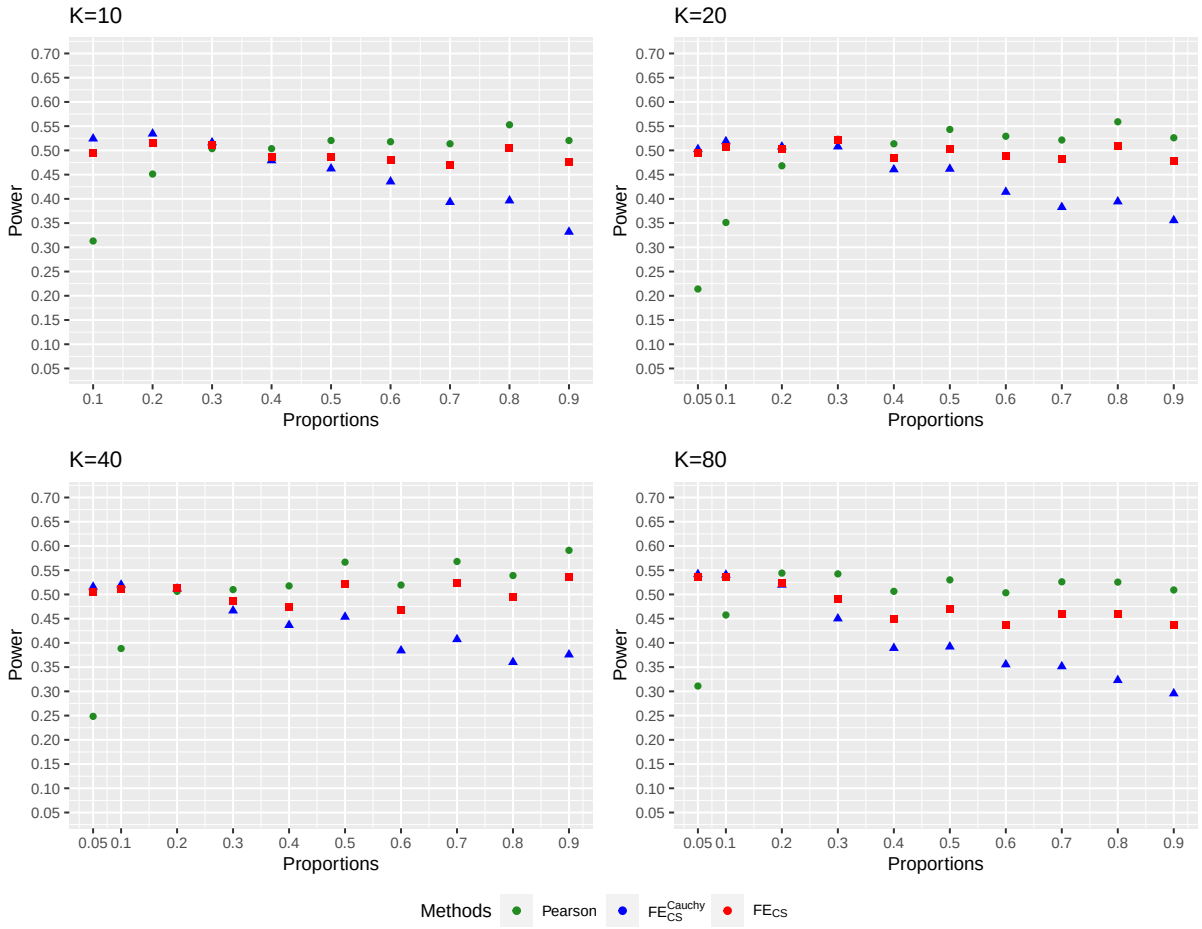


Figure A8: Statistical power of FE_{CS}, $\text{FE}_{\text{CS}}^{\text{Cauchy}}$, and Pearson at significance level $\alpha = 0.01$ across varying frequencies of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

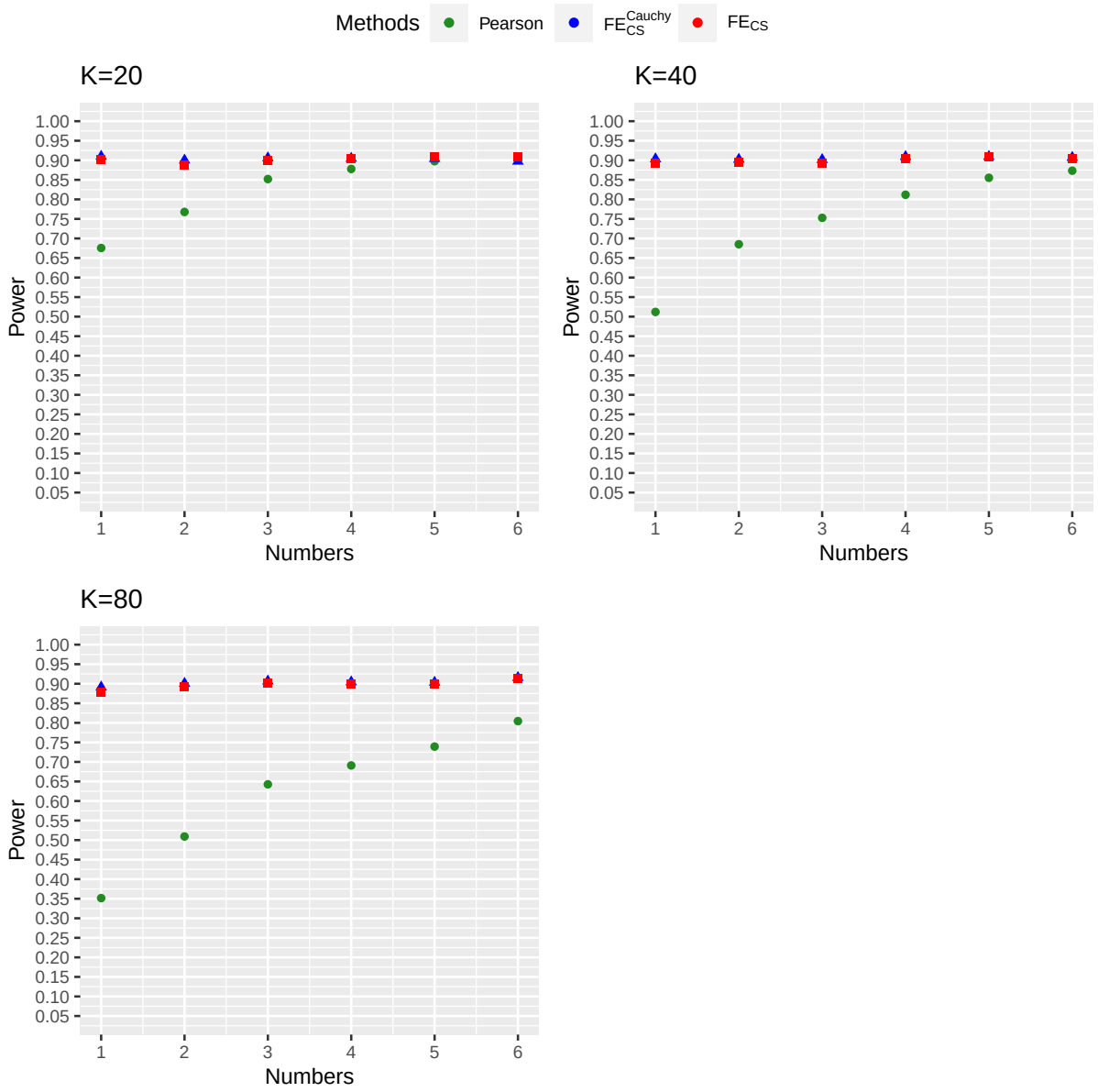


Figure A9: Statistical power of FE_{CS}, $\text{FE}_{\text{CS}}^{\text{Cauchy}}$, and Pearson at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

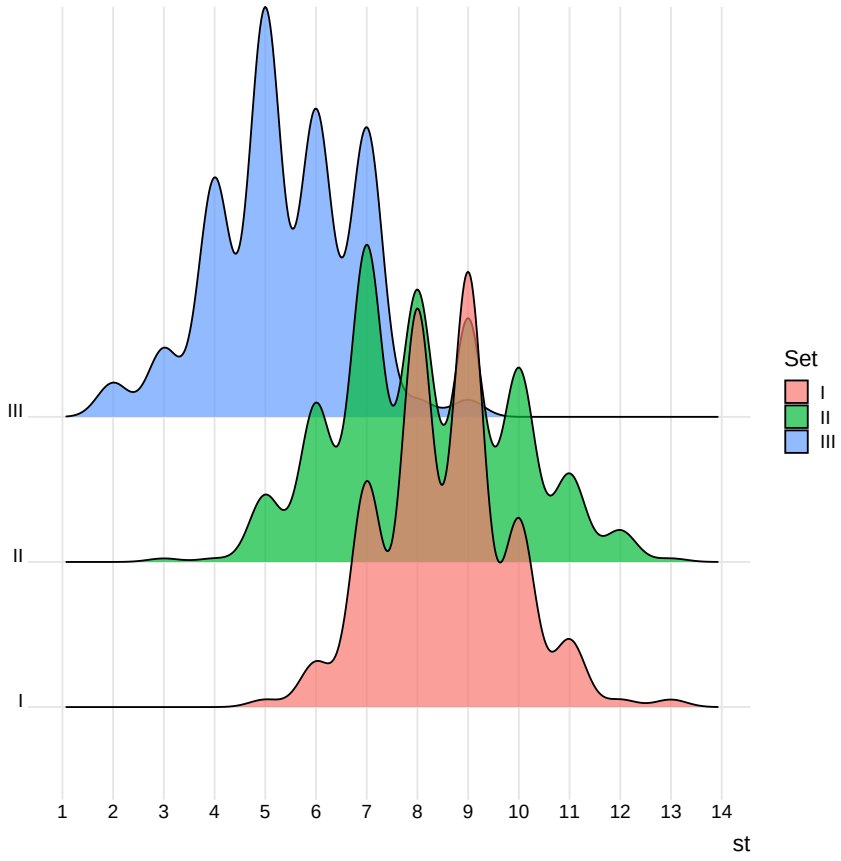


Figure A10: Distributions of numbers of p-values $p_{jk} \leq 0.05$ of each gene j in gene Categories I, II, and III in Figure 2.5(a).

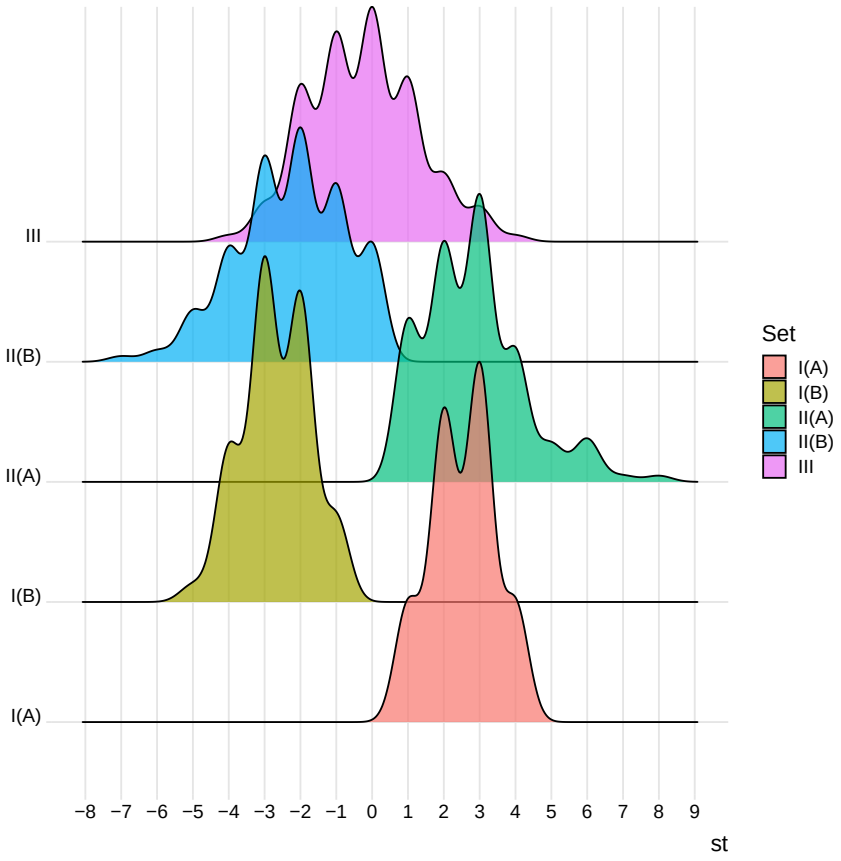


Figure A11: Distributions of quantities $S_{\text{sign},j} = \sum_{k=1}^{16} \text{sign}(\beta_{\text{age},jk}) \text{I}_{\{\min\{\tilde{p}_{jk}^L, \tilde{p}_{jk}^R\}\}}$ each gene j in Categories I(A), I(B), II(A), II(B), and III in Figure 2.6.

Table A2: Up-regulated/down-regulated age-related pathways detected in one-sided design by FE_{CS} with significance level $p \leq 0.01$. The reference columns of the 2 tables list literature that supports the relationships between the pathways and aging/early development processes.

(a): Pathways by up-regulated genes

Pathways	p-values	References
Phagosome Maturation	0.0005	Vieira et al. (2002)
Glutathione Redox Reactions I	0.00085	Mandal et al. (2015); Erden-İnal et al. (2002)
Tryptophan Degradation III (Eukaryotic)	0.0006	Van der Goot and Nollen (2013)
FAT10 Cancer Signaling Pathway	0.0041	Canaan et al. (2014); Aichem and Groettrup (2016)
Isoleucine Degradation I	0.0058	Canfield and Bradshaw (2019); Salcedo et al. (2021)
Glutamine Biosynthesis I	0.0065	Meynial-Denis (2016); Canfield and Bradshaw (2019)
Histamine Biosynthesis	0.0065	Mazurkiewicz-Kwilecki and Nsonwah (1989); Terao et al. (2004)
Tumor Microenvironment Pathway	0.0060	Mori et al. (2018); Sandiford et al. (2018)
Glutaryl-CoA Degradation	0.0065	Porcellini et al. (2007)
Valine Degradation I	0.0079	Canfield and Bradshaw (2019); Salcedo et al. (2021)
Androgen Signaling	0.0047	He et al. (2018); Rey (2021); Zhou et al. (2015)

(b): Pathways by down-regulated genes

Pathways	p-values	References
EIF2 Signaling	0.00001	Ma et al. (2013)
Remodeling of Epithelial Adherens Junctions	0.0019	Parrish (2017)
Tight Junction Signaling	0.00028	Parrish (2017); Ren et al. (2014)
NER (Nucleotide Excision Repair, Enhanced Pathway)	0.0087	Maynard et al. (2009)

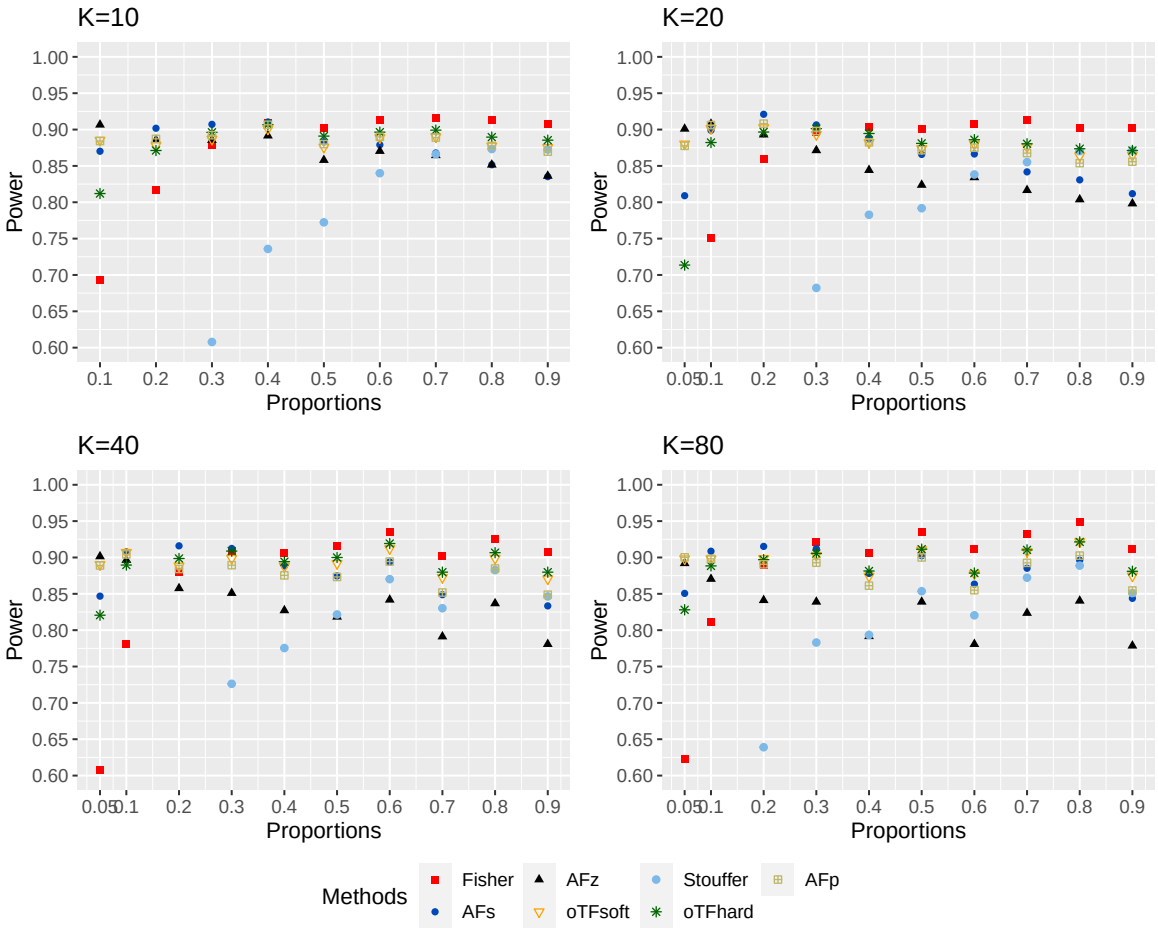


Figure A12: Statistical power of Fisher, AFs, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer at significance level $\alpha = 0.01$ across varying proportions of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each proportion ℓ/K and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

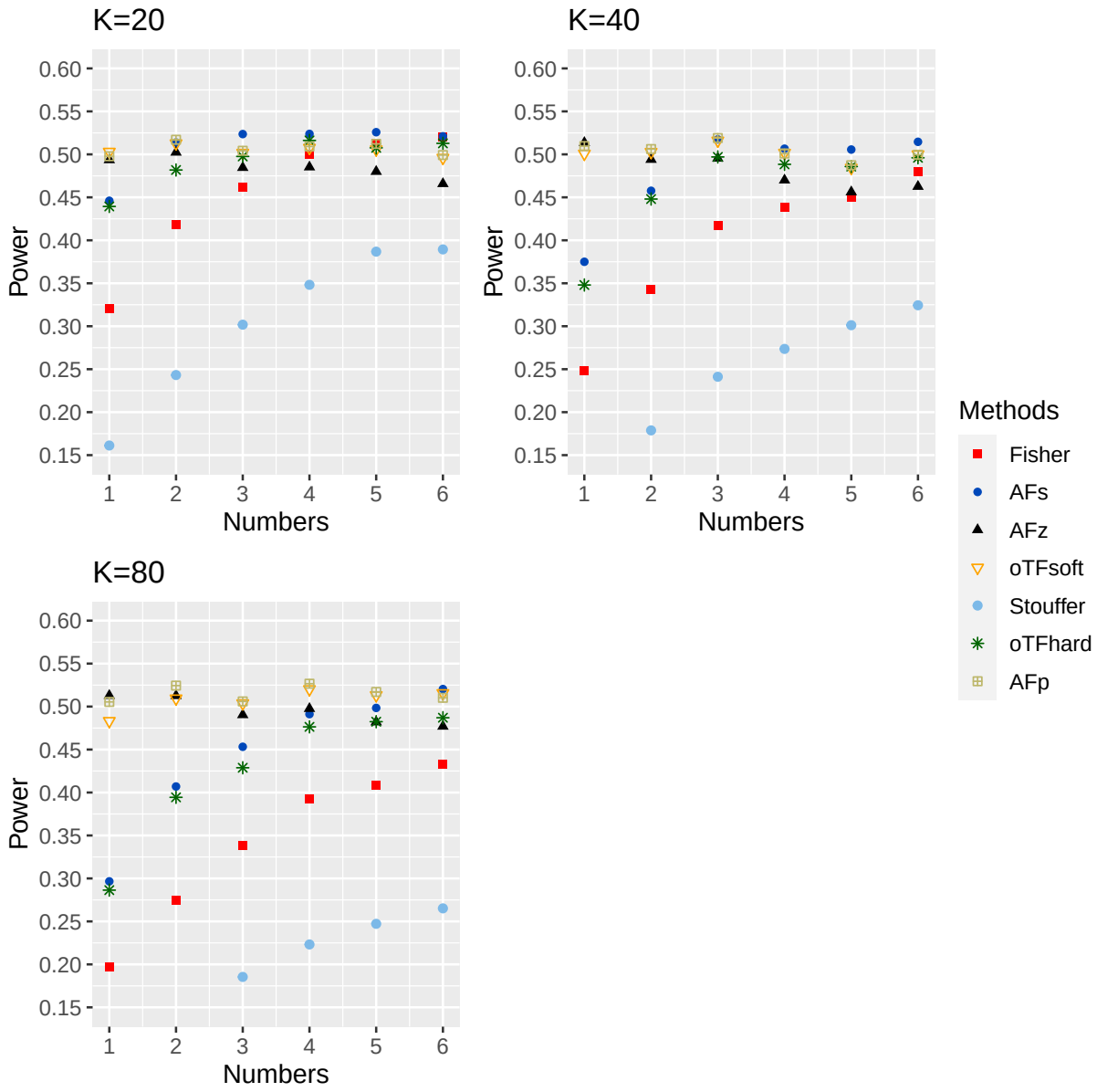


Figure A13: Statistical power of Fisher, Stouffer, and 5 modified Fisher's methods at significance level $\alpha = 0.05$ across varying numbers of true signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible compared to the scale of the mean power and hence omitted.

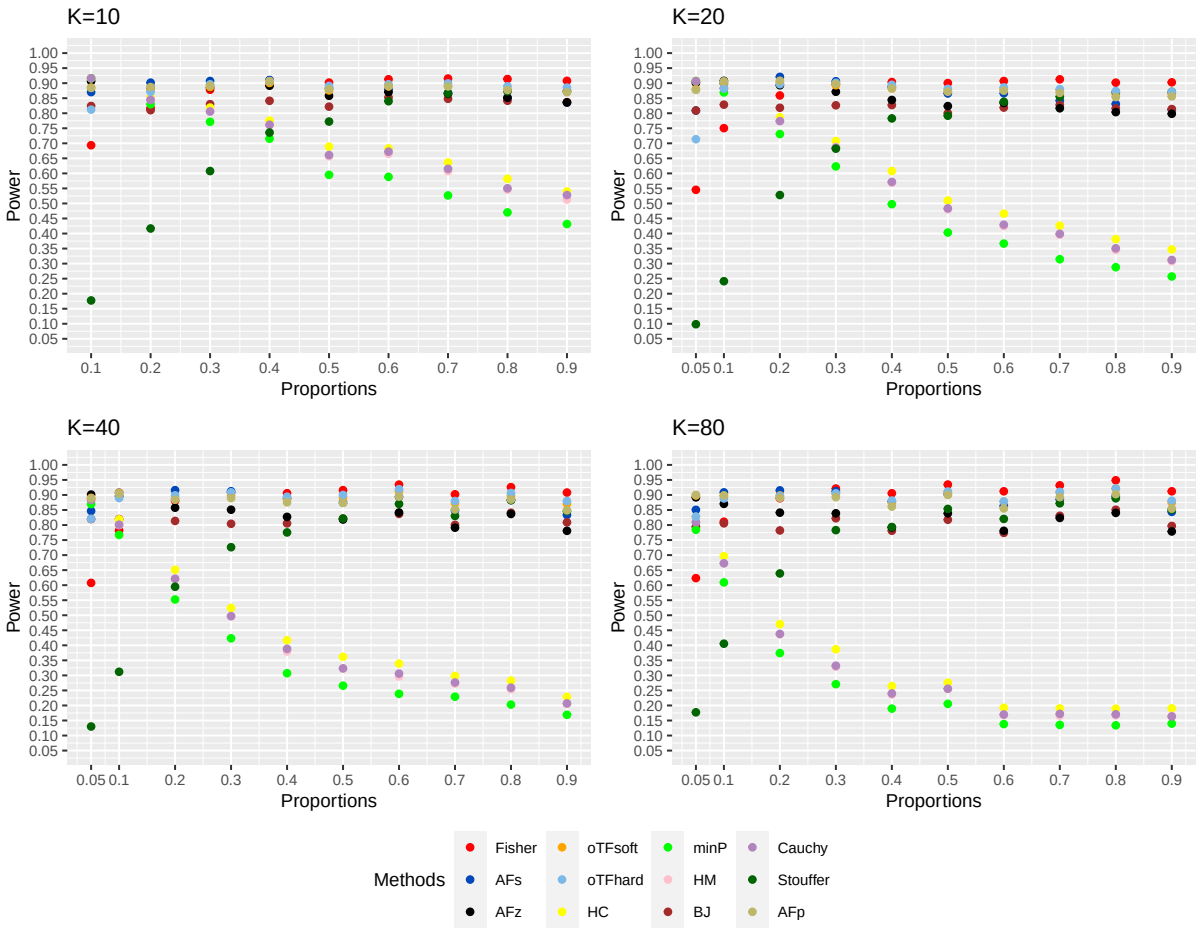


Figure A14: Statistical power of Fisher, AFs, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer at significance level $\alpha = 0.01$ across varying proportions of signals $\ell/K = 0.05, 0.1, 0.2, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ/K and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

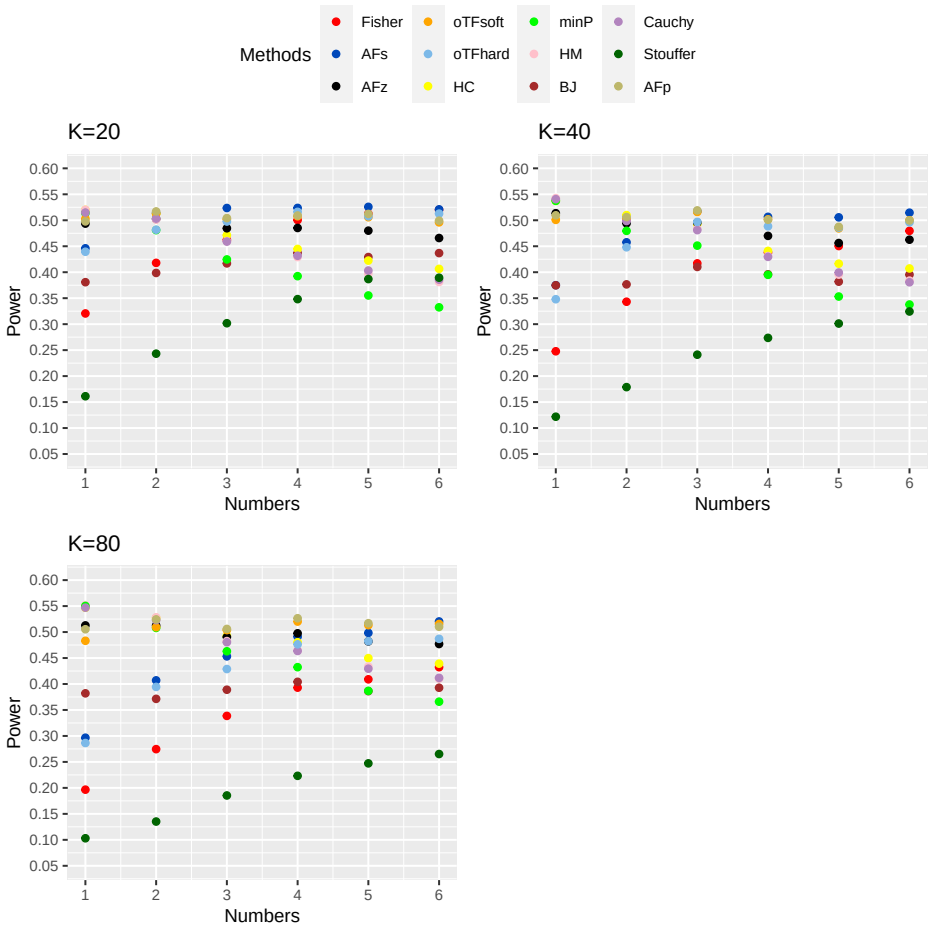


Figure A15: Statistical power of Fisher, AFs, AFp, AFz, oTFsoft, oTFhard, HC, minP, HM, BJ, Cauchy (CA), and Stouffer at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, 3, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

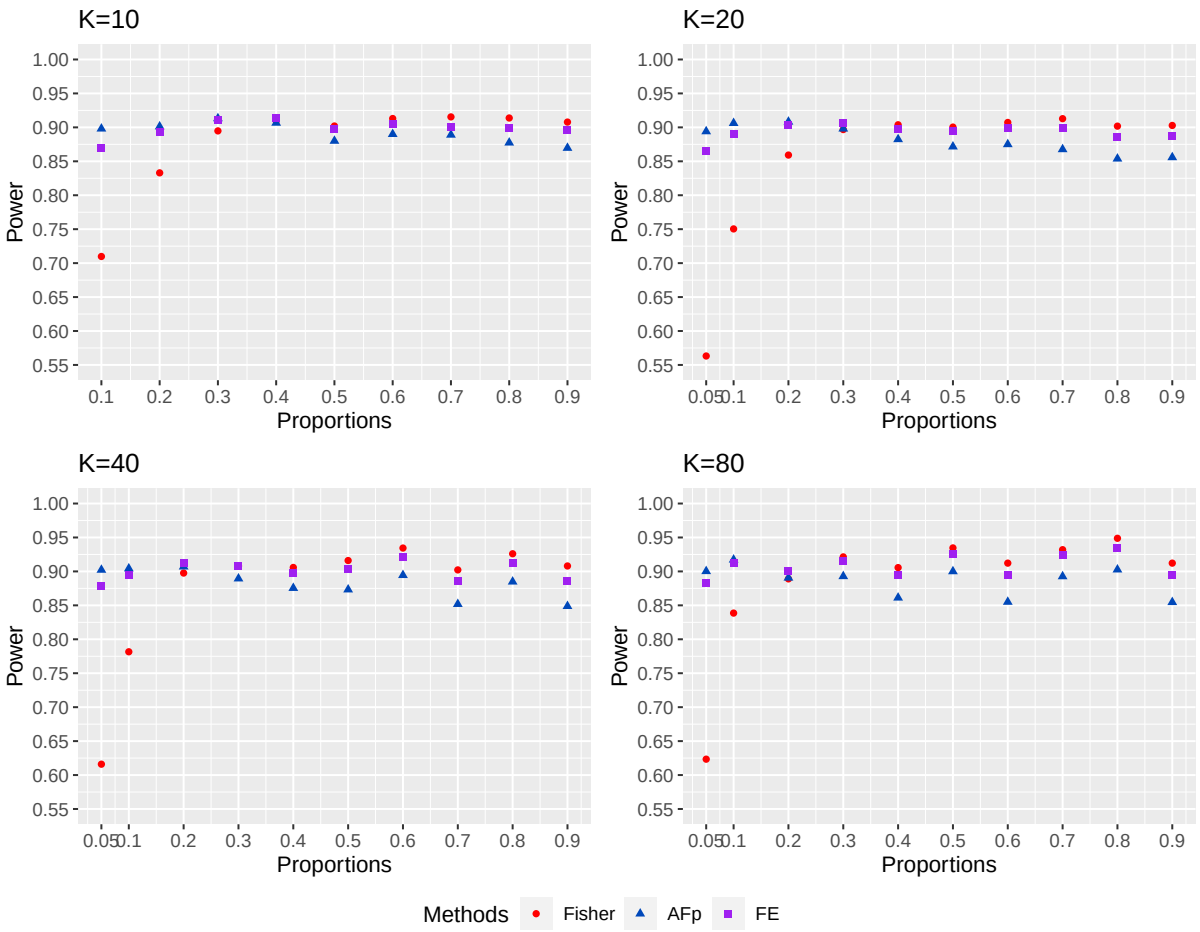


Figure A16: Statistical power of FE, Fisher, and AFp at significance level $\alpha = 0.01$ across varying proportions of signals $\ell/K = 0.05, 0.1 \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ/K and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

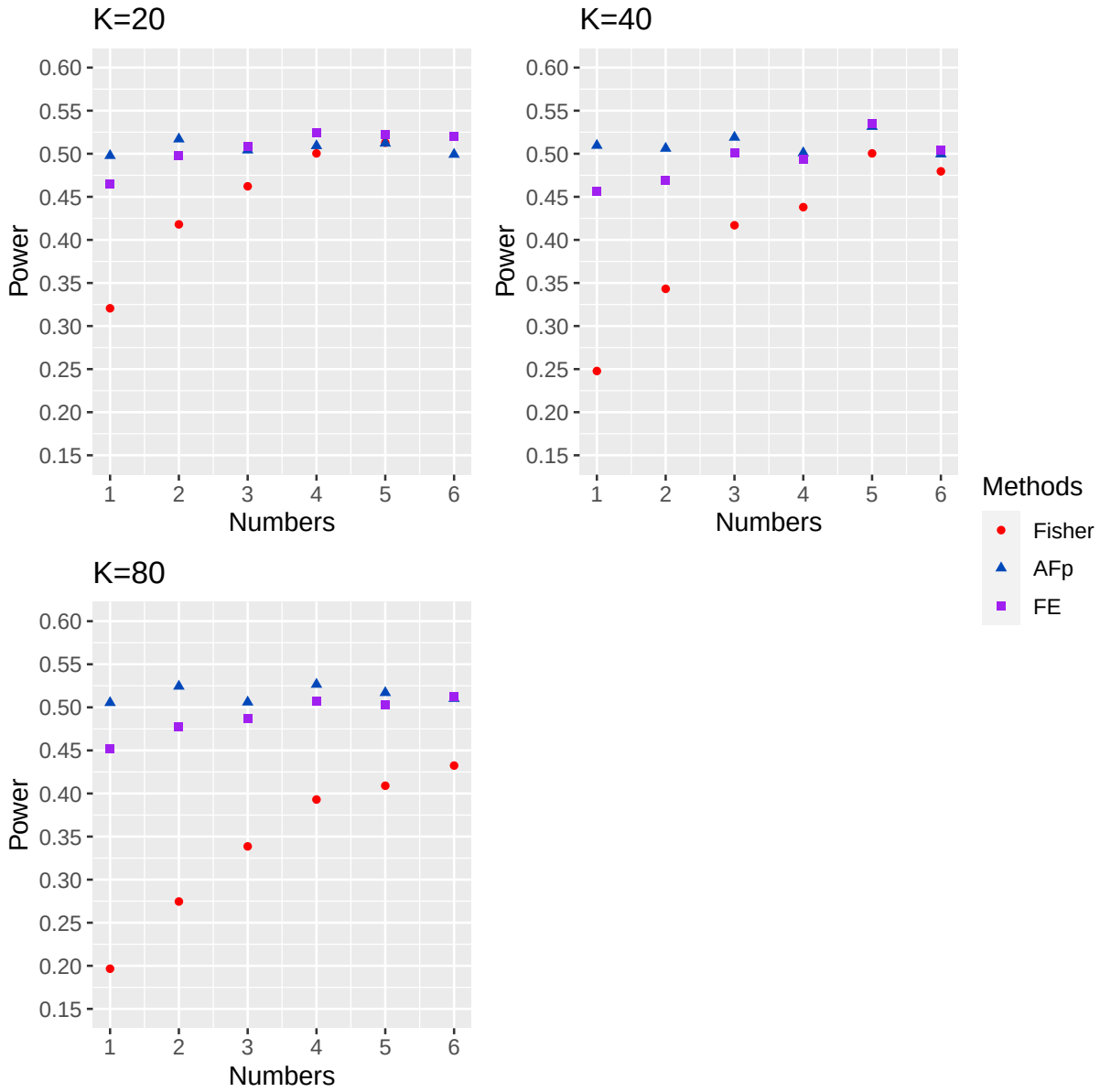


Figure A17: Statistical power of FE, Fisher, and AFp at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

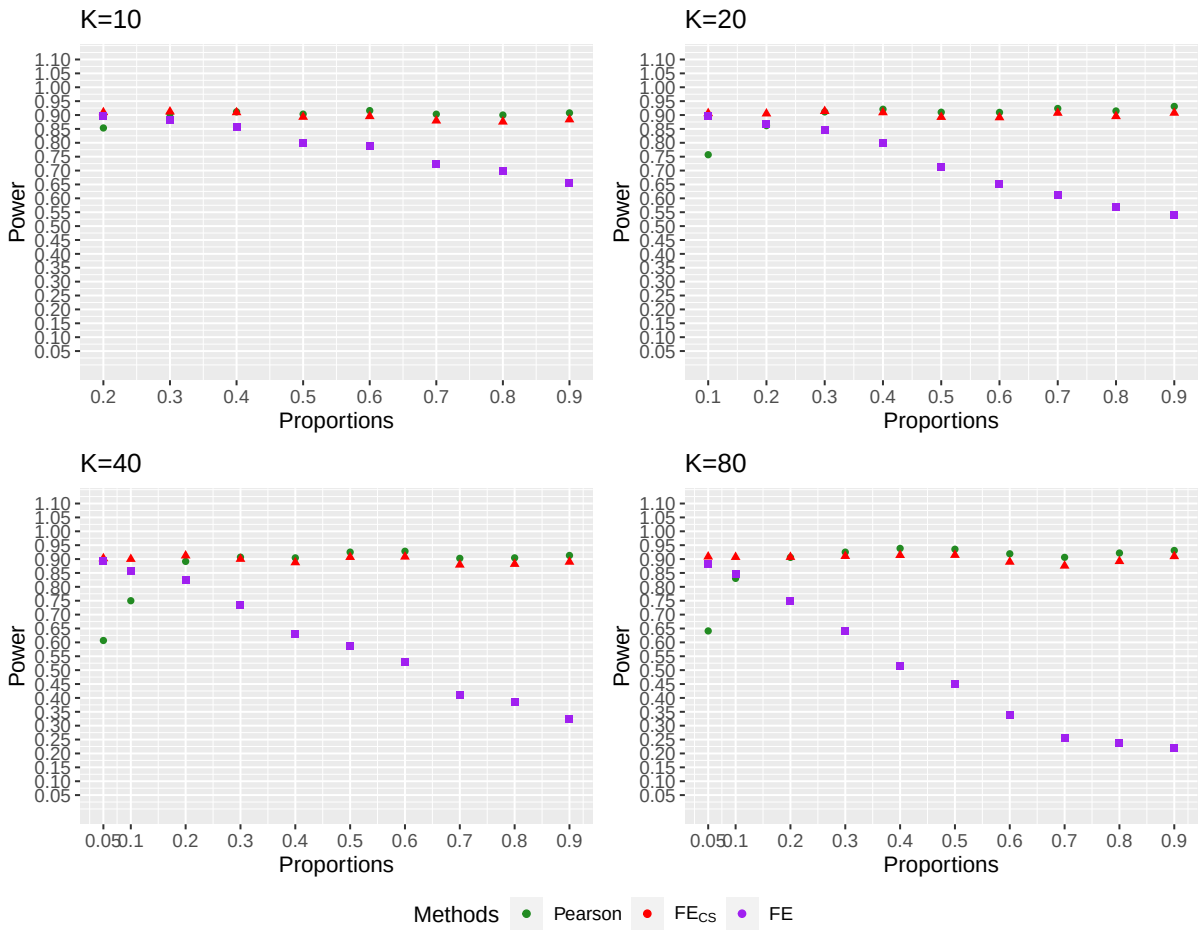


Figure A18: Statistical power of FE, FE_{CS}, and Pearson at significance level $\alpha = 0.01$ across varying proportions of signals $\ell/K = 0.05, 0.1, \ldots, 0.9$ and varying numbers of combined p-values $K = 10, 20, 40, 80$. For each ℓ/K and K, we choose the smallest μ_0 such that the best performer has at least 0.9 statistical power. The standard errors are negligible and hence omitted.

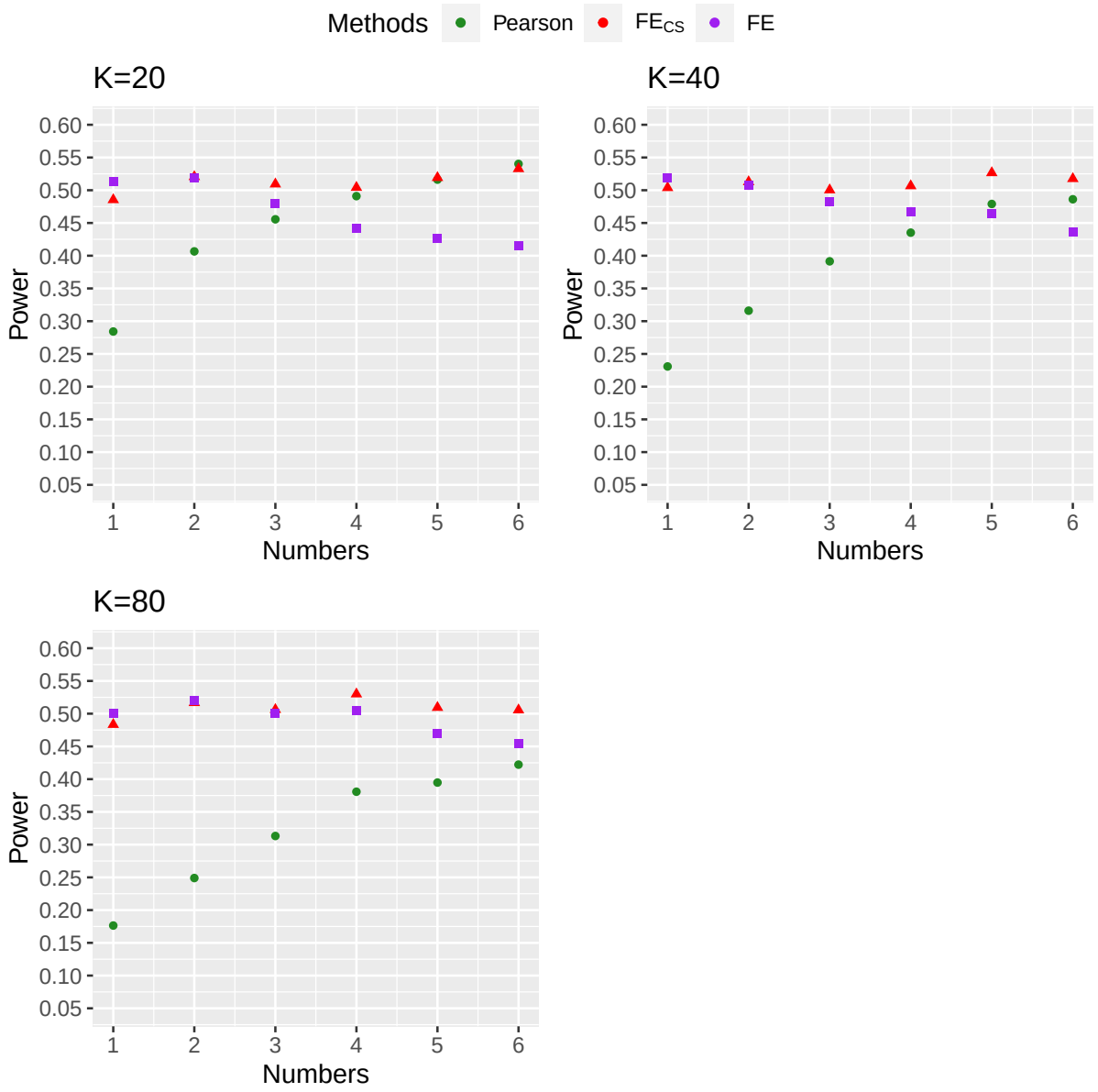


Figure A19: Statistical power of FE, FE_{CS}, and Pearson at significance level $\alpha = 0.05$ across varying numbers of signals $\ell = 1, 2, \ldots, 6$ and varying numbers of combined p-values $K = 20, 40, 80$. For each ℓ and K, we choose the smallest μ_0 such that the best performer has at least 0.5 statistical power. The standard errors are negligible and hence omitted.

Appendix B Supplementary Materials for Chapter 3

B.1 Technical Arguments

B.1.1 Proof of Theorem 3.1

We need 3 lemmas to prove Theorem 3.1. Lemmas B1 and B2 build connection between $p_i = 2(1 - \Phi(|Y_i|))$ and Y_i^2.

Lemma B1 (Vershynin (2018)). *For all $t > 0$, we have:*

$$1 - \Phi(t) \leq \min\left\{\frac{1}{\sqrt{2\pi}}\frac{1}{t}, \frac{1}{2}\right\} e^{-\frac{t^2}{2}}.$$

Lemma B2 (Birgé (2001)). *Let $X \sim \chi_d^2(\nu)$, then for all $x > 0$:*

$$\mathbb{P}[X \geq (d+\nu) + 2\sqrt{(d+2\nu)x} + 2x] \leq \exp(-x) \tag{B1}$$

$$\mathbb{P}[X \leq (d+\nu) - 2\sqrt{(d+2\nu)x}] \leq \exp(-x). \tag{B2}$$

Lemma B3 provides lower bound of $-\log Y_{(i)}$ for $i \geq n - s + 1$.

Lemma B3. *Let $Y_1, \ldots, Y_n$ be independent and identically distributed random variables with the CDF F. If $\eta = F(\varrho) < 1$ for some constant $\varrho > 0$ and $k \leq n^{1-\beta}$ for some $0 < \beta < 1$, then we have:*

$$\mathbb{P}(Y_{(n-k+1)} > \exp\{-(\log n)^{\frac{1}{2}}\}) \to 1 \text{ as } n \to +\infty.$$

of Lemma B3. We begin the proof by considering the CDF of $Y_{(n-k+1)}$ under the null and plug in $\gamma_n = \exp\{-(\log n)^{\frac{1}{2}}\}$ with a sufficiently large n:

$$\begin{aligned}
F_{Y_{(n-k+1)}}(\gamma_n) &= \sum_{j=n-k+1}^{n} \binom{n}{j} F^j(\gamma_n)\left(1-F(\gamma_n)\right)^{n-j} \\
&\leq \sum_{j=n-k+1}^{n} \binom{n}{n-j} F^j(\gamma_n) \\
&\leq \sum_{j=n-k+1}^{n} \binom{n}{n-j} F^j(\varrho) \\
&\leq \eta^{n-k+1} \sum_{j=0}^{k-1} \binom{n}{j} \\
&\leq \eta^{n-k+1} \left(\frac{en}{k-1}\right)^{k-1}
\end{aligned}$$

where the last quantity converges to zero as $k = O(n^{1-\beta})$. □

We prove Theorem 3.1 using the above 3 lemmas.

Proof of Theorem 3.1. We prove that if we pick $C^{(n)} = 2\beta n^{1-\beta}\left(1 + 2/\sqrt{\log\log n}\right)\log n$ as the critical value, both type I and type II errors of $T(s)$ go to zero as n diverges.

The proof of Theorem 3.1 is organized in two parts. In the first part, we prove that the probability of $T(s)$ greater than $C^{(n)}$ goes to 0 as n diverges. In the second part, we show that when $\|\theta\|_2^2 \geq n^{1-\beta}C^{(0)}\log n$, the probability of $T(s)$ greater than $C^{(n)}$ goes to one.

For the first part, note that we can rewrite $T(s)$ in the following form:

$$T(s) = \sum_{i=1}^{s} -2\log p_{(i)} = \sup_{|\mathcal{I}|=s} \sum_{i=1}^{n} -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}}.$$

Note that $-2\log(p_i)$ follows a chi-squared distribution under the null and hence $\sum_{i=1}^{n} -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}}$ follows a chi-squared distribution with degrees of freedom $2s$ for a given $\mathcal{I}$. In the following steps

we use Lemma B2 to bound $\sum_{i=1}^{n} -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}}$ and finally provide a upper bound for $T(s)$. We consider the following inequalities:

$$\begin{aligned}
\mathbb{P}(T(s) \geq 2s + 2\sqrt{2sx} + 2x) &= \mathbb{P}(\sup_{|\mathcal{I}|=s} \sum_{i=1}^{n} -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}} \geq 2s + 2\sqrt{2sx} + 2x) \\
&\leq \binom{n}{s} \mathbb{P}(\chi^2_{2s} \geq 2s + 2\sqrt{2sx} + 2x) \\
&\leq \binom{n}{s} \exp(-x) \\
&\leq \left(\frac{en}{s}\right)^s \exp(-x) \\
&\leq \exp(\beta s \log n + s - x). \qquad \text{(B3)}
\end{aligned}$$

The second inequality is due to (B1) in Lemma B2 and the third is due to the Stirling's formula.

Picking $x = s\beta\left(1 + 1/\sqrt{\log n}\right)\log n$, as n goes to infinity, we have

$$(B3) = \exp\left\{s - s\beta\sqrt{\log n}\right\} \to 0.$$

Note that this choice of x leads to:

$$2s + 2\sqrt{2sx} + 2x < C^{(n)},$$

for sufficiently large n. This implies the probability that $T(s)$ is greater than $C^{(n)}$ goes to 0, hence the type I error converges to zero.

For the second part, without loss of generality, we assume all the non-zero entries of θ are among the first s μ_i's ($i = 1, \ldots, s$). The following arguments show that the probability that $T(s)$ is smaller than $C^{(n)}$ goes to zero as long as $\|\theta\|_2^2 \geq n^{1-\beta}C^{(0)}\log n$. Note that we have:

$$\begin{aligned}
T(s) &= \sum_{i=1}^{s} -2\log p_{(i)} = \sum_{i=1}^{s} -2\log(2(1 - \Phi(|Y|_{(n-i+1)}))) \\
&\geq \sum_{i=1}^{s} |Y|^2_{(n-i+1)} + 2\sum_{i=1}^{s} \log(|Y|_{(n-i+1)}) - 2s\log\sqrt{\pi/2} \\
&\geq \sum_{i=1}^{s} Y_i^2 + 2s\log(|Y|_{(n-s+1)}) - 2s\log\sqrt{\pi/2} \\
&\geq \sum_{i=1}^{s} Y_i^2 + 2s\log(\max\{Y_{(n-s+1)}, 0\}) - 2s\log\sqrt{\pi/2}, \qquad \text{(B4)}
\end{aligned}$$

where $|Y|_{(n-i+1)}$ denotes the $(n-i+1)$-th smallest value of $|Y_i|$'s and $Y_{(n-i+1)}$ denotes the $(n-i+1)$-th smallest value of Y_i's. The first inequality is due to Lemma B1. Note that $\sum_{i=1}^{s} Y_i^2$ follows a chi-squared distribution with degrees of freedom s and non-central parameter $v = \sum_{i=1}^{s} \mu_i^2 \geq n^{1-\beta} C^{(0)} \log n$ under the alternative. Applying (B2) in Lemma B2 and picking the corresponding $x = \sqrt{\log n}$, we have:

$$\begin{aligned}
&\mathbb{P}\Big(\sum_{i=1}^{s} Y_i^2 \leq (s+v) - 2\sqrt{(s+2v)} (\log n)^{1/4} \Big) \\
&\leq \exp(-(\log n)^{1/2}) \to 0,
\end{aligned}$$

Let $\Lambda = \{Y_{s+1}, \ldots, Y_n\}$ be the subset of $Y_1, \ldots, Y_n$. Note that $Y_{s+1}, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$. Let $Y'_{(n-2s+1)}$ be the $(n-2s+1)$-th smallest value from Λ. Applying Lemma B3, we obtain

$$\begin{aligned}
&\lim_{n \to +\infty} \mathbb{P}\big(2s \log(\max\{Y_{(n-s+1)}, 0\}) > -2s\sqrt{\log(n-s)} \big) \\
&\geq \lim_{n \to +\infty} \mathbb{P}\big(Y'_{(n-2s+1)} > \exp(-\sqrt{\log(n-s)}) \big) = 1.
\end{aligned}$$

Combining the above arguments, one can show that with $v = \sum_{i=1}^{s} \mu_i^2 \geq n^{1-\beta} C^{(0)} \log n$ and $C^{(0)} > 2\beta$

$$\begin{aligned}
&\lim_{n \to +\infty} \mathbb{P}\Big(\sum_{i=1}^{s} Y_i^2 + 2s \log(\max\{Y_{(n-s+1)}, 0\}) - 2s \log \sqrt{\pi/2} \\
&> (s+v) - 2\sqrt{(s+2v)} (\log n)^{1/4} - 2s(\log(n-s))^{1/2} - 2s \log \sqrt{\pi/2} \Big) = 1,
\end{aligned}$$

indicating the probability that $T(s)$ is smaller than $C^{(n)} = 2\beta n^{1-\beta}(1 + 2/\sqrt{\log \log n}) \log n$ converges to zero, hence the type II error goes to zero as n diverges. □

B.1.2 Proof of Theorem 3.2

Similar to the proof of Theorem 3.1, the proof of Theorem 3.2 is structured in two parts. In the first part we show that searching across the candidate set $\mathcal{S}$ will not substantially inflate type I error of AFg. For the second part, we show that if $\|\theta\|_2^2 \geq n^{1-\beta}C^{(0)}\log n$, we can always find a candidate truncation point s_i, such that $T(s_i)$ will be greater than C_k with probability going to one.

Proof of Theorem 3.2. For the first part of proof, denote by $|\mathcal{S}| = M+1$ the number of elements in the candidate set $\mathcal{S}$, we have

$$\begin{aligned} M+1 &\leq \log_{1+1/\log n}(\sqrt{n}) + 2 \\ &= \frac{\log n}{2\log(1+1/\log n)} + 2 \\ &\leq (1/2)(\log n)^2 + (1/2)\log n + 2. \end{aligned} \tag{B5}$$

The order of $|\mathcal{S}|$ is much smaller than n, which is critical to control the type I error. Define the event:

$$A_i(x_{s_i}) = \Big\{ \sup_{|\mathcal{I}_i|=s_i} \sum_{j\in\mathcal{I}_i} -2\log(p_j)\mathrm{I}_{\{j\in\mathcal{I}_i\}} \geq 2s_i + 2\sqrt{2s_i x_{s_i}} + 2x_{s_i} \Big\}.$$

Note that for each $i = 0, \ldots, M$, we have

$$T(s_i) = \sum_{j=1}^{s_i} -2\log p_{(j)} = \sup_{|\mathcal{I}_i|=s_i} \sum_{j\in\mathcal{I}_i} -2\log(p_j)\mathrm{I}_{\{j\in\mathcal{I}_i\}}.$$

Also note that each $\sum_{j\in\mathcal{I}_i} -2\log(p_j)\,\mathrm{I}_{\{j\in\mathcal{I}_i\}}$ follows a chi-squared distribution with degrees of freedom $2s_i$ under the null. Then we have

$$\begin{aligned} &\mathbb{P}\big(\cup_{i=0}^{M} A_i(x_{s_i})\big) \\ &\leq \sum_{i=0}^{M} \binom{n}{s_i} \mathbb{P}\Big(\chi^2_{2s_i} \geq 2s_i + 2\sqrt{2s_i x_{s_i}} + 2x_{s_i}\Big) \\ &\leq \sum_{i=0}^{M} \binom{n}{s_i} \exp(-x_{s_i}) \\ &\leq \sum_{i=0}^{M} \Big\{\frac{en}{s_i}\Big\}^{s_i} \cdot \exp(-x_{s_i}) \\ &= \sum_{i=0}^{M} \exp\left\{s_i\log(n/s_i) + s_i - x_{s_i}\right\}, \end{aligned} \tag{B6}$$

where we apply (B1) to the second inequality. Picking $x_{s_i} = s_i \left(1 + 1/\sqrt{\log\log n}\right) \log\left(n/s_i\right) = s_i(1+\delta_n)\log\left(n/s_i\right)$ for $i = 0, \ldots, M$ and combining (B5) with (B6), we have:

$$(B6) \leq ((1/2)(\log n)^2 + (1/2)\log n + 2) \cdot \exp\left(1 - \log n/(2\sqrt{\log\log n})\right) \to 0.$$

Note that the choice of x_{s_i} leads to:

$$2s_i + 2\sqrt{2s_i x_{s_i}} + 2x_{s_i} < C_i,$$

for each $i = 0, \ldots, M$ when n is sufficiently large. Thus, under null we have:

$$\mathbb{P}\big(\cup_{i=0}^{M}\left\{T(s_i) > C_i\right\}\big) \to 0,$$

implying that the type I error is well controlled.

In the second part, we show that we can find at least one s_i such that $T(s_i) > C_i$ with probability goes to one, which implies that the type II error converges to zero as n diverges. Again, we assume all the non-zero entries of θ are among the first s μ_i's ($j = 1, \ldots, s$) without loss of generality.

Let $0 < i^* < \log_{1+1/\log n}(n^{\frac{1}{2}})$ for $s = \lceil n^{1-\beta} \rceil$ such that:

$$s_{i^*-1} \leq s \leq s_{i^*}.$$

Then we have:

$$T(s_{i^*}) = \sum_{j=1}^{s_{i^*}} -2\log p_{(j)} \geq \sum_{j=1}^{s} -2\log p_{(j)} \geq \sum_{j=1}^{s} -2\log p_j.$$

By similar arguments in the second part of proof of Theorem 3.1, as long as $\|\theta\|_2^2 \geq n^{1-\beta}C^{(0)}\log n$, we have:

$$\lim_{n\to\infty} \mathbb{P}(T(s_{i^*}) > n^{1-\beta}C^{(0)}\log n) \geq \lim_{n\to\infty} \mathbb{P}\big(\sum_{i=1}^{s} -2\log p_j > n^{1-\beta}C^{(0)}\log n\big) = 1.$$

Meanwhile, note that for the corresponding critical values C_{i^*}, we have

$$\begin{aligned} C_{i^*} &= 2s_{i^*}(1+2\delta_n)\log(n/s_{i^*}) \\ &\leq 2s(1+1/\log n)(1+2\delta_n)\log\left(n/s\right). \end{aligned}$$

Note that as n diverges, $2s(1+1/\log n)(1+2\delta_n)\log\left(n/s\right) < n^{1-\beta}C^{(0)}\log n$. Hence the probability that $T(s_{i^*})$ is greater than C_{i^*} goes to one as long as $\|\theta\|_2^2 \geq n^{1-\beta}C^{(0)}\log n$. □

B.1.3 Proof of Theorem 3.3

Proof of Theorem 3.3. In order to prove Theorem 3.3, we prove (i) and (ii) under Setup 3.1 with $\frac{1}{2} < \beta < 1$ in Section 3.3. As shown in Theorem 3.2, the sum of type I and type II errors of AFg goes to zero as $L_n \geq C^{(0)} \cdot n^{1-\beta} \log n$. Hence the proof of (i) is done.

We continue on the proof of (ii), which is essentially to find the lower bound of the following minimax risk $\mathcal{R}_{0,\beta}(L_n)$ and show that it is greater than $1 - \varepsilon$ for a properly chosen a_ε and all $L_n \leq a_\varepsilon \lambda_n$.

$$\mathcal{R}_{0,\beta}(L_n) = \inf_{\varphi} \{\mathbb{P}_0(\varphi = 1) + \sup_{\theta \in \Theta_{0,\beta}(L_n)} \mathbb{P}_\theta(\varphi = 0)\},$$

For the characterization of lower bound of $\mathcal{R}_{0,\beta}(L_n)$, a standard scheme is to reduce it to quantifying the "distance" between two probability measures that are respectively associated with the null and alternative parameter spaces. This argument is firstly considered by Le Cam (LeCam, 1973). More precisely, we consider the following lemma as a special case of Le Cam's technique by considering the chi-squared divergence between a mixture measure associated with $\Theta_{0,\beta}(L_n)$ and $\mathbb{P}_0$, the probability measure associated with the null.

Lemma B4 (Tsybakov (2008))**.** *Let μ be a probability measure on the alternative parameter space $\Theta_{0,\beta}(L_n)$, denote by $\mathbb{P}_\mu$ the mixture probability measure:*

$$\mathbb{P}_\mu = \int_{\Theta_{0,\beta}(L_n)} P_\theta \mu(d\theta).$$

Then we have:

$$\mathcal{R}_{0,\beta}(L_n) \geq 1 - \sqrt{\chi^2\left(\mathbb{P}_\mu, \mathbb{P}_0\right)},$$

where $\chi^2(\cdot, \cdot)$ is the chi-squared divergence defined as follows:

$$\chi^2\left(\mathbb{P}', \mathbb{P}\right) = \int \left(d\mathbb{P}'/d\mathbb{P}\right)^2 d\mathbb{P} - 1.$$

Here $\mathbb{P}'$ and $\mathbb{P}$ are two mutually absolute continuous probability measures.

The argument in Lemma B4 is rather standard. The key is to carefully find the least favorable prior on the union of the null and the alternative so that it is impossible to distinguish them. In our case, one needs to properly choose the μ and derive a tight enough upper bound of $\chi^2\left(\mathbb{P}_\mu, \mathbb{P}_0\right)$. Let μ_ρ be the uniform distribution on the set of $\theta \in \Theta_{0,\beta}(L_n)$ such that $\|\theta\|_0 = s$ and all the nonzero entries equal some $\rho > 0$. From the extensive literature on this problem (e.g., Baraud (2002)), we consider the following result developed in Collier et al. (2017), which chooses $\mu = \mu_\rho$ and derives the upper bound of $\chi^2\left(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0\right)$. To our best knowledge, this is among the sharpest results on the lower bound of $\mathcal{R}_{0,\beta}(L_n)$

Lemma B5 (Collier et al. (2017))**.** *For all $\rho > 0$ and $1 \le s \le n$, we have:*

$$\chi^2\left(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0\right) \le \left(1 - \frac{s}{n} + \frac{s}{n} e^{\rho^2}\right)^s - 1,$$

where $\mathbb{P}_{\mu_\rho}$ represents the mixture probability measure defined by μ_ρ.

We then complete the proof of Theorem 3.3 by proving (ii) using the above two lemmas. Let $s' = \lfloor n^{1-\beta} \rfloor$ with $1/2 < \beta < 1$ and $A \in (0,1)$ be some constant, where $\lfloor \cdot \rfloor$ is the floor operator that finds the the largest integer that smaller than $\lfloor x \rfloor$. Let $\rho = \sqrt{A \log n}$ and hence $L_n = A s' \log n$. Note that $\rho^2 = A \log n$. By Lemma B5, we have

$$\begin{aligned}\chi^2\left(\mathbb{P}_{\mu_\rho}, \mathbb{P}_0\right) &\le \left(1 - \frac{s'}{n} + \frac{s'}{n} n^A\right)^{s'} - 1 \le \exp\left\{(s')^2 (n^A - 1)/n\right\} - 1 \\ &\le \exp\left\{n^{1-2\beta+A}\right\} - 1.\end{aligned}$$

Applying Lemma B4, we have

$$\mathcal{R}_{0,\beta}(L_n) \ge 1 - \sqrt{\exp\left\{n^{1-2\beta+A}\right\} - 1}.$$

For any $\varepsilon \in (0,1)$, as long as $A \le (2\beta - 1) + \log\log(1+\varepsilon^2)/\log n$, we have $\mathcal{R}_{0,\beta}(L_n) \ge 1 - \varepsilon$. For a sufficiently large n, we may pick $A = (2\beta - 1)/2$ to satisfy this condition. □

B.1.4 Proof of Theorem 3.4

The idea of the proof is similar to that in the proof of Theorem 3.1. In addition, we need the following lemma to control the deviation of Studentized statistics from a standard normal random variable:

Lemma B6 (Delaigle et al. (2011))**.** *Consider independent and identically distributed mean-zero random variables $X_1, ..., X_m$ with $E(|X_i|^4) \leq B, E(X_i^2) = 1$ and $\gamma = E(X_i^3) < +\infty$. Let $B > 1$ be a constant and $T = \sqrt{m}\bar{X}/L_X$ with $\bar{X} = (1/m)\sum_{i=1}^m X_i$ and $L_X = \sqrt{(1/m)\sum_{i=1}^m (X_i - \bar{X})^2}$. Then*

$$\frac{\mathbb{P}\left(T > x\right)}{1 - \Phi(x)} = \exp\Big\{ -\frac{1}{6} m^{-1/2}\left(2x^3\right)\gamma \Big\}\Big\{1 + \Gamma(m, x)\big((1 + |x|)m^{-1/2} + (1 + |x|)^4 m^{-1}\big)\Big\},$$

uniformly in x satisfying $0 \leq x \leq Bm^{1/4}$ as $m \to \infty$, where the function Γ is bounded in absolute value by a finite, positive constant $C_1(B)$ (which depends on B only).

Proof of Theorem 3.4. Without loss of generality, we assume $\sigma^2 = 1$. The proof is structured in two parts similarly to the proofs of Theorems 3.1 and 3.2. For the first part, recall that we can rewrite the test statistic as:

$$T(s) = \sum_{i=1}^{s} -2\log p_{(i)} = \sup_{|\mathcal{I}|=s} \sum_{i=1}^{n} -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}}.$$

Note that under null, the random variables $Y_{ij}, i = 1, \ldots, n, j = 1, \ldots, m$ are independent and identically distributed with zero means. Denote $Z_i = \sqrt{m}\bar{Y}_i$ and $L_i^2 = \frac{1}{m}\sum_{j=1}^m \left(Y_{ij} - \bar{Y}_i\right)^2 = \frac{1}{m}\sum_{j=1}^m Y_{ij}^2 - \bar{Y}_i^2$. Then by Markov's inequality and Marcinkiewicz–Zygmund inequality (note that $D = \max\{6\eta + \varepsilon, 4\}$ implies $\frac{D}{2} \geq 2$, and $\mathbb{E}|Y_{ij}|^D \leq B_0 < \infty$), using arguments similar to Theorem 3 in **?**, we have:

$$\mathbb{P}\left(|Z_i| > t_1\right) \leq \frac{C_1(D, B_0)}{t_1^D} \tag{B7}$$

$$\mathbb{P}\Big(\big|(1/m)\sum_{j=1}^m Y_{ij}^2 - 1\big| > t_2\Big) \leq \frac{C_2(D, B_0)\mathbb{E}\big\{\big|Y_{ij}^2 - 1\big|^{\frac{D}{2}}\big\}}{m^{\frac{D}{4}} t_2^{\frac{D}{2}}} \leq \frac{C_3(D, B_0)}{m^{\frac{D}{4}} t_2^{\frac{D}{2}}}, \tag{B8}$$

where $C_1(D, B_0)$, $C_2(D, B_0)$ and $C_3(D, B_0)$ are some absolute constants that only depend on D and B_0.

We consider two events, denoted by $B_1 = \cup_i\{|Z_i| > m^{\frac{1}{6}}\}$ and $B_2 = \cup_i\{|\frac{1}{m}\sum_{j=1}^m Y_{ij}^2 - 1| > \frac{1}{\log m}\}$. For the two events, applying inequalities (B7) and (B8), we have as m diverges:

$$\mathbb{P}\left(B_1 = \cup_i\left\{|Z_i| > m^{\frac{1}{6}}\right\}\right) \leq \frac{nC_1(D,B_0)}{m^{\frac{D}{6}}} = \frac{C_1(D,B_0)}{m^{\frac{D}{6}-\eta}} \to 0$$
$$\mathbb{P}\Big(B_2 = \cup_i\Big\{\Big|\frac{1}{m}\sum_{j=1}^m Y_{ij}^2 - 1\Big| > \frac{1}{\log m}\Big\}\Big) \leq \frac{nC_3(D,B_0)(\log m)^{\frac{D}{2}}}{m^{\frac{D}{4}}} = \frac{C_3(D,B_0)(\log m)^{\frac{D}{2}}}{m^{\frac{D}{4}-\eta}} \to 0,$$

which imply that $\mathbb{P}\left(B_1^c \cap B_2^c\right) \to 1$. Under $B_1^c \cap B_2^c$, one can bound all the T_i's simultaneously. In addition, note that under event B_1^c, we also have $\left|\bar{Y}_i\right| < m^{-\frac{1}{3}}$ and hence $1 - m^{-\frac{2}{3}} - \frac{1}{\log m} \leq L_i^2 \leq 1 + \frac{1}{\log m}$. Hence under event $B_1^c \cap B_2^c$, we have

$$|T_i| < (1 + a_m)m^{\frac{1}{6}} \text{ for all } i = 1, \ldots, n,$$

where $a_m \to 0$ as m diverges.

Note that each p-value $p_i = 2(1 - \Phi(|T_i|))$ $(i = 1, \ldots, n)$ is not exact as T_i does not follow $\mathrm{N}(0,1)$ under the null in general. Here we use Lemma B6 to bound the deviation of p_i from the exact p-value $\tilde{p}_i$. Indeed, under event $B_1^c \cap B_2^c$, applying Lemma B6 twice for both $\mathbb{P}(T > x)$ and $\mathbb{P}(T < -x)$ respectively, we have:

$$\begin{aligned}
\sum_{i=1}^s -2\log p_{(i)} &= \sup_{|\mathcal{I}|=s}\sum_{i=1}^n -2\log(p_i)\mathrm{I}_{\{i\in\mathcal{I}\}} \\
&\leq \sup_{|\mathcal{I}|=s}\sum_{i=1}^n \Big\{ -2\log(\tilde{p}_i)\mathrm{I}_{\{i\in\mathcal{I}\}} + \frac{1}{3}m^{-1/2}\left|2\hat{t}_i\right|^3|\gamma|\mathrm{I}_{\{i\in\mathcal{I}\}} \\
&\quad + 2\log\left(1 + \Gamma'(m,\hat{t}_i)\left\{(1+|\hat{t}_i|)m^{-1/2} + (1+|\hat{t}_i|)^4 m^{-1}\right\}\right)\mathrm{I}_{\{i\in\mathcal{I}\}}\Big\} \qquad \text{(B9)} \\
&= \sup_{|\mathcal{I}|=s}\sum_{i=1}^n -2\log(\tilde{p}_i)\mathrm{I}_{\{i\in\mathcal{I}\}} + sC(\gamma,B), \qquad \text{(B10)}
\end{aligned}$$

where $\Gamma'(m,x)$ is some function bounded by a finite, positive constant $C_1'(B)$ depending on B only, and $C(\gamma,B)$ is some constant that depends only on γ and B. By truncating $|T_i|$'s uniformly above by some quantity in the order of $O(m^{\frac{1}{6}})$, we not only meet the conditions of Lemma B6, but also bound the major deviation term $\frac{1}{3}m^{-1/2}(|2\hat{t}_i|^3|\gamma|)$ in (B9) by some constant. For the term $\sup_{|\mathcal{I}|=s}\sum_{i=1}^n -2\log(\tilde{p}_i)\mathrm{I}_{\{i\in\mathcal{I}\}}$ in (B10), note that as $\tilde{p}_i$ is the exact p-value and follows $\mathrm{Unif}(0,1)$ under the null, we have $\sum_{i=1}^n -2\log(\tilde{p}_i)\mathrm{I}_{\{i\in\mathcal{I}\}} \sim \chi_{2s}^2$ for any $\mathcal{I}$ with $|\mathcal{I}| = s$ under the null. Replacing p_i by $\tilde{p}_i$ and applying the same argument of (B3) in the proof of Theorem 3.1, we can

obtain that $(B10) \leq 2n^{1-\beta}\beta\left(1+2/\sqrt{\log\log n}\right)\log n = C^{(n)}$ as n and m diverge. Finally, by a union bound argument we achieve $\mathbb{P}\left(T(s) \leq C^{(n)}\right) \to 1$ as n and m diverge. Therefore, the type I error is still well-controlled if we use $C^{(n)}$ as the critical value.

For the second part, we show the type II error is controlled by using $C^{(n)}$ as the critical value. Without loss of generality, we assume all the non-zero entries of θ are among the first μ_i's ($i = 1, \ldots, s$). Then we have:

$$\|\theta\|_2^2 = \sum_{i=1}^{s} \mu_i^2 \geq n^{1-\beta} C^{(0)} \log n/m. \tag{B11}$$

Let $|T|_{(n-i+1)}$ be the $(n-i+1)$-th smallest value of $|T_i|$'s and $T_{(n-i+1)}$ be the $(n-i+1)$-th smallest value of T_i's, then we have:

$$\begin{aligned}
T(s) &= \sum_{i=1}^{s} -2\log\left(2\left(1-\Phi\left(|T|_{(n-i+1)}\right)\right)\right) \\
&\geq \sum_{i=1}^{s} |T|_{(n-i+1)}^2 + 2\sum_{i=1}^{s} \log\left(|T|_{(n-i+1)}\right) - 2s\log\sqrt{\pi/2} \\
&\geq \sum_{i=1}^{s} T_i^2 + 2\sum_{i=1}^{s} \log\left(|T|_{(n-i+1)}\right) - 2s\log\sqrt{\pi/2} \\
&\geq \sum_{i=1}^{s} T_i^2 + 2s\log(\max\{T_{(n-s+1)}, 0\}) - 2s\log\sqrt{\pi/2}.
\end{aligned} \tag{B12}$$

Inequality (B12) is similar to inequality (B4) in the proof of Theorem 3.1. The only difference is that we replace all the Y_i's by T_i's. Similar to the idea in the proof of Theorem 3.1, the following arguments show that $\sum_{i=1}^{s} T_i^2$ is greater than $\sum_{i=1}^{s} \mu_i^2$ and the effect of other terms in (B12) are negligible.

We first introduce some notations. Let $Y_{ij}^* = Y_{ij} - \mu_i$, then we have:

$$\begin{aligned}
\bar{Y}_i^* &= \frac{1}{m}\sum_{j=1}^{m}(Y_{ij}-\mu_i) = \frac{1}{m}\sum_{j=1}^{m} Y_{ij}^* = \frac{Z_i}{\sqrt{m}} - \mu_i \\
L_i^2 &= \frac{1}{m}\sum_{j=1}^{m}\left(Y_{ij}-\mu_i-\left(\bar{Y}_i-\mu_i\right)\right)^2 = \frac{1}{m}\sum_{j=1}^{m}\left(Y_{ij}^*-\bar{Y}_i^*\right)^2 = \frac{1}{m}\sum_{j=1}^{m} Y_{ij}^{*2} - \bar{Y}_i^{*2}.
\end{aligned} \tag{B13}$$

Consider events $B_1^* = \cup_i \big\{|\bar{Y}_i^*| > (B/\log m)^{\frac{1}{2}}\big\}$ and $B_2^* = \cup_i \big\{|(1/m)\sum_{j=1} Y_{ij}^{*2} - 1| > 1/(\log m)^{1/2}\big\}$, similar to the arguments for (B7) and (B8), by Markov' inequality and Marcinkiewicz–Zygmund inequality, we have:

$$\mathbb{P}\big(B_1^* = \cup_i\{|\bar{Y}_i^*| > \Big(\frac{B}{\log m}\Big)^{\frac{1}{2}}\}\big) \leq \frac{nC_4(D,B_0)(\log m)^{\frac{D}{2}}}{B^{\frac{D}{2}} m^{\frac{D}{2}}} = \frac{C_4(D,B_0)(\log m)^{\frac{D}{2}}}{B^{\frac{D}{2}} m^{\frac{D}{2}-\eta}} \to 0 \quad \text{and}$$

$$\mathbb{P}\big(B_2^* = \cup_i\big\{|\frac{1}{m}\sum_{j=1}^{m} Y_{ij}^{*2} - 1| > \frac{1}{\log m}\big\}\big) \leq \frac{nC_5(D,B_0)(\log m)^{\frac{D}{2}}}{m^{\frac{D}{4}}} = \frac{C_5(D,B_0)(\log m)^{\frac{D}{2}}}{m^{\frac{D}{4}-\eta}} \to 0,$$

implying that $\mathbb{P}((B_1^{*c} \cap B_2^{*c})) \to 1$. Here $C_4(D,B_0)$ and $C_5(D,B_0)$ are some absolute constants that only depend on D and B_0. Under event $B_1^{*c} \cap B_2^{*c}$, the range of all the $\bar{Y}_i^*$'s and $\frac{1}{m}\sum_{j=1}^{m} Y_{ij}^{*2}$'s are upper bounded simultaneously. Hence we can bound all the L_i's:

$$L_i^2 \leq 1 + \frac{1+B}{\log m} \text{for all } i = 1, \ldots, n.$$

We first bound the second term in (B12) by arguments similar to the one in the proof in Theorem 3.1. Let $\Lambda' = \{T_{s+1}, \ldots, T_n\}$ be the subset of $T_1, \ldots, T_n$. Note that $\mu_i = 0$ for $i = s+1, \ldots, n$, hence $T_{s+1}, \ldots, T_n$ are independent and identically distributed. Denote by $T'_{(n-2s+1)}$ the $(n-2s+1)$-th smallest value from Λ'. By Lemma B6, for sufficiently large m, there exists a positive constant $\varrho_0 > 0$, such that

$$F_{T_i}(\varrho_0) = 1 - \mathbb{P}\left(T_i > \varrho_0\right) = 1 - (1 - \Phi(\varrho_0))(1 - C_m(\varrho_0, \gamma, B)) < 1,$$

for each $T_i \in \Lambda'$. Here the function $C_m(\varrho_0, \gamma, B)$ depends only on m, ϱ_0, γ, B and decreases to zero as m increases. Then by Lemma B3, we have:

$$\begin{aligned} &\lim_{m,n\to+\infty} \mathbb{P}\left(2s\log(\max\{T_{(n-s+1)}, 0\}) > -2s\sqrt{\log(n-s)}\right) \\ &\geq \lim_{m,n\to+\infty} \mathbb{P}\left(2s\log(T'_{(n-2s+1)}) > -2s\sqrt{\log(n-s)}\right) = 1. \end{aligned} \tag{B14}$$

Hence we are able to bound the second term in (B12). We then consider the first term in (B12). Note that:

$$\begin{aligned}\sum_{i=1}^{s} T_i^2 &= \sum_{i=1}^{s}(Z_i-\sqrt{m}\mu_i)^2/L_i^2 + 2\sum_{i=1}^{s}(Z_i-\sqrt{m}\mu_i)\sqrt{m}\mu_i/L_i^2 + m\sum_{i=1}^{s}\mu_i^2/L_i^2 \\ &\geq \sum_{i=1}^{s}\frac{(Z_i-\sqrt{m}\mu_i)^2}{1+(1+B)(\log m)^{-1}} + 2\sum_{i=1}^{s}\frac{(Z_i-\sqrt{m}\mu_i)\sqrt{m}\mu_i}{1+(1+B)(\log m)^{-1}} + \frac{m\sum\mu_i^2}{1+(1+B)(\log m)^{-1}} \\ &\geq 2\sum_{i=1}^{s}\frac{(Z_i-\sqrt{m}\mu_i)\sqrt{m}\mu_i}{1+(1+B)(\log m)^{-1}} + \frac{m\sum\mu_i^2}{1+(1+B)(\log m)^{-1}}. \end{aligned} \tag{B15}$$

For the first term in (B15), note that $Z_i-\sqrt{m}\mu_i = (1/\sqrt{m})\sum_{j=1}^{m}(Y_{ij}-\mu_i) = (1/\sqrt{m})\sum_{j=1}^{m}Y_{ij}^*$, hence we have:

$$\sum_{i=1}^{s}\frac{(Z_i-\sqrt{m}\mu_i)\sqrt{m}\mu_i}{1+(1+B)(\log m)^{-1}} = \sum_{i=1}^{s}\sum_{j=1}^{m}\frac{\mu_i Y_{ij}^*}{1+(1+B)(\log m)^{-1}}.$$

Since $\mu_i Y_{ij}^*$'s are zero-mean independent random variables, by Chebyshev's inequality, we have:

$$\mathbb{P}\Big(|\sum_{i=1}^{s}\sum_{j=1}^{m}\mu_i Y_{ij}^*| > t\Big) \leq \frac{m\sum_{i=1}^{s}\mu_i^2}{t^2}.$$

Consider event $B_3^* = \big\{|\sum_{i=1}^{s}\sum_{j=1}^{m}\mu_i Y_{ij}^*| > \sqrt{m\sum_{i=1}^{s}\mu_i^2}(\log n)^{1/4}\big\}$, we have $\mathbb{P}(B_3^{*c}) \to 1$. Then under event $B_1^{*c}\cap B_2^{*c}\cap B_3^{*c}$, we obtain

$$(B15) \geq \frac{m\sum_{i=1}^{s}\mu_i^2}{1+(1+B)(\log m)^{-1}}\Big(1-\frac{(\log n)^{\frac{1}{4}}}{\sqrt{m\sum_{i=1}^{s}\mu_i^2}}\Big). \tag{B16}$$

Under event $B_1^{*c}\cap B_2^{*c}\cap B_3^{*c}$, combining (B14), (B16) and the assumption (B11), for sufficiently large m and n, we have $(B12) > C^{(n)}$. Finally, note that $\mathbb{P}\left(B_1^{*c}\cap B_2^{*c}\cap B_3^{*c}\right) \to 1$, yielding that the type II error is well-controlled. □

B.1.5 Proof of Theorem 3.5

We combine the arguments and ideas in the proofs of Theorems 3.2 and 3.4. The proof is structured in two parts similar to the previous proofs.

Proof of Theorem 3.5. Type 1 error control:

Without loss of generality, we assume $\sigma^2 = 1$. We still consider the events B_1 and B_2 defined in the proof of Theorem 3.4. By the same arguments in the proof of Theorem 3.4, we have $\mathbb{P}\left(B_1^c \cap B_2^c\right) \to 1$. Hereafter we consider the case under event $B_1^c \cap B_2^c$. Notice that in the proof of Theorem 3.2, it is already shown that $|\mathcal{S}| = M + 1 \leq (1/2)(\log n)^2 + (1/2)\log n + 2$. Similar to argument (B10) in the proof of Theorem 3.4, under event $B_1^c \cap B_2^c$, for every s_i we have:

$$T(s_i) \leq \sup_{|\mathcal{I}_i|=s_i} \sum_{j\in\mathcal{I}_i} -2\log(\tilde{p}_j)\mathrm{I}_{\{j\in\mathcal{I}_i\}} + s_i C(\gamma, B), \tag{B17}$$

where all the $\tilde{p}_j$'s independently follow Unif$(0,1)$ and $C(\gamma, B)$ are some constant depending only on γ and B, as defined in the proof of Theorem 3.4. Therefore, to finish the proof of the first part, it is sufficient to show that $\sup_{|\mathcal{I}_i|=s_i} \sum_{j\in\mathcal{I}_i} -2\log(p_j)\mathrm{I}_{\{j\in\mathcal{I}_i\}} + s_i C(\gamma, B) \leq C_i$ for $i = 0, \ldots, M$ with probability going to one. Indeed, consider similar events $A_i(x_{s_i})$ defined in the proof of Theorem 3.2 except that we replace p_j by the exact p-value $\tilde{p}_j$, following the argument (B6) in proof of Theorem 3.2, we obtain:

$$\mathbb{P}\big(\cup_{i=0}^{M} A_i(x_{s_i})\big) \to 0,$$

where we pick $x_{s_i} = s_i\left(1 + 1/\sqrt{\log\log n}\right)\log\left(n/s_i\right) = s_i\left(1+\delta_n\right)\log\left(n/s_i\right)$ for $i = 0, \ldots, M$. Then under event $B_1^c \cap B_2^c \cap \big(\cup_{i=0}^{M} A_i(x_{s_i})\big)^c$, for sufficiently large n, we have

$$\sup_{|\mathcal{I}_i|=s_i} \sum_{j\in\mathcal{I}_i} -2\log(\tilde{p}_j)\mathrm{I}_{\{j\in\mathcal{I}_i\}} + s_i C(\gamma, B) \leq 2s_i + 2\sqrt{2s_i x_{s_i}} + 2x_{s_i} + s_i C(\gamma, B) \leq C_i.$$

The first part is done, noticing that $\mathbb{P}\big(B_1^c \cap B_2^c \cap \big(\cup_{i=0}^{M} A_i(x_{s_i})\big)^c\big) \to 1$.

For the second part, without loss of generality, we still assume all the non-zero entries of θ are the first s μ_i's ($i = 1, \ldots, s$). Hence we have:

$$\sum_{i=1}^{s} \mu_i^2 \geq n^{1-\beta} C^{(0)} \log n/m.$$

Denote $0 < i^* < \log_{1+1/\log n}(n^{\frac{1}{2}})$ such that:

$$s_{i^*-1} < s < s_{i^*}. \tag{B18}$$

Hence we have:

$$T(s_{i^*}) = \sum_{j=1}^{s_{i^*}} -2\log p_{(j)} \geq \sum_{j=1}^{s} -2\log p_{(j)} \geq \sum_{j=1}^{s} -2\log\left(2\left(1-\Phi\left(|T|_{(n-j+1)}\right)\right)\right).$$

Consider the same events B_1^*, B_2^* and B_3^* in the proof of Theorem 3.4, by the same arguments we have $\mathbb{P}\left(B_1^{*c} \cap B_2^{*c} \cap B_3^{*c}\right) \to 1$. Hence under event $B_1^{*c} \cap B_2^{*c} \cap B_3^{*c}$, combining (B12), (B14), and (B16) in the proof of Theorem 3.4, we have

$$\begin{aligned} T(s_{i^*}) &\geq \sum_{j=1}^{s} -2\log\left(2\left(1-\Phi\left(|T|_{(n-j+1)}\right)\right)\right) \\ &\geq \frac{m\sum_{i=1}^{s}\mu_i^2}{1+(1+B)(\log m)^{-1}}\left(1-\frac{(\log n)^{\frac{1}{4}}}{\sqrt{m\sum_{i=1}^{s}\mu_i^2}}\right) - 2s\sqrt{\log(n-s)} - 2s\log\sqrt{\pi/2}. \end{aligned} \tag{B19}$$

Also note that for sufficiently large m and n,

$$\begin{aligned} C_{i^*} &= 2s_{i^*}(1+2\delta_n)\log\left(n/s_{i^*}\right) \\ &\leq 2s\left(1+1/\log n\right)(1+2\delta_n)\log\left(n/s\right) \\ &< \frac{m\sum_{i=1}^{s}\mu_i^2}{1+(1+B)(\log m)^{-1}}\left(1-\frac{(\log n)^{\frac{1}{4}}}{\sqrt{m\sum_{i=1}^{s}\mu_i^2}}\right) - 2s\sqrt{\log(n-s)} - 2s\log\sqrt{\pi/2}. \end{aligned}$$

Hence the result follows. □

B.2 Null Distribution of RTP Statistics

In this section we introduce two propositions that characterizing the behavior of RTP statistic $T(\ell)$ under the null.

Proposition B1 (Nagaraja (2006))**.** *For independent p-values $p_1, \ldots, p_n$, Let $T(\ell) = \sum_{i=1}^{\ell} -2 \log p_{(i)}$ with $1 < \ell < n$, then under the null $p_1, \ldots, p_n \overset{i.i.d.}{\sim}$ Unif$(0,1)$, where Unif$(0,1)$ denotes the uniform distribution on the interval (0,1),*

$$T(\ell) \sim U + \sum_{i=0}^{n-\ell-1} U_i.$$

Here U and $U_1, \ldots U_{n-\ell-1}$ are independent distributed random variables such that $U \sim \chi^2_{2\ell}$ and $U_i \sim$ GAM $(2\ell/(n-i), 1)$ for $i = 0, \ldots, n-\ell-1$, where GAM(a, b) denotes the gamma distribution with shape and rate parameters a and b.

Although one can derive the closed form of CDF of $T(\ell)$ under the null (Proposition B2), it is numerically unstable due to term (A) that involves repeated integration and catastrophic cancellation of alternating signed terms. Figure B1 shows the signed $\log_{10}$-scaled magnitude of the smallest and the largest terms in (A), with $n = 20, 30, 40, 50$, and $\ell = 4, 5$.

Proposition B2 (Nagaraja (2006))**.** *Under the same conditions of Proposition B1, for RTP test statistic $T(\ell)$ such that $1 < \ell < n$, we have*

$$\mathbb{P}\left(T(\ell) > t\right) = \underbrace{\sum_{j=1}^{n-\ell} w_j \exp\Big\{-\frac{c_j t}{2c_{n-\ell+1}}\Big\} \frac{1}{(\ell-1)!} \int_0^{t/2} \exp\left\{y d_j\right\} y^{\ell-1} dy}_{(A)}$$
$$+ \sum_{j=0}^{\ell-1} \exp\{-t/2\} \frac{\left(\frac{t}{2}\right)^j}{j!}, \quad \textit{where}$$
$$c_j = n - j + 1,$$
$$d_j = \frac{c_j}{c_{n-\ell+1}} - 1,$$
$$w_j = \prod_{k=1; k\neq j}^{n-\ell} \frac{n-k+1}{j-k}.$$

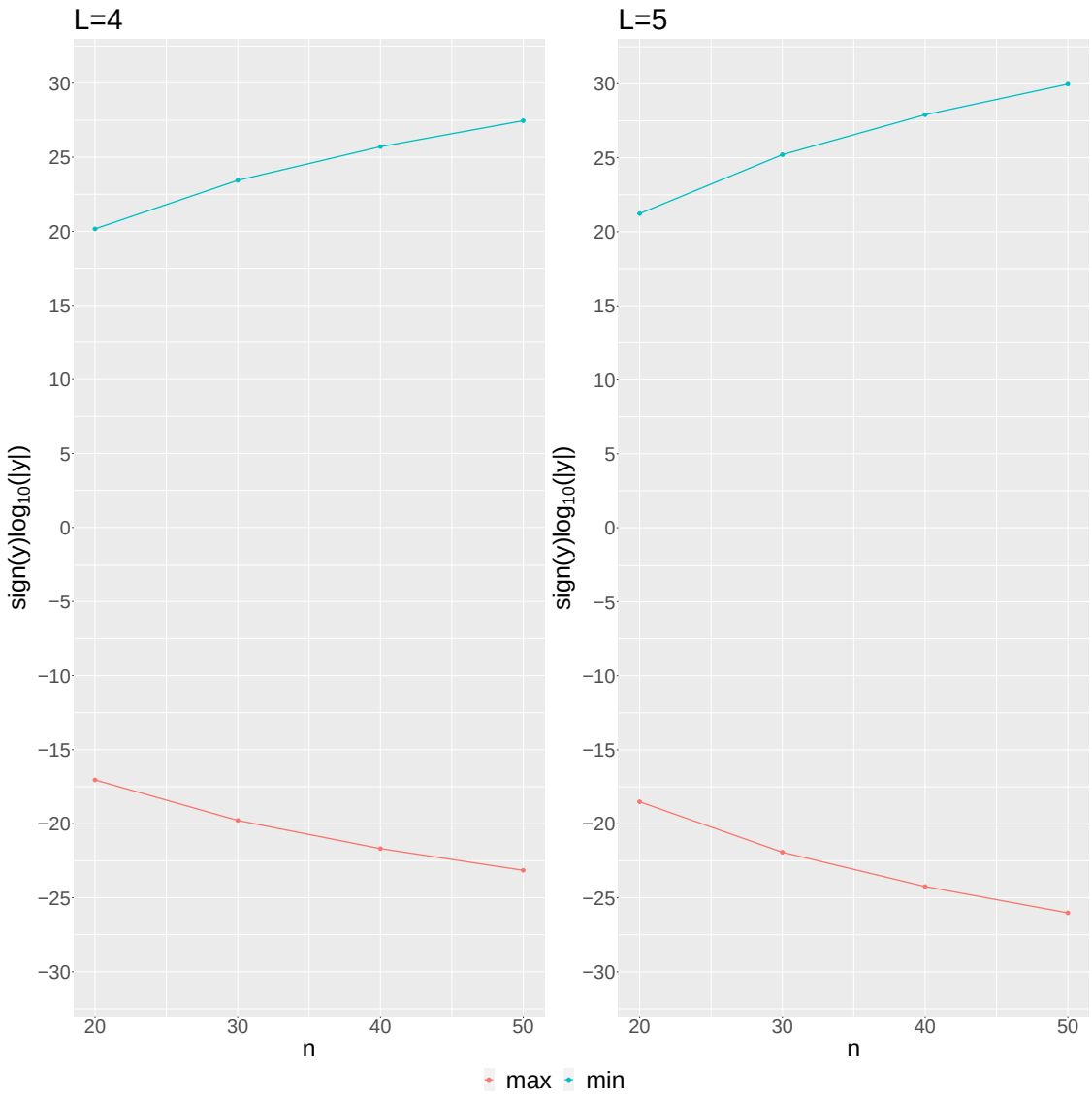


Figure B1: The signed $\log_{10}$-scaled magnitude of the smallest and largest terms in (A) in Proposition B2 with $n = 20, 30, 40, 50$ and $\ell = 4$ and 5, and $t = \mathbb{E}(T(\ell))$. Note the scale of the magnitude of the extreme terms is far greater than 1, while the summation of the terms in (A) falls into the range $[0, 1]$.

B.3 Fast Computation for AFg and RTP

In this section, we derive the fast-computing algorithm of AFg based on the cross-entropy method by De Boer et al. (2005). Denote by $U(s_k)$ the random variable follows the same distribution of $T(s_k)$ under the null. For more stable numeric performance, we consider the following equivalent form of AFg in the following two subsections:

$$T_{\text{AFg}} = \max_{s_k \in \mathcal{S}} -\log \mathbb{P}\left(U(s_k) > t(s_k)\right),$$

where $t(s_k)$ is the observation of $T(s_k)$.

B.3.1 The Efficient Sampling Method via Cross-Entropy

In this subsection, we introduce the efficient sampling method adapted from De Boer et al. (2005). Let $T_{\text{AFg}} = T(p_1, \ldots, p_n)$ be the AFg test statistic and $\hat{t}$ be the observation of the test statistic. Let $\mathbf{X}_1$, $\mathbf{X}_2, \ldots$, $\mathbf{X}_N$ be N independent and identically distributed random vectors that follow $\mathrm{N}_n(0, I_{n\times n})$. Denote $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})'$ for $i = 1, \ldots, n$ and g_0 as the density function of $\mathrm{N}_n(0, I_{n\times n})$. We reformulate T_{AFg} as:

$$T(\mathbf{X}_i) = T\left(2(1 - \Phi(|X_{i1}|)), \ldots, 2(1 - \Phi(|X_{in}|))\right).$$

Denote by $U(\mathbf{X}_i)$ the random variable that follows the same distribution of $T(\mathbf{X}_i)$ under the null. The goal of our algorithm is to estimate the upper tail probability of $U(X_i)$ (one-sided p-value of $T(\mathbf{X}_1)$):

$$\mathcal{P} = \mathbb{P}(U(\mathbf{X}_1) > \hat{t}).$$

One straightforward way is to use simple Monte Carlo:

$$\hat{\mathcal{P}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{I}_{\left\{T(\boldsymbol{X}_i) \geq \hat{t}\right\}}.$$

However, when $\hat{t}$ is extremely large, such that the event $\{T(\mathbf{X}_i) > \hat{t}\}$ is rare, using the above strategy requires extremely large N, which is not computationally feasible. An alternative way is

to use importance sampling: we instead draw $\mathbf{X}_1, \ldots, \mathbf{X}_N$ from an importance sampling density g, and estimate $\mathcal{P}$ by

$$\hat{\mathcal{P}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{\{T(\boldsymbol{X}_i) \geq \gamma\}} \frac{g_0(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i)},$$

where g is properly chosen to allow the event $\{T(\mathbf{X}_i) > \hat{t}\}$ to happen more frequently. One can show the best choice of g is $g^*(x) = \mathrm{I}_{\{T(x) \geq \hat{t}\}} g_0(x)/\mathcal{P}$. It is impossible to obtain $g^*(x)$ as $\mathcal{P}$ is unknown in practice.

In order to pick a proper g for the real practice, De Boer et al. (2005) propose an adaptive procedure (the cross-entropy method) to choose g from a family of densities $g(\cdot;\theta)$ with parameter θ, and choose the θ that optimizes the following stochastic program, which is essentially to find θ that minimizes the empirical Kullback–Leibler divergence between $g^*(\cdot)$ and $g(\cdot,\theta)$:

$$\hat{\theta} = \max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathrm{I}_{\left\{T(\mathbf{X}_i) \geq \hat{t}\right\}} W(\mathbf{X}_i;\theta') \log g(\mathbf{X}_i;\theta),$$

where $W(\mathbf{X}_i;\theta) = \frac{g_0(\mathbf{X}_i)}{g(\mathbf{X}_i;\theta')}$ with θ' as any reference parameter. Here $X_1, \ldots, X_N$ are sampled from $g(\cdot,\theta')$. The choice of the family of g is critical for a good finite-sample performance of the cross-entropy method. Since AFg is a test statistic with heavy-tailed null distribution where a small fraction of extreme p-values can dominate the statistic, we choose the production of the following Gaussian mixture distributions as $g(\cdot,\theta)$:

$$\frac{1}{j+1}\mathrm{N}(0,\theta) + \frac{j}{j+1}\mathrm{N}(0,1) \text{ for } j = 1, \ldots, n.$$

By drawing each $\mathbf{X}_{ij}$ $(j = 1, \ldots, n)$ from different normal mixture distributions, the probability that $\mathbf{X}_{ij}$ to be draw from $\mathrm{N}(0,\theta)$ depends on j, which leads to that only a small fraction of $\mathbf{X}_{ij}$'s can be extremely large and hence only a small fraction of p-values $p_{ij} = 2(1 - \Phi(|X_{ij}|))$ $(j = 1, \ldots, n)$ are extremely small. With the above choice of the family of densities $g(\cdot;\theta)$ and criteria to choose proper θ, we present the following efficient importance sampling algorithm adapted from De Boer et al. (2005):

Algorithm B1. Algorithm for efficient estimation of tail probability of AFg.

Input: N, $\hat{t}$ and ρ.

Step 1. Set t=1, $\theta_0 = 1$ and $\widehat{\varphi}_0 = -\infty$.

Step 2. Generate random sample $\mathbf{X}_1$,...,$\mathbf{X}_N$, where each $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})'$ and all the $\mathbf{X}_{ij}$'s (j=1,. . .,n) are independently sampled from the Gaussian mixture distributions:

$$\frac{1}{j+1} N(0, \theta_{t-1}) + \frac{j}{j+1} N(0,1) \text{ for } j = 1, \ldots, n.$$

Denote the density function of the product of the Gaussian mixture distributions as $g(\cdot; \theta_{t-1})$. Calculate and sort $T_s(\mathbf{X}_i)$ for $i = 1, \ldots,$ N from the smallest to the largest, denoted as $T_{s(1)} \leq T_{s(2)} \leq \ldots \leq T_{s(N)}$. Let $\widehat{\varphi}_t := T_{s(\lceil (1-\rho)N \rceil)}$, if this is less than $\hat{t}_s$. Otherwise set $\widehat{\varphi}_t = \hat{t}_s$.

Step 3. Update θ_t:

$$\theta_t = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{T(\mathbf{X}_i) \geq \hat{\varphi}_t\}} W(\mathbf{X}_i, \theta_{t-1}) \log(g(\mathbf{X}_i; \theta, \lambda)),$$

where $W(\mathbf{X}_i, \theta_{t-1}) = \frac{g_0(\mathbf{X}_i)}{g(\mathbf{X}_i; \theta_{t-1})}$. Recall f is the density function of $\mathrm{N}_n(0, I_{n\times n})$.

Step 4. If $\hat{\varphi}_t = \hat{t}$ then proceed to step 5, otherwise set $t = t + 1$ and back to step 2.

Step 5. Resample $\mathbf{X}_1, \ldots \mathbf{X}_N$ from $g(\cdot; \theta_{T_0})$, where T_0 denotes the final number of iterations, then

$$\hat{\mathcal{P}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\left\{T(\mathbf{X}_i\} \geq \hat{t}\right\}} W(\mathbf{X}_i, \theta_{T_0}).$$

Here we choose the $N = 10^5$ and $\rho = 0.01$ in practice. Table B1 shows Coefficients of variation for estimating tail probability $\mathcal{P}$ using algorithm B1 under varying combinations of n and $\hat{t}$.

Remark B1. For AFg, there is no closed form of $\mathbb{P}(U(s_k) > t(s_k))$ for $1 < i < n$. In the following Section B.3.2, we present another version of efficient sampling method to build the reference library for estimating $\mathbb{P}(T(s_k) > t(s_k))$, the tail probability of $T(s_k)$ with a given observation $t(s_k)$.

Table B1: Coefficients of variation for estimating tail probability $\mathcal{P}$ using algorithm B1 with $n = 1000, 1500, 2000, 2500$ and observations $\hat{t} = 2, 4, \ldots, 10$ ($\hat{t} = 10$ corresponds to $\hat{\mathcal{P}}$ around 10^{-4}) based on 30 times repeated simulations.

	$\hat{t}$ =2	$\hat{t}$ =4	$\hat{t}$ =6	$\hat{t}$ =8	$\hat{t}$ =10
$n = 1000$	0.007	0.012	0.028	0.075	0.195
$n = 1500$	0.007	0.011	0.041	0.076	0.150
$n = 2000$	0.007	0.012	0.035	0.087	0.178
$n = 2500$	0.007	0.015	0.023	0.077	0.173

B.3.2 Algorithm for the Construction of the Reference Library of RTP

In this subsection, we introduce the cross-entropy method to construct the reference library for estimating $-\log(\mathbb{P}(U(s_k) > t(s_k)))$, the minus logarithm transformation of the upper tail probability of RTP statistic $U(s_k)$ with any given observation $t(s_k)$. The general idea of the algorithm is to find a collection of points $(t^{(1)}(s_k), \varphi_k^{(1)}), \ldots, (t^{(p)}(s_k), \varphi_k^{(p)})$, where $t^{(1)}(s_k) <, \ldots, < t^{(p)}(s_k)$ is a collection of quantiles that are a broad range of the upper tail probability of $U(s_k)$, and $\varphi_{k\ell}$'s are corresponding minus logarithm transformation of the estimated upper tail probabilities of $U(s_k)$ given $t^{(\ell)}(s_k)$ for $\ell = 1, \ldots, p$. We then fit an increasing spline function $sp(t(s_k))$ representing the relationship between $t(s_k)$ and $\varphi_k = -\log(\mathbb{P}(U(s_k) > t(s_k)))$ using the points $(t^{(\ell)}(s_k), \varphi_k^{(\ell)})(\ell = 1, \ldots, p)$ to build the reference library for the upper tail probability of $U(s_k)$. The hybrid procedure to find suitable collection of points $(t^{(\ell)}(s_k), \varphi_k^{(\ell)})$ $(\ell = 1, \ldots, p)$ is as the follows.

For the points $(t^{(\ell)}(s_k), \varphi_k^{(\ell)})$ $(\ell = 1, \ldots, g)$ that are supposed to cover the less stringent upper tail of $U(s_k)$ ($\mathbb{P}(U(s_k) > t(s_k)) > 0.01$), we first prespecify the values of $\varphi_k^{(1)} <, \ldots, < \varphi_k^{(g)}$. For example, in practice, we let $g = 91$ and let $(\varphi_k^{(1)}, \ldots, \varphi_k^{(91)})$=$(-\log(0.95), -\log(0.94), \ldots,$ $-\log(0.1), -\log(0.05), -\log(0.04), \ldots, -\log(0.01))$. We then sample a 10^5 Monte Carlo sample for $U(s_k)$ and find the corresponding quantiles $t^{(\ell)}(s_k)$ given $\varphi_k^{(\ell)}$ $(\ell = 1, \ldots, g)$.

For the points $(t^{(\ell)}(s_k), \varphi_k^{(\ell)})$ $(\ell = g + 1, \ldots, p)$ that are supposed to cover the more extreme

upper tail probability of $U(s_k)$ ($\mathbb{P}(U(s_k) > t(s_k)) \leq 0.01$). By Proposition B1, note that under the null

$$U(s_k) \sim \chi^2_{2s_k} + \sum_{i=1}^{n-s_k} \frac{s_k}{n-i+1}\chi^2_2,$$

which leads to $\mathrm{E}(U(s_k)) = 2s_k + \sum_{i=1}^{n-s_k} \frac{2s_k}{n-i+1}$ and $\mathrm{Var}(U(s_k)) = 4s_k + \sum_{i=1}^{n-s_k} \frac{4s_k}{n-i+1}$. We choose the values of $t^{(g)}(s_k)$ and $t^{(p)}(s_k)$ by letting $\frac{t^{(g+1)}(s_k)-\mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_1}$ and $\frac{t^{(p)}(s_k)-\mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_2}$, where η_{q_h} ($h = 1, 2$) are quantiles such that $1 - \Phi(\eta_{q_h}) = q_h$. Here we choose $q_1 = 0.01$ and $q_2 = 10^{-20}$. Other quantiles' values are determined by letting $\log t^{(\ell+1)}(s_k) - \log t^{(\ell)}(s_k) = \frac{\log t^{(p)}(s_k) - \log t^{(g+1)}(s_k)}{p-g}$ for $\ell = g+2, \ldots, p-1$. We then plug the collection of quantiles $t^{(g+1)}(s_k), \ldots, t^{(p)}(s_k)$ into algorithm B3 based on the cross-entropy method by De Boer et al. (2005) and calculate $\varphi_k^{(g+1)} = -\log \hat{\mathbb{P}}(U(s_k) > t^{(g+1)}(s_k)), \ldots, \varphi_k^{(p)} = -\log \hat{\mathbb{P}}(U(s_k) > t^{(p)}(s_k))$. We summarize the above procedure as the following Algorithm B2:

Algorithm B2. Algorithm for the construction of the reference library of RTP.

Input: $\varphi_k^{(1)} <, \ldots, \varphi_k^{(g)}$; g, p, q_1 and q_2.

Step 1. Sample a 10^5 Monte Carlo sample for $U(s_k)$ and find the corresponding quantiles $t^{(\ell)}(s_k)$ given $\varphi_k^{(\ell)}$ ($\ell = 1, \ldots, g$).

Step 2. Determine the values of $t^{(g+1)}(s_k)$ and $t^{(p)}(s_k)$ such that $\frac{t^{(g+1)}(s_k)-\mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_1}$ and $\frac{t^{(p)}(s_k)-\mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_2}$. Determine the values of the other quantiles by letting $\log t(s_k)^{(\ell+1)} - \log t(s_k)^{(\ell)} = \frac{\log t^{(p)}(s_k) - \log t(s_k)^{(g+1)}}{p-g}$ for $\ell = g+2, \ldots, p-1$. Calculate $\varphi_k^{(\ell)}$ given $t^{(\ell)}(s_k)$ ($\ell = g+1, \ldots, p$) using algorithm B3.

Step 3. Fit an increasing spline function $sp(t(s_k))$ based on the collection of points $(t^{(\ell)}(s_k), \varphi_k^{(\ell)})$ ($\ell = 1, \ldots, p$)

For our case, we choose $g = 91$, $p = 391$, $q_1 = 0.01$, $q_2 = 10^{-20}$ and $(\varphi_k^{(1)}, \ldots, \varphi_k^{(91)})$ =$(-\log(0.95), -\log(0.94), \ldots, -\log(0.1), -\log(0.05), -\log(0.04), \ldots, -\log(0.01))$.

The cross-entropy method to estimate the upper tail probability of $U(s_k)$ is similar to the one used for AFg. Again, let $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_N$ be N independent and identically distributed random vectors that follow $\mathrm{N}_n(0, I_{n\times n})$. Denote $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})'$ for $i = 1, \ldots, n$ and g_0 as the density function of $\mathrm{N}_n(0, I_{n\times n})$. Then we denote $T_{s_k}(\mathbf{X}_i)$ as follows:

$$T_{s_k}(\mathbf{X}_i) = \sum_{i=1}^{s_k} -2\log p_{(j)},$$

where $p_{(j)}$ is the j-th smallest p-value among $p_1 = 2\left(1 - \Phi(|\mathbf{X}_{i1}|)\right), \ldots, p_n = 2\left(1 - \Phi(|\mathbf{X}_{in}|)\right)$ $(j = 1, \ldots, n)$. we summary the algorithm that estimates $\mathbb{P}(U(s_k) > f)$ for any f as the following Algorithm B3:

Algorithm B3. Cross-entropy method for RTP.

Input: N, f, ρ, τ.

Step 1. Set t=1, $\theta_0 = 1$ and $\hat{f}_0 = -\infty$.

Step 2. Generate random sample $\mathbf{X}_1$,...,$\mathbf{X}_N$, where each $\mathbf{X}_i = (\mathbf{X}_{i1}, \ldots, \mathbf{X}_{in})'$ and all the $\mathbf{X}_{ij}$'s $(j = 1, \ldots, n)$ are independently sampled from the Gaussian mixture distributions:

$$\frac{1}{j^\tau + 1} N\big(0, \theta^{\frac{1}{\sqrt{\log(j+2)}}}\big) + \frac{j^\tau}{j^\tau + 1} N(0, 1) \text{ for } j = 1, \ldots, n.$$

We calculate the p-values to combine by two-sided z-score test $p_1 = 2(1 - \Phi(\mathbf{X}_{i1})), \ldots, p_n = 2(1 - \Phi(\mathbf{X}_{in}))$. And τ is determined by

$$\sum_{j=1}^{n} \frac{1}{j^\tau + 1} = s_k.$$

Denote the density function of the product of the Gaussian mixture distributions as $g(\cdot; \theta_{t-1})$. Calculate and sort $T_{s_k}(\mathbf{X}_i)$ for $i = 1, \ldots,$ N from the smallest to the largest, denoted as $T_{s_k(1)} \leq T_{s_k(2)} \leq \ldots \leq T_{s_k(N)}$. Let $\hat{f}_t := T_{s_k(\lceil (1-\rho)N \rceil)}$, if this is less than f. Otherwise set $\hat{f}_t = f$.

Step 3. Update θ_t:

$$\theta_t = \arg\max_{\theta} \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{T_{s_k}(\mathbf{X}_i) \geq \hat{f}_t\}} W(\mathbf{X}_i, \theta_{t-1}) \log(g(\mathbf{X}_i; \theta, \lambda)),$$

where $W(\mathbf{X}_i, \theta_{t-1}) = \frac{g_0(\mathbf{X}_i)}{g(\mathbf{X}_i; \theta_{t-1})}$. Recall g_0 is the density function of $\mathrm{N}_n(0, I_{n\times n})$.

Step 4. If $\hat{f}_t = f$ then proceed to step 5, otherwise set $t = t + 1$ and back to step 2.

Step 5. Resample $\mathbf{X}_1, \ldots \mathbf{X}_N$ from $g(\cdot; \theta_{T_0})$, where T_0 denotes the final number of iterations, then estimate $\hat{\mathcal{P}} = \mathbb{P}(U(s_k) \geq f)$ by:

$$\hat{\mathcal{P}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{I}_{\{U(s_k)(\mathbf{X}_i) \geq f\}} W(\mathbf{X}_i, \theta_{T_0}).$$

In practice, we choose $N = 3 \times 10^4$ and $\rho = 0.01$. One can note that there are two differences between Algorithm B3 and Algorithm B1. First, For Algorithm B3, there is an extra parameter τ to characterize the Gaussian mixture distributions for the generation of $\mathbf{X}_1, \ldots, \mathbf{X}_N$. The reason is that when s_k is relatively small compared to n, the total number of p-values to combine, the null distribution of $T(s_k)$ is relatively heavy-tailed, and a small fraction of extreme p-values can dominate the behavior of $U(s_k)$. While as s_k increases, the behavior of $T(s_k)$ is more and more similar to the behavior of Fisher's combination test, that is, the tail of the null distribution of $T(s_k)$ gets lighter and lighter, and hence a larger and larger proportion of extreme p-values are needed to have an impact on the behavior of $T(s_k)$, and a smaller τ is needed. The second difference is that the variance of the alternative component of the Gaussian mixtures decreases as j increases.

We then investigate the stability of Algorithm B3. For $s_1, s_5, s_{15}, s_{25}, s_M$ under each $n = 1000, 1500, 2000, 2500$, we estimate $\varphi_k^{(1)} = -\log \hat{\mathbb{P}}(U(s_k) > t^{(1)}(s_k)), \ldots, \varphi_k^{(30)} = -\log \hat{\mathbb{P}}(U(s_k) > t^{(30)}(s_k))$, where $\frac{t^{(1)}(s_k) - \mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_1}$ with $q_1 = 0.01$, $\frac{t^{(30)}(s_k) - \mathrm{E}(U(s_k))}{(\mathrm{Var}(U(s_k)))^{1/2}} = \eta_{q_2}$ with $q_2 = 10^{-15}$ (recall $1 - \Phi(\eta_{q_h}) = q_h$ for $h = 1, 2$), and $\log t^{(\ell+1)}(s_k) - \log t^{(\ell)}(s_k) = \frac{\log t^{(30)}(s_k) - \log t^{(1)}(s_k)}{30-1}$ for $1 < \ell < 30$. Coefficients of variation for each $\varphi_k^{(\ell)}$ are calculated based on 30 times repeated simulations. Table B2 shows the maximum coefficients of variation for $s_1, s_5, s_{15}, s_{25}, s_M$ under each n (Here M M is the smallest integer such that $\log n \, (1 + 1/\log n)^{M-1} \geq n/\log n$.).

Table B2: Maximum coefficients of variation for $s_1, s_5, s_{15}, s_{25}, s_M$ under each n based on 30 times repeated simulations.

	s_1	s_5	s_{15}	s_{25}	s_M
n=1000	0.068	0.066	0.059	0.020	0.024
n=1500	0.073	0.072	0.071	0.035	0.031
n=2000	0.074	0.080	0.065	0.037	0.041
n=2500	0.083	0.070	0.081	0.081	0.024

Appendix C for Chapter 4

C.1 Technical Arguments: Proof of Theorems

Before showing the technical arguments, we first define some notations and introduce the concept of asymptotically tailed independence, where the latter plays a key role in the proofs of Theorems 4.1 and 4.2.

Definition C1 (Chen and Yuen (2009))**.** Two nonnegative non-identically distributed random variables Y_1 and Y_2 with distributions F_1 and F_2, respectively, are said to be asymptotically tailed independent if

$$\lim_{t\to\infty} \frac{P(Y_1 > t, Y_2 > t)}{\bar{F}_1(t) + \bar{F}_2(t)} = 0, \tag{C1}$$

where $\bar{F}_i = 1 - F_i(t)$ denotes the survival function of Y_i for each $i = 1, 2$.

It suffices to show the asymptotically tailed independence by showing $P(Y_1 > t|Y_2 > t) = o(1)$ or $P(Y_2 > t|Y_1 > t) = o(1)$, or equivalently, $P(Y_1 > t, Y_2 > t) = o(P(Y_1 > t))$ or $o(P(Y_2 > t))$.

More generally, two real-valued random variables, Y_1 and Y_2, are said to be asymptotically independent if the relation (C1) holds with (Y_1, Y_2) in the numerator being replaced by (Y_1^+, Y_2^+),(Y_1^+, Y_2^-), (Y_1^-, Y_2^+), where $Y_i^+ = \max(Y_i, 0)$ and $Y_i^- = \max(-Y_i, 0)$ for i =1, 2.
In this case, one can show that to prove Y_1 and Y_2 are asymptotically tailed independent, it suffices to prove that $P(Y_i^+ > t, Y_j^+ > t)$, $P(Y_i^+ > t, Y_j^- > t)$, $P(Y_i^- > t, Y_j^+ > t)$ are all $o(P(Y_1 > t))$ or $o(P(Y_2 > t))$.

C.1.1 Proof of Theorem 4.1

Before proving Theorem 4.1, first we introduce two lemmas, Lemmas C1 and C2.

Lemma C1. *If X_1 and X_2 are bivariate standard normally distributed with correlation $-1 < \rho < 1$, then $|X_1|$ and $|X_2|$ are asymptotically tailed independent.*

Proof. Use the upper bound for upper tailed probability of the bivariate standard normal random

variables. $P(X_1 > t, X_2 > t) \leq \Phi(-t)\Phi(-\theta t)(1+\rho)$ for $t > 0$ and $\rho > 0$, where $\theta = \sqrt{\frac{1-\rho}{1+\rho}}$ (Willink, 2005). We first assume $\rho > 0$. When $\rho < 0$, let $Z_2 = -X_2$. Then X_1 and Z_2 are bivariate standard normally distributed with correlation $\rho > 0$ and $P(|X_1| > t, |X_2| > t) = P(|X_1| > t, |Z_2| > t)$. So it suffices to prove the case of $\rho > 0$. Now we consider the case where $\rho > 0$,

$$
\begin{aligned}
&P(|X_1| > t, |X_2| > t) \\
&\leq P(X_1 > t, X_2 > t) + P(-X_1 > t, -X_2 > t) + P(X_1 > t, -X_2 > t) + P(-X_1 > t, X_2 < t) \\
&= I + II + III + IV.
\end{aligned}
$$

For I, we have $I = P(X_1 > t, X_2 > t) \leq \Phi(-t)\Phi(-\theta t)(1+\rho) = o(P(X_1 > t))$. For II, we note $II = I$ (X_1 and X_2 are bivariate standard normal random variables, so their joint pdf are symmetric around 0). For III, first let $X_2 = c_1 X_1 + c_2 Z$, where $c_1 > 0$ (because $\rho > 0$) and $c_2 > 0$ and Z is a standard normal random variable independent of X_1. Then we have

$$
\begin{aligned}
P(X_1 > t, -X_2 > t) &= P(X_1 > t, -c_1 X_1 - c_2 Z > t) \\
&= P(X_1 > t, -c_2 Z > t + c_1 X_1) \\
&\leq P(X_1 > t, -c_2 Z > t + c_1 t) \\
&= P(X_1 > t) P(-c_2 Z > t + c_1 t) = o(P(|X_1| > t)).
\end{aligned}
$$

We then further note $IV = III$ since X_1 and X_2 are exchangeable. Combine all the results, we have $P(|X_1| > t, |X_2| > t) = o(P(|X_1| > t))$. □

Remark C1. From Willink's upper bound for the bivariate normal random variables, it is clear that when ρ is close to 1, we can see the "asymptotically tailed independence phenomenal" only when t is extremely large.

Lemma C2 (Chen and Yuen (2009))**.** *If $U_1, \ldots, U_n \in R_{-\gamma}$ are asymptotically tailed independent random variables with CDFs $F_1, \ldots, F_n$, respectively; then $P(U_1 + \ldots + U_n > t) \sim \sum_{i=1}^{n} \bar{F}_i(t)$.*

Proof of Theorem 4.1. First we assume the transformation $g(p)$ is nonnegative. Since $U_i \in R_{-\gamma}, \forall i = 1, \ldots, n$, by Lemma C2, it suffices to prove $U_1, \ldots, U_n$ are pairwise asymptotically tailed independent. Here we have

$$\begin{aligned} P(U_i > t | U_j > t) &= P(g(p_i) > t | g(p_j) > t) \\ &= P(|X_i| > t^* | |X_j| > t^*) \sim o((P(|X_i| > t^*)) = o(P(U_i > t)). \end{aligned} \tag{C2}$$

Note that $t^* \to \infty$ as $t \to \infty$. The second equality is because $g(p)$ and $2(1 - \Phi(|X|))$ are both monotone decreasing and continuous. $P(|X_i| > t^* | |X_j| > t^*) \sim o((P(|X_i| > t^*)) = o(P(U_i > t))$ is because of Lemma C1. Therefore, $U_1, \ldots, U_n$ are pairwise asymptotically tailed independent and we complete the proof. When the transformation $g(p)$ is not nonnegative, see Remark C2 for detailed proof. □

Remark C2. As described in the proof, we prove Theorem 4.1 by assuming the transformation $g(p)$ is nonnegative. In fact, it can be easily extended to real-valued transformation $g(p)$. In order to prove the asymptotically tailed independence for the general case, it suffices to prove that $P(U_i^+ > t, U_j^+ > t)$, $P(U_i^+ > t, U_j^- > t)$, $P(U_i^- > t, U_j^+ > t)$ are all $o(P(U_i > t))$ or $o(P(U_j > t))$ as $t \to \infty$.

First for any $t > 0$, $P(U_i^+ > t, U_j^+ > t) = P(U_i > t, U_j > t)$. We can show that $P(U_i > t, U_j > t) = o(P(U_i > t))$ with the same argument as in (C2). Therefore $P(U_i^+ > t, U_j^+ > t) = o(P(U_i > t))$. It remains to prove $P(U_i^+ > t, U_j^- > t) = o(P(U_i > t))$ since $P(U_i^- > t, U_j^+ > t) = o(P(U_j > t))$ can be proved similarly.

First we have $P(U_i^+ > t, U_j^- > t) = P(U_i > t, -U_j > t) = P(U_i > t, U_j < -t)$ for $\forall t > 0$. It suffices to show the result hold for the condition (A2.1) in Theorem 4.1, otherwise for the alternative condition (A2.2), since U_j is bounded below, we have $P(U_j^- > t) = P(U_j < -t) = 0$ for large enough t, which immediately implies $P(U_i^+ > t, U_j^- > t) = 0$. Now we consider the condition (A2.1), where $g(p)$ is continuous and strictly decreasing for $0 < p < 1$. Note that for any large fixed t, there exist a corresponding large fixed value s_1 and a small fixed value s_2, such that

$$\begin{aligned} \{U_i > t\} &= \{|X_i| > s_1\} \\ \{U_j < -t\} &= \{|X_j| < s_2\}. \end{aligned}$$

Because X_i and X_j are bivariate normal distributed with correlation $|\rho_{ij}| \neq 1$, we let $X_i = C_1 Z + C_2 X_j$, where C_1 and C_2 are some constants, $Z \overset{D}{\sim} N(0,1)$ and independent of X_j, and then applying similar trick in the proof of Lemma C1:

$$\begin{aligned}
P(U_i^+ > t, U_j^- > t) &= P(|X_i| > s_1, |X_j| < s_2) \\
&\leq P(|C_1 Z| + |C_2 X_j| > s_1, |X_j| < s_2) \\
&\leq P(|C_1 Z| > s_1 - |C_2| s_2, |X_j| < s_2) \\
&= P(|C_1 Z| > s_1 - |C_2| s_2) P(|X_j| < s_2) = o(P(|X_j| < s_2)) = o(P(U_j^- > t))
\end{aligned}$$

note $P(U_j^- > t) = O(P(U_j > t))$ by the balance condition (A3). Hence we complete the proof.

C.1.2 Proof of Theorem 4.2

Proof of Theorem 4.2. First we prove $w_i U_i$ and $w_j U_j$ for $\forall m+1 \leq i < j \leq n$ are asymptotically tailed independent, where the corresponding $|\rho_{ij}| < 1$ for $\forall m+1 \leq i < j \leq n$. As discussed in the Remark C2 for Theorem 4.1, without loss of generality, we can assume both U_i and U_j are nonnegative random variables. Suppose $w_i \leq w_j$:

$$\begin{aligned}
P(w_i U_i > t | w_j U_j > t) &= \frac{P(w_i U_i > t, w_j U_j > t)}{P(w_j U_j > t)} \\
&\leq \frac{P(w_j U_i > t, w_j U_j > t)}{P(w_j U_j > t)} \to 0.
\end{aligned}$$

The last line is because U_i and U_j $\forall m+1 \leq i < j \leq n$ are asymptotically tailed independent which were already proved in Theorem 4.1.

Suppose $w_i > w_j$:

$$\begin{aligned}
P(w_i U_i > t | w_j U_j > t) &= \frac{P(w_i U_i > t, w_j U_j > t)}{P(w_j U_j > t)} \\
&\leq \frac{P(w_i U_i > t, w_i U_j > t)}{P(w_j U_j > t)} \\
&= \frac{P(w_i U_i > t, w_i U_j > t)}{P(\frac{w_j}{w_i} w_i U_j > t)} \\
&\sim \frac{P(w_i U_i > t, w_i U_j > t)}{(\frac{w_j}{w_i})^\gamma P(w_i U_j > t)} \to 0
\end{aligned}$$

The last line is because U_i and U_j $\forall m+1 \leq i < j \leq n$ are asymptotically tailed independent and also because the distribution of $w_i U_j$ has a regularly varying tail with index γ.
Hence we have

$$\frac{P(w_i U_i > t, w_j U_j > t)}{P(w_i U_i > t) + P(w_j U_j > t)} \leq P(w_i U_i > t | w_j U_j > t) \to 0.$$

Therefore, $w_i U_i$ and $w_j U_j$ $\forall m+1 \leq i < j \leq n$ are asymptotically tailed independent.

Second, we consider the case with extreme correlation $|\rho_{ij}| = 1$. In this case, $X_1 = ... = X_m$ with probability 1 and hence $U_1 = ... = U_m$ with probability 1. Therefore, it suffice to show that $(\sum_{i=1}^{m} w_i)U_1$ and $w_j U_j$, for $\forall m+1 \leq j \leq n$, are asymptotically tailed independent, since $\rho_{ij} = 1$ or -1 for $1 \leq i < j \leq m$.

This can be easily proved by the following inequality:

$$P\left(\left(\sum_{i=1}^{m} w_i\right) U_1 > t | w_j U_j > t\right) \leq \sum_{i=1}^{m} P(w_i U_1 > t/m | w_j U_j > t) \to 0.$$

Therefore,

$$\begin{aligned}
P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) &= P(\sum_{i=1}^{n} w_i U_i > t) \\
&= P\left(\left(\sum_{i=1}^{m} w_i\right) U_1 + \sum_{i=m+1}^{n} w_i U_i > t\right) \\
&\sim \left(\sum_{i=1}^{m} w_i\right)^{\gamma} P(U_1 > t) + \sum_{i=m+1}^{n} w_i^{\gamma} P(U_i > t) \\
&= \left[\left(\sum_{i=1}^{m} w_i\right)^{\gamma} + \sum_{i=m+1}^{n} w_i^{\gamma}\right] P(U_1 > t).
\end{aligned}$$

The third line is because $(\sum_{i=1}^{m} w_i)\, U_1$ and $w_j U_j$, $\forall m+1 \leq j \leq n$, are asymptotically tailed independent and because of Lemma C2 and the property of regularly varying tailed random variables. □

C.1.3 Proof of Theorem 4.3

Before proving Theorem 4.3, we first introduce two lemmas for the proof. Lemma C3 is the combination of Theorem 4.2 and Theorem 4.3 in Davis (1983). Below are the conditions for Lemma C3:

(B1): Let $U_1^*, \ldots, U_{n^*}^*, \ldots$ stationary sequence of regularly varying random variables with index $0 < \gamma \leq 1$ and with common distribution function F^*.

(B2): Let $G^*(t) = P(|U_1^*| > t)$. The distribution of U_1^* satisfies the balance condition; that is, $\frac{1-F^*(t)}{G^*(t)} \to p$ and $\frac{F^*(-t)}{G^*(t)} \to q$ as $t \to \infty$, where $0 \leq p \leq 1$. and $p + q = 1$.

In addition to conditions $(B1)$ and $(B2)$, there are three additional conditions (D), (D') and (D'') given in Davis (1983), all of which are assumptions for dependent structure of $U_1^*, \ldots, U_{n^*}^*$, and are required for Lemma C3. For the details of conditions (D), (D') and (D''), see Davis (1983). We do not provide details of these conditions because they are very technical but obviously satisfied in Theorem 4.3, as shown in the proof of Theorem 4.3.

Lemma C3 (Davis (1983))**.** *Suppose conditions (B1), (B2), (D), (D') and (D") hold. For* $0 < \gamma \leq 1$ *we have*

$$\frac{\sum_{i=1}^{n^*} U_i^* - b_{n^*}}{a_{n^*}} \to_d S_\gamma^*,$$

where S_γ^* *is a random variable;* a_{n^*} *is a term such that* $n^* G^*(a_{n^*} x) \to x^{-\gamma}$ *for* $0 < \gamma \leq 1$ *as* $n^* \to \infty$ *and* $x > 0$*;* b_{n^*} *is defined as follows*

$$b_{n^*} = \begin{cases} 0, & 0 < \gamma < 1, \\ n^* \int_{-a_{n^*}}^{a_{n^*}} x dF^*(x), & \gamma = 1. \end{cases}$$

The following lemma describes the order of a_{n^*} and b_{n^*} given that some of the conditions of Theorem 4.3 are satisfied.

Lemma C4. *If* G^**,* F^* *and* U_i^* *for* $i = 1, \ldots, n$ *satisfy conditions for Lemma C3 and conditions (A3) and (C2), we have*

$$a_{n^*} = O((n^*)^{1/\gamma} L_{n^*}) \text{ for } 0 < \gamma \leq 1$$
$$b_{n^*} = O(n^* L_{n^*}) \text{ for } \gamma = 1,$$

where L_{n^*} is the power function of $\log n^*$.

Proof. First, we prove $a_{n^*} = O((n^*)^{1/\gamma} L_{n^*})$ for $0 < \gamma \leq 1$. Suppose $a_{n^*} \neq O((n^*)^{1/\gamma} L_{n^*})$. Then for any $k > 0$, there exits an arbitrary large n^*, such that $a_{n^*} > (n^*)^{\frac{1}{\gamma}} (\log n^*)^k$. Hence we have

$$\begin{aligned} n^* G^*(a_{n^*} x) &\leq n^* G^* \left((n^*)^{\frac{1}{\gamma}} \log n^*)^k x \right) \\ &\leq \frac{C n^* \left(\log \left((n^*)^{\frac{1}{\gamma}} (\log n^*)^k x \right) \right)^h}{\left((n^*)^{\frac{1}{\gamma}} (\log n^*)^k x \right)^\gamma} \\ &= \frac{C}{x^\gamma} \cdot \frac{\left(\frac{1}{\gamma} \log(n^*) + k \log \log n^* + \log x \right)^h}{(\log n^*)^{k\gamma}}, \end{aligned} \tag{C3}$$

where C and h are some fixed constants. The second inequality is due to conditions (A3) and (C2). Indeed, given the two conditions, we have $G^*(t) \overset{\text{(i)}}{\leq} C \bar{F}^*(t) \overset{\text{(ii)}}{\leq} \frac{C(\log(t))^h}{t^\gamma}$, where (i) is due to balance condition (A3) and (ii) is due to condition (C2). By choosing k such that $k\gamma > h$, we have (C3) $\to 0$ for $\forall x > 0$, which immediately leads to contradiction since by definition of a_{n^*} we have $n^* G^*(a_{n^*} x) \to \frac{1}{x^\gamma}$.

Then we prove $b_{n^*} = O(n^* L_{n^*})$ for $\gamma = 1$. Since conditions (A3) and (C2) hold, we can choose a large enough constant M, such that,

$$\begin{aligned} &\bar{F}^*(t) \leq \frac{(\log(t))^h}{t} \text{ for } \forall t > M. \\ &F^*(-t) \leq c \bar{F}^*(t), \end{aligned}$$

where c and h are some fixed constants. By the definition of b_{n^*}, we have

$$\begin{aligned} b_{n^*} &= n^* \int_{-a_{n^*}}^{a_{n^*}} x dF^*(x) \\ &= \underbrace{n^* \int_{-a_{n^*}}^{-M} x dF^*(x)}_{I} + \underbrace{n^* \int_{-M}^{0} x dF^*(x)}_{II} + \underbrace{n^* \int_{0}^{M} x dF^*(x)}_{III} + \underbrace{n^* \int_{M}^{a_{n^*}} x dF^*(x)}_{IV}. \end{aligned}$$

For II and III, we have $II \leq n^* \int_{-M}^{0} M dF^*(x) \leq n^* M = O(n^*)$ and $III \leq n^* \int_{0}^{M} M dF^*(x) \leq n^* M = O(n^*)$. For I, we have

$$I = n^* \int_{-a_{n^*}}^{-M} x dF^*(x) = \underbrace{n^*(-M)F(-M) + n^* a_{n^*} F(-a_{n^*})}_{\text{(i)}} - \underbrace{n^* \int_{-a_{n^*}}^{-M} F^*(x)dx}_{\text{(ii)}},$$

where (i) is $O\left(n^* L_{n^*}\right)$. This is because by condition (A3) we have $n^* a_{n^*} F(-a_{n^*}) \leq a_{n^*} c n^* \bar{F}^*(a_{n^*}) \leq c_1 a_{n^*} n^* G^*(a_{n^*}) = O\left(n^* L_{n^*}\right)$, where the last equality is due to the fact that $n^* G^*(a_{n^*} x) \to \frac{1}{x}$ for any $x > 0$ and $a_{n^*} = O(n^* L_{n^*})$ when $\gamma = 1$. For (ii), we have

$$\begin{aligned}
\text{(ii)} = n^* \int_{-a_{n^*}}^{-M} F^*(x)dx = n^* \int_{M}^{a_{n^*}} F^*(-y)dy &\leq n^* \int_{M}^{a_{n^*}} c\bar{F}^*(y)dy \\
&\leq n^* \int_{M}^{a_{n^*}} c\frac{(\log y)^h}{y}dy \\
&= O\left(n^*(\log(a_{n^*}))^{h+1}\right) = O(n^* L_{n^*}).
\end{aligned}$$

Hence we have $I = O(n^* L_{n^*})$. For IV, we have

$$\begin{aligned}
|IV| = \left| n^* \int_{M}^{a_{n^*}} x dF^*(x) \right| &= \left| n^* \int_{M}^{a_{n^*}} x d(1 - \bar{F}^*(x)) \right| = \left| n^* \int_{M}^{a_{n^*}} x d\bar{F}^*(x) \right| \\
&= \left| n^* a_{n^*} \bar{F}_{a_{n^*}} - n^* M \bar{F}^*(M) - n^* \int_{M}^{a_{n^*}} \bar{F}^*(x)dx \right| \\
&\leq \left| n^* a_{n^*} \bar{F}_{a_{n^*}} \right| + \left| n^* M \bar{F}^*(M) \right| + \left| n^* \int_{M}^{a_{n^*}} \bar{F}^*(x)dx \right| \\
&\leq \left| n^* \int_{M}^{a_{n^*}} \frac{(\log(x))^h}{x}dx \right| + O(n^* L_{n^*}),
\end{aligned}$$

where the last inequality is due to the fact $n^* a_{n^*} \bar{F}^*(a_{n^*}) \leq c_1 a_{n^*} n^* G^*(a_{n^*}) = O(n^* L_{n^*})$ given condition (A3) and definition of a_{n^*}. Also note that $\left| n^* \int_{M}^{a_{n^*}} \frac{(\log(x))^h}{x}dx \right| = \left| n^* \frac{(\log a_{n^*})^{h+1}}{h+1} - n^* \frac{(\log M)^{h+1}}{h+1} \right| = O(n^* L_{n^*})$. Hence we have $IV = O(n^* L_{n^*})$ and further $b_{n^*} = O(n^* L_{n^*})$.

□

Remark C3. Lemma C3 and Lemma C4 suggest that for the regularly varying tailed random variables $U_1^*, \ldots U_{n^*}^*$ with index $0 < \gamma \leq 1$, $\sum_{i=1}^{n^*} U_i^* = O(n^{*1/\gamma} L_{n^*})$. For example, for CA test, its corresponding $a_{n^*} = \frac{2n^*}{\pi}$ and $b_{n^*} = 0$; for HM test, $a_{n^*} = n^*$ and $b_{n^*} = n^* \ln(n^*)$; for BC_η test ($\eta = 1/\gamma$, $0 < \gamma < 1$), $a_{n^*} = (n^*)^{1/\gamma}$. The distribution of S_γ^* is dependent on γ and described in detail in Theorem 4.2 and Theorem 4.3 in Davis (1983). For the purpose of this paper, we only need to use the order of $\sum_{i=1}^{n} U_i^*$,which is $O_p((n^*)^{1/\gamma} L_{n^*})$ $(0 < \gamma \leq 1)$.

The following Lemmas C5 and C6 are useful when characterizing the lower bound of $g(p)$.

Lemma C5 (ratio inequality of Mill)**.** *For any $x > 0$,*

$$\frac{x}{\phi(x)} \leq 1/(1 - \Phi(x)) \leq \frac{x}{\phi(x)} \frac{1 + x^2}{x^2},$$

where $\Phi(x)$ and $\phi(x)$ are CDF and pdf of the standard normal distribution, respectively.

Lemma C6. *If conditions (A2), (A3) and (C2) hold, then we have the following two inequalities for the transformation $g(p)$.*
There exist $p_1 > 0, C_1 > 0, k \geq 0$ such that for $0 < p < p_1$

$$g(p) \geq \frac{C_1}{p^{1/\gamma} |\ln(p)|^k}.$$

and there exist $p_2 > 0, C_2 > 0$, $k \geq 0$ such that for $p_2 < p < 1$

$$g(p) \geq \frac{-C_2 |\ln(1-p)|^k}{(1-p)^{1/\gamma}}.$$

Proof. To prove the first statement, let $t = g(p)$ and by condition (A2), $g(p)$ is strictly decreasing for small enough p, hence $g^{-1}(t)$ exists for large enough t and is also strictly decreasing. Note for any large fixed t, we have $F(t) = P(g(p) \leq t) = P(p \geq g^{-1}(t)) = 1 - g^{-1}(t)$, hence $\bar{F}(t) = g^{-1}(t)$ for large enough t and further $g(p) = \bar{F}^{-1}(p)$ for small enough p, where we have $\bar{F}^{-1}(\bar{F}(t)) = t$ for large enough t. We now prove the first statement by contradiction, assume for

any $k > 0$, there exists an arbitrary small p such that $g(p) = \bar{F}^{-1}(p) < \frac{1}{p^{1/\gamma}|\log p|^k}$, which leads to the following contradiction:

$$\begin{aligned} t = \bar{F}^{-1}\left(\bar{F}(t)\right) &\leq \bar{F}^{-1}\left(\frac{1}{t^{\gamma}|\log t|^{h}}\right) \\ &< \frac{\left(t^{\gamma}|\log t|^{h}\right)^{\frac{1}{\gamma}}}{\left|\log\left(t^{-\gamma}|\log t|^{-h}\right)\right|^{k}} \\ &= \frac{t}{|\log t|^{-\frac{h}{\gamma}}\left(\gamma\log t + h\log\log t\right)^{k}} < t \text{ by choosing large enough k,} \end{aligned}$$

where $h \geq 0$ is some fixed constant. The first inequality is due to condition (C2) and that $\bar{F}^{-1}(p)$ is strictly decreasing for small enough p. The second inequality is due to our assumption $g(p) = \bar{F}^{-1}(p) < \frac{1}{p^{1/\gamma}|\log p|^k}$ for an arbitrary small p. Given this contradiction, the proof of the first statement is completed.

We then prove the second statement. First note that when $g(p)$ is bounded below, then the statement is trivial. Since condition (A2) holds for $g(p)$, we only need to prove the statement when $g(p)$ is strictly decreasing for $0 < p < 1$, because it is trivial for the case $g(p)$ is bounded below and one can note $\frac{-C_2|\ln(1-p)|^k}{(1-p)^{1/\gamma}} \to -\infty$ as p goes to one.

Now we consider the case where $g(p)$ is strictly decreasing for $0 < p < 1$. In this case, by similar arguments when we prove the first statement, we denote $t = g(p)$ again and easily note that $g^{-1}(t)$ exists and further $g(p) = \bar{F}^{-1}(p)$ for $0 < p < 1$, where $\bar{F}^{-1}(\bar{F}(-t)) = -t$.

We now prove the second statement by contradiction. Given the previously defined notations, by assuming for any $k > 0$, there exists an arbitrary small p such that $\bar{F}^{-1}(p) < -C_2\frac{|\log(1-p)|^k}{(1-p)^{1/\gamma}}$, we derive the following contradiction:

$$\begin{aligned} -t = \bar{F}^{-1}(\bar{F}(-t)) &= \bar{F}^{-1}(1 - F(-t)) \\ &\leq \bar{F}^{-1}\left(1 - c_3\frac{|\log t|^{h}}{t^{\gamma}}\right) \\ &< -C_2\frac{\left|\log c_3\frac{|\log t|^{h}}{t^{\gamma}}\right|^{k}}{\left(c_3\frac{|\log t|^{h}}{t^{\gamma}}\right)^{1/\gamma}} \\ &= -t \times C_2\frac{|\log c_3 - \gamma\log t + h\log\log t|^{k}}{c_3(\log t)^{\frac{h}{\gamma}}} < -t \text{ by choosing large enough } k. \end{aligned}$$

The first inequality is due to the fact that $\bar{F}^{-1}(p)$ is strictly decreasing and the inequality $F(-t) < c_3\bar{F}(t) \leq c_3\frac{|\log t|^h}{t^\gamma}$ for large enough t and some constants $c_3 > 0$ and $h \geq 0$, which can be proved given conditions (A3) and (C2) hold. The second inequality is due to our assumption $\bar{F}^{-1}(p) < -C_2\frac{|\log(1-p)|^k}{(1-p)^{1/\gamma}}$. Given this contradiction, the proof of the second statement is completed. □

Remark C4. One can show that some common transformations $g(p)$ previously discussed satisfy the inequalities above. Indeed, the Box-Cox transformation $g(p) = \frac{1}{p^{1/\gamma}}$ satisfies condition (C2). For Cauchy's method, since the corresponding transformation $g(p) = \tan\{(0.5-p)\pi\}$ satisfies $\lim_{p\to 0}\frac{g(p)}{1/p} = \frac{1}{\pi}$ and $\lim_{p\to 1}\frac{g(p)}{\frac{-1}{\pi(1-p)}} = 1$, it also satisfies condition (C2). For the truncated Cauchy method, since $g(p) = \tan\{(0.5-p)\pi\}$ when $p \leq 1-\delta$, again we have $\lim_{p\to 0}\frac{g(p)}{1/p} = \frac{1}{\pi}$. Also note when $p > 1-\delta$, $g(p) = \tan\{(\delta-0.5)\pi\}$, hence $\lim_{p\to 1}\frac{g(p)}{\frac{-1}{(1-p)}} = 0$. Therefore, the transformation for the truncated Cauchy method also satisfies condition (C2).

Proof of Theorem 4.3. For this theorem, we only consider $0 < \gamma \leq 1$. Since $\boldsymbol{X}$ has banded correlation matrix (condition $(C1)$), we can split $U_1, \ldots, U_n$ into d_0+1 groups. Because we are only looking for the order of asymptotic distribution of $\sum_{i=1}^n U_i$, we can assume n is a multiple of d_0+1 and let $\frac{n}{d_0+1} - 1 = n_0$. Let the divided d_0+1 groups be
$\{U_1, U_{(d_0+1)+1}, \ldots, U_{(d_0+1)n_0+1}\}$; $\{U_2, U_{(d_0+1)+2}, \ldots, U_{(d_0+1)n_0+2}\}$; $\ldots$;
$\{U_{d_0+1}, U_{(d_0+1)+d_0+1}, U_{(d_0+1)n_0+d_0+1}\}$. For the ith group, the random variables
$\{U_i, U_{(d_0+1)+i}, \ldots, U_{(d_0+1)n_0+i}\}$ are independent and identically distributed and hence are stationary. Also, they are random variables with regularly varying tails with index γ that satisfy conditions $(A2)$ and $(A3)$. Thus conditions $(B1)$ and $(B2)$ hold. In addition, since they are independent, it is obvious conditions (D), (D') and (D'') in Davis (1983) for dependent structure hold. Let $S_i = \sum_{j=0}^{n_0} U_{j(d_0+1)+i}, i = 1, \ldots, d_0+1$. Since d_0 is fixed, by applying Lemmas C3 and C4, we obtain that S_i is $O_p(n^{1/\gamma}L_n)$. Therefore, $T(\boldsymbol{X}) = \sum_{i=1}^n U_i = S_1 + \ldots + S_{d_0+1}$ is also $O_p(n^{1/\gamma}L_n)$. Hence now it suffices to prove that under the alternative hypothesis H_a, $\frac{T(\boldsymbol{X})}{n^{1/\gamma}L_n}$ converges to ∞

with probability 1. Note that,

$$\begin{aligned}
T(\boldsymbol{X}) &= \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} g(2(1-\Phi(|X_i|))) \\
&= \sum_{i\in S} g(2(1-\Phi(|X_i|))) + \sum_{i\in S^c} g(2(1-\Phi(|X_i|))) \\
&= \sum_{i\in S} g(2(1-\Phi(|X_i|))) + O_p(n^{1/\gamma}L_n) \\
&\geq g(2(1-\Phi(\max_{i\in S}|X_i|))) + (n^\beta - 1)g(2(1-\Phi(\min_{i\in S}|X_i|))) + O_p(n^{1/\gamma}L_n),
\end{aligned}$$

where $S = \{i : \mu_i \neq 0\}$ and S^c is the complementary index set of S. The equality in the third line is due to Lemmas C3 and C4. We claim that if the second term $(n^\beta - 1)g(2(1-\Phi(\min_{\{i\in S\}}|X_i|)))$ in the last line is negative, its magnitude is $o_p(n^{1/\gamma})$.

Let $\epsilon_n > 0$ and $\epsilon_n \to 0$ as $n \to \infty$. We have

$$\begin{aligned}
P(\min_{i\in S}|X_i| < \epsilon_n) &\leq \sum_{i\in S} P(|X_i| < \epsilon_n) = n^\beta P(|X_i| < \epsilon_n) \\
&= n^\beta\{\Phi(\mu_0+\epsilon_n) - \Phi(\mu_0-\epsilon_n)\} \leq 2\phi(\mu_0-\epsilon_n)n^\beta\epsilon_n \leq n^\beta\epsilon_n.
\end{aligned}$$

Apply Lemma C6 we have for small value of $\epsilon_n > 0$,

$$g(2(1-\Phi(\epsilon_n))) \geq \frac{-C^1|\log(2\Phi(\epsilon_n)-1)|^k}{(2\Phi(\epsilon_n)-1)^{1/\gamma}}. \tag{C4}$$

Note that $2\Phi(\varepsilon_n) - 1 = 2(\Phi(\varepsilon_n) - \Phi(0)) = 2(\phi(0)\varepsilon_n + o(\varepsilon_n)) = \varepsilon_n(1+o(1))$, then we have

$$\begin{aligned}
|\log(2\Phi(\epsilon_n)-1)|^k &= |\log(\varepsilon_n(1+o(1)))|^k \leq 2^k|\log\varepsilon_n|^k \\
(2\Phi(\epsilon_n)-1)^{\frac{1}{\gamma}} &= (\varepsilon_n(1+o(1)))^{\frac{1}{\gamma}} \geq 2^{-\frac{1}{\gamma}}\varepsilon_n^{\frac{1}{\gamma}}.
\end{aligned}$$

Then for the right hand side of (C4), we have

$$(C4) \geq -\frac{2^kC^1|\log\varepsilon_n|^k}{2^{-\frac{1}{\gamma}}\varepsilon_n^{\frac{1}{\gamma}}} = -C^0\frac{|\ln(\epsilon_n)|^k}{\epsilon_n^{1/\gamma}},$$

where $C^1 > 0, C^0 > 0$ are constants. Now we let $\epsilon_n = n^{\beta_0-1}$, where $\beta < \beta_0 < 1/2$. Then we have $P(\min_{i\in S}|X_i| < \epsilon_n) \leq n^\beta\epsilon_n = n^{\beta+\beta_0-1} = o(1)$ and $n^\beta g(2(1-\Phi(\epsilon_n))) \geq -C^0n^{\beta-(\beta_0-1)(1/\gamma)}|\ln(n^{\beta_0-1})|^k$. So we prove that $(n^\beta - 1)g(2(1-\Phi(\min_{\{i\in S\}}|X_i|)))$ is $o_p(n^{1/\gamma})$.

Then it suffices to prove that $\frac{g(2(1-\Phi(\max_{i\in S}|X_i|)))}{n^{1/\gamma}L_n}$ converges to ∞ with probability 1. Let $S_+ = \{i \in S, \mu_i > 0\}$. Denote $X_i = \mu_0 + Z_i$ for $i \in S_+$, where $\mu_0 = \sqrt{2\tau \log n}$ and $Z_i \overset{D}{\sim} N(0,1)$. Without loss of generality we assume $|S_+| \geq s/2$. Under the assumption of banded correlation for $X_1, \ldots, X_n$, it follows from Lemma 6 in Cai et al. (2014) that $\max_{i\in S_+} Z_i \geq \sqrt{2\log|S_+|} + o_p(1)$. Then we have $\max_{i\in S}|X_i| \geq \max_{i\in S_+}|X_i| \geq \mu_0 + \max_{i\in S_+} Z_i \geq \mu_0 + \sqrt{2\log|S_+|} + o_p(1)$. Hence we have

$$
\begin{aligned}
&g(2(1-\Phi(\max|X_i|))) \geq \frac{C_1}{(2(1-\Phi(\max|X_i|)))^{\frac{1}{\gamma}}|\log(2(1-\Phi(\max|X_i|)))|^k} \\
&\geq \frac{C_1}{(1-\Phi(\max|X_i|))^{\frac{1}{\gamma}-\delta}} + o_p(1) \\
&\geq C_2 \max_{i\in S}|X_i|^{\frac{1}{\gamma}-\delta}\exp\{(\frac{1}{\gamma}-\delta)\max_{i\in S}|X_i|^2/2\} + o_p(1) \\
&\geq C_2(\sqrt{2\log|S_+|}+\mu_0)^{\frac{1}{\gamma}-\delta}\exp\{(\frac{1}{\gamma}-\delta)(\log|S_+| + \mu_0^2/2 + \mu_0\sqrt{2\log|S_+|})\} + o_p(1) \\
&\geq \exp\{(\frac{1}{\gamma}-\delta)(\log|S_+| + \mu_0^2/2 + \mu_0\sqrt{2\log|S_+|})\} + o_p(1) \\
&\geq C_3\exp\{(\frac{1}{\gamma}-\delta)(\beta\log(n) + \tau\log(n) + \sqrt{2\tau\log(n)}\sqrt{2\beta\log(n) - 2\log(2)})\} + o_p(1) \\
&\geq C_3\exp\{(\frac{1}{\gamma}-\delta)(\beta\log(n) + \tau\log(n) + \sqrt{2\tau\log(n)}\sqrt{2\beta\log(n)} - \sqrt{2\log(2)})\} + o_p(1) \\
&\geq C_3\exp\{(\frac{1}{\gamma}-\delta)(\beta\log(n) + \tau\log(n) + \sqrt{2\tau\log(n)}\sqrt{2\beta\log(n)})\} + o_p(\exp(\frac{1}{\gamma}-\delta)\sqrt{2\tau\log(n)}) \\
&\geq C_3\exp\{(\frac{1}{\gamma}-\delta)(\log(n)(\sqrt{\beta}+\sqrt{\tau})^2)\} + o_p(\exp\sqrt{2\tau\log(n)}) \\
&= C_3 n^{(\frac{1}{\gamma}-\delta)(\sqrt{\tau}+\sqrt{\beta})^2} + o_p(\exp\sqrt{2\tau\log(n)}).
\end{aligned}
$$

Note that δ in the second line is a small positive number. The inequality in the first line is due to Lemma C6; the inequality in the second line is because $|\log(p)|^k$ is smaller than $p^{-\delta}$ for any positive number δ when p is small and because $\max|X_i|$ goes to infinity with probability 1; the inequality in the third line is due to Lemma C5. Since $\sqrt{\tau}+\sqrt{\beta} > 1$, we can choose δ so small that $(\frac{1}{\gamma}-\delta)(\sqrt{\tau}+\sqrt{\beta})^2 > \frac{1}{\gamma}$. Therefore, the proof is complete. □

Remark C5. When $\gamma \leq 1$ and $0 < \beta < 1/4$ (the strong sparsity situation), the detection boundary for test statistic $T(\boldsymbol{X})$ is optimal.

Remark C6. For $T_{n,\boldsymbol{w}} = \sum_{i=1}^{n} w_i g(p_i)$ under the conditions of Theorem 4.3 and $\boldsymbol{w} \in R_+^n$ and $\sum_{i=1}^{n} w_i = n$, if $\max_i w_i \leq (\log n)^{\eta_1}$ and $\min_i w_i \geq 1/(\log n)^{\eta_2}$ for some fixed constants $\eta_1, \eta_2 > 0$, then the result of Theorem 4.3 can be easily extended to $T_{n,\boldsymbol{w}}(\boldsymbol{X})$. Indeed, combining the arguments in the proof of Theorem 4.3 and conditions on the weights $\boldsymbol{w}$, one can show that under the null, $T_{n,\boldsymbol{w}}(\boldsymbol{X}) \leq \max_i w_i \sum_{i=1}^{n} g(p_i) \leq O_p(n^{1/\gamma} L_n)(\log n)^{\eta_1} = O_p(n^{1/\gamma} L_n)$. Similarly, under the alternative, one can show that $T_{n,\boldsymbol{w}}(\boldsymbol{X}) \geq \min_i w_i \sum_{i=1}^{n} g(p_i) \geq$
$O_p(n^{\frac{1}{\gamma}(\sqrt{\tau}+\sqrt{\beta})^2} L_n)(\log n)^{-\eta_2} = O_p(n^{\frac{1}{\gamma}(\sqrt{\tau}+\sqrt{\beta})^2} L_n)$. Then the result follows.

C.2 Results Related to Truncated Cauchy Method (CA$^{\text{tr}}$)

C.2.1 Truncated Cauchy: a Remedy for Large Negative Penalty Issue in Cauchy

As a simple remedy of the large negative penalty issue in Cauchy (discussed in Sections 4.2.1, 4.4.3 and 4.5), we propose the truncated Cauchy test (CA$^{\text{tr}}$) that truncates any of the n p-values greater than $1-\delta$ to be $1-\delta$. Recall the statistic of CA$^{\text{tr}}$ can be written as:

$$T_{CA^{tr}} = \sum_{i=1}^{n} \tan\left(\pi\left(\frac{1}{2} - p_i\right)\right) 1(p_i < 1-\delta) + \tan\left(\pi\left(\delta - \frac{1}{2}\right)\right) 1(p_i \geq 1-\delta).$$

The theorems introduced in Section 4.3 imply that CA$^{\text{tr}}$ enjoys almost the same advantages of the Cauchy method in terms of type I error control and power for the detection of weak and sparse signals. Indeed, like Cauchy's method, Theorems 1 and 2 ensure that we can approximate the null distribution of CA$^{\text{tr}}$ under dependence using its null distribution under independence assumption of p_i, $i = 1, \ldots, n$. The test statistic of CA$^{\text{tr}}$ no longer follows the standard Cauchy distribution under the null and independence assumption. To deal with the computational issue of the truncated Cauchy method, we propose a hybrid strategy, which uses approximation by generalized central limit theorem (GCLT) in general but switches to an efficient importance sampling procedure by cross-entropy parameter selection when n is small ($n < 25$) and the targeted size is large ($\alpha \geq 5 \times 10^{-3}$).

Below we first show that when n is sufficiently large, we can apply generalized central limit theorem (GCLT) from Shintani and Umeno (2018) to approximate the null distribution of $T_{CA^{tr}}$.

Proposition C1. *Assume that $p_1, \ldots, p_n$ independently and identically follow $Unif(0,1)$. Let $\nu_\delta = \tan\left(\pi(\delta - \frac{1}{2})\right)$, $f_{1n} = \int_{\nu_\delta}^{+\infty} \frac{\cos(x/n)}{(1+x^2)}$, $f_{2n} = \int_{\nu_\delta}^{+\infty} \frac{\sin(x/n)}{(1+x^2)}$ and*

$$\theta_n = \arctan\left(\frac{\delta \sin(\nu_\delta/n) + ((1-\delta)/\pi) f_{2n}}{\delta \cos(\nu_\delta/n) + ((1-\delta)/\pi) f_{1n}}\right).$$

Then we have:

$$\frac{T_{CA^{tr}} - n^2\theta_n}{n} \xrightarrow{d} S(1, 1, \frac{1}{2}, 0),$$

where $S(\alpha, \beta, \gamma, \mu)$ is a stable distribution with parameters $\alpha = 1, \beta = 1, \gamma = \frac{1}{2}$ and $\mu = 0$, which is defined with its characteristic function as:

$$S(x; \alpha, \beta, \gamma, \mu) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi(t) e^{-ixt} \, \mathrm{d}t,$$

with $\phi(t) = \exp\{i\mu t - \gamma^\alpha |t|^\alpha (1 - i\beta \operatorname{sgn}(t) w(\alpha, t))\}$ and

$$w(\alpha, t) = \begin{cases} \tan(\pi\alpha/2) & \text{if } \alpha \neq 1 \\ -2/\pi \log|t| & \text{if } \alpha = 1 \end{cases}.$$

Remark C7. Proposition C1 can be obtained by simple calculation using formula (4) in Shintani and Umeno (2018). Table C4 examines the approximation performance of GCLT for small n and varying size α. The result shows satisfying accuracy when $\alpha < 5 \times 10^{-3}$. When $\alpha \geq 5 \times 10^{-3}$, GCLT needs larger n to perform well (roughly $n \geq 25$). As a result, we develop an efficient importance sampling procedure for this scenario. Briefly, Proposition C2 below gives narrow upper and lower bounds for the tail probability of truncated Cauchy. By applying the framework proposed by De Boer et al. (2005) for estimating rare event probability, we develop a cross-entropy procedure to search within the narrow bounds for a high-precision approximation for the tail probability of the truncated Cauchy. Details of the efficient importance sampling are shown in Section C.2.3. Table C4 further shows the accurate calculation of the importance sampling with affordable computing when $n < 25$. In summary, when calculating p-values for CA$^{\text{tr}}$, to balance the computing and performance, we propose to set $\delta = 0.01$ (discussed in Section C.2.4) and use GCLT approximation when $\alpha < 5 \times 10^{-3}$ or $n \geq 25$. When $\alpha \geq 5 \times 10^{-3}$ and $n < 25$, importance sampling will be used. In Section 4.4.3 and Section 4.5, we will demonstrate the superior performance of truncated Cauchy over Cauchy using simulations and a real application.

Specifically, it avoids the large negative penalty issue of the Cauchy method but still enjoys similar robust properties for type I error control under dependency and power for detecting weak and sparse signals.

Proposition C2. *Assume that $p_1, \ldots, p_n$ independently and identically follow $Unif(0,1)$. Let $1-\delta$ be the truncation point of the truncated Cauchy test with $0 < \delta < \frac{1}{2}$. The upper tail probability of the null distribution of the truncated Cauchy method satisfies:*

$$P\left(X_1 \geq t\right) \leq P\left(\frac{T_{CA^{tr}}}{n} > t\right) \leq P\left(X_1 \geq t\right)(1+\delta)^n,$$

where $t > 0$ and X_1 is a Cauchy distributed random variable.

C.2.2 Proof of Proposition C2

Proof. Define the following random variables,

$$Y_i = X_i 1(X_i \geq \nu_\delta) + \nu_\delta 1(X_i < \nu_\delta) \;\; i = 1, \ldots, n.$$

Here $X_i's$ identically and independently follow the standard Cauchy distribution, and recall that $\nu_\delta = \tan\left(\pi(\delta - \frac{1}{2})\right)$ for $0 < \delta < \frac{1}{2}$. Define index set $\mathcal{I} = \{k : X_k < \nu_\delta\}$ and let $m = |\mathcal{I}|$, the cardinality of $\mathcal{I}$. Then under the null, we can rewrite the upper tail probability of the truncated Cauchy method's test statistic in the following form:

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t\right) = \sum_{j=0}^{n} P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\; m = j\right).$$

Given the above equivalent form, the tail probability can be divided into the two parts below, which will be bounded in the following proof:

$$I = P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\; m = 0\right),$$
$$II = \sum_{j=1}^{n} P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\; m = j\right).$$

For I, we have

$$I = P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\; X_1, \ldots, X_n \geq \nu_\delta\right) \leq P\left(\frac{1}{n}\sum_{i=1}^{n} X_i \geq t\right) = P\left(X_1 \geq t\right).$$

For II, note that for the terms $P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\ m=j\right)$ for $j=1,\ldots,n-1$, we have

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\ m=j\right) = \binom{n}{j} P\left(\frac{1}{n-j}\sum_{i=1}^{n-j} X_i \geq \frac{nt-j\nu_\delta}{n-j}\ , m=j\right)$$
$$\leq \binom{n}{j} \left(P\left(X_n < \nu_\delta\right)\right)^j P\left(X_1 \geq t\right).$$

Since $t>0$ and $\nu_\delta<0$, $P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t,\ m=n\right)=0$. Hence by the binomial theorem,

$$P\left(\frac{1}{n}\sum_{i=1}^{n} Y_i \geq t\right) = I + II \leq P(X_1 \geq t) + \sum_{j=1}^{n} \binom{n}{j} \left(P\left(X_n < \nu_\delta\right)\right)^j P\left(X_1 \geq t\right)$$
$$\leq P\left(X_1 \geq t\right)\left(1 + P(X_n < \nu_\delta)\right)^n.$$

Notice $\tan\left(\pi(\frac{1}{2}-p)\right)$ follows the standard Cauchy distribution under the null, hence $P(X_n < \nu_\delta) = \delta$, then the result follows. □

C.2.3 The Cross-Entropy Method (CE) for CA$^{\text{tr}}$.

This subsection introduces the cross-entropy method (CE) to build the library of the null reference distribution of the truncated Cauchy method (CA$^{\text{tr}}$).

Let $\boldsymbol{X}_1$, $\boldsymbol{X}_2$, …, $\boldsymbol{X}_N$ be N independent and identically distributed random vectors that follow $N_n(0, I_{n\times n})$. Denote $\boldsymbol{X}_i = (X_{i1},\ldots,X_{in})'$ for $i=1,\ldots,n$ and f as the density function of $N_n(0, I_{n\times n})$. We consider an equivalent version of CA$^{\text{tr}}$:

$$S_{CA^{tr}}(\boldsymbol{X}_i) = \frac{1}{n}\sum_{j=1}^{n} \tan\left(\pi\left(\Phi(X_{ij}) - \frac{1}{2}\right)\right) 1\left(\delta < \Phi(X_{ij})\right) + \tan\left(\pi\left(\delta - \frac{1}{2}\right)\right) 1\left(\delta \geq \Phi(X_{ij})\right),$$

where Φ is the CDF of standard normal distribution. $S_{CA^{tr}}$ has the same distribution as the null distribution of $T_{CA^{tr}}/n$, so building the null reference distribution of $T_{CA^{tr}}$ is equivalent to building that of $S_{CA^{tr}}$.

For our proposed method, we first set-up the range of the significant levels of interest, which is denoted as $[\alpha_{\text{min}}, \alpha_{\text{max}}]$. For example, if we are interested in the significant levels $\alpha = 0.05$, 0.01, 1×10^{-3} and 1×10^{-4}, the range of the significant levels is $[1\times10^{-4}\ , 0.05]$. For each significance

level α, define the corresponding quantile φ_α satisfying $\varphi_\alpha = P(S_{CA^{tr}} > \varphi_\alpha) = \alpha$. We then calculate the range of φ_α using Proposition C2, denoted as

$$[\varphi_{\min}, \varphi_{\max}] = \left[F^{-1}_{\text{Cauchy}}\left(1-\alpha_{\max}\right), F^{-1}_{\text{Cauchy}}\left(1-\alpha_{\min}/(1+\delta)^n\right)\right], \tag{C5}$$

where F^{-1}_{Cauchy} is the inverse CDF of the standard Cauchy distribution and n is the total number of p-values to combine, which is also the dimension of $\boldsymbol{X}_i$. That is, for any significant level α between $\alpha_{\min}$ and $\alpha_{\max}$, $\varphi_{\min} \leq \varphi_\alpha \leq \varphi_{\max}$.

We then choose m points between $\varphi_{\min}$ and $\varphi_{\max}$, such that $\varphi_1 = \varphi_{\min}$, $\varphi_m = \varphi_{\max}$ and $\log\varphi_{k+1} - \log\varphi_k = \frac{\log(\varphi_{\max})-\log(\varphi_{\min})}{m}$ for $k = 1, \ldots, m-1$. For each φ_k, we then apply the method adapted from De Boer et al. (2005) to estimate $\hat{p}_k = P\left(S_{CA^{tr}} \geq \varphi_k\right)$ for $k = 1,\ldots, m$, which will be described in Algorithm S2. We then fit an increasing spline function $sp(\varphi)$ representing the relationship between φ_α and $-\log(\alpha)$ using points $(\varphi_k, -\log(\hat{p}_k))$ for $k = 1, \ldots, m$ to build the reference distribution of $S_{CA^{tr}}$. Users may enter the observed test statistic value $T_{CA^{tr},obs}$ to get the corresponding p-value $\exp\left(-sp(T_{CA^{tr},obs}/n)\right)$. The reason we fit an increasing spline function for φ_α and $-\log(\alpha)$ instead of φ_α and α is to make the fitting procedure numerically stable.

We summarize the above steps in the following Algorithm S1:

Algorithm S1:

1. Calculate $[\varphi_{\min}, \varphi_{\max}]$ using Proposition C2 given $[\alpha_{\min}, \alpha_{\max}]$, the range of significant levels of interest.

2. Choose m points between $\varphi_{\min}$ and $\varphi_{\max}$, such that $\varphi_1 = \varphi_{\min}$, $\varphi_m = \varphi_{\max}$ and $\log\varphi_{k+1} - \log\varphi_k = \frac{\log(\varphi_{\max})-\log(\varphi_{\min})}{m}$ for $k = 1, \ldots, m-1$.

3. For each φ_k (k=1,. . ., m), set $\varphi = \varphi_k$, let $\hat{p}$ be the output of Algorithm S2 with input φ, $N = 10^5$ and $\rho = 0.01$. Let $\hat{p}_k = \hat{p}$.

4. Fit an increasing spline function for φ_α and $-\log(\alpha)$ using points $(\varphi_k, -\log(\hat{p}_k)), k = 1\ldots, m$ to build the reference distribution of $S_{CA^{tr}}$.

We then show Algorithm S2, the method adapted from De Boer et al. (2005) to estimate $\hat{p}_k = P\left(S_{CA^{tr}} \geq \varphi_k\right)$ for $k = 1, \ldots, m$ in step 3 in Algorithm S1:

Algorithm S2 (adapted algorithm from De Boer et al. (2005)):

Input: N, φ and ρ.

1. Set t=1, $\lambda_0 = 1/(n+1)$, $\theta_0 = 1$ and $\widehat{\varphi}_0 = -\infty$.

2. Generate random samples $\boldsymbol{X}_1$,...,$\boldsymbol{X}_N$, where each $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{in})'$ and all the X_{ij}'s (j=1,. . .,n) are independently sampled from the Gaussian mixture distribution:

$$\lambda_{t-1} N(0, \theta_{t-1}) + (1 - \lambda_{t-1}) N(0, 1).$$

Denote the corresponding joint density function as $g(\cdot; \theta_{t-1}, \lambda_{t-1})$. Calculate and sort $S_{CA^{tr}}(\boldsymbol{X}_i)$ for $i = 1, \ldots, n$ from the smallest to the largest, denoted as $S_{CA^{tr}(1)} \leq S_{CA^{tr}(2)} \leq \ldots \leq S_{CA^{tr}(N)}$. Let $\widehat{\varphi}_t := S_{CA^{tr}(\lceil (1-\rho)N \rceil)}$, if this is less than φ. Otherwise set $\hat{\varphi}_t = \varphi$.

3. Update θ_t and λ_t:

$$(\theta_t, \lambda_t) = \arg\max_{\theta, \lambda} \frac{1}{N} \sum_{i=1}^{N} 1\left(S_{CA^{tr}}(\boldsymbol{X}_i) \geq \hat{\varphi}_t\right) W\left(\boldsymbol{X}_i, \theta_{t-1}, \lambda_{t-1}\right) \log(g(\boldsymbol{X}_i; \theta, \lambda)),$$

where $W\left(\boldsymbol{X}_i, \varphi_{t-1}, \lambda_{t-1}\right) = \frac{f(\boldsymbol{X}_i)}{g(\boldsymbol{X}_i; \theta_{t-1}, \lambda_{t-1})}$. Recall f is the density function of $N_n(0, I_{n\times n})$.

4. If $\hat{\varphi}_t = \varphi$ then proceed to step 5, otherwise set $t = t + 1$ and back to step 2.

5. Resample $\boldsymbol{X}_1, \ldots \boldsymbol{X}_N$ from $g(\cdot; \theta_T, \lambda_T)$, where T denotes the final number of iterations, then estimate $\hat{p} = P\left(S_{CA^{tr}} \geq \varphi\right)$ by:

$$\hat{p} = \frac{1}{N} \sum_{i=1}^{N} 1\left(S_{CA^{tr}}\left(\boldsymbol{X}_i\right) \geq \varphi\right) W\left(\boldsymbol{X}_i, \theta_T, \lambda_T\right).$$

The key of Algorithm S2 is to choose $g(\cdot; \theta, \lambda)$ in the form of a mixture of Gaussian distributions. The reason is that CA$^{\text{tr}}$'s behavior is dominated by extremely small p-values, and often a tiny fraction of extremely small p-values can lead to an extreme value of CA$^{\text{tr}}$.

C.2.4 Choice of the Value of δ

For selection of δ, conceptually δ should be large enough so that it avoids the large negative penalty issue in Cauchy. But for computational purpose, it cannot be too large so approximation by our fast-computing procedures is accurate. To balance the computing and performance, we recommend to set $\delta = 0.01$. The reason is as follows. Besides almost exactly empirical statistical power of CA$^{\text{tr}}$ with varying $\delta = 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1$ under the same simulation settings in Section 4.4.2 (Tables C2 and C1), as shown in Table C3, under the same simulation setting in Section 4.4.3, with all the type I errors controlled under the size of tests, the power of CA$^{\text{tr}}$ with $\delta = 0.01$ is very close to the power of CA$^{\text{tr}}$ with $\delta = 0.05$ and $\delta = 0.1$, and much greater than the power of CA$^{\text{tr}}$ with $\delta \leq 0.005$, indicating selection of $\delta \geq 0.01$ can sufficiently alleviate the large negative penalty issue. In the meantime, also note that for $n = 25$ (we recommend to use CE algorithm for CA$^{\text{tr}}$ when $n < 25$, discussed in Remark C7 and Section C.2.5), we have $(1 + 0.01)^{25} \approx 1.28$, while $(1 + 0.05)^{25} - 1 \approx 3.38$ and $(1 + 0.05)^{25} - 1 \approx 10.83\%$. Hence compared to $\delta = 0.05$ or $\delta = 0.1$, the choice of $\delta = 0.01$ can lead to a much more narrower range of φ_α derived from Equation (C5) for the CE algorithm. Furthermore, by Proposition C2, $\delta = 0.01$ also provides more accurate approximation when one wishes to use standard Cauchy to approximate the null distribution of CA$^{\text{tr}}$. Based on all consideration above, we recommend and set $\delta = 0.01$ throughout the paper.

C.2.5 Performance Benchmark of GCLT and CE

We choose $\delta = 0.01$ for the CA$^{\text{tr}}$ method, and sample 10^8 $T_{CA^{tr}}$ samples by Monte-Carlo sampling for different total numbers of p-values n =2, 3, 4, 5, 10, 15, 20, 25. For GCLT, we obtain the asymptotic quantiles η_q of $\frac{T_{CA^{tr}} - n^2\theta_n}{n}$ with q =0.9, 0.95, 0.99, 0.995, 0.999, 0.9995 and 0.9999 from the stable distribution $S(1, 1, \frac{1}{2}, 0)$ defined in Proposition C1, which correspond to the significant levels α=0.1, 0.05, 0.01, 10^{-3}, 5×10^{-3}, 5×10^{-4} and 10^{-4}. Note that quantities $n\eta_q + n^2\theta_n$ are the corresponding asymptotic quantiles for $T_{CA^{tr}}$ (definition of θ_n is in Proposition C1) for each combination of q and n. Then we plug all the $n\eta_q + n^2\theta_n$'s into the corresponding Monte-Carlo samples to calculate the empirical upper tail probability.

For the CE method, for each fixed n, we set range of significant levels of interest to be

Table C1: Mean uncorrected power for tests CA, HM and CA^{tr} (truncated CA) with $\delta = 0.1, 0.05, 0.01, 0.005$,$0.001, 0.0005, 0.0001$ across correlation $\rho = 0, 0.3, 0.6$,$0.9, 0.99$, $n = 100$, and proportion of signals $s/n = 5\%, 10\%, 20\%$. The standard error is far less than the mean power and hence not shown here.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
	CA	0.753	0.633	0.521	0.396	0.350
	HM	0.753	0.633	0.522	0.397	0.349
	δ =0.0001	0.753	0.633	0.521	0.396	0.350
	δ =0.0005	0.753	0.633	0.521	0.397	0.350
5%	δ =0.001	0.753	0.633	0.521	0.397	0.350
	δ =0.005	0.753	0.633	0.522	0.397	0.350
	δ =0.01	0.753	0.633	0.522	0.397	0.350
	δ =0.05	0.753	0.633	0.522	0.397	0.349
	δ =0.1	0.753	0.633	0.522	0.397	0.349
	CA	0.873	0.693	0.536	0.374	0.322
	HM	0.873	0.693	0.536	0.375	0.321
	δ =0.0001	0.873	0.693	0.536	0.374	0.322
	δ =0.0005	0.873	0.693	0.536	0.374	0.321
10%	δ =0.001	0.873	0.693	0.536	0.375	0.321
	δ =0.005	0.873	0.693	0.536	0.375	0.321
	δ =0.01	0.873	0.693	0.536	0.375	0.321
	δ =0.05	0.873	0.693	0.536	0.375	0.321
	δ =0.1	0.873	0.693	0.536	0.375	0.321
	CA	0.957	0.742	0.545	0.357	0.302
	HM	0.957	0.741	0.546	0.357	0.301
	δ =0.0001	0.957	0.742	0.546	0.357	0.301
	δ =0.0005	0.957	0.742	0.546	0.357	0.301
20%	δ =0.001	0.957	0.742	0.546	0.357	0.301
	δ =0.005	0.957	0.742	0.546	0.357	0.301
	δ =0.01	0.957	0.742	0.546	0.357	0.301
	δ =0.05	0.957	0.742	0.546	0.357	0.301
	δ =0.1	0.957	0.742	0.546	0.357	0.301

Table C2: Mean corrected power for tests CA, HM and CA$^{\text{tr}}$ (truncated CA) with $\delta = 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001$ across correlation $\rho = 0, 0.3, 0.6, 0.9, 0.99$, $n = 100$, and proportion of signals $s/n = 5\%, 10\%, 20\%$. The standard error is far less than the mean power and hence not shown here.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
	CA	0.753	0.615	0.495	0.387	0.347
	HM	0.753	0.616	0.496	0.389	0.347
	δ =0.0001	0.753	0.615	0.495	0.388	0.347
	δ =0.0005	0.753	0.615	0.495	0.388	0.347
5%	δ =0.001	0.753	0.615	0.495	0.388	0.347
	δ =0.005	0.753	0.616	0.495	0.388	0.347
	δ =0.01	0.753	0.616	0.495	0.388	0.347
	δ =0.05	0.753	0.616	0.495	0.389	0.347
	δ =0.1	0.753	0.616	0.496	0.389	0.347
	CA	0.873	0.675	0.508	0.365	0.318
	HM	0.873	0.675	0.509	0.367	0.318
	δ =0.0001	0.873	0.675	0.508	0.365	0.318
	δ =0.0005	0.873	0.675	0.508	0.365	0.318
10%	δ =0.001	0.873	0.675	0.508	0.366	0.318
	δ =0.005	0.873	0.675	0.509	0.366	0.318
	δ =0.01	0.873	0.675	0.509	0.366	0.318
	δ =0.05	0.873	0.675	0.509	0.366	0.318
	δ =0.1	0.873	0.675	0.509	0.366	0.318
	CA	0.957	0.723	0.516	0.347	0.298
	HM	0.957	0.724	0.518	0.348	0.298
	δ =0.0001	0.957	0.723	0.517	0.347	0.298
	δ =0.0005	0.957	0.724	0.517	0.347	0.298
20%	δ =0.001	0.957	0.724	0.517	0.348	0.298
	δ =0.005	0.957	0.724	0.517	0.348	0.298
	δ =0.01	0.957	0.724	0.517	0.348	0.298
	δ =0.05	0.957	0.724	0.517	0.348	0.298
	δ =0.1	0.957	0.724	0.517	0.348	0.298

Table C3: Mean proportion of rejection of CA, HM and CA$^{\text{tr}}$ (truncated CA) with $\delta = 0.1, 0.05, 0.01, 0.005, 0.001, 0.0005, 0.0001$ across $\rho = 0.2, 0.3$, under the same simulation setting in Section 4.4.3. The standard errors are far less than the mean proportion and hence omitted.

ρ_{11}	Methods/Cutoffs	0.05	0.01	0.005	0.001	5×10^{-4}	10^{-4}
	CA	0.333	0.202	0.147	0.0582	0.0399	0.0135
	HM	0.864	0.525	0.380	0.154	0.107	0.0355
	$\delta = 0.0001$	0.457	0.320	0.251	0.131	0.0937	0.0350
	$\delta = 0.0005$	0.606	0.422	0.328	0.148	0.106	0.0354
ρ_{11} =0.2	$\delta = 0.001$	0.685	0.463	0.353	0.151	0.107	0.0355
	$\delta = 0.005$	0.821	0.516	0.375	0.154	0.107	0.0356
	$\delta = 0.01$	0.849	0.523	0.379	0.154	0.107	0.0355
	$\delta = 0.05$	0.867	0.527	0.381	0.154	0.107	0.0356
	$\delta = 0.1$	0.867	0.527	0.381	0.154	0.107	0.0355
	CA	0.431	0.428	0.419	0.356	0.309	0.190
	HM	1.000	0.992	0.971	0.822	0.717	0.439
	$\delta = 0.0001$	0.928	0.916	0.898	0.778	0.689	0.433
	$\delta = 0.0005$	0.989	0.977	0.955	0.813	0.711	0.438
ρ_{11} =0.3	$\delta = 0.001$	0.996	0.985	0.963	0.818	0.714	0.439
	$\delta = 0.005$	1.000	0.991	0.970	0.822	0.717	0.439
	$\delta = 0.01$	1.000	0.992	0.971	0.822	0.717	0.439
	$\delta = 0.05$	1.000	0.992	0.972	0.823	0.717	0.439
	$\delta = 0.1$	1.000	0.992	0.972	0.823	0.717	0.439

Table C4: Approximated tail probability of CA^{tr} with $\delta = 0.01$ by generalized central limit theory (GCLT) and our proposed cross-entropy method (CE) evaluated at total number of studies n =2, 3, 4, 5, 10, 15, 20, 25 and 30.

	α	n=2	n=3	n=4	n=5	n=10	n=20	n=25	n=30
GCLT	0.1	0.0870382	0.0864065	0.0886247	0.0875406	0.0891782	0.0947768	0.0954372	0.0956934
	0.05	0.0451842	0.0450655	0.0457294	0.0454963	0.0461855	0.0479721	0.0482332	0.0483709
	0.01	0.0096584	0.0096709	0.0097111	0.0097127	0.0097765	0.0098790	0.0098943	0.0099051
	5×10^{-3}	0.0049012	0.0049070	0.0049185	0.0049215	0.0049389	0.0049683	0.0049715	0.0049748
	1×10^{-3}	0.0009948	0.0009955	0.0009963	0.0009973	0.0009977	0.0009985	0.0009980	0.0009987
	5×10^{-4}	0.0004977	0.0004990	0.0004994	0.0004991	0.0004995	0.0004996	0.0004995	0.0004995
	1×10^{-4}	0.0000998	0.0001001	0.0001000	0.0000999	0.0000999	0.0000999	0.0000998	0.0000999
CE	0.1	0.0998460	0.1002799	0.0999711	0.1000699	0.1001043	0.0999127	0.1001925	0.1002092
	0.05	0.0499530	0.0499556	0.0500164	0.0498986	0.0499906	0.0499466	0.0501307	0.0499883
	0.01	0.0099885	0.0100282	0.0100509	0.0099742	0.0099965	0.0100538	0.0099953	0.0099947
	5×10^{-3}	0.0049928	0.0049927	0.0050023	0.0050125	0.0049725	0.0050595	0.0049952	0.0049998
	1×10^{-3}	0.0010037	0.0009876	0.0009994	0.0009928	0.0009923	0.0010062	0.0010034	0.0009999
	5×10^{-4}	0.0005022	0.0004936	0.0005003	0.0005042	0.0004982	0.0005030	0.0004997	0.0004943
	1×10^{-4}	0.0001002	0.0000990	0.0000989	0.0000998	0.0000976	0.0001007	0.0001010	0.0000994

$[0.01, 10^{-5}/(1.01)^n]$, using Algorithm S2 to build the reference library and estimate the quantiles η_q corresponding to the significant levels α=0.1, 0.05, 0.01, 10^{-3}, 5×10^{-3}, 5×10^{-4} and 10^{-4}.

Table C4 shows the mean empirical upper tail probability of the two methods for each combination of n and 1-q from 30 times repeated simulations.

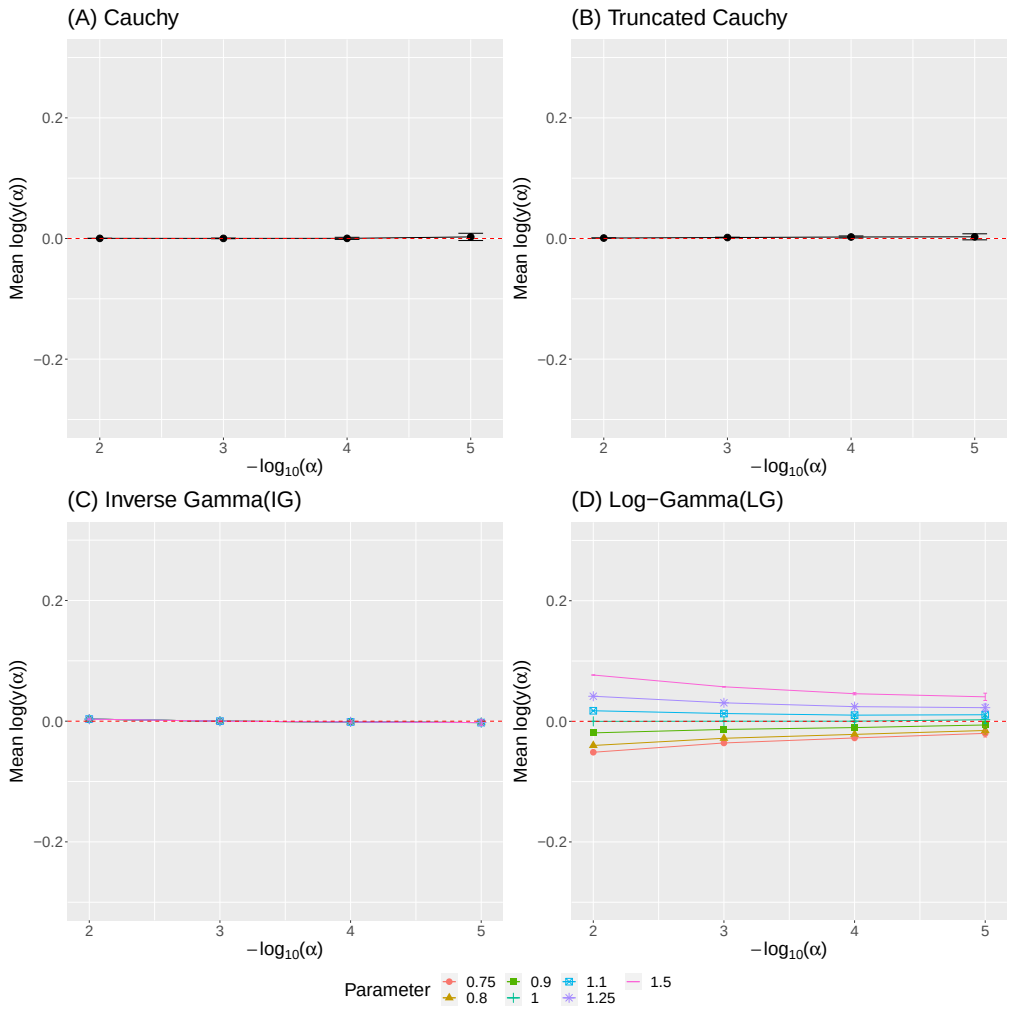


Figure C1: All the sub-figures represent the mean logarithm of ratio $\frac{3P(U>t_\alpha)}{P(T_{3,w}(X)>t_\alpha)}$ $(y(\alpha))$ across different significance levels α for correlation level $\rho = 1$ for 4 different methods, Cauchy, Truncated Cauchy, Inverse Gamma and log-Gamma distributions. We set the shape parameter of the inverse Gamma distribution and the rate parameter of log-Gamma distribution to be 1. We further set the scale parameter of inverse Gamma and shape parameter of log-Gamma distribution to be 0.75, 0.8, 0.9, 1, 1.1, 1.25, and 1.5. The x-axis is the negative logarithm of significance level α to base 10 where α is set to be $10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$. The red dash line is the reference line $y = 0$.

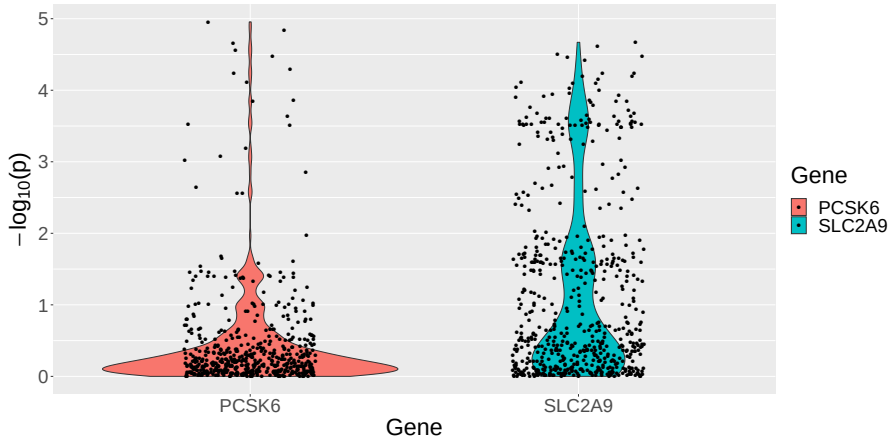


Figure C2: Jitter plots of p-values for SNPs in genes SLC29A9 (left) and PCS29A9 (right).

Bibliography

Abrahamson, I. G. (1967). Exact Bahadur efficiencies for the Kolmogorov-Smirnov and Kuiper one-and two-sample statistics. *The Annals of Mathematical Statistics*, 38(5):1475–1490.

Aichem, A. and Groettrup, M. (2016). The ubiquitin-like modifier FAT10 in cancer development. *The International Journal of Biochemistry and cell biology*, 79:451–461.

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011). Global testing under sparse alternatives: Anova, multiple comparisons and the higher criticism. *The Annals of Statistics*, 39(5):2533–2556.

Bahadur, R. R. (1967a). An optimal property of the likelihood ratio statistic. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 13–26.

Bahadur, R. R. (1967b). Rates of convergence of estimates and test statistics. *The Annals of Mathematical Statistics*, 38(2):303–324.

Bahadur, R. R. et al. (1960). Stochastic comparison of tests. *The Annals of Mathematical Statistics*, 31(2):276–295.

Baraud, Y. (2002). Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, pages 577–606.

Barnett, I., Mukherjee, R., and Lin, X. (2017). The generalized higher criticism for testing snp-set effects in genetic association studies. *Journal of the American Statistical Association*, 112(517):64–76.

Barnett, I. J. and Lin, X. (2014). Analytical p-value calculation for the higher criticism test in finite-d problems. *Biometrika*, 101(4):964–970.

Begum, F., Ghosh, D., Tseng, G. C., and Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic acids research*, 40(9):3777–3784.

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.

Berk, R. H. and Jones, D. H. (1978). Relatively optimal combinations of test statistics. *Scandinavian Journal of Statistics*, pages 158–162.

Berk, R. H. and Jones, D. H. (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Probability Theory and Related Fields*, 47(1):47–59.

Birgé, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series*, pages 113–133.

Brown, M. B. (1975). 400: A method for combining non-independent, one-sided tests of significance. *Biometrics*, pages 987–992.

Cai, T. T., Liu, W., and Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 349–372.

Canaan, A., DeFuria, J., Perelman, E., Schultz, V., Seay, M., Tuck, D., Flavell, R. A., Snyder, M. P., Obin, M. S., and Weissman, S. M. (2014). Extended lifespan and reduced adiposity in mice lacking the FAT10 gene. *Proceedings of the National Academy of Sciences*, 111(14):5313–5318.

Canfield, C.-A. and Bradshaw, P. C. (2019). Amino acids in the regulation of aging and aging-related diseases. *Translational Medicine of Aging*, 3:70–89.

Cavalier, L. and Tsybakov, A. (2002). Sharp adaptation for inverse problems with random noise. *Probability Theory and Related Fields*, 123(3):323–354.

Chen, Y., Liu, P., Tan, K. S., and Wang, R. (2021). Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Statistica Sinica*. Also available as "https://www3.stat.sinica.edu.tw/preprint/SS-2021-0071_Preprint.pdf".

Chen, Y. and Yuen, K. C. (2009). Sums of pairwise quasi-asymptotically independent random variables with consistent variation. *Stochastic Models*, 25(1):76–89.

Chen, Z., Yang, W., Liu, Q., Yang, J. Y., Li, J., and Yang, M. Q. (2014). A new statistical approach to combining p-values using gamma distribution and its application to genome-wide association study. *BMC Bioinformatics*, 15:1–7.

Collier, O., Comminges, L., Tsybakov, A. B., et al. (2017). Minimax estimation of linear and quadratic functionals on sparsity classes. *The Annals of Statistics*, 45(3):923–958.

Davis, R. A. (1983). Stable limits for partial sums of dependent random variables. *The Annals of Probability*, pages 262–269.

De Boer, P.-T., Kroese, D. P., Mannor, S., and Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research*, 134(1):19–67.

Delaigle, A., Hall, P., and Jin, J. (2011). Robustness and accuracy of methods for high dimensional data analysis based on Student's t-statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):283–301.

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

Dudbridge, F. and Koeleman, B. P. (2003). Rank truncated product of p-values, with application to genomewide association scans. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 25(4):360–366.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology*, 80(2):351–363.

Erden-İnal, M., Sunal, E., and Kanbak, G. (2002). Age-related changes in the glutathione redox system. *Cell Biochemistry and Function*, 20(1):61–66.

Fang, Y., Chang, C., and Tseng, G. C. (2023+a). Heavy-tailed distribution for combining dependent p-values with asymptotic robustness. *Statistica Sinica*. Also available as "https://www3.stat.sinica.edu.tw/ss_newpaper/SS-2022-0046_na.pdf".

Fang, Y., Chang, C., and Tseng, G. C. (2023+b). On p-value combination of independent and non-sparse signals: asymptotic efficiency and Fisher ensemble. *Statistica Sinica*. Also available as "https://www3.stat.sinica.edu.tw/ss_newpaper/SS-2022-0261_na.pdf".

Farrah, T. E., Dhillon, B., Keane, P. A., Webb, D. J., and Dhaun, N. (2020). The eye, the kidney, and cardiovascular disease: old concepts, better tools, and new horizons. *Kidney International*, 98(2):323–342.

Fisher, R. A. (1992). Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer.

Goeman, J. J., Rosenblatt, J. D., and Nichols, T. E. (2019). The harmonic mean p-value: Strong versus weak control, and the assumption of independence. *Proceedings of the National Academy of Sciences of the United States of America*, 116(47):23382.

Goldenshluger, A. and Tsybakov, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *The Annals of Statistics*, 29(6):1601–1619.

Goldie, C. M. and Klüppelberg, C. (1998). Subexponential distributions. *A practical Guide to Heavy Tails: Statistical Techniques and Applications*, pages 435–459.

Grossman, C. J. (1985). Interactions between the gonadal steroids and the immune system. *Science*, 227(4684):257–261.

Guerra, R. and Goldstein, D. R. (2016). *Meta-analysis and combining information in genetics and genomics*. Chapman and Hall/CRC.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *The Annals of Statistics*, 38(3):1686–1732.

He, Y., Hooker, E., Yu, E.-J., Wu, H., Cunha, G. R., and Sun, Z. (2018). An indispensable role of androgen receptor in wnt responsive cells during prostate development, maturation, and regeneration. *Stem Cells*, 36(6):891–902.

Heard, N. (2021). Standardized partial sums and products of p-values. *Journal of Computational and Graphical Statistics*, (just-accepted):1–22.

Heard, N. A. and Rubin-Delanchy, P. (2018). Choosing between methods of combining p-values. *Biometrika*, 105(1):239–246.

Hoh, J., Wille, A., and Ott, J. (2001). Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Research*, 11(12):2115–2119.

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.

Huo, Z., Tang, S., Park, Y., and Tseng, G. (2020). P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. *Bioinformatics*, 36(2):524–532.

Jager, L. and Wellner, J. A. (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics*, 35(5):2018–2053.

Janková, J., Shah, R. D., Bühlmann, P., and Samworth, R. J. (2020). Goodness-of-fit testing in high dimensional generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(3):773–795.

Jorde, L. B. and Wooding, S. P. (2004). Genetic variation, classification and "race". *Nature Genetics*, 36(11):S28–S33.

Kanehisa, M. and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Karamata, J. (1933). Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, 61:55–62.

Knight, J. and Nigam, Y. (2017). Anatomy and physiology of ageing 5: the nervous system. *Nursing Times*, 113(6):55–58.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent p-values. *Statistics & Probability Letters*, 60(2):183–190.

Krämer, A., Green, J., Pollard Jr, J., and Tugendreich, S. (2014). Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics*, 30(4):523–530.

Kuo, C.-L. and Zaykin, D. V. (2011). Novel rank-based approaches for discovery and replication in genome-wide association studies. *Genetics*, 189(1):329–340.

Landfield, P. W., Waymire, J., and Lynch, G. (1978). Hippocampal aging and adrenocorticoids: quantitative correlations. *Science*, 202(4372):1098–1102.

LeCam, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pages 38–53.

Leposavić, G. M. and Pilipović, I. M. (2018). Intrinsic and extrinsic thymic adrenergic networks: sex steroid-dependent plasticity. *Frontiers in Endocrinology*, 9:13.

Li, J. and Siegmund, D. (2015). Higher criticism: p-values and criticism. *The Annals of Statistics*, 43(3):1323–1350.

Li, J. and Tseng, G. C. (2011). An adaptively weighted statistic for detecting differential gene expression when combining multiple transcriptomic studies. *The Annals of Applied Statistics*, 5(2A):994–1019.

Li, Q., Hu, J., Ding, J., and Zheng, G. (2014). Fisher's method of combining dependent statistics using generalizations of the gamma distribution with applications to genetic pleiotropic associations. *Biostatistics*, 15(2):284–295.

Lipták, T. (1958). On the combination of independent tests. *Magyar Tud Akad Mat Kutato Int Kozl*, 3:171–197.

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association*, 66(336):802–806.

Littell, R. C. and Folks, J. L. (1973). Asymptotic optimality of Fisher's method of combining independent tests II. *Journal of the American Statistical Association*, 68(341):193–194.

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019). ACAT: a fast and powerful p-value combination method for rare-variant analysis in sequencing studies. *The American Journal of Human Genetics*, 104(3):410–421.

Liu, Y. and Xie, J. (2019). Accurate and efficient p-value calculation via Gaussian approximation: a novel monte-carlo method. *Journal of the American Statistical Association*, 114(525):384–392.

Liu, Y. and Xie, J. (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association*, 115(529):393–402.

Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21.

Ma, T., Trinh, M. A., Wexler, A. J., Bourbon, C., Gatti, E., Pierre, P., Cavener, D. R., and Klann, E. (2013). Suppression of eIF2α kinases alleviates Alzheimer's disease–related plasticity and memory deficits. *Nature Neuroscience*, 16(9):1299–1305.

Maleki, F., Ovens, K., Hogan, D. J., and Kusalik, A. J. (2020). Gene set analysis: challenges, opportunities, and future research. *Frontiers in Genetics*, 11:654.

Mandal, P. K., Saharan, S., Tripathi, M., and Murari, G. (2015). Brain glutathione levels–a novel biomarker for mild cognitive impairment and Alzheimer's disease. *Biological Psychiatry*, 78(10):702–710.

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753.

Marees, A. T., de Kluiver, H., Stringer, S., Vorspan, F., Curis, E., Marie-Claire, C., and Derks, E. M. (2018). A tutorial on conducting genome-wide association studies: Quality control and statistical analysis. *International Journal of Methods in Psychiatric Research*, 27(2):e1608.

Maynard, S., Schurman, S. H., Harboe, C., de Souza-Pinto, N. C., and Bohr, V. A. (2009). Base excision repair of oxidative DNA damage and association with cancer and aging. *Carcinogenesis*, 30(1):2–10.

Mazurkiewicz-Kwilecki, I. and Nsonwah, S. (1989). Changes in the regional brain histamine and histidine levels in postmortem brains of Alzheimer patients. *Canadian Journal of Physiology and Pharmacology*, 67(1):75–78.

Meynial-Denis, D. (2016). Glutamine metabolism in advanced age. *Nutrition Reviews*, 74(4):225–236.

Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*, volume 99. Eindhoven University of Technology Eindhoven, The Netherlands.

Mori, H., Cardiff, R. D., and Borowsky, A. D. (2018). Aging mouse models reveal complex tumor-microenvironment interactions in cancer progression. *Frontiers in Cell and Developmental Biology*, 6:35.

Mosteller, F. and Bush, R. R. (1954). *Selected quantitative techniques*. Addison-Wesley.

Mudholkar, G. S. and George, E. O. (1979). The logit method for combining probabilities. In *Symposium on Optimizing Methods in Statistics*, pages 345–366. Academic Press New York.

Nagaraja, H. N. (2006). Order statistics from independent exponential random variables and the sum of the top order statistics. In *Advances in Distribution Theory, Order Statistics, and Inference*, pages 173–185. Springer.

Normand, S.-L. T. (1999). Meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, 18(3):321–359.

Okbay, A., Baselmans, B. M., De Neve, J.-E., Turley, P., Nivard, M. G., Fontana, M. A., Meddens, S. F. W., Linnér, R. K., Rietveld, C. A., Derringer, J., et al. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics*, 48(6):624–633.

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

Parrish, A. R. (2017). The impact of aging on epithelial barriers. *Tissue Barriers*, 5(4):e1343172.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, pages 379–410.

Poole, W., Gibbs, D. L., Shmulevich, I., Bernard, B., and Knijnenburg, T. A. (2016). Combining dependent p-values with an empirical adaptation of Brown's method. *Bioinformatics*, 32(17):i430–i436.

Porcellini, E., Calabrese, E., Guerini, F., Govoni, M., Chiappelli, M., Tumini, E., Morgan, K., Chappell, S., Kalsheker, N., Franceschi, M., et al. (2007). The hydroxy-methyl-glutaryl CoA reductase promoter polymorphism is associated with Alzheimer's risk and cognitive deterioration. *Neuroscience Letters*, 416(1):66–70.

Raghavachari, N. and Garcia-Reyero, N. (2018). Overview of gene expression analysis: transcriptomics. In *Gene Expression Analysis*, pages 1–6. Springer.

Ren, W.-Y., Wu, K.-F., Li, X., Luo, M., Liu, H. C., Zhang, S. C., and Hu, Y. (2014). Age-related changes in small intestinal mucosa epithelium architecture and epithelial tight junction in rat models. *Aging Clinical and Experimental Research*, 26(2):183–191.

Rey, R. A. (2021). The role of androgen signaling in male sexual development at puberty. *Endocrinology*, 162(2).

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

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *The Annals of Mathematical Statistics*, pages 220–238.

Salcedo, C., Andersen, J. V., Vinten, K. T., Pinborg, L. H., Waagepetersen, H. S., Freude, K. K., and Aldana, B. I. (2021). Functional metabolic mapping reveals highly active branched-chain amino acid metabolism in human astrocytes, which is impaired in iPSC-derived astrocytes in Alzheimer's disease. *Frontiers in Aging Neuroscience*, 13.

Sandiford, O. A., Moore, C. A., Du, J., Boulad, M., Gergues, M., Eltouky, H., and Rameshwar, P. (2018). Human aging and cancer: role of miRNA in tumor microenvironment. *Exosomes, Stem Cells and MicroRNA*, pages 137–152.

Savage, I. R. (1969). Nonparametric statistics: a personal review. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 107–144.

Schumacher, B., Van Der Pluijm, I., Moorhouse, M. J., Kosteas, T., Robinson, A. R., Suh, Y., Breit, T. M., Van Steeg, H., Niedernhofer, L. J., Van Ijcken, W., et al. (2008). Delayed and accelerated aging share common longevity assurance mechanisms. *PLoS Genetics*, 4(8):e1000161.

Shabalin, A. A., Tjelmeland, H., Fan, C., Perou, C. M., and Nobel, A. B. (2008). Merging two gene-expression studies via cross-platform normalization. *Bioinformatics*, 24(9):1154–1160.

Shah, R. D. and Bühlmann, P. (2018). Goodness-of-fit tests for high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(1):113–135.

Shao, Q.-M., Wang, Q., et al. (2013). Self-normalized limit theorems: A survey. *Probability Surveys*, 10:69–93.

Shintani, M. and Umeno, K. (2018). Super generalized central limit theorem—limit distributions for sums of non-identical random variables with power laws. *Journal of the Physical Society of Japan*, 87(4):043003.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, pages 397–420. Springer.

Song, C., Min, X., and Zhang, H. (2016). The screening and ranking algorithm for change-points detection in multiple samples. *The Annals of Applied Statistics*, 10(4):2102.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100(16):9440–9445.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A., and Williams Jr, R. M. (1949). The American soldier: Adjustment during army life. *Studies in Social Psychology in World War II*, 5.

Su, Y.-C., Gauderman, W. J., Berhane, K., and Lewinger, J. P. (2016). Adaptive set-based methods for association testing. *Genetic Epidemiology*, 40(2):113–122.

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., et al. (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.

Sun, R. and Lin, X. (2020). Genetic variant set-based tests using the generalized berk–jones statistic with application to a genome-wide association study of breast cancer. *Journal of the American Statistical Association*, 115(531):1079–1091.

Svishcheva, G. R., Belonogova, N. M., Zorkoltseva, I. V., Kirichenko, A. V., and Axenovich, T. I. (2019). Gene-based association tests using gwas summary statistics. *Bioinformatics*, 35(19):3701–3708.

Terao, A., Steininger, T. L., Morairty, S. R., and Kilduff, T. S. (2004). Age-related changes in histamine receptor mRNA levels in the mouse brain. *Neuroscience Letters*, 355(1-2):81–84.

Tippett, L. H. C. et al. (1931). The methods of statistics. *The Methods of Statistics*.

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

Tsybakov, A. B. (2008). *Introduction to nonparametric estimation*. Springer Science & Business Media.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121.

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

Van der Goot, A. T. and Nollen, E. A. (2013). Tryptophan metabolism: entering the field of aging and age-related pathologies. *Trends in Molecular Medicine*, 19(6):336–344.

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press.

Vieira, O. V., Botelho, R. J., and Grinstein, S. (2002). Phagosome maturation: aging gracefully. *Biochemical Journal*, 366(3):689–704.

Visscher, P. M., Brown, M. A., McCarthy, M. I., and Yang, J. (2012). Five years of gwas discovery. *The American Journal of Human Genetics*, 90(1):7–24.

Vovk, V., Wang, B., and Wang, R. (2021). Admissible ways of merging p-values under arbitrary dependence. *Annals of Statistics*.

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

Willink, R. (2005). Bounds on the bivariate normal distribution function. *Communications in Statistics-Theory and Methods*, 33(10):2281–2297.

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

Wilson, D. J. (2019b). Reply to Held: When is a harmonic mean p-value a Bayes factor? *Proceedings of the National Academy of Sciences*, 116:5857–5858.

Wilson, D. J. (2020). Generalized mean p-values for combining dependent tests: comparison of generalized central limit theorem and robust risk analysis. *Wellcome Open Research*, 5.

Won, S., Morris, N., Lu, Q., and Elston, R. C. (2009). Choosing an optimal method to combine p-values. *Statistics in Medicine*, 28(11):1537–1553.

Xu, G., Lin, L., Wei, P., and Pan, W. (2016). An adaptive two-sample test for high-dimensional means. *Biometrika*, 103(3):609–624.

Yang, X., Kui, L., Tang, M., Li, D., Wei, K., Chen, W., Miao, J., and Dong, Y. (2020). High-throughput transcriptome profiling in drug and biomarker discovery. *Frontiers in Genetics*, 11:19.

Yu, K., Li, Q., Bergen, A. W., Pfeiffer, R. M., Rosenberg, P. S., Caporaso, N., Kraft, P., and Chatterjee, N. (2009). Pathway analysis by adaptive combination of p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 33(8):700–709.

Zahn, J. M., Poosala, S., Owen, A. B., Ingram, D. K., Lustig, A., Carter, A., Weeraratna, A. T., Taub, D. D., Gorospe, M., Mazan-Mamczarz, K., et al. (2007). AGEMAP: a gene expression database for aging in mice. *PLoS Genetics*, 3(11):e201.

Zaykin, D. V. (2011). Optimally weighted z-test is a powerful method for combining probabilities in meta-analysis. *Journal of Evolutionary Biology*, 24(8):1836–1841.

Zaykin, D. V., Zhivotovsky, L. A., Czika, W., Shao, S., and Wolfinger, R. D. (2007). Combining p-values in large-scale genomics experiments. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry*, 6(3):217–226.

Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H., and Weir, B. S. (2002). Truncated product method for combining p-values. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 22(2):170–185.

Zhang, H., Jin, J., and Wu, Z. (2020a). Distributions and power of optimal signal-detection statistics in finite case. *IEEE Transactions on Signal Processing*, 68:1021–1033.

Zhang, H., Tong, T., Landers, J., and Wu, Z. (2020b). TFisher: A powerful truncation and weighting procedure for combining p-values. *The Annals of Applied Statistics*, 14(1):178–201.

Zhang, H. and Wu, Z. (2022). The generalized Fisher's combination and accurate p-value calculation under dependence. *Biometrics*.

Zhou, Y., Bolton, E. C., and Jones, J. O. (2015). Androgens and androgen receptor signaling in prostate tumorigenesis. *Journal of Molecular Endocrinology*, 54(1):R15–R29.