2023

ADDITIVE P-VALUE COMBINATION TEST

Xing Ling
Michigan Technological University, xling@mtu.edu

ADDITIVE P-VALUE COMBINATION TEST

By

Xing Ling

A DISSERTATION

Submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

In Statistics

MICHIGAN TECHNOLOGICAL UNIVERSITY

2023

This dissertation has been approved in partial fulfillment of the requirements for the Degree of DOCTOR OF PHILOSOPHY in Statistics.

Department of Mathematical Sciences

Dissertation Advisor:	*Dr. David Hemmer*
Committee Member:	*Dr. Iosif Pinelis*
Committee Member:	*Dr. Qiuying Sha*
Committee Member:	*Dr. Kui Zhang*
Committee Member:	*Dr. Chee-Wooi Ten*
Department Chair:	*Dr. Jiguang Sun*

Contents

List of Figures

List of Tables

Preface

This dissertation is submitted for the degree of Doctor of Philosophy at Michigan Technological University. It contains published, completed papers, some important preparations, and achievements for future publications completed by the author. The research is to the best of my knowledge and original, except where references are made to previous work. Part of this work contains previously published material. Xing Ling led all of the work under the supervision of Dr. Yeonwoo Rho from the Department of Mathematical Sciences at Michigan Technological University.

The first chapter *Stable Combination Test* was published in Statistica Sinica in 2022. Xing Ling and Dr. Yeonwoo Rho designed the research. Xing Ling developed the `R` codes and performed statistical analyses. Xing Ling and Dr. Yeonwoo Rho wrote the manuscript. The second chapter *Unifying Additive p-value Combination Tests with Regularly Varying Tails* is in preparation for future publication. Xing Ling and Dr. Yeonwoo Rho designed the research. Xing Ling developed the `R` codes and performed statistical analyses. Xing Ling and Dr. Yeonwoo Rho wrote the manuscript. Dr. Sangyoon J. Han from the Department of Biomedical Engineering of Michigan Technological University provided the dataset for the case study and wrote the data description. The third chapter *A simple remedy for the multiplicity problem in rolling window Granger causality tests* is in preparation for future publication. Xing Ling and Dr. Yeonwoo Rho designed the research. Xing Ling developed the `R` codes, performed statistical analyses, and wrote the manuscript. Dr. Yeonwoo Rho provided comments for editing the manuscript.

Acknowledgments

At Michigan Technological University, I have obtained valuable experience in doing scientific research by using the tool of mathematics and statistics and also gained some practical experience in teaching. I would like to express my sincerest gratitude to the Department of Mathematical Science for giving me the opportunity to pursue a graduate degree. I would like to express my deepest gratitude to all the people who have helped me, including but not limited to my academic advisor and course instructors.

I would like to express my sincerest thanks and appreciation to the committee members, Prof. David Hemmer, Prof. Iosif Pinelis, Prof. Qiuying Sha, Prof. Kui Zhang, and Prof. Chee-Wooi Ten for their support and valuable advice. It is a great honor to have these outstanding scholars be part of the committee.

I would like to show my gratitude to Prof. Yeonwoo Rho for advising my previous work and teaching me a lot in my career. I always admire her enthusiasm for theoretical research work. Her passion and knowledge have guided me the most in my Ph.D. study. Without Prof. Yeonwoo Rho's guidance, I cannot imagine I can make the work presented in this dissertation. I would like to thank Prof. Sangyoon J. Han for collaborating with me and providing appropriate help.

I would like to thank Xuewei Cao who brought the papers of Liu and Xie [2020], Li et al. [2011], and Van der Sluis et al. [2013] to my attention. Thanks to Dr. Stan Pounds, Dr. John P. Nolan, and the participants of the 2022 Joint Statistical Meeting for giving me valuable advice when I presented the work of the third chapter at a poster presentation.

I would like to show my gratitude to the department chair Prof. Jiguang Sun, the graduate program director Prof. Melissa Keranen, the University Ombuds Prof. Susanna D. Peters JD, and Prof. Debra Charlesworth from the Graduate School for their valuable support and conflict resolutions. I'd also thank Prof. Qinghui Cheng,

Prof. Zhenlin Wang, and Prof. Ruihong Zhang for giving me advice and help.

List of Abbreviations

CCT	Cauchy Combination Test
FCT	Fréchet Combination Test
FDR	False Discovery Rate
FWER	Family-Wise Error Rate
GCLT	Generalized Central Limit Theorem
GMP	Generalized Mean p-value
HMP	Harmonic Mean p-value
IDEMV	Infections Disease Equity Market Volatility Tracker
OLS	Ordinary Least Square
SCT	Stable Combination Test
UCT	Unified Combination Test

Abstract

This dissertation includes four Chapters. A brief description of each chapter is organized as follows.

In Chapter 1, some developments on multiple hypotheses tests are introduced. Some preliminaries about the definition and the assumption are included.

In Chapter 2, a Stable Combination Test is proposed to combine p-values from multiple hypotheses tests. We show the proposed method controls the family-wise error rate at the target level and maintains asymptotically optimal power even when the elementary p-values from the individual hypotheses are dependent.

In Chapter 3, a deeper dig into additive p-value combination test is performed. A common idea behind some existing combination tests including the Stable Combination Test is extracted and a unified framework is proposed. The tails of the combined test statistics in this framework can be approximated by stable distribution. The tests in this framework are proven to have a well-controlled family-wise error rate and non-trivial power.

In Chapter 4, we illustrate the usefulness of the proposed unified framework by capturing the dynamic structure instabilities of Granger causality in a vector autoregression model. The p-value combination tests in the framework are easy to implement, robust to dependence, and have comparable performance to the bootstrap technique.

Chapter 1

Introduction

Multiplicity has been a long-standing issue in areas of applied statistics, such as clinical trials, DNA microarray experiments, and functional magnetic resonance imaging studies, where testing a large number of multiple hypotheses is necessary. Achieving higher power, while controlling error rate, has been one of the most important missions in the multiplicity area. There are two widely used approaches to control the overall type I error rate. One is the family-wise error rate (FWER) approach, where the probability of making at least one type I error among all hypotheses is controlled. The other is the false discovery rate (FDR) approach, where the expected proportion of false discoveries out of all rejections is controlled. FDR-controlling procedures usually have higher powers at the cost of increments of size than FWER-controlling procedures because the FDR is equivalent to the FWER when all underlying hypotheses are true but smaller otherwise [Benjamini and Hochberg, 1995]. Many methods have been proposed under either FWER or FDR frameworks. In this dissertation, we mainly focus on controlling the FWER of a global hypothesis.

Let $H_1, \ldots, H_n$ be individual null hypotheses and the corresponding alternative hypotheses be $H_1^c, \ldots, H_n^c$. Suppose the corresponding individual p-values are $p_1, \ldots, p_n$. The global null hypothesis, defined as $\boldsymbol{H_0} = \bigcap_{i=1}^n H_i$, is true when all H_is are true. The global alternative hypothesis, defined as $\boldsymbol{H_a} = \bigcup_{i=1}^n H_i^c$, is true if there is at least one false null hypothesis.

A simple and widely used approach to control the FWER is the Bonferroni procedure, where the global null hypothesis is rejected if $\min p_i n$ is less than the significance level. However, Bonferroni is so conservative that it has been criticized for its low power when a large number of tests are undertaken or tests are strongly positively correlated. See O'Brien [1984], Moran [2003], Dmitrienko et al. [2009] among others. Simes [1986] improved the Bonferroni procedure in the sense of making use of the ordered information of the underlying p-values instead of just the smallest p-value. The combined p-value is defined as $\min p_{(i)} n/i$, where $p_{(i)}$ is the ith ordered underlying p-value. Simes [1986] proved that the size is exactly the significance level when the underlying tests are independent. It is noticeable that Simes' procedure controls the FWER in a weak sense under the global null hypothesis, thus it cannot be directly used to test the individual null hypotheses, whereas the Bonferroni procedure can control the FWER in the strong sense regardless of the composition of the true and false underlying hypotheses [Zhang et al., 2013]. In order to provide statements of individual hypotheses, Hochberg [1988] proposed a step-up version of the Simes' test which have strong control of the FWER.

Many methods are proposed under the condition of independent underlying p-values. However, in practical applications dependency among the underlying individual tests is another key element to consider. Most of the above-mentioned methods often suffer from low powers and size distortion as a consequence of the fact that many heavily dependent underlying tests carry similar information as the fewer "effective" tests [Meinshausen et al., 2011]. Bootstrap, permutation, or other resampling techniques have been used to provide correct sizes since they don't make specific assumptions about the true joint distribution. For example, Westfall and Young [1993] proposed two step-wise permutation-based procedures, "min-P" and "max-T", to estimate the adjusted p-value and thus control the FWER strongly without modeling the dependence among underlying individual tests. However, it involved an extra condition to control the FWER in the strong sense, called "subset pivotality", to greatly reduce the number of intersection hypotheses in the closed family. The subset pivotality condition assumes that the joint distribution of the p-values under the global null hypothesis is identical to the joint distribution of the p-values under any subsets of the global null hypothesis [Westfall and Young, 1993]. This condition is been portrayed as too stringent when the number of hypotheses is large [Romano et al., 2008]. Furthermore, Rempala and Yang [2013] argued that the permutation distribution depends on data. When the underlying hypotheses about the data are not all true, the permutation distribution may have little in common with the true state of the data, thus the size is not necessarily properly controlled. Moreover, resampling-based methods are computationally burdensome in the analysis of massive data. The improvements might be too minor compared with the additional computation cost [Ventura et al., 2004].

In recent years, computationally efficient methods with higher powers have been proposed. Inspired by Simes [1986], Li et al. [2011] and Van der Sluis et al. [2013] extended Simes' procedure to address the correlations among individual tests. They estimated the effective number of independent p-values by considering the eigenvalues of the correlation matrix of the p-values. Since the correlation matrix of p-values is unobservable, these methods propose to take advantage of external information, like the correlation structure of phenotypes. However, these methods cannot be used if such external information is not available.

The aforementioned methods can be understood as variants of the ordered p-value method, where the combined p-value is based on the rank information of individual p-values. There is another approach of combining p-values based on additive methods, where the magnitude of p-values is important rather than the rank information [Henning and Westfall, 2015]. There are many methods associated with additive combinations. For example, Fisher's combination test [Fisher, 1992] stated that $-2\sum_{i=1}^{n}\log p_i$ follows a χ^2_{2n} distribution. Stouffer's Z-score method [Stouffer et al., 1949] stated that

$1/\sqrt{n}\sum_{i=1}^{n}\Phi^{-1}(1-p_i)$ follows a standard normal distribution. Rüschendorf [1982] showed that the arithmetic mean of p-values should be adjusted by a constant factor of two. Mattner [2010] claimed that the geometric mean should be multiplied by the mathematical constant, e. Vovk and Wang [2020] define a so-called merging function as the generalized mean p-value adjusted by a factor.

This dissertation focuses on additive p-value combination tests when the underlying p-values are allowed to be dependent. Additive combination tests utilize the fact that a p-value under a null hypothesis is uniformly distributed. Each individual p-value is transformed into a new random variable, which is then further linearly combined to form a combined test statistic. The function used to transform p-values in the first step characterizes the method. One of the key steps associated with a combined test statistic is figuring out its null distribution. This step is simpler if the transformation function is chosen in a way that the distribution of the transformed p-values under the null is closed under addition. For example, if the ith null hypothesis is true, p_i would more or less follow a uniform distribution, and $-2\log p_i$ follows a gamma distribution with shape parameter one and scale parameter two. If tests are independent, Fisher's statistic, $-2\sum_{i=1}^{n}\log p_i$, follows a gamma distribution with shape parameter n and scale parameter two. Other examples of distributions that are closed under addition include Pareto distribution utilized in Wilson [2019, 2020], Loggamma distribution used by Wilson [2019], Cauchy distribution investigated by Liu and Xie [2020], and stable distribution.

Stable distribution is a family of distributions that are closed under linear combination. A distribution is said to be *stable* if a linear combination of two independent and identical distributed random variables has the same type of distribution up to location and scale parameters [Nolan, 2020]. There are multiple parameterizations for stable laws and we follow Nolan [2020]'s 1-parameterization. According Definition 1.5 in Nolan [2020], a random variable W is $\boldsymbol{S}(\alpha, \beta, \gamma, \delta)$ if W has a characteristic function

$$E[\exp{(iuW)}] = \begin{cases} \exp\left\{-\gamma^\alpha |u|^\alpha \left[1 - i\beta \tan(\frac{\pi\alpha}{2})(\operatorname{sign} u)\right] + i\delta u\right\} & \alpha \neq 1 \\ \exp\left\{-\gamma |u| \left[1 + i\beta \frac{2}{\pi}(\operatorname{sign} u) \log|u|\right] + i\delta u\right\} & \alpha = 1, \end{cases}$$

where $u \in (-\infty, \infty)$, $i = \sqrt{-1}$, and $\operatorname{sign} u$ is the sign function which takes value -1 if $u < 0$, 0 if $u = 0$, and 1 if $u > 0$. There are four parameters: an index of stability $\alpha \in (0, 2]$, a skewness parameter $\beta \in [-1, 1]$, a scale parameter $\gamma > 0$, and a location parameter $\delta \in (-\infty, \infty)$. When the distribution is standardized, i.e., when scale $\gamma = 1$, and location $\delta = 0$, the symbol $\boldsymbol{S}(\alpha, \beta)$ is an abbreviation for $\boldsymbol{S}(\alpha, \beta, 1, 0)$. We write $F(x|\alpha, \beta, \gamma, \delta) = \Pr(W < x)$, the distribution function of a stable random variable W with parameters α, β, γ, and δ. Similarly, when the distribution is standardized, $F(x|\alpha, \beta)$ is short for $F(x|\alpha, \beta, 1, 0)$.

In order to adapt the additive p-value combination tests, we assume the following assumption throughout this dissertation,

Assumption 1.1. *Under the global null hypothesis $\boldsymbol{H_0}$, p_is are uniformly distributed for all $i = 1, \ldots, n$ on $(0, 1)$.*

Remark 1.1. Assumption 1.1 is satisfied if the individual tests are exact tests with continuous test statistics. If individual tests are based on asymptotic results or on discrete statistics, the sample sizes should be large enough to satisfy this assumption approximately. However, this assumption can be relaxed to some non-uniform p-values as long as individual tests are more conservative than the nominal level. In this case, a combined p-value is also conservative, which means that the combined p-value can control the size correctly. See Remark 5 and Corollary 2 in Liu and Xie [2020] for more details. For brevity, the rest of this dissertation assumes uniform p-values under the null.

Throughout this dissertation, we write $g(x) \sim h(x)$ as $x \to \infty$ to indicate $\lim_{x\to\infty} \frac{g(x)}{h(x)} = 1$. The symbol $\mathbb{R}$ indicates the set of all real numbers. The symbol $A \setminus B$ indicates $A \cap B^C$. For instance, $\mathbb{R} \setminus \{0\}$ is the set of all real numbers except 0. $\Gamma(\cdot)$ is the gamma function.

Chapter 2

Stable Combination Test

Abstract

In this chapter, a stable combination test is proposed as a natural extension of Cauchy combination tests by Liu and Xie [2020]. Similarly to the Cauchy combination test, the stable combination test is simple to compute, enjoys good sizes, and has asymptotically optimal powers even when the individual tests are not independent. This finding is supported both in theory and in finite samples.

Keywords: Additive combination test; multiple hypothesis testing; stable distribution.

2.1 Introduction

Liu and Xie [2020] proposed the Cauchy combination test (CCT) which originated from the observation that the standard Cauchy distribution is closed under convex combinations. What makes a Cauchy distribution an attractive candidate for a combination test is that this relationship holds even when the random variables are dependent [Pillai and Meng, 2016]. In CCT, the individual p-values are transformed into standard Cauchy random variables, and a convex combination of these Cauchy random variables is used as a combined test statistic. The critical values are taken from a standard Cauchy distribution. Liu and Xie [2020] proved that the CCT controls the size well if the significance level is small and has asymptotically optimal powers under sparse alternatives. The CCT is fast to compute and robust to various forms of dependence structures.

This chapter is motivated by the fact that there is a wide class of distributions that is closed under addition – strictly stable distributions. In fact, the Cauchy distribution is also a part of the strictly stable distribution family. This observation naturally leads to a stable combination test (SCT), which is an extension of the CCT. In the SCT, a stable distribution function is used in the transformation step. It is well-known that a linear combination of independent stably distributed random variables is still stable. In this chapter, we show that the SCT statistic is also stably distributed asymptotically even when the underlying p-values are dependent, and therefore, can control the error rate successfully when the number of tests, n, is large enough. We also prove that the SCT has asymptotically optimal powers under sparse alternatives as long as n is large enough. Our simulation results also suggest that our method is robust to dependent structures in finite samples.

This chapter is organized as follows. In Section 2.2, we summarize the CCT and introduce our SCT method. The sizes and powers of the SCT are explored in theory. In Section 2.3, simulation results are provided to demonstrate favorable sizes and powers of the SCT in finite samples. Section 2.4 concludes. Technical proofs are relegated to Appendix A.

2.2 Method

Inspired by Pillai and Meng [2016]'s finding that a sum of dependent Cauchy random variables is still a Cauchy random variable, Liu and Xie [2020] proposed to combine p-values based on a Cauchy distribution. The p-values are first transformed into standard Cauchy random variables and then a weighted sum is taken. The test statistic is defined as the weighted sum

$$T_n(\boldsymbol{p}) = \sum_{i=1}^{n} w_i \tan[\pi(0.5 - p_i)],$$

where w_is are nonnegative weights and $\sum_{i=1}^{n} w_i = 1$. If the individual p-values are independent or perfectly dependent, it is straightforward that the test statistic follows the standard Cauchy distribution under the global null hypothesis. One of the main contributions of Liu and Xie [2020] is to relax this condition. The tail probability of the test statistic $T_n(\boldsymbol{p})$ is approximately the same as that of a standard Cauchy distribution when p_is satisfy bivariate Gaussian copula condition. In order to prove the above statement, Liu and Xie [2020] assumed the p-values are calculated from Z-tests. That is, $p_i = 2[1 - \Phi(|X_i|)]$ from the ith two-sided Z-test, with X_i follows $N(\mu_i, 1)$, $\mathrm{E}[(X_1, \ldots, X_n)^T] = (\mu_1, \ldots, \mu_n)^T = \boldsymbol{\mu}$ and $\mathrm{Cov}[(X_1, \ldots, X_n)^T] = \Sigma$. They also assume that $(X_i, X_j)^T$ for $i \neq j$ are pairwise bivariate normally distributed. In this case, the test statistic can be rewritten as

$$T_n(\boldsymbol{p}) = T_n(\boldsymbol{X}) = \sum_{i=1}^{n} w_i \tan\{\pi[2\Phi(|X_i|) - 1.5]\}.$$

They proved that $T_n(\boldsymbol{X})$ has the same tail probability as a standard Cauchy random variable if $\boldsymbol{\mu} = \boldsymbol{0}$. This means that the combined p-value can be derived from a standard Cauchy distribution. Liu and Xie [2020] also proved that if $\boldsymbol{\mu} \neq \boldsymbol{0}$ and if $\boldsymbol{\mu}$ satisfies the sparse alternative assumption with large enough signals, this test has an asymptotically optimal power since $T_n(\boldsymbol{X}) \to \infty$ with probability one.

Inspired by the fact that a Cauchy distribution is a special case of a stable distribution, we propose a Stable Combination Test (SCT). Let $W_i = F^{-1}(1 - p_i|\alpha, \beta)$ for $i = 1, \ldots, n$, where $F(\cdot|\alpha, \beta)$ is the distribution function of a standardized stable random variable with stability parameter α and skewness parameter β. The function F^{-1} indicates the quantiles of F, defined by $F^{-1}(p|\alpha, \beta) = \inf\{x \in \mathbb{R} : F(x|\alpha, \beta) \geq p\}$. Though there are no closed forms for stable distribution functions except for Normal, Cauchy, and Lévy distributions, stable quantiles can still be approximated numerically. We define our test statistic as follows:

$$T_n(\boldsymbol{p}) = a_n \sum_{i=1}^{n} w_i W_i, \tag{2.2.1}$$

where $w_i > 0$ is the nonnegative weight imposed on ith test with $\sum_{i=1}^n w_i = 1$ and $a_n = \left(\sum_{j=1}^n w_j^\alpha\right)^{-1/\alpha}$ is the normalizing factor.

We consider the stability parameters $0 < \alpha < 2$. If $\alpha \neq 1$, the skewness parameter ranges $-1 < \beta \leq 1$. If $\alpha = 1$, only $\beta = 0$ is considered. This is to ensure that $F(\cdot|\alpha, \beta)$ is strictly stable. A distribution is called strictly stable if the sum of i.i.d. random variables from this distribution follows the same distribution up to a normalizing factor without requiring a centering factor. Since $F^{-1}(1 - p_i|\alpha, \beta)$ follows $\boldsymbol{S}(\alpha, \beta)$, $\sum_{i=1}^n w_i F^{-1}(1 - p_i|\alpha, \beta)$ also follows $\boldsymbol{S}(\alpha, \beta)$ up to a normalizing factor if $\boldsymbol{S}(\alpha, \beta)$ is strictly stable. This motivates our definition of $T_n(\boldsymbol{p})$ for $\alpha \neq 1$ with the normalizing factor $\left(\sum_{j=1}^n w_j^\alpha\right)^{-1/\alpha}$. However, when $\alpha = 1$, $\boldsymbol{S}(1, \beta)$ is no longer strictly stable unless $\beta = 0$. When $\alpha = 1$, and $\beta = 0$, $\boldsymbol{S}(1, 0)$ is a standard Cauchy distribution. Note that a naive extension of the CCT to different α and β, $\sum_{i=1}^n w_i F^{-1}(1-p_i|\alpha, \beta)$, would not work without considering the normalizing factor a_n.

Remark 2.1. The test statistic $T_n(\boldsymbol{p})$ can still be defined for $\alpha = 1$ and $\beta \neq 0$ if an additional centering factor $\frac{2}{\pi}\beta \sum_{j=1}^n w_j \log w_j$ is considered. However, this direction will not be elaborated in this paper for the following reasons: (i) it had relatively poor sizes and powers in our unreported simulation, (ii) the requirements for the power proof need to be stronger if this case is included, and (iii) the computation for $F^{-1}(\cdot|1, \beta)$ is unstable if $\beta \neq 0$. For these reasons, we only consider $\beta = 0$ when $\alpha = 1$ for the rest of this dissertation.

The rest of this section addresses that our SCT statistic (2.2.1) is also approximately stably distributed under the global null hypothesis, even when the underlying p-values are not independent. This makes it possible to construct a test that can control the FWER. We also prove that this test has asymptotically optimal powers under alternatives.

2.2.1 Size

Under Assumption 1.1, observe that $1 - p_i$ is uniformly distributed under the global null hypothesis $\boldsymbol{H_0}$ for $1 \leq i \leq n$, therefore, $W_i = F^{-1}(1 - p_i|\alpha, \beta)$ is identically distributed with marginal distribution $\boldsymbol{S}(\alpha, \beta)$. If the individual tests are independent, it is trivial that the test statistic

$$T_n(\boldsymbol{p}) \overset{d}{=} W_0, \tag{2.2.2}$$

where W_0 follows a stable distribution $\boldsymbol{S}(\alpha, \beta)$. This can be seen by simple computations using the property that the sums of α-stable random variables are still α-stable; see Proposition 1.4 and equation (1.7) in Nolan [2020].

However, if W_i are not independent, there is no exact relationship as in (2.2.2). Instead, an asymptotic relationship can be established when the number of tests, n, is large enough. For instance, Jakubowski and Kobus [1989] showed that a dependent sum of stable random variables is also asymptotically stable. In this paper, we adapt Jakubowski and Kobus [1989]'s Theorems 4.1 and 4.2 to establish that $T_n(\boldsymbol{p})$ converges to W_0 under some dependence assumptions in Theorem 2.1 below. Owing to this theorem, type I errors of SCTs can be controlled as long as n is large enough.

The following Assumptions 2.1, 2.2, and 2.3 are adapted from equations (4.4), (4.8), and (4.5) of Jakubowski and Kobus [1989], respectively.

Assumption 2.1. *Let $A \subset \mathbb{R} \setminus \{0\}$ be a finite union of disjoint intervals of the form $(a, b]$ that do not contain 0, and A^c be the complementary set of A. The sequence $\{W_i\}_{i=1}^n$ satisfies*

$$\sup_{1 \leq p < q < r \leq n} \left| \Pr\left(\bigcap_{p \leq i \leq r} (a_n w_i W_i \in A^c) \right) - \right.$$
$$\left. \Pr\left(\bigcap_{p \leq i \leq q} (a_n w_i W_i \in A^c) \right) \Pr\left(\bigcap_{q \leq i \leq r} (a_n w_i W_i \in A^c) \right) \right| \to 0$$

for every A as $n \to \infty$.

Assumption 2.2. *The sequence $\{W_i\}_{i=1}^n$ is ρ-mixing with $\sum_{j=1}^{\infty} \rho(2^j) < +\infty$. A sequence $\{X_i\}_{i=1}^n$ is called ρ-mixing if*

$$\rho(m) = \sup_{1 \leq i \leq j \leq n} \sup \left\{ |corr(f,g)| : f \in \mathcal{L}^2(\mathcal{F}_i^j), g \in \mathcal{L}^2(\mathcal{F}_{j+m}^{\infty}) \right\} \xrightarrow[m \to \infty]{} 0,$$

where $\mathcal{F}_i^j$ is the σ-field generated by $(X_i, \ldots, X_j)$ and $\mathcal{L}^2(\mathcal{F}_i^j)$ be the space of square-integrable, $\mathcal{F}_i^j$-measurable random variables.

Assumption 2.3. *Let $\Delta(r)$ be an arbitrary division of the set $\{1, 2, \ldots, n\}$ into r segments, $0 = m_0 \leq m_1 \leq \cdots \leq m_r = n$. For every $\varepsilon > 0$, the sequence $\{W_i\}_{i=1}^n$ satisfies*

$$\lim_{r \to \infty} \limsup_{n \to \infty} \inf_{\Delta(r)} \sum_{q=1}^{r} \sum_{m_{q-1} < i < j \leq m_q} \Pr\left(a_n w_i |W_i| > \varepsilon, a_n w_j |W_j| > \varepsilon\right) = 0.$$

The first two assumptions, Assumptions 2.1 and 2.2, mainly concern the long-range dependence. Assumption 2.1 basically assumes asymptotic long-range independence and will be used to address the convergence in distribution of our test statistic with $0 < \alpha < 1$. Assumption 2.2 is a ρ-mixing condition, which will be used for $1 \leq \alpha < 2$.

Assumption 2.3 limits the amount of short-range dependence by assuming that large values cannot be clustered in a small segment [Beirlant et al., 2006]. In our setting under the global null, Assumption 2.3 means that at most one W_i can have a large absolute value within a small neighborhood. If W_is are independent, this condition can be easily satisfied. However, this condition may not hold if the short-range dependence is too strong.

Remark 2.2. Note that long-range and short-range dependencies make the best sense either when there is a natural order among the individual tests or when the tests are independent. This situation is not too unusual in practice. For instance, any sequential testing, including testing a sequence of genes, would fall within this category.

Suppose Assumptions 2.1 and 2.3 or 2.2 and 2.3 are satisfied. Define an i.i.d. sequence $\{\tilde{W}_i\}$ that has the same marginal distribution as $\{W_i\}$. Note that $\tilde{T}_n(\boldsymbol{p}) =$

$a_n \sum_{i=1}^n w_i \tilde{W}_i$ follows a $\boldsymbol{S}(\alpha, \beta)$ distribution for any n, using a similar argument as in (2.2.2). By applying Theorems 4.1 and 4.2 of Jakubowski and Kobus [1989], our test statistic converges in distribution to $\boldsymbol{S}(\alpha, \beta)$. This observation leads to the following Theorem 2.1.

Theorem 2.1. *Let W_0 be a random variable that follows $\boldsymbol{S}(\alpha, \beta)$, where $0 < \alpha < 2$ and $-1 \leq \beta \leq 1$. Assume one of the following conditions:*

(i) $0 < \alpha < 1$ and Assumptions 2.1 and 2.3 hold.

(ii) $1 \leq \alpha < 2$ and Assumptions 2.2 and 2.3 hold.

Then, if the global null hypothesis $\boldsymbol{H_0}$ is true,

$$T_n(\boldsymbol{p}) = a_n \sum_{i=1}^n w_i W_i \xrightarrow{d} \boldsymbol{S}(\alpha, \beta)$$

as $n \to \infty$.

Based on Theorem 2.1, the global null hypothesis is rejected at significance level s if $T_n(\boldsymbol{p}) > t_s$, where t_s is the upper s quantile of $\boldsymbol{S}(\alpha, \beta)$.

Remark 2.3. A stable distribution with parameters $\alpha = 1$ and $\beta = 0$ is a Cauchy distribution. In this case, our SCT is equivalent to CCT [Liu and Xie, 2020]; i.e.

$$T_n(\boldsymbol{p}) = \sum_{i=1}^n w_i \tan[\pi(0.5 - p_i)].$$

Liu and Xie [2020]'s method is robust to dependencies among the underlying p-values, similar to ours. While our theorems for the SCT do cover include the CCT, our technical settings are slightly different from those of Liu and Xie [2020]. The first difference lies in the forms of dependencies allowed in assumptions. Assumption C.1 of Liu and Xie [2020] assumed that every pair of test statistics of individual tests is bivariate normal. While the p-values follow a uniform distribution marginally, their pairwise dependencies are modeled through bivariate Gaussian copulas. On the contrary, our assumptions do not require Gaussian copulas. Instead, we control long-range and short-range dependencies. This means that our assumptions require a structure of dependence such as a natural order. Our assumptions also require relatively weaker dependencies, whereas Liu and Xie [2020] did not impose any restrictions on the strength of dependencies. The second difference is that Liu and Xie [2020]'s test is controlled only when the significance level s is small enough, while ours would work for any s. This is because Liu and Xie [2020]'s Theorems 1 and 2 concern the right tail probabilities only. They showed that the right tail probability of the test statistic is approximately the same as the right tail probability of a Cauchy random variable only when the significance level s is small enough. By contrast, the type I error of the SCT can be controlled at any significance level as long as n is large enough. This is because the in-distribution convergence result in our Theorem 2.1 is much stronger than the right tail convergence results in Theorems 1 and 2 in Liu and Xie [2020]. The last difference is that our result holds only when the number of individual tests, n, is large enough, while Liu and Xie [2020]'s Theorem 1 showed that the CCT can control the size also when n is fixed.

Remark 2.4. The form of our SCT statistic also resembles Stouffer's Z-score [Stouffer et al., 1949]. A stable distribution with tail parameter $\alpha = 2$ is a normal distribution no matter what the skewness parameter β is. In this case, our SCT test statistic is equivalent to Stouffer's Z-score; i.e.

$$T_n(\boldsymbol{p}) = \frac{1}{\sqrt{\sum_{j=1}^{n} w_j^2}} \sum_{i=1}^{n} w_i \Phi^{-1}(1 - p_i).$$

Stouffer's Z-score [Stouffer et al., 1949, Mosteller and Bush, 1954] method was designed for independent hypotheses. Abelson [2012] found that Stouffer's test is more sensitive to consistent departures from the null hypothesis than Fisher's method for independent tests. Although Kim et al. [2013] found that Stouffer's test works well in the analysis of large-scale microarray data for dependent tests, and the form of our SCT statistic can cover Stouffer's Z-score, we do not include $\alpha = 2$ in our proof for Theorem 2.1. This is because the simulation results in Section 2.3 show that Stouffer's

Z-score always performs worse than the SCTs with $\alpha < 2$. In particular, Stouffer's Z-score tends to severely over-reject under strong dependencies. Their size-adjusted powers are always dominated by the other choices of αs. Accordingly, even though it is not impossible that Stouffer's Z-score still works in dependent cases, we do not pursue this direction in theory.

2.2.2 Power

In this section, we prove that our SCT test statistics have asymptotically optimal powers under sparse alternative hypotheses. We consider a similar setting to the one in Liu and Xie [2020]. Let $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ be the collection of test statistics, where X_i corresponds to the i-th individual test. Suppose X_i marginally follows a normal distribution. Denote $\mathrm{E}[\boldsymbol{X}] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}) = \boldsymbol{\Sigma}$. Without loss of generality, we assume $\boldsymbol{\Sigma}$ is the correlation matrix, i.e., each X_i has variance 1. The p-value for i-th two-sided test is $p_i = p(X_i) = 2[1 - \Phi(|X_i|)]$. The global null hypothesis is specified as $\boldsymbol{H_0} : \boldsymbol{\mu} = \boldsymbol{0}$ versus the global alternative hypothesis $\boldsymbol{H_a} : \boldsymbol{\mu} \neq \boldsymbol{0}$.

Assumption 2.4. *Let $S = \{1 \leq i \leq n : \mu_i \neq 0\}$, the collection of indices for which the individual null hypotheses H_is are false. Let $S_+ = \{1 \leq i \leq n : \mu_i > 0\}$, and assume $|S_+| \geq |S|/2$ without loss of generality.*

1. *The p-values in S^c follow a uniform distribution and $\{W_i\}_{i \in S^c}$ satisfy the requirements in Theorem 2.1 with $a_{|S^c|}$.*

2. *The number of elements in S is n^{k_0} with $0 < k_0 < 0.5$.*

3. *The magnitude for all nonzero μ_i is the same. For $i \in S$, $|\mu_i| = \mu_0 = \sqrt{2r \log n}$, and $\sqrt{r} + \sqrt{k_0} > \max\{\sqrt{\alpha}, 1\}$.*

4. *There exists a positive constant c_0 such that $\min_{i=1}^{n} w_i \geq c_0 n^{-1}$. The sum of weights $\sum_{j \in S} w_j = n^{k_0 - 1}$.*

Part 1 of Assumption 2.4 requires the p-values in the set S^c satisfy Assumptions 2.1 and 2.3 if $0 < \alpha < 1$ or satisfy Assumption 2.2 and 2.3 if $1 \leq \alpha < 2$. Under this condition, the contribution of p-values in the set S^c to the test statistic is bounded. Part 2 requires a sparse alternative, which is commonly taken in the multiple testing field. Part 3 controls the strength of signals. The magnitude of the nonzero signals should be large enough to ensure the test statistic is arbitrarily large. Part 4 helps keep the contribution of $\max_{i \in S} p_i$ under control. Note that p_i can still be close to 1 even when $i \in S$. In this case, $F^{-1}(1 - p_i)$ can be negative, possibly leading to a less powerful test. This assumption is to guarantee that such p_is would not affect the power of the test asymptotically.

Theorem 2.2. *Consider $0 < \alpha < 2$ and $-1 < \beta \leq 1$. Under Assumption 2.4, for any significance level s, the power of the SCT converges to 1 as $n \to \infty$:*

$$\lim_{n \to \infty} \Pr[T_n(\boldsymbol{p}) > t_s] = 1,$$

where t_s is the upper s-quantile of stable distribution $\boldsymbol{S}(\alpha, \beta)$, i.e., $F(t_s | \alpha, \beta) = 1 - s$.

The proof is attached in Appendix A.2.

For Theorem 2.2, we no longer consider $\beta = -1$. The powers of small βs tend to be dominated by other βs given the same α, making it not worth considering $\beta = -1$ for a powerful test. This is because the left tail becomes heavier as β gets closer to -1, which prevents a test from having better powers. See Section 2.3 for a related discussion.

2.3 Simulation Results

In this section, we explore the size, raw power, and size-adjusted power of the SCT in finite samples in a similar setting as Liu and Xie [2020]. A collection of test scores, $\boldsymbol{X}$, is drawn from $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. All diagonal elements of the covariance matrix $\boldsymbol{\Sigma}$ are set as 1. There are four models for the covariance matrix $\boldsymbol{\Sigma}$ considered to represent different dependence structures. Model 1 is the scheme where the individual tests are independent. In Models 2, 3, and 4, the off-diagonal entries of the covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})$ are functions of ρ.

1. Independent. The correlation between each pair of underlying test scores is zero, i.e., $\mathbf{\Sigma} = I_n$.
2. AR(1) correlation. The correlation between a pair of underlying test scores decays exponentially fast as their distances increase; $\sigma_{ij} = \rho^{|i-j|}$.
3. Exchangeable structure. The correlation between each pair of underlying test scores $\sigma_{ij} = \rho$ for all $i \neq j$.
4. Polynomial decay. The correlation between the ith and jth test scores, σ_{ij}, is set to be $\frac{1}{0.7+|i-j|^{\rho}}$. It should be noted that the correlation is a decreasing function of ρ, unlike Models 2 and 3 above.

The simulation is conducted in `R`. We use `qstable` function in `stabledist` package [Wuertz et al., 2016] to calculate quantiles of stable distributions. We truncate too-small and too-large p-values at 10^{-6} and $1 - 10^{-6}$, respectively. This is to avoid technical issues involved with too large quantiles in absolute values in the `qstable` function. The number of Monte Carlo replications is 1000. The number of individual tests in each Monte Carlo replication is 40 ($n = 40$). The significance level is set to be 5%. The parameter ρ that governs the strength of the dependencies is set to be $0.2, 0.4, 0.6$, or 0.8. Note that larger ρ implies stronger dependencies in Models 2 and 3 and weaker dependencies in Model 4. For the SCT, all combinations of $\alpha = 0.1, 0.3, \ldots, 1.9$ and $\beta = -0.8, -0.6, \ldots, 1$ are considered in addition to $(\alpha, \beta) = (1, 0)$, which is equivalent to the CCT. We also consider Stouffer's Z-score, which would correspond to the SCT with $\alpha = 2$ and $\beta = 0$. Note that although Stouffer's Z-score can be written in the SCT form, Stouffer's Z-score is not a part of the SCT family we consider in our paper. The test statistics are calculated as equation (2.2.1) with equal weights $w_i = 1/n$.

When calculating the sizes, data are generated from a multivariate normal distribution with mean $\boldsymbol{\mu} = \mathbf{0}$. For powers, a sparse alternative hypothesis with $k_0 = 0.43$ and $r = 0.54$ is considered. This means that we set $\mu_i = \sqrt{2r \log n} \approx 2$ to randomly chosen $[n^{k_0}] = [n^{0.43}] = 4$ indices in each replication. Note that $(\sqrt{0.43} + \sqrt{0.54})^2 \approx 1.93 > \max(\alpha, 1)$ for all αs considered in our simulation, satisfying Part 3 of Assumption 2.4. For raw powers, the cutoff values are taken directly from the corresponding stable distributions. For the size-adjusted powers, 1000 Monte Carol replications are first drawn under the global null hypothesis. Combined test statistics are calculated for each Monte Carlo replication. The simulation-based cutoff for each method is determined as the 95% quantile of the 1000 test statistics. After that, another set of 1000 Monte Carlo replications is drawn under the sparse alternative. The number of test statistics that are greater than the simulation-based cutoffs are counted to compute the size-adjusted powers.

Figures 2.1-2.4 present the sizes, raw powers, and size-adjusted powers under the four models. Colored lines indicate the proportion of rejections of the SCT with different αs and βs. Red dots indicate the CCT, which corresponds to the SCT with $\alpha = 1$ and $\beta = 0$. Black dots indicate Stouffer's Z-scores. The black solid lines in the size plots represent the nominal significance level 0.05.

Figure 2.1 presents the sizes, raw powers, and size-adjusted powers when the underlying tests are independent. All methods considered in our simulations, including Stouffer's Z-score, are supposed to work fine in this case. The Stouffer's Z-score and the SCT with $\alpha = 1.7$ and $\beta = -0.8$ have the best size, 0.05. However, these two methods are not necessarily the best due to their relatively low powers. In particular, Stouffer's Z-score is the lowest in both raw and size-adjusted powers. The SCT with $\alpha = 1.7$ and $\beta = -0.8$ also has relatively low raw and size-adjusted powers. In this independent scenario, the SCT with $\alpha \geq 1$ and $\beta \geq 0$ tends to have higher powers without losing the size control. In particular, most SCT methods including the CCT controls the size quite successfully, although there is a slight tendency of under-rejection. In terms of powers of the SCT, when $\alpha < 1$, β does not seem to affect the powers much, whereas when $\alpha > 1$ the powers have an increasing trend as β increases. It is noticeable the CCT tends to have better powers than the SCTs with $\alpha < 1$, keeping the sizes similar. However, when $\alpha > 1$ and $\beta > 0$, the SCT performs better in general than the CCT, both in size and power. In particular, when $\alpha = 1.5, 1.7$ and $\beta = 1$, the SCT performs best with well-controlled sizes and the highest size-adjusted powers.

Models 2-4 represent dependent cases. In these cases, the SCT works better than Stouffer's Z-score. When tests are dependent, Stouffer's Z-score is not supported in our theorems. See Remark 2.4. Stouffer's Z-score is the weakest in our simulations as well, as can be seen in Figures 2.2-2.4. In Models 2-4, Stouffer's Z-score tends to over-reject, and this tendency gets worse as the dependency gets stronger. Stouffer's Z-score also tends to have much lower powers than the SCT methods. While it sometimes has decent raw powers (e.g., Model 3 with $\rho = 0.6$ and 0.8), these powers are inflated due to their higher sizes. Their size-adjusted powers are consistently low in settings.

As for the behavior of the SCT family including the CCT, it seems that different sets of α and β work better in different situations. There is a tendency that the stronger the dependency is, the more oversized larger αs and the more undersized smaller αs. Smaller αs are generally required to keep the sizes under control for stronger dependencies. However, too small αs may result in too conservative tests. In general, the SCTs with $\alpha \geq 1$ paired with larger βs tend to have well-controlled sizes and better powers for models with weaker or no dependencies, whereas $\alpha \leq 1$ paired with larger βs tend to have better performances when dependencies are stronger.

Figure 2.2 presents the sizes and powers in Model 2 where tests are correlated with AR(1) correlation structures with different ρs. With weaker dependencies ($\rho = 0.2$ or 0.4), the SCT with $\alpha \geq 1$ has the best size-adjusted powers and well-controlled sizes. In particular, when $\rho = 0.2$, SCT with $\alpha = 1.7$ and $\beta = 1$ works the best with sizes less than 0.05 and the highest raw powers and size-adjusted powers. When $\rho = 0.4$, the SCT with $\alpha = 1.3$ and $\beta = 0.6$ works the best with a size very close to the target value and the highest size-adjusted power. In Model 2 with higher dependencies ($\rho = 0.6$ or 0.8), the SCT with $\alpha < 1$ has the best size-adjusted powers and well-controlled sizes. For instance, when $\rho = 0.6$, SCTs with $\alpha = 0.9$ have the highest size-adjusted powers and under-controlled sizes. When $\rho = 0.8$, the SCT with $\alpha = 0.1$ has the highest size-adjusted power with the size under control. The effect of β is not as much. In general, $\beta = 1$ works reasonably well for all $\alpha < 1$.

Figure 2.3 presents the sizes and powers of Model 3 where tests are correlated with exchangeable correlation structures with different ρs. The dependencies in Model 3 are stronger than those of Model 2 given the same ρ. As a result, smaller αs tend to work better in this case compared to the Model 2 cases. In Model 3, when the dependency is relatively weak with $\rho = 0.2$, the SCT with $\alpha = 1.1$ and $\beta = 1$ works best with size 0.048, raw powers 0.459, and size-adjusted powers 0.469. When the dependency is moderate or strong in Model 3, the SCT with α close to 0 and β close to 1 tends to have the best size-adjusted powers. In particular, when $\rho = 0.4, 0.6,$ or 0.8, the SCT with $\alpha = 0.1$ and $\beta = 1$ has the greatest size-adjusted powers and controlled sizes.

Figure 2.4 presents the results of Model 4 where tests are correlated as polynomial decayed correlation structures with different ρs. Model 4's dependencies are even stronger than those of Model 3 in general. Therefore, smaller αs, compared to other models, tend to produce better sizes and powers. Note that unlike Models 2 and 3, the larger ρ is, the weaker the dependencies are in this case. For all four ρs considered, $\alpha < 1$ and larger βs tend to control sizes better with higher powers. In particular, when $\rho = 0.4, 0.6, 0.8$ in Model 4, $\alpha = 0.5$ or 0.7 and $\beta = 1$ have the best size-adjusted powers and under-controlled sizes. Under the strongest dependency setting with $\rho = 0.2$, the SCTs with $\alpha = 0.1, 0.3$ and $\beta = 1$ produce the highest size-adjusted powers. In terms of raw powers when $\rho = 0.2, 0.4$ or 0.8, the SCTs with $\alpha = 0.9$

and $\beta = 1$ work the best with under-controlled sizes and largest raw powers. When $\rho = 0.6$, the SCT with $\alpha = 0.9$ and $\beta = 0.8$ has the highest raw power.

The different performances of different αs and βs seem to be strongly connected to how heavily the tails of the transformed p-values are. The right tail probability of a stable random variable is $P(W_0 > x) \sim c_\alpha(1+\beta)x^{-\alpha}$ as $x \to \infty$, where $c_\alpha = \sin(\pi\alpha/2)\Gamma(\alpha)/\pi$ and W_0 is a stable random variable that follows $\boldsymbol{S}(\alpha, \beta)$. This right tail approximation holds for all $0 < \alpha < 2$ and $-1 < \beta \leq 1$. It is noteworthy that the right tail probability is an increasing function in β and a decreasing function in α for large enough x. This means that the smaller the α is and the larger the β is, the stably transformed p-values W_i used for our combined test statistic in equation (2.2.1) has heavier right tails. It seems that this heavier right tail is particularly useful when the dependence is strong for size control. One interesting observation is that the heavier tail seems to affect in different ways under the null and the alternative. Under the null, the heavier tails lead to rejecting less, often correcting the over-rejection behavior under stronger dependencies. This is also the cause of under-rejection when α is too small. Under the alternative, the effects of heavier tails vary depending on the source. The heavier tail induced by larger βs usually leads to rejecting more, leading to better raw powers. On the contrary, the heavier tail behavior due to smaller αs on powers is not monotone. The raw powers increase as α increases, with a peak at around $\alpha = 1.5$ or 1.7, and then rapidly decreases. The only exception to the above observation on powers is Model 3 with $\rho = 0.8$. In this case, the powers decrease as β increases when $\alpha > 1$, and the powers tend to increase as α increases.

The power behaviors due to αs are somewhat consistent with the size behavior. However, the increasing powers as β increases cannot be explained by the heavier right tails. This behavior as well as the under-rejection for very small αs may be explained by the left tail behavior. Our SCT is an additive combination test where its summands may be negative. When the p-values are too close to 1, the stably transformed p-values take large negative values, which might reduce the chance of detecting the false null hypothesis when added to the test statistic. The large p-values are connected to the left tail probability of a stable distribution, which approximately has a power law $\Pr(W_0 < -x) \sim c_\alpha(1-\beta)x^{-\alpha}$ for large positive x when $-1 \leq \beta < 1$. When $\beta = 1$, $\Pr(W_{0,1} < -x) < P(W_0 < -x)$ for any $\beta < 1$. This means that the left tail probability of W_0 is a decreasing function in β for all $-1 \leq \beta \leq 1$ as well as in α, unlike the right tail probability. As a result, the smaller αs and βs are, the heavier the left tails are. Since heavier left tails may result in the loss of powers, the powers could decrease as α and β decrease. This is indeed consistent with our observations in powers in most cases.

In addition, notice that the effect of α on both tails is exponential whereas that of β is only linear. In particular, for small αs, the effect of β is not as noticeable. This is because the effect of α dominates over the effect of β in these cases. On the contrary, the effect of βs is more noticeable both in sizes and powers when α is relatively large. This is because the changes in the tail probabilities due to α are dominated by that of β for αs closer to 2.

It is also noteworthy that the SCT's behavior is somewhat consistent in all models. In particular, the exchangeable structure in Model 3 does not satisfy our long-range independence assumption in Assumptions 2.1 and 2.2, unlike the other two dependent models, Models 2 and 4. The fact that the SCT's finite sample behavior in Model 3 was similar to that of Models 2 and 4 implies that the SCT can in fact be applied to a wider range of conditions.

In summary, the SCT can control the sizes in finite samples for all four models when $0 < \alpha \leq 1$ even under strong dependencies, unlike Stouffer's Z-score. However, when $1 < \alpha < 2$, sizes tend to be substantially inflated under moderate and strong dependencies in finite samples. The size behaviors can be explained by how heavy the right tails of the transforming stable distributions are. In general, the heavier right tails seem to help control the size against strong dependencies. The heavier right tails are obtained when α is small and β is larger. This explains why smaller α and larger

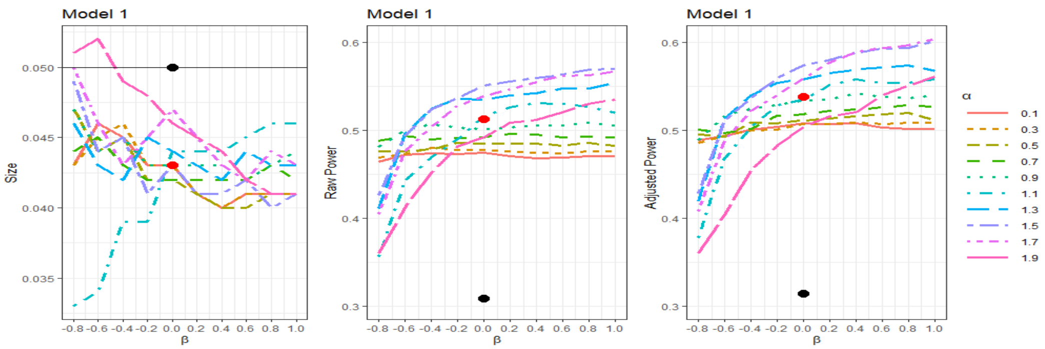

Figure 2.1: Sizes, raw powers, and size-adjusted powers of Model 1.

βs are preferred in the strong dependency case.

The powers of the SCTs tend to decrease as the dependency gets stronger. In general, the SCTs with $\alpha > 1$ and large β tend to have the best powers under no or weak dependencies, whereas the SCTs with $\alpha \leq 1$ and large β have the best sizes and powers under moderate and strong dependencies. The powers are affected by how heavy left tails of the transforming stable distributions. In general, larger αs and βs lead to lighter left tails, which often result in better powers.

Based on this simulation results, we recommend using the SCT with $\alpha \approx 1.5$ and $\beta = 1$ if the dependence is suspected to be relatively low, and using the SCT with $0.5 \leq \alpha \leq 1$ and $\beta \geq 0$ when the individual tests are suspected to be strongly dependent. If there is no knowledge of the strength of the dependencies, we recommend using either the CCT or the SCT with $\alpha \approx 0.9$ and $\beta = 1$, which lead to the best size-controlled tests without losing too much power in most cases in our simulation.

2.4 Discussion

In this chapter, we formulated an additive combination test based on stable distributions. The individual p-values are first transformed into stably distributed random variables and then their weighted sum is considered. This weighted sum still has a stable distribution, making it possible to construct a test for the global null hypothe-

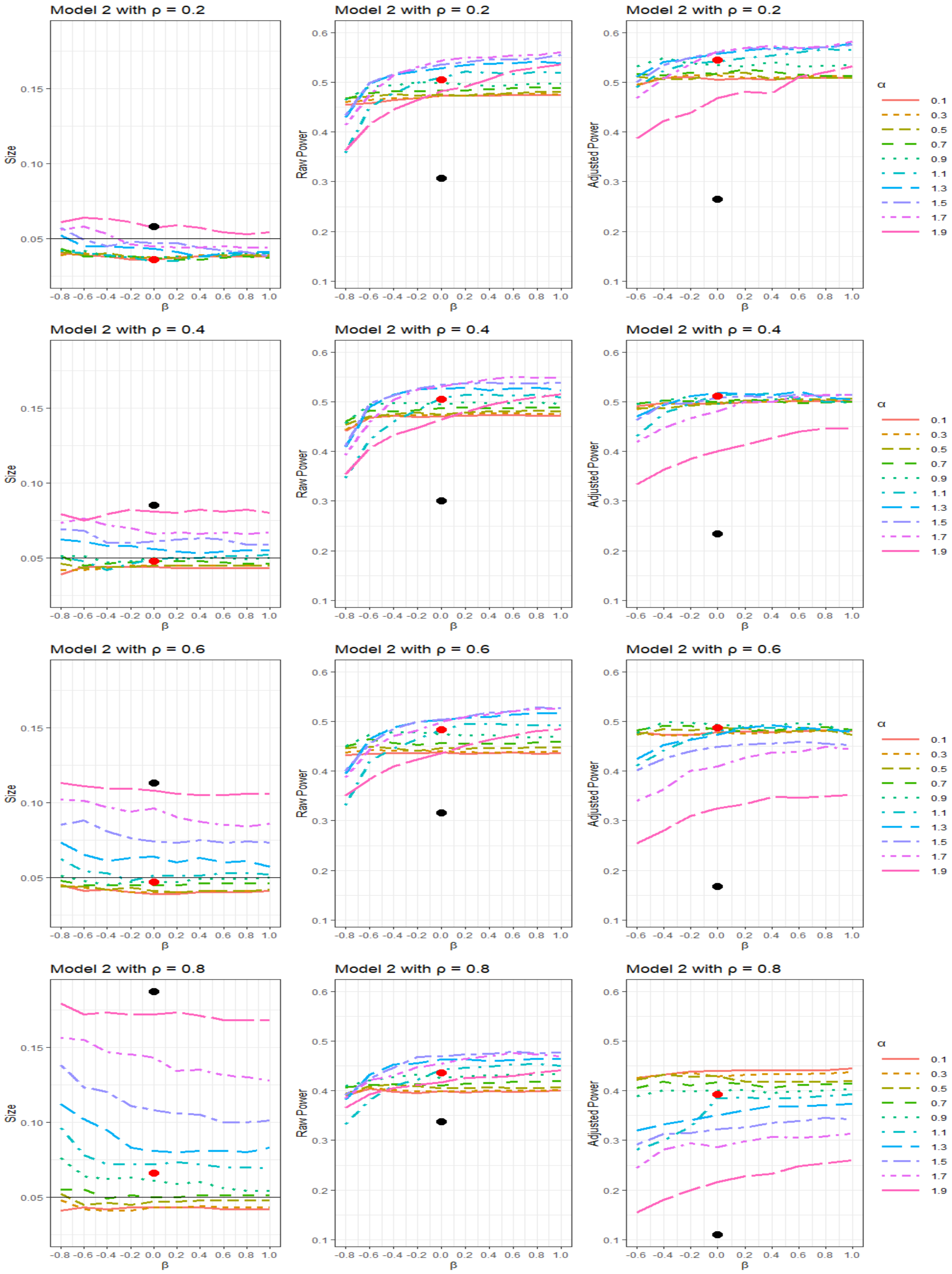

Figure 2.2: Sizes, raw powers, and size-adjusted powers of Model 2.

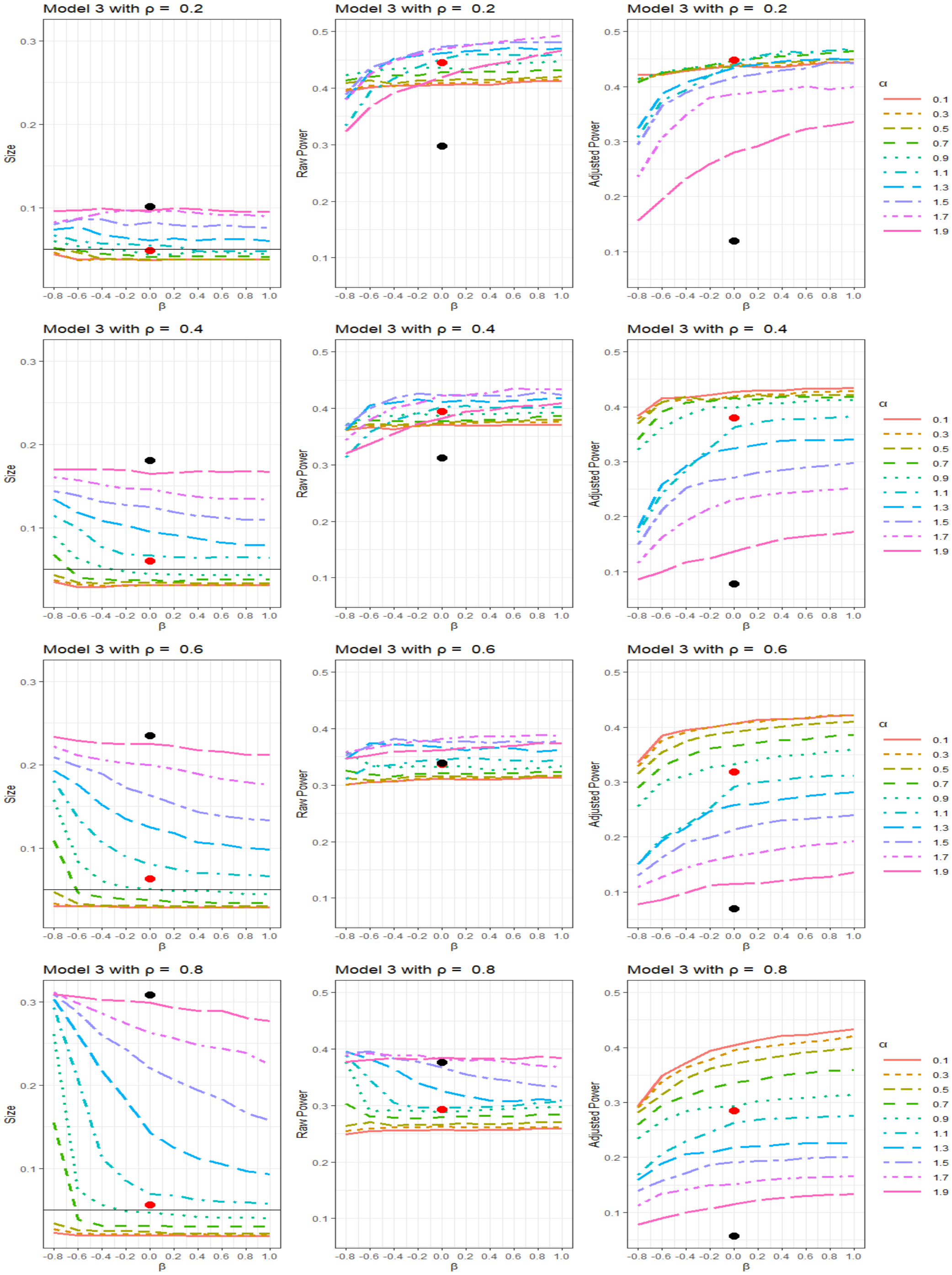

Figure 2.3: Sizes, raw powers, and size-adjusted powers of Model 3.

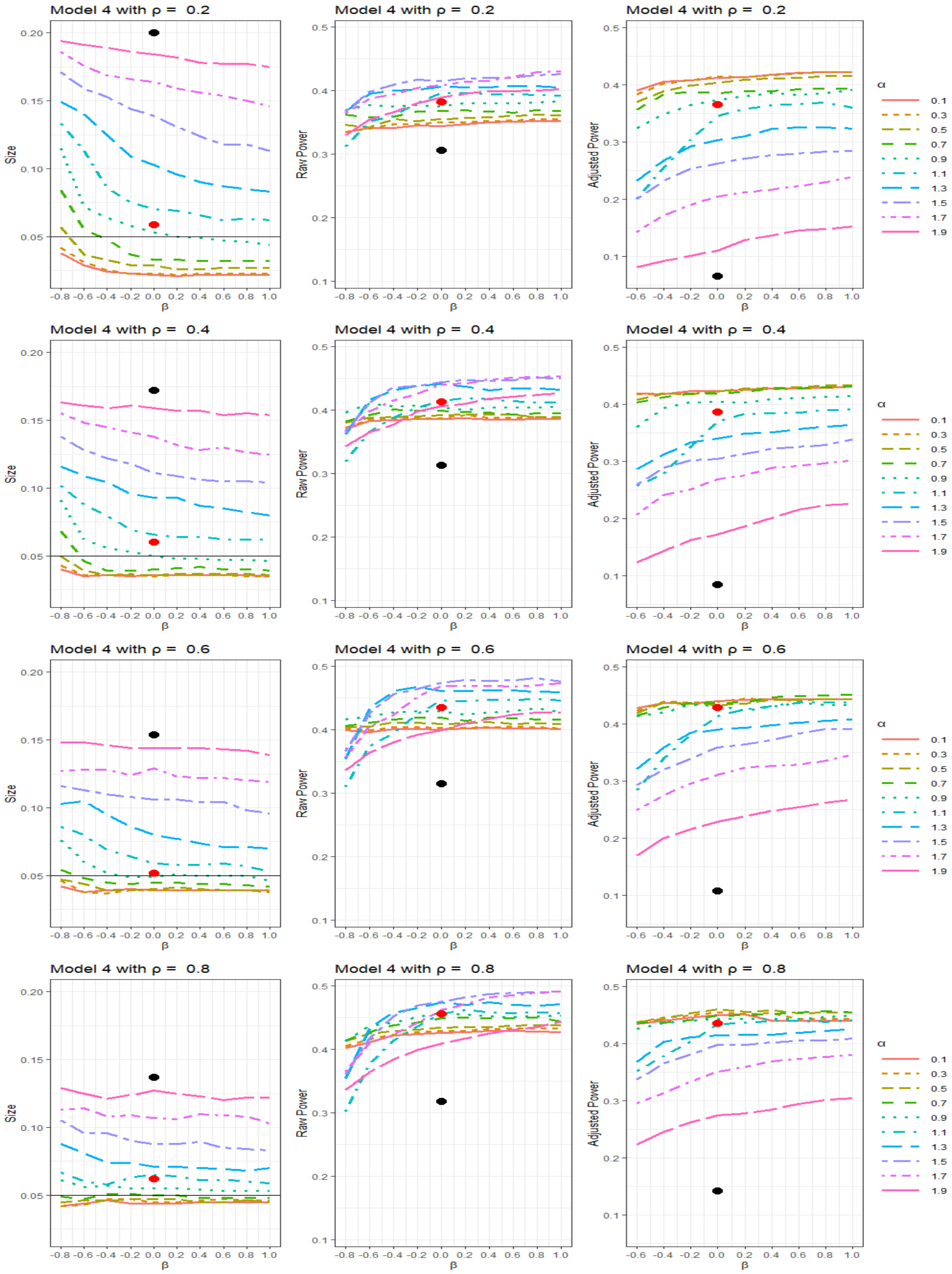

Figure 2.4: Sizes, raw powers, and size-adjusted powers of Model 4.

sis. This method can be considered as an extension of the Cauchy combination test, which is based only on the Cauchy distributed random variables because the Cauchy distribution is also a stable distribution. Similarly to Liu and Xie [2020]'s result, our test is robust to some forms of dependencies among individual p-values. We proved that this new test can successfully control the size and has asymptotically optimal power, which is further confirmed in simulations.

Chapter 3

Unified Combination Test with Regularly Varying Tails

Abstract

This chapter proposes a unifying framework for the additive p-value combination methods such as the harmonic mean p and the Cauchy combination test. These methods can be understood as convex combinations of transformed p-values, with a normalizing factor. We prove that the tails of combined statistics can be approximated by stable distributions, as long as the transformation functions are regularly varying. The asymptotic behaviors of tests in this class depend mainly on the variation exponents of their transformation functions. These tests are proven to have asymptotic optimal powers. The finite sample performances are demonstrated with a discussion on choices of stability parameter and transformation function. An application to biomedical engineering data is also presented.

Keywords: Additive combination test; hypothesis testing; multiplicity; regularly varying; stable distribution.

3.1 Introduction

This chapter also focuses on the additive p-value combination methods with the purpose to control the FWER under dependencies. Some recently proposed additive p-value combination methods are proven to be robust to weak or strong dependence structures among the p-values. These methods include the Cauchy combination test (CCT) developed by [Liu and Xie, 2020], harmonic mean p-value (HMP) proposed by Wilson [2019], the generalized mean p-values (GMP) proposed by Wilson [2020], and the stable combination test (SCT) proposed by Ling and Rho [2022]. The CCT transforms individual p-values under the null hypotheses to a standard Cauchy random variable, and then their weighted sum has the same tail probability as a standard Cauchy random variable. The SCT extends the CCT to the stable combination family. The HMP utilizes the fact that the reciprocal of a uniformly distributed random variable follows a stable distribution. The generalized central limit theorem (GCLT) is then applied to obtain the distribution of the average of the reciprocals, i.e., the harmonic mean. The GMP extends the harmonic mean into the generalized mean.

The aforementioned methods in fact share similar ideas: they are built upon the fact that a convex combination of regularly varying distribution is also regularly varying. This observation motivates a unified framework of additive p-value combination tests, which enables closer comparisons. In our unified framework, we find that the major differences between these methods come from two factors: the variation exponents and the form of the transformation functions. We prove that methods in this unified framework can use the stable distribution as their null distributions. They can control the FWER at the desired level and enjoy asymptotically optimal powers. In particular, when the variation exponent is chosen to be 1, the size can be controlled no matter how strong the dependencies among p-values are. When the variation exponent is not 1, we may gain higher power only when an asymptotic tail independence condition is met. Our simulation confirms our theoretical results.

The rest of this chapter is organized as follows. In Section 3.2, we briefly summarize existing additive combination tests and propose a unified framework based on the closure property of regularly varying random variables. We show that, under an asymptotic tail independence assumption, the FWER of these tests approximates the target value. Under some conditions on the global alternative assumptions, the unified framework has asymptotically optimal power. In Section 3.3, the choices of unified framework parameters are discussed and some simulation results are provided. A case study is illustrated in Section 3.4, as well as guidance on how to determine if the data meets the tail independence condition. Section 3.5 concludes. Heavy technical work is relegated to Appendix B. In this chapter, we continue following the notations introduced in Chapter 2.

3.2 A Unified Framework

In this section, we first briefly review recent developments in p-value combination methods that are closely connected: Wilson [2019], Liu and Xie [2020], Wilson [2020], and SCT proposed in Chapter 2. Inspired by their similar constructions and performances, we propose a unified p-value combination test.

Wilson [2019]'s HMP was originally inspired by Bayesian model averaging. They start from a weighted average of maximized likelihood ratios, noting that it resembles a model-averaged Bayes factor. They further observe that each maximized likelihood can be approximated by inverses of its p-value if the degree of freedom of the chi-squared null distribution is two and based on Wilk's theorem [Wilks, 1938]. Motivated by this observation, Wilson [2019] proposed the harmonic combined p-value, imitating the mean Bayes factor:

$$\mathring{p} = \left(\sum_{i=1}^{n} w_i p_i^{-1} \right)^{-1},$$

where w_is indicate weights of individual p-values such that $\sum_{i=1}^{n} w_i = 1$. The limiting distribution of $\mathring{p}$ can be approximated by the generalized central limit theorem (GCLT) in Uchaikin and Zolotarev [2011]. Given Assumption 1.1, p_i is uniformly distributed for $i = 1, \ldots, n$ so that the reciprocal $1/p_i$ follows a logGamma distribution with scale and shape parameters 1, LogGamma(1, 1). The tail probabilities of a LogGamma(1, 1) satisfy the power-law conditions of the GCLT (equations (2.5.17) and (2.5.18) of Uchaikin and Zolotarev [2011]). For the right tail, $\Pr\left(p_i^{-1} > x\right) \sim x^{-1}$ as $x \to \infty$ for all $i = 1, \ldots, n$. It is trivial that the left tail satisfies the power-law condition since $\Pr\left(p_i^{-1} < -x\right) = 0$ for all $x < 0$. Therefore, by the GCLT in Uchaikin and Zolotarev [2011], the sum of independent $1/p_i$s converges weakly to a stable distribution with stability parameter $\alpha = 1$ and skewness $\beta = 1$ up to a scale and location shift.

Wilson [2020] extended the HMP to generalized mean p-value (GMP). The key idea is that the tail probabilities of p_i^r satisfy the power-law conditions not only when $r = -1$, which corresponds to the HMP, but also for all $r < 0$. If p_i follows a uniform distribution between 0 and 1 and $r < 0$, the right tail probability of p_i^r is $\Pr(p_i^r > x) = x^{1/r}$ for all $x \geq 0$ and the left tail probability $\Pr(p_i^r < x) = 0$ for all $x < 0$. Therefore, the power-law tail probability conditions of GCLT from Uchaikin and Zolotarev [2011] hold for all $r < 0$ under the global null hypothesis as long as Assumption 1.1 approximately holds with independent p_is with large n. Assuming the independence of p_is, Wilson [2020] derived that $\frac{1}{b_{r,n}}(p_1^r + \ldots + p_n^r - c_{r,n})$ converges to a stable distribution with stability parameter $\alpha = \min(-1/r, 2)$ and skewness parameter $\beta = 1$, i.e., $\boldsymbol{S}(\min(-1/r, 2), 1)$. Here, $c_{r,n}$ and $b_{r,n}$ are nonrandom numbers that only depend on r and n, representing centering and scaling coefficients, respectively, as provided in Table 1 of Wilson [2020]. In particular, notice that $\alpha = 2$ when $-0.5 \leq r < 0$. A stable distribution with stability parameter $\alpha = 2$ represents a normal distribution. This special case can be understood as a result of the usual central limit theorem without the need for the GCLT, since when $-0.5 \leq r < 0$, p_i^r is not heavy-tailed and has a finite variance. Therefore, the threshold of a GMP can be written as a function of a stable distribution quantile. With significance level s, Wilson [2020] proposed to reject the global null hypothesis $\boldsymbol{H_0}$ if the GMP is smaller than the threshold $\Psi_{r,n}(s) = \left\{\frac{c_{r,n}+b_{r,n}F^{-1}(1-s|\alpha,1)}{n}\right\}^{1/r}$. Though the GCLT requires the underlying p-values to be independent, Wilson argued that the independence assumption of the GCLT can be relaxed to the Davis-Resnick condition,

$$\Pr(p_j < \epsilon | p_i < \epsilon) \to 0 \text{ as } \epsilon \to 0, \text{for all i} \neq \text{j}, \tag{3.2.1}$$

when $r < 0$, utilizing Lemma 2.1 of Davis and Resnick [1996]. Wilson [2020] also argued that when $r < -0.5$ and when significance level $s \to 0$, the threshold takes a simpler form. This can be shown using the fact that p_i^r is regularly varying. In this case, the threshold of the GMP can be shown to be $\frac{s}{n^{1+1/r}}$. This threshold is consistent with that from the GCLT.

As discussed in Chapter 2, Liu and Xie [2020] proposed the Cauchy combination test (CCT) independently from Wilson's work under a bivariate normal copula condition. We extended CCT into the stable combination test (SCT) in Chapter 2 under some mixing conditions.

The aforementioned methods have close connections. First, they all combine multiple p-values into a global one. In particular, Wilson [2019, 2020] combined the individual p-values using the harmonic mean and generalized mean, respectively. Liu and Xie [2020] and the method proposed in Chapter 2 shared the idea to transform individual p-values into a stable random variable and took their convex combination. Second, the null distribution of the combined p-values follows a stable distribution. Last but not least, their finite sample performances are quite similar as well, as observed in their papers. These methods all work well under dependencies as well.

Motivated by their similarities, we present a unified framework that can be viewed as a generalization of Liu and Xie [2020], Wilson [2019, 2020], Ling and Rho [2022]'s approaches. We note the transformation functions in Wilson [2019, 2020], Liu and Xie [2020], and Ling and Rho [2022] can be generalized to a family of functions with regularly varying tail behavior: Both the survival function of a Stable distribution (except Gaussian distribution) and a power law function can be understood as regularly varying functions. These similarities naturally suggest a unified framework based on regularly random variables.

Before defining the unified test statistic, we introduce the definition of a regularly varying function. A positive measurable function $h : (0, \infty) \to (0, \infty)$ is called *regularly varying* at infinity with index $a \in (-\infty, \infty)$ if for all $x > 0$,

$$\lim_{t \to \infty} \frac{h(xt)}{h(t)} = x^a,$$

where a is called the exponent of variation [Karamata, 1933]. If $a = 0$, this function is called slowly varying and generically denoted as $L(x)$.

The following Assumption 3.1 specifies the property of transformation functions based on regular variation to guarantee a size-controlled test.

Assumption 3.1. *Assume that $\phi(x) : (0, 1) \to (-\infty, \infty)$ is continuous and decreasing in x. Let the tail probabilities of $\phi(p)$ be regularly varying with variation exponent $-\alpha$, where p is a uniform random variable between 0 and 1 and $0 < \alpha < 2$. That is,*

$$\Pr\left[\phi(p) > x\right] \sim q_1 x^{-\alpha} L(x) \tag{3.2.2}$$

and

$$\Pr\left[\phi(p) < -x\right] \sim q_2 x^{-\alpha} L(x) \tag{3.2.3}$$

as $x \to \infty$, where $q_1, q_2 \in [0, 1]$, $q_1 + q_2 = 1$, and $L(x)$ is a slowly varying function. We further assume $L(x)$ has a finite limit l_ϕ, i.e. $\lim_{x \to \infty} L(x) = l_\phi$.

Assumption 3.1 consists of two parts: the function ϕ has to be continuous and decreasing in p, and tail probabilities of $\phi(p)$ should be regularly varying at both tails. The former ensures that the transformation preserves most order information in the p-values. For instance, the smallest p-value is most likely from an alternative hypothesis if the global null is false. With the decreasing condition on ϕ, $\max_{1 \le i \le n} \phi(p_i) = \phi(\min_{1 \le i \le n} p_i)$ carries the same information as the smallest p-value. The latter condition on the regularly varying tails of ϕ is closely related to the extreme value theory. This condition is similar to that of the GCLT, where both tails are required to have a power law behavior.

However, there are some transformation functions that cannot satisfy Assumption 3.1, and therefore, cannot be included in our unified framework. For instance, the transformation function used in the Fisher combination test, $-\log(p)$, which follows a Gamma distribution, is not regularly varying because $\Pr\left[-\log(p) > x\right] = e^{-x}$ for a uniform random variable p between 0 and 1. Another example is Stouffer's Z-score test [Stouffer et al., 1949, Mosteller and Bush, 1954]. The transformation function for the Stouffers Z-score test is $\Phi^{-1}(1-p)$, which follows a normal distribution. However, the tails of a normal distribution are not regularly varying because $\Pr[\Phi^{-1}(1-p) > x] = 1 - \Phi(x) \sim \frac{e^{-x^2/2}}{x\sqrt{2\pi}}$ [Nolan, 2020, Feller, 1968].

The test statistic of the unified framework is defined as a weighted sum of $\phi(p_i)$s

$$T_n(\boldsymbol{p}) = a_n \sum_{i=1}^{n} w_i \phi(p_i) - b_n, \tag{3.2.4}$$

where $\phi(\cdot)$ is a transformation function satisfying Assumption 3.1, w_i is a nonnegative weight imposed on the ith test, and $\sum_{i=1}^{n} w_i = 1$. With $0 < \alpha < 2$ as defined in Assumption 3.1, the normalizing constant $a_n = \left(\sum_{j=1}^{n} w_j^{\alpha}\right)^{-1/\alpha}$ is chosen such that $\sum_{i=1}^{n} \Pr\left[w_i|\phi(p_i)| > a_n^{-1}\right] \to l_\phi$. The shifting factor b_n is defined as follows,

$$b_n = \begin{cases} 0 & 0 < \alpha < 1, \\ n\mathrm{E}\sin[\phi(p_1)/n] & \alpha = 1, \\ a_n\mathrm{E}[\phi(p_1)] & 1 < \alpha < 2. \end{cases} \tag{3.2.5}$$

It is well-known that if the underlying p-values are independent [Nolan, 2020, Janson, 2011], as $n \to \infty$,

$$T_n(\boldsymbol{p}) \xrightarrow{d} \boldsymbol{S}(\alpha, \beta, \gamma, 0), \tag{3.2.6}$$

where $\beta = q_1 - q_2$, and $\gamma = \left[\frac{\pi l_\phi}{2\sin(\pi\alpha/2)\Gamma(\alpha)}\right]^{1/\alpha}$.

The unified combination test (UCT) statistic in equation (3.2.4) covers a wide range of methods in the literature. We present four examples of transformation functions that satisfy Assumption 3.1, and thus belongs to our unified framework.

Example 3.1. Let ϕ be a quantile function of a stable random variable, $\phi(p) = F^{-1}(1 - p|\alpha, \beta)$ as the stable combination test (SCT) introduced in Chapter 2. This function is decreasing in p and follows a stable distribution if p is uniformly distributed. The two tail probabilities are $\Pr[\phi(p) > x] \sim x^{-\alpha}\left(\frac{1+\beta}{2}\right)L_1(x)$ and $\Pr[\phi(p) < -x] \sim x^{-\alpha}\left(\frac{1-\beta}{2}\right)L_1(x)$ as $x \to \infty$, where $L_1(x) = \frac{2}{\pi}\sin\left(\frac{\pi\alpha}{2}\right)\Gamma(\alpha)$ is a constant that depends only on α, satisfying the regularly varying condition in Assumption 3.1. In this case, $q_1 = (1 + \beta)/2$, $q_2 = (1 - \beta)/2$, and $l_\phi = \frac{2}{\pi}\sin\left(\frac{\pi\alpha}{2}\right)\Gamma(\alpha)$. If $\alpha = 1$ and $\beta = 0$, $\phi(p)$ is a Cauchy random variable, making the CCT [Liu and Xie, 2020] a special case of this class of transformation functions.

Example 3.2. Let $\phi(p) = p^{-1/\alpha}$ be a power function with $0 < \alpha < 2$. The function $\phi(\cdot)$ is decreasing, and $\phi(p)$ has a Pareto distribution with index α if p is uniformly distributed. The right tail probability is regularly varying with $\Pr(p^{-1/\alpha} > x) \sim x^{-\alpha}$, and $\Pr(p^{-1/\alpha} < -x) = 0$ as $x \to \infty$. Therefore, Assumption 3.1 is satisfied with $q_1 = 1, q_2 = 0$, and $l_\phi = 1$. With this choice of ϕ, our unified framework is equivalent to the GMP in Wilson [2020]. The HMP in Wilson [2019] is also included in our framework with $\alpha = 1$.

Example 3.3. Let $\phi(p) = [-\log(1 - p)]^{-1/\alpha}$. This function $\phi(\cdot)$ is decreasing and follows a standard Fréchet distribution with shape parameter α if p is uniformly distributed. The corresponding survival function is $\Pr[\phi(p) > x] = 1 - e^{-x^{-\alpha}} \sim x^{-\alpha}$ as $x \to \infty$ and $\Pr[\phi(p) < -x] = 0$. We refer to this case as Fréchet Combination Test (FCT). The FCT satisfies Assumption 3.1 with $q_1 = 1, q_2 = 0$, and $l_\phi = 1$. Note that q_1, q_2, and l_ϕ are the same for the GMP and FCT. This means that the two methods are approximately equivalent, which can also be seen in the finite sample simulations in Section 3.3.

Example 3.4. Let $\phi(p) = p^{-1} + 1$. The function $\phi(\cdot)$ is decreasing, and $\phi(p)$ follows a $F(2, 2)$ distribution for a uniform random variable p. The two tail probabilities can be approximated as $\Pr[\phi(p) > x] = (x - 1)^{-1} \sim x^{-1}$ and $\Pr[\phi(p) < -x] = 0$ when $x > 0$ as $x \to \infty$. Assumption 3.1 is satisfied with $q_1 = 1, q_2 = 0$, and $l_\phi = 1$, similarly to the GMP and FCT.

The rest of this section addresses that our test statistic (3.2.4) can control the family-wise error rate under the null and has nontrivial power under some alternatives when individual p-values are allowed to be dependent.

3.2.1 Size

Recall that $T_n(\boldsymbol{p})$ is approximately distributed as a stable distribution in equation (3.2.6) when p_i are independent. In Chapter 2 we extended this result to weakly dependent p_i's when $\phi(1-p)$ is a stable quantile function, assuming that $\{p_i\}_{i=1}^n$ have a natural order. In this chapter, we do not require this natural order or weak dependence. Instead, an asymptotic tail independence condition is assumed. The following assumption is adapted from Lemma 3.1 of Jessen and Mikosch [2006].

Assumption 3.2. *For $i = 1, \ldots, n$, let p_i marginally follow a uniform distribution and w_i be a fix positive weight such that $\sum_{i=1}^n w_i = 1$. For any $i \neq j$,*

$$\begin{aligned}\lim_{t\to\infty} \frac{\Pr\left[w_i\phi(p_i) > t, w_j\phi(p_j) > t\right]}{\Pr\left[w_1\phi(p_1) > t\right]} &= \lim_{t\to\infty} \frac{\Pr\left[w_i\phi(p_i) < -t, w_j\phi(p_j) > t\right]}{\Pr\left[w_1\phi(p_1) > t\right]} \\ &= \lim_{t\to\infty} \frac{\Pr\left[w_i\phi(p_i) < -t, w_j\phi(p_j) < -t\right]}{\Pr\left[w_1\phi(p_1) > t\right]} = 0.\end{aligned} \tag{3.2.7}$$

This asymptotic tail independence condition assumes that the tail probability of a pair-wise joint distribution decays faster than that of the marginal. Under this condition, it is rare to observe two simultaneous extreme events than one extreme event. The asymptotic tail independence condition can be considered as an extension of the Davis-Resnick condition in equation (3.2.1), on which Wilson [2020] relies. Note that the Davis-Resnick condition works only for nonnegative regularly varying random variables, whereas our Assumption 3.2 allows for negative values. Assumption 3.2 allows a wide range of dependence structures, for example, jointly normal distribution and jointly Farlie-Gumbel-Morgenstern distribution [Geluk and Tang, 2009].

While Assumption 3.2 does not require weak dependence nor a natural order, it seems that we no longer have the in-distribution convergence to a stable distribution. Instead, the following theorem proves that the right tail probability of $T_n(\boldsymbol{p})$ in (3.2.4) can be approximated by that of a stable random variable.

Theorem 3.1. *Suppose Assumptions 1.1, 3.1, and 3.2 hold, then under the global null hypothesis $\boldsymbol{H_0}$, the right tail probability of the unified test statistic $T_n(\boldsymbol{p})$ can be approximated by that of a stable random variable. That is,*

$$\lim_{t\to\infty} \frac{\Pr[T_n(\boldsymbol{p}) > t]}{\Pr(W_0 > t)} = 1, \tag{3.2.8}$$

where W_0 is a stable random variable that follows $\boldsymbol{S}(\alpha, \beta, \gamma, 0)$ with stability index $0 < \alpha < 2$, skewness parameter $\beta = q_1 - q_2 = \lim_{t\to\infty} \frac{\Pr[\phi(\cdot)>t]-\Pr[\phi(\cdot)<-t]}{\Pr[|\phi(\cdot)|>t]}$, location parameter 0, and scale parameter $\gamma = \left[\frac{l_\phi \pi}{2\sin(\pi\alpha/2)\Gamma(\alpha)}\right]^{1/\alpha}$.

The proof is attached in Appendix B.1. This theorem implies that the multiple tests with rejection region $\{T_n(\boldsymbol{p}) > t_s\}$ can control the Type I error at significance level s, where the cutoff value t_s is determined by the upper s-level quantile of a stable distribution $\boldsymbol{S}(\alpha, \beta, \gamma, 0)$. This holds as long as t_s is large enough, or in other words, as long as the significance level s is small enough. Owing to this theoretical result, the combined p-value of $\boldsymbol{H_0}$ is calculated as the probability of observing as or more extreme test statistic from the corresponding stable distribution, i.e. $p_c = 1 - F[T_n(\boldsymbol{p})|\alpha, \beta, \gamma, 0]$.

Remark 3.1. When individual tests are perfectly correlated, i.e., when all p_i's are equal, the tail independence assumption in Assumption 3.2 does not hold. Under perfect correlation, $\phi(p_i) = \phi(p_1)$ for all $i = 1, \ldots, n$ and the test statistic takes a simpler form, $T_n(\boldsymbol{p}) = a_n\phi(p_1) + b_n$. In this case, only $\alpha = 1$ can control the Type I error. Let W_0 have the distribution as defined in Theorem 3.1. By Assumption 3.1, as $t \to \infty$ and n fixed,

$$\frac{\Pr[T_n(\boldsymbol{p}) > t]}{\Pr(W_0 > t)} \sim \frac{\left(\frac{t+b_n}{a_n}\right)^{-\alpha}}{t^{-\alpha}} \sim \left(\sum_{j=1}^{n} w_j^\alpha\right)^{-1}.$$

Since the function w_j^α is decreasing in α, $\left(\sum_{j=1}^n w_i^\alpha\right)^{-1}$ is smaller than 1 if $0 < \alpha < 1$, equals 1 if $\alpha = 1$, and is greater than 1 if $1 < \alpha < 2$. Therefore, the UCT will under-reject when $\alpha < 1$ and over-reject when $\alpha > 1$. The further α is away from 1, the larger the type I error rate is away from the target value.

Remark 3.2. Liu and Xie [2020] proved the same result as our Theorem 3.1 for a special case when $\phi(1-p)$ is a Cauchy quantile function. Assumption C.1 of Liu and Xie [2020] assumed that the test scores of individual tests are bivariate normal. Bivariate normal random variables are always asymptotically tail independent as long as they are not perfectly correlated. Therefore, excluding the perfectly correlated case, our Assumption 3.2 is more general than the bivariate normal assumption in Liu and Xie [2020]. Our Theorem 3.1 extends Theorem 1 of Liu and Xie [2020]to wider choices of α and transformation function $\phi(\cdot)$.

Remark 3.3. If we add an assumption that $\phi(p)$ is nonnegative, the UCT has a valid multiple-level test. This can be shown by the closed testing procedure arguments in Wilson [2019, 2020]. Let $R \subset \{1, 2, \dots, n\}$ be a collection of indexes, R^c be its complement set, and t_s be the critical value derived from the upper s-quantile of stable distribution $\boldsymbol{S}(\alpha, 1, \gamma, 0)$. The multiple-level test,

$$\text{reject } \bigcap_{\mathrm{i} \in \mathrm{R}} \mathrm{H_i} \;\text{ if }\; T_{|R|}(\boldsymbol{p}) \geq \frac{a_{|R|}}{a_n} t_s + \frac{a_{|R|}}{a_n} b_n - b_{|R|},$$

is valid since

$$T_n(\boldsymbol{p}) = \frac{a_n}{a_{|R|}} T_{|R|}(\boldsymbol{p}) + \frac{a_n}{a_{|R|}} b_{|R|} + a_n \sum_{i \in R^c} w_i \phi(p_i) - b_n \geq t_s.$$

when ϕ takes nonnegative values only. This multiple-level test states that if the subset R is significant then the whole set is also significant. When $\alpha \neq 1$, the term $\frac{a_{|R|}}{a_n} b_n - b_{|R|} = 0$, therefore the form is simpler that $T_{|R|}(\boldsymbol{p}) \geq \frac{a_{|R|}}{a_n} t_s$. When $\alpha = 1$, this term is less than 0. HMP has a simpler form without the shifting factor-related term because HMP has an approximate multilevel test threshold, whereas our formula is exact. See section 3.C of the Supporting Information of Wilson [2019]. Note that a wide range of ϕ functions are nonnegative. For instance, among the four examples of ϕ listed above, if $\phi(p)$ is Pareto, Fréchet, $F(2,2)$, or stable with $0 < \alpha < 1$ and $\beta = 1$, ϕ is nonnegative. However, if $\phi(p)$ is stable with $\alpha \geq 1$ or $\beta \neq 1$, ϕ can take negative values, and the multiple-level test may not exist.

Remark 3.4. The proof of Theorem 3.1 in the supplementary material Section B.1 requires three approximations: closure property of $\phi(p_1), \ldots, \phi(p_n)$ (see Lemma 3.1 of Jessen and Mikosch [2006]), tail probability approximation of $\phi(p_i)$ in equation (3.2.2), and tail probability approximation of a stable random variable W_0 in equation (B.1.1). All of these approximations require the critical value to be large enough.

3.2.2 Power

In the following, we show that the UCT has an asymptotically nontrivial power. We consider a sparse alternative similar to Chapter 2, where only a relatively small number of hypotheses are false.

Assumption 3.3. *Let S be the collection of indices for which the individual null hypotheses H_is are false.*

1. *The number of elements in S is assumed to be n^{k_0} with $0 < k_0 < 0.5$.*

2. *There exists a positive constant $c_0 \leq 1$ such that $\min_{i=1}^n w_i \geq c_0/n$.*

3. *There exists a constant $\varepsilon_1 > 0$ such that, as $n \to \infty$*

$$\Pr\left[\phi\left(\min_{i \in S} p_i\right) \leq M_1 c_0^{-1} n^{m_1}\right] \longrightarrow 1, \tag{3.2.9}$$

   where $m_1 = \varepsilon_1 + \max(1/\alpha, 1)$ and $M_1 > 0$ is a constant.

4. *If $\phi(\cdot)$ is always nonnegative, no additional assumption is required for the p-values in set S^c. If $\phi(\cdot)$ takes both negative and positive values, we have the following additional conditions:*

- *There exist a constant* $0 < \varepsilon_2 < 1/\alpha - k_0$ *such that as* $n \to \infty$

$$\Pr\left[\phi\left(\max_{i\in S} p_i\right) < M_2 c_0 n^{m_2}\right] \longrightarrow 1, \tag{3.2.10}$$

  where $m_2 = 1/\alpha - \varepsilon_2 - k_0 > 0$ *and* $M_2 < 0$ *is a constant.*

- *The p-values in set* S^c *follow a uniform distribution and satisfy the asymptotic tail independent condition in Assumption 3.2.*

Part 1 of Assumption 3.3 is for the sparse alternative, which is commonly taken in the multiple testing field. Part 2 imposes a condition on the minimum weight and helps find the upper bound of the normalizing factor a_n. Part 3 controls the minimum p-value in set S to be small enough as in equation (3.2.9). This condition guarantees that there will be at least one test statistic that is large enough when the null hypothesis is false, contributing to the nontrivial power. Part 4 is necessary only when $\phi(\cdot)$ are allowed to be negative. When $\phi(p_i)$ is always nonnegative for all i, a large p_i doesn't affect the power. On the contrary, if $\phi(\cdot)$ can be negative, $\phi(p_i)$ may take a large negative value when p_i is too close to 1. In this case, too large p_i may contribute heavily to the UCT statistic, harming the power. The conditions on the $\max_{i\in S} p_i$ in equation (3.2.10) and p-values in set S^c ensure their contributions are negligible compared to $\min_{i\in S} p_i$ and guarantee nontrivial power of the UCT, even when $\phi(\cdot)$ can take negative values.

These conditions on the magnitudes of minimum and maximum p-values are not too strong. Without loss of generality, assume that ϕ^{-1} exists. From equations (3.2.2) and (3.2.3), the tail behaviours of $\phi^{-1}(\cdot)$ can be obtained: as $x \to \infty$,

$$\phi^{-1}(x) \sim q_1 x^{-\alpha} L(x) \text{ and } \phi^{-1}(-x) \sim 1 - q_2 x^{-\alpha} L(x).$$

Therefore, the conditions in equations (3.2.9) and (3.2.10) are equivalent to, as $n \to \infty$,

$$\Pr\left[\min_{i\in S} p_i \leq q_1 l_\phi \left(M_1 c_0^{-1}\right)^{-\alpha} n^{-\varepsilon_1\alpha - \max(\alpha,1)}\right] \longrightarrow 1, \tag{3.2.11}$$

and

$$\Pr\left[\max_{i\in S} p_i \leq 1 - q_2 l_\phi (-M_2 c_0)^{-\alpha} n^{\alpha(\varepsilon_2+k_0)-1}\right] \longrightarrow 1. \tag{3.2.12}$$

Since $\varepsilon_2 < 1/\alpha - k_0$, $n^{\alpha(\varepsilon_2+k_0)-1} \to 0$, and therefore $1 - n^{\alpha(\varepsilon_2+k_0)-1} \to 1$. Given the form of $\phi(\cdot)$, we know the values of q_1, q_2, l_ϕ, and thus the conditions can be simplified. The conditions for different transformation functions are the same up to some constant adjustment given the same α. These conditions on the magnitudes of minimum and maximum p-values are not too strong. In particular, it can be shown that equation (3.2.11) is comparable with the assumptions of Liu and Xie [2020], while equation (3.2.12) is slightly stronger. See supplementary material B.3 for more detail.

The condition on $\min_{i\in S} p_i$ in Equation (3.2.9) is crucial for achieving nontrivial power. While providing the condition on $\min(p_i)$ in an order of n makes it simple to understand, this condition can be relaxed to provide a tighter bound. From the proof of Theorem 3.2 in the Supplementary Material B.2 one can see that this bound is tight except for $\alpha > 1$. This stems from setting the lower bound $\min(n^{1-1/\alpha}, 1)$ of a_n, which is tight when $\alpha \leq 1$ and can be improved when $\alpha > 1$. Without taking the lower bound of a_n in the proof, equation (3.2.9) can be replaced by the following:

$$\Pr\left[\phi\left(\min_{i\in S} p_i\right) \leq \frac{M_1 n^{1+\varepsilon_1}}{c_0 a_n}\right] \longrightarrow 1. \tag{3.2.13}$$

With a properly defined $\phi^{-1}(\cdot)$, equation (3.2.13) is equivalent to

$$\Pr\left[\min_{i\in S} p_i \leq q_1 l_\phi \left(\frac{c_0}{M_1}\right)^{\alpha} \left(\frac{a_n}{n^{1+\varepsilon_1}}\right)^{\alpha}\right] \longrightarrow 1. \tag{3.2.14}$$

An interesting observation is that equation (3.2.14) holds easier if $q_1 l_\phi \left(\frac{c_0}{M_1}\right)^{\alpha} \left(\frac{a_n}{n^{1+\varepsilon_1}}\right)^{\alpha}$ is larger. Understanding this behaviour would guide how to choose $\phi(\cdot)$ and α. In particular, q_1, l_ϕ, and α are solely determined by the choice of the transformation function $\phi(\cdot)$ and can be controlled by the user. In the following three remarks, we discuss the influence of each component, q_1, l_ϕ, and α, on the power of UCT.

Remark 3.5. Equation (3.2.14) suggests that larger q_1 would lead to a higher power of UCT. This was indeed the case in Ling and Rho [2022]'s simulation for SCT. For SCT, $q_1 = (1+\beta)/2 < 1$ if $\beta < 1$. The largest possible q_1 is 1 and is achieved when $\beta = 1$. Ling and Rho [2022] reported that SCT with $\beta = 1$ does outperform $\beta < 1$ case for the same α. Examples 3.2–3.4 all have the largest possible $q_1 = 1$.

Remark 3.6. The larger l_ϕ could also lead to a higher power of UCT. Examples 3.2–3.4 all have $l_\phi = 1$. For SCT, $l_\phi = \frac{2}{\pi}\sin\left(\frac{\pi\alpha}{2}\right)\Gamma(\alpha)$, which is a decreasing function in α. In fact, $\lim_{\alpha\downarrow 0} l_\phi = 1$, and $l_\phi < 1$ for all $\alpha > 0$. This suggests that SCT may be slightly inferior to other choices of transformation functions in finite samples, which is indeed consistent in our simulations in Section 3.3.

Remark 3.7. The effect of α is more complicated since a_n depends on α. There are two terms to consider: $\left(c_0 {M_1}^{-1}\right)^\alpha$ and $\left(a_n n^{-(1+\varepsilon_1)}\right)^\alpha$. Since $M_1 > 1$ and $0 < c_0 \leq 1$, $\left(c_0 {M_1}^{-1}\right)^\alpha$ is a decreasing function in $\alpha > 0$. Since $a_n \leq n^{1-1/\alpha}$ for any α and n, $a_n n^{-(1+\varepsilon_1)} < 1$, and therefore, $\left(a_n n^{-(1+\varepsilon_1)}\right)^\alpha$ would also be a decreasing function in α if a_n is fixed, which would have suggested that higher power would be achieved if α is chosen closer to 0. However, this is not the case because a_n depends on α. In fact, $a_n = (\sum w_i^\alpha)^{-1/\alpha}$ is increasing in α when $\alpha > 1$, decreasing when $\alpha < 1$, and is equal to 1 when $\alpha = 1$.

An interesting observation from equation (3.2.11) is that the smaller $\min_{i\in S} p_i$ is required for larger α, given $\alpha \leq 1$. For a simpler argument, let's assume equal weights. In this case, $a_n = n^{1-1/\alpha}$ and $\min p_i$ is bounded from above by $q_1 l_\phi \left(M_1 c_0^{-1}\right)^{-\alpha} n^{-\varepsilon_1\alpha - 1}$, which decreases as α increases. This suggests that this condition is easier to achieve if α is smaller, closer to 0.

Similarly, the larger q_1, l_ϕ, and c_0 are, the more likely the equation (3.2.11) is satisfied, given the same n, M_1, and ε_1. For q_1, Examples 3.1–3.4 all have the largest $q_1 = 1$, except for SCT with $\beta < 1$. As pointed out in Chapter 2, the SCT with $\beta < 1$ was dominated by $\beta = 1$ in their finite sample simulations. Larger c_0 means that the weights are more equally spread across p_is, which may contribute to a higher power.

The following theorem ensures that the power of a test that belongs to the unified framework goes to one as $n \to \infty$.

Theorem 3.2. *Under Assumption 3.3, assume the weights* $\sum_{i \in S} w_i = n^{k_0 - 1}$ *and* $\sum_{i=1}^{n} w_i = 1$. *For any significant level s, the power of the unified framework goes to 1 as follows:*

$$\lim_{n \to \infty} \Pr\left[T_n(\boldsymbol{p}) > t_s\right] = 1, \tag{3.2.15}$$

where t_s *is the upper s-quantile of the stable distribution* $\boldsymbol{S}(\alpha, \beta, \gamma, 0)$ *with* $0 < \alpha < 2$, $\beta = q_1 - q_2$, *and* $\gamma = \left[\frac{l_\phi \pi}{2 \sin(\pi\alpha/2)\Gamma(\alpha)}\right]^{1/\alpha}$.

The proof is attached in the Appendix B.2.

3.3 Simulation studies and selection of parameters

This section presents simulation results to examine the size and power of the UCT for various choices of tail indexes α and transformation functions $\phi(\cdot)$ under different levels of dependence strengths. We adopt similar simulation settings in Liu and Xie [2020] and Chapter 2. Three values of sample sizes are examined; $n = 40, 100$, and 300. The number of Monte Carlo replications is 10^4 when the nominal significance level s is 0.05 and $5 * 10^5$ when $s = 10^{-4}$. In each replication $\boldsymbol{X} = (X_1, \ldots, X_n)$ are drawn from an n-variate normal distribution $N_n(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \boldsymbol{0}$ under $\boldsymbol{H_0}$ and $\boldsymbol{\mu} \neq \boldsymbol{0}$ under $\boldsymbol{H_a}$. Let Φ be the distribution function of the standard normal distribution, the p-values $p_i = 2 - 2\Phi(X_i)$ are computed from the Z-score test statistics X_i. The number of signals is $n^{0.2}$ for each n under $\boldsymbol{H_a}$. All signals have the same strength $\mu_a = \sqrt{1.8 \log n}$. The elements of the covariance matrix $\boldsymbol{\Sigma}$ are set to follow a decaying structure where $\mathrm{Cov}(X_i, X_j) = \rho^{|i-j|}$ with $0 \leq \rho < 1$ for $i, j = 1, \ldots, n$. Size and raw power are calculated as the percentage of rejections of the test statistics using the quantile of the reference stable distribution as the critical value. For all settings, we consider equal weights for the UCT methods; that is, $w_i = 1/n$, and $a_n = (\sum_{i=1}^{n} w_i^{\alpha})^{-1/\alpha} = n^{1-1/\alpha}$.

The size-adjusted powers are calculated to make the comparison fair to different sizes. The simulation-based critical values are the upper s-quantiles of the simulated UCT statistics drawn under the null hypothesis. The quantile of the reference stable distribution is calculated by `qEstable` in `FMStable` package [Robinson, 2012] in `R` when $\beta = 1$. When $\beta \neq 1$, the `qstable` function in `stabledist` package [Wuertz et al., 2016] is utilized. `qEstable` is more accurate and quicker than `qstable` but the former only works for $\beta = 1$. In the simulation of SCT, too large and too small p-values are truncated at $1 - 10^{-6}$ and 10^{-6} respectively to avoid any technical issue involved with too large quantiles in absolute values.

When computing the UCT statistics, we need to find the form for b_n in equation (3.2.5). In particular, finding b_n for $\alpha = 1$ may be tricky since it involves finding expected values for a sine function. This can either be done by simply simulating $\sin[\phi(p_i)/n]$ and taking their averages. An alternative approach is using well-known approximations. In our simulations, we used $b_n = \log n + 1 - 2c_E$ for FCT [Janson, 2011], and $b_n = \log n + 1 - c_E$ for HMP [Wilson, 2019, Zaliapin et al., 2005], where $c_E \approx 0.57721$ is the Euler's constant.

We examine the performances of the UCT based on their size, raw power, and size-adjusted power. The effect of different tail indexes α under various strengths in correlations ρ are discussed in section 3.3.1. The effect of different transformation functions ϕ is discussed in section 3.3.2.

3.3.1 The effect of tail index

In this subsection, we explore the effect of choices of α under various dependence strengths. The effect of α is similar across different choices of transformation function $\phi(\cdot)$ in our unreported simulation. In this subsection, we only present the result with the FCT to save space.

Recall that the FCT has tail probability $\Pr[\phi(p_i) > x] \sim x^{-\alpha}$ satisfying Assumption 3.1. The Fréchet distribution has expectation $\Gamma(1 - 1/\alpha)$ when $1 < \alpha < 2$.

Therefore, the FCT statistic has the following form in equal weights,

$$T_n(\boldsymbol{p}) = \begin{cases} n^{-1/\alpha} \sum_{i=1}^n \phi(p_i) & 0 < \alpha < 1, \\ n^{-1} \sum_{i=1}^n \phi(p_i) - (\log n + 1 - 2c_E) & \alpha = 1, \\ n^{-1/\alpha} \sum_{i=1}^n \phi(p_i) - n^{1-1/\alpha}\Gamma(1 - 1/\alpha) & 1 < \alpha < 2. \end{cases}$$

Our multivariate normal setting satisfies Assumption 1.1 and 3.2. Thus, the FCT statistic $T_n(\boldsymbol{p})$ asymptotically has the same tail probability as $\boldsymbol{S}(\alpha, 1, \gamma, 0)$ with $\gamma = [2/\pi \sin(\pi\alpha/2)\Gamma(\alpha)]^{-1/\alpha}$ by Theorem 3.1.

Figures 3.1 and 3.2 present finite sample sizes, raw powers, and size-adjusted powers of the FCT when the significance levels are $s = 0.05$ and $s = 10^{-4}$, respectively. Figures C.1 and C.2 in supplementary material present the same figures with significance levels $s = 0.01$ and $s = 0.005$, providing more details of the effect of different significance levels. Tail indexes $\alpha = 0.1, 0.3, 0.5, 0.7, 1, 1.3, 1.5, 1.7, 1.9$ are considered in the x-axis. The y-axis represents the percentage of rejections simulations. The black horizontal line is the nominal significance level. Different strengths of correlations are considered with $\rho = 0, 0.2, 0.4, 0.6, 0.8$, and varying sample sizes $n = 40, 100, 300$.

The first rows of Figures 3.1 and 3.2 demonstrate how different choices of αs affect the sizes. When the dependence is strong ($\rho = 0.8$), $\alpha = 1$ seems to be the only reasonable choice. This is somewhat expected because, as mentioned in Remark 3.1, the UCT methods cannot control the size under perfect correlation ($\rho = 1$) unless $\alpha = 1$. It seems that this behavior is also consistent when the dependence is strong; when $\rho = 0.8$, the FCT tends to under-reject for $\alpha < 1$ and over-reject for $\alpha > 1$. This result was consistently observed for all UCT methods in our unreported simulations. On the contrary, if $\alpha = 1$, the FCT (and all other UCT) controls the size well for any n and significance level s. Therefore, under the presence of a strong correlation, we recommend $\alpha = 1$.

However, when the dependence is weak or moderate ($\rho = 0, 0.2, 0.4, 0.6$), other choices of α may achieve higher powers, while controlling sizes reasonably. As can be seen from the first row of Figures 3.1 and 3.2, the FCT controls the size well with most αs, as long as α is not too close to 2. When α is close to 2 ($\alpha = 1.7, 1.9$), the FCT tends to under-reject with significance level $s = 0.05$ under weak or moderate dependence. In this case, the larger α is, the more severe the size distortion is. This is explained by the slower convergence to the stable tail probability limit when α is close to 2. Recall we require t/a_n to go to infinity in the proof of Theorem 3.1, where t is the critical value and $a_n = (\sum_{i=1}^{n} w_i^{\alpha})^{-1/\alpha}$ is the normalizing constant. Since the critical value depends on the significance level, the value of the significance level s at which the requirement is satisfied depends on the magnitude of a_n. In particular, when $\alpha > 1$, a_n is an increasing function both in α and n. This means that larger t is required if α is very close to 2 when the sample size n is held constant. In contrast when $\alpha < 1$, the large n does not affect the requirement of t because $a_n < 1$ according to Lemma A.3 in Appendix A. We can indeed check this behavior from the simulation. Compare the first rows of Figures 3.1 and 3.2, except for the $\rho = 0.8$ case. We can see that the under-rejection behavior is no longer observed in 3.2 with significance level 10^{-4}. Figures C.1 and C.2 in Appendix C present the same plots with significance levels 0.01 and 0.005, which demonstrate that the under-rejection behavior gradually reduces as the significance level s decreases, or equivalently, as the critical value t increases. Another interesting observation is that as the sample size n increases, the size distortion of large α gets worse. This phenomenon is also due to the fact that a smaller significance level is required for a larger sample size if $\alpha > 1$.

The second and third rows of Figures 3.1–3.2 present the raw and size-adjusted powers, respectively. From the 3rd row of Figure 3.1 and 3.3, the size-adjusted power increases as n increases, as expected by Theorem 3.2. An interesting observation is that the raw and size-adjusted powers seem to slightly increase as α increases. In Figure 3.1, the raw power seems to drastically decrease when $\alpha > 1.5$. This, again, is due to the too-small t/a_n ratio mentioned above. In the proof of Theorem 3.2, we still need large enough t/a_n. This can indeed be double-checked by Figure 3.2, where the drastic decrease in raw power is no longer observed, thanks to the much smaller significance level $s = 0.0001$. Another driver of this phenomenon is that a stronger condition is required when alpha is large. In equation (3.2.11), $\phi^{-1}(M_1 c0^{-1} n^{m_1})$ relies on α negatively. The larger α is, the smaller value of minimum p-value is required. In our simulation, we use the same set of p-values for varying αs. However, the smallest p-value in our simulation is not small enough to ensure the Part 3 of Assumption 3.3 when α is close to 2. Even when the significance level is relatively large $s = 0.05$, it seems that $\alpha > 1$ may still offer better good size-adjusted powers than $\alpha = 1$, as long as α stays away from 2 reasonably, say $\alpha = 1.5$.

In the following we look more closely into the performance of the FCT under weak dependence where $\rho = 0, 0.05, 0.1, 0.15$, and 0.2 when the significance level is 0.05. Unlike the strong correlation cases, there are no inflated sizes with any $\alpha \in (0, 2)$ in the 1st row of Figure 3.3. We still have severe under-rejection when α is closer to 2 because t/a_n is too small, as mentioned above. The conservative size and decreasing raw power when $1 < \alpha < 2$ is due to the inaccurate approximation when the significance level is not small enough as discussed above. When we reduce the significance level, the peak moves to a larger α. Under 5% significance level in our simulation settings, $\alpha = 1.3$ produces the highest raw powers when $n = 40$ and 100, and $\alpha = 1.5$ does when $n = 300$. Under 1% and 0.5% significance level in our simulation settings, $\alpha = 1.7$ produces the best raw power for all n. In a weak dependence structure, if a higher significance level as 5% is required, $\alpha = 1.5$ produces the best power, and if a lower significance level as 0.5% is required, $\alpha = 1.7$ achieves a higher power. However, when $\alpha \to 2^-$, stable distribution approaches the normal distribution and the regularly varying tail behavior is only evident on extreme tails. As a consequence, we do not recommend α larger than 1.5 unless the significance level is extremely tiny.

In conclusion, if the significance level is not small enough, $\alpha = 1$ is the best choice under a strong dependence structure and choosing $\alpha > 1$ helps achieve higher powers

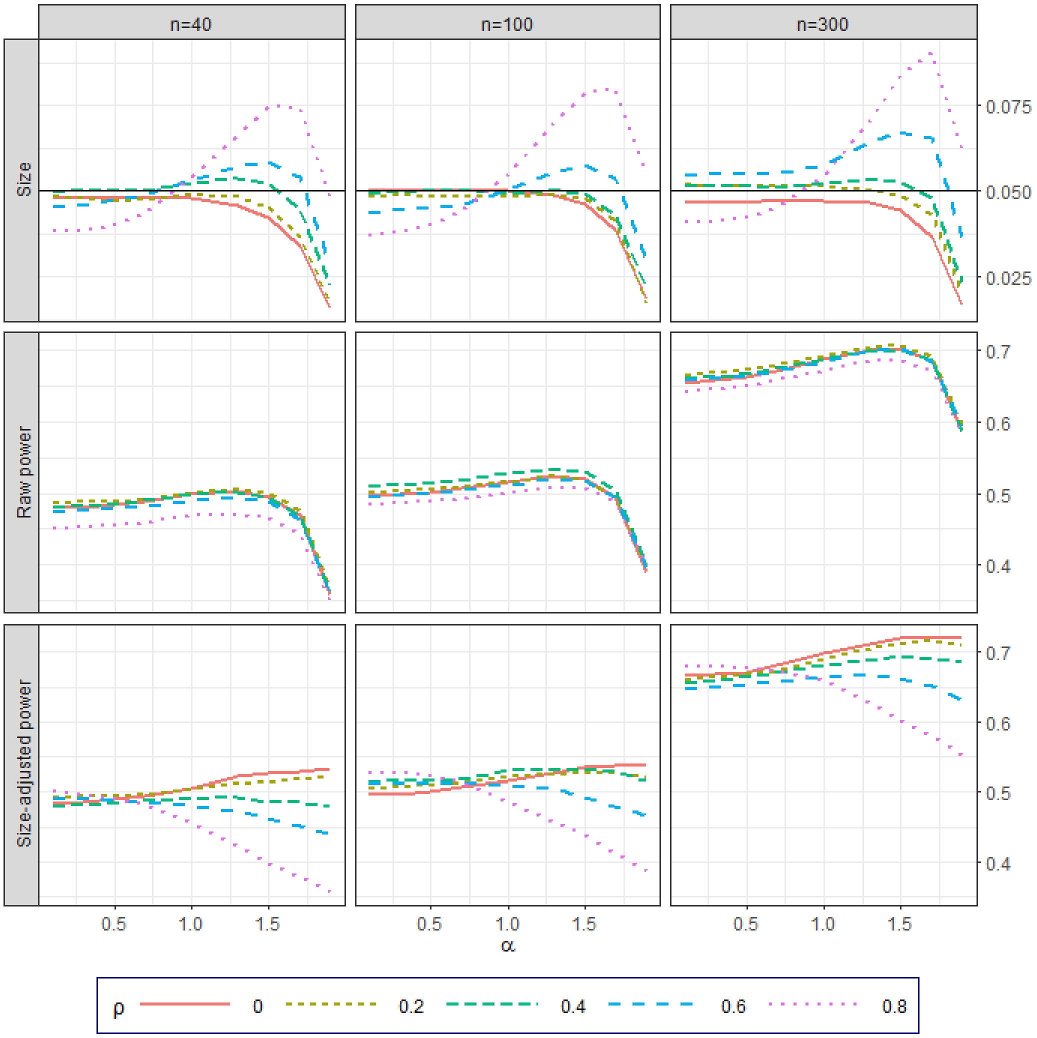

Figure 3.1: Size, raw power, and size-adjusted powers of the FCT at the nominal significance level 5%.

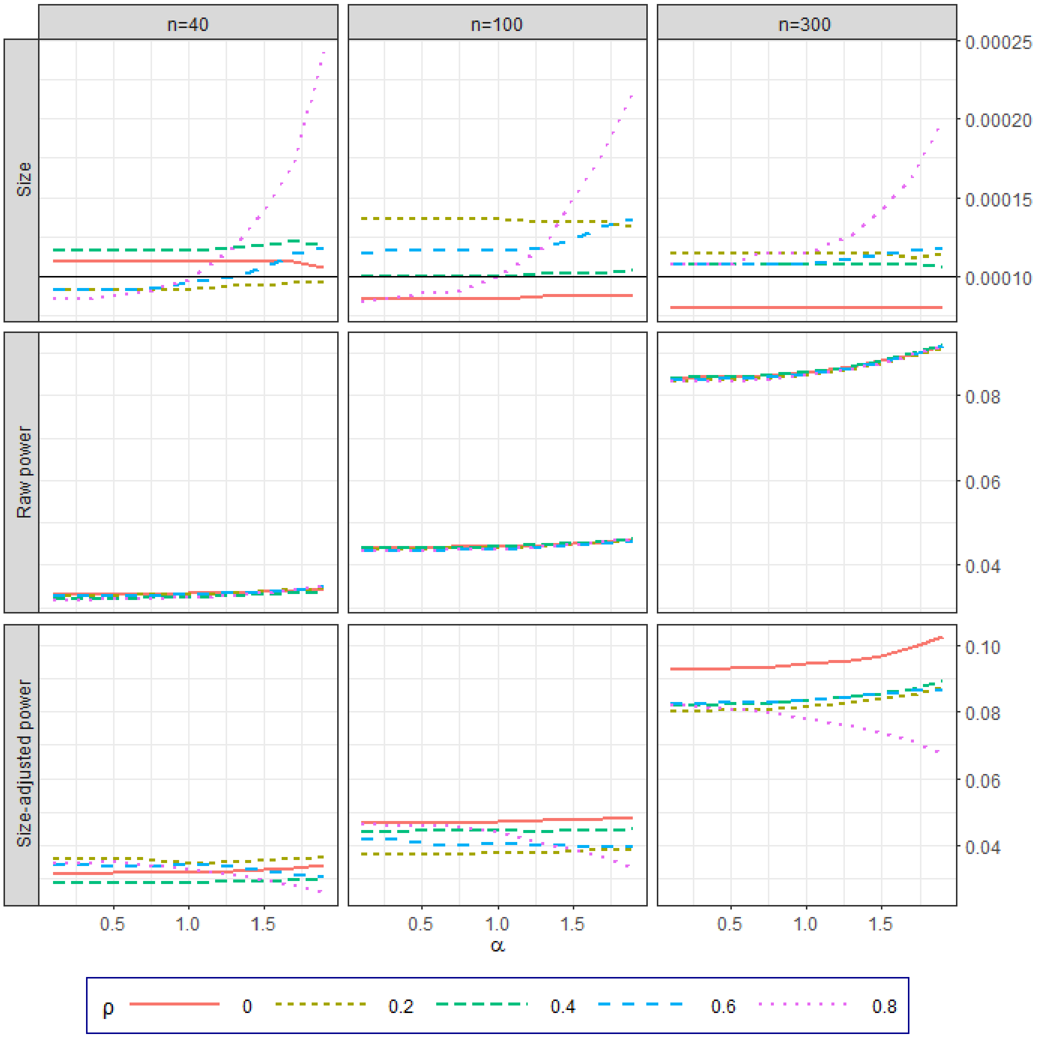

Figure 3.2: Size, raw power, and size-adjusted powers of the FCT at the nominal significance level 10^{-4}.

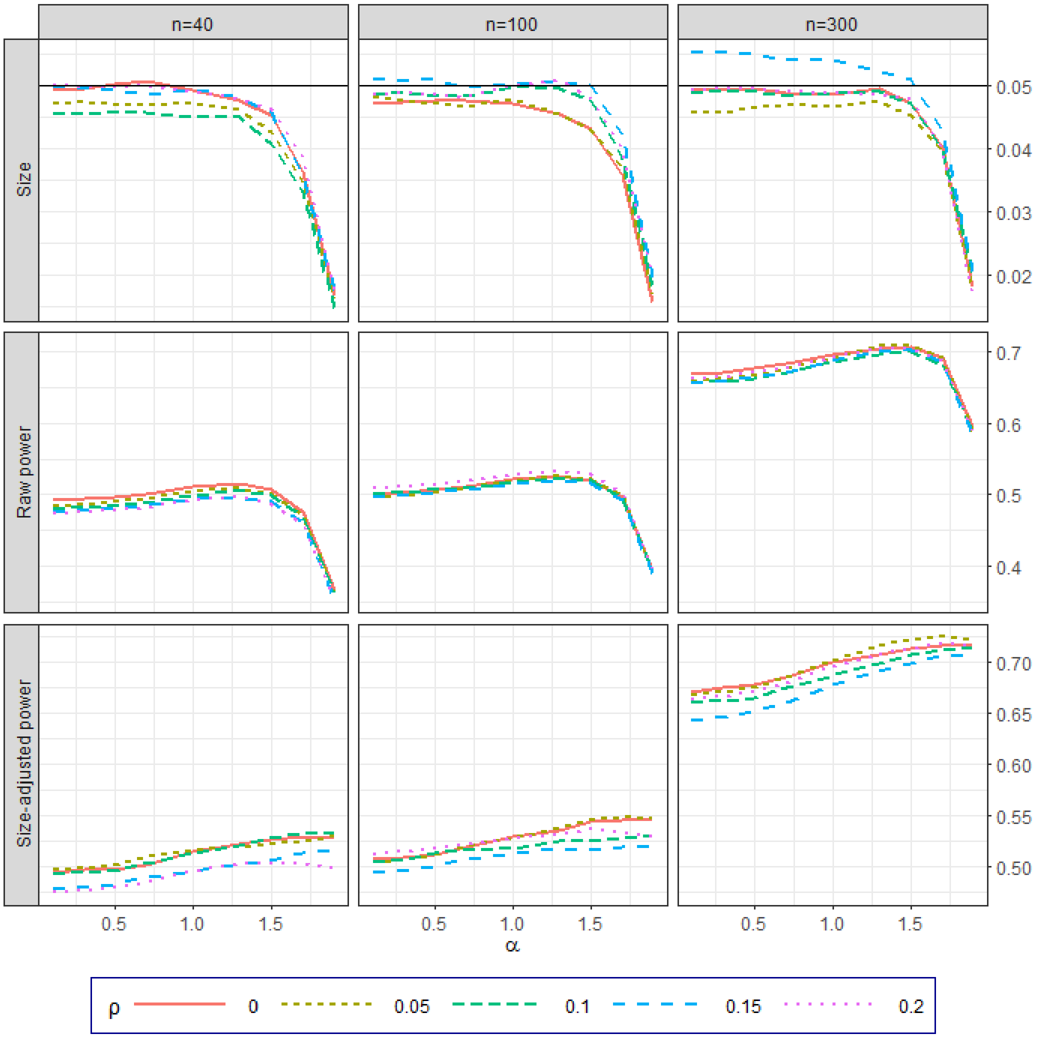

Figure 3.3: Size, raw power, and size-adjusted powers of the FCT with varying weak correlation at the nominal significance level 5%.

under a weak dependence structure. These suggestions are consistent with the results in the previous Chapter.

3.3.2 The effect of transformation function

In this subsection, we consider the performance of different transformation functions ϕ. When any ϕ takes only positive values, the test has better size and power under dependence. Moreover, the test can control the FWER in a strong sense in this case.

Along with our recommendation of $\alpha = 1$ under strong dependence and $\alpha = 1.5$ under weak dependence, we compare the following six methods in Figure 3.4: FCT with $\alpha = 1, 1.5$, GMP with $\alpha = 1, 1.5$, SCT with $(\alpha, \beta) = (1, 0)$ and $(\alpha, \beta) = (1.5, 1)$. The x-axis is the correlation ranging from 0 to 1, and the y-axis represents the percentage of rejections in 10^4 simulations. The black horizontal line indicates the nominal significance level 5%. In legend, “FCT1” refers to FCT with α =1, “FCT1.5” refers to FCT with $\alpha = 1.5$, “GMP1” refers GMP with $\alpha = 1$, which is the same as the HMP, “GMP1.5” refers GMP with $\alpha = 1.5$, “SCT1” refers to SCT with $(\alpha, \beta) = (1, 0)$, which is the same as CCT, and “SCT1.5” represents SCT with $(\alpha, \beta) = (1.5, 1)$. This result is reported only in graphs for brevity, but their corresponding numerical values are available in Appendix C. The FCT and GMP have similar performance despite the value of α because $p \approx -\log(1 - p)$ when p is small.

When dependence is weak, for example, $\rho < 0.3$, all methods control the size well, and especially the SCT with $(\alpha, \beta) = (1.5, 1)$ produces the highest raw power. When $\rho < 0.75$, the size-adjusted powers of these six methods are close, and thus suggest that the unified framework is robust to dependence. When dependence is strong, only tests with $\alpha = 1$ successfully control the size. Among them, the “SCT1” has a higher size than “FCT1” and “GMP1” when the correlation is high. The “SCT1” (i.e. CCT) is less out-performed than the other two methods because it takes both positive and negative values and it may potentially suffer from a large negative penalty when

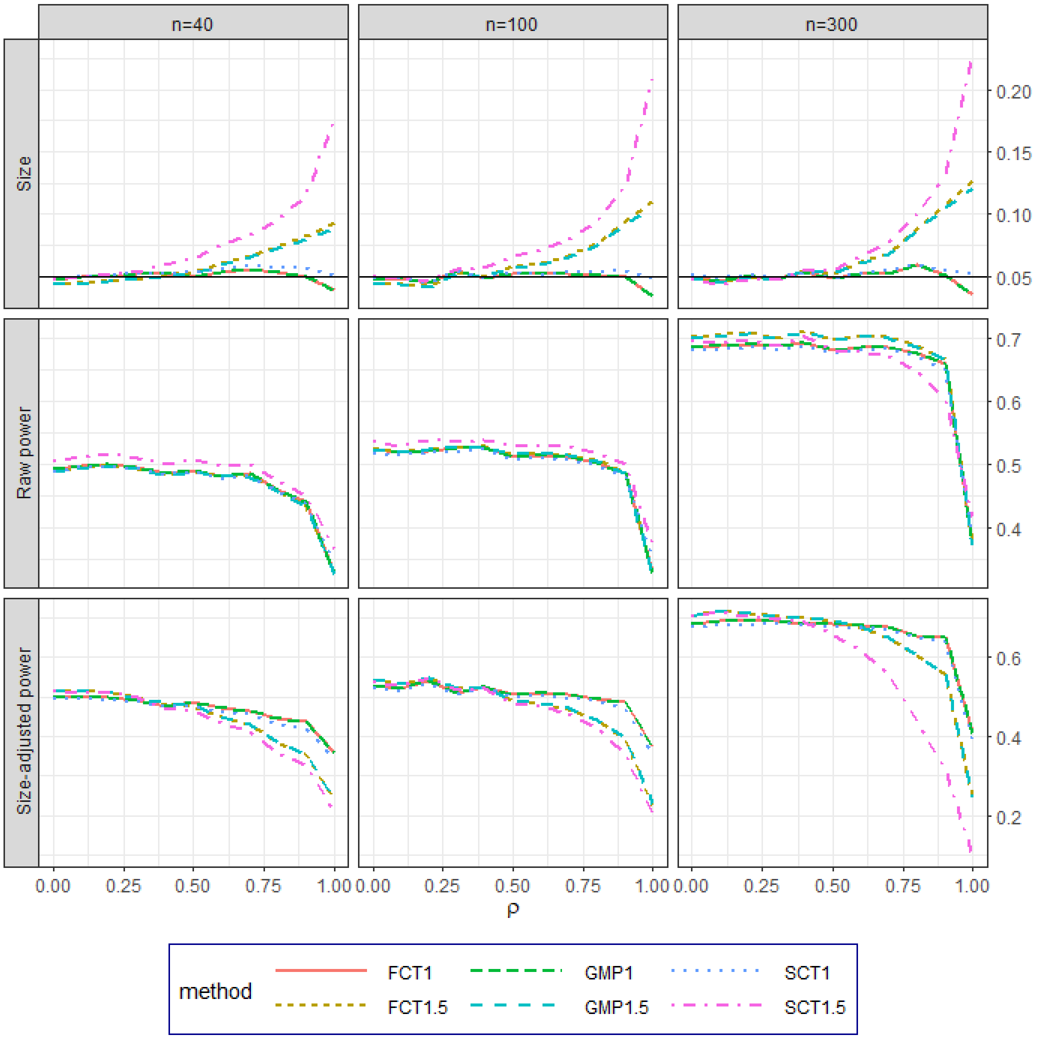

Figure 3.4: Size, raw power, and size-adjusted power of six methods.

$p_i \to 1^-$ even though under Part 4 of Assumption 3.3.

3.4 Application

We demonstrate applications with a set of paired time series of fluorescence intensity in this section. The data were captured from live-cell imaging of the two fluorescently-tagged molecules, i.e., Talin-GFP and Vinculin-TMR, which are involved in cell adhesion to the extracellular matrix [Humphries et al., 2007, Geiger et al., 2009]. As focal adhesions, puncta-like molecular complexes where Talin and Vinculin are part of, slide over time, from the time-lapse images, point-like sources were detected via 2D Gaussian fitting of the Talin-GFP channel and tracked using linear assignment problem [Han et al., 2021]. From the x- and y-locations of each trajectory, fluorescence intensities of both Talin and Vinculin channels were measured and stored in a separate array. While how Talin binds Vinculin *in vitro* is well-understood via biochemistry, how they are recruited and associated with each other in live cells has been unclear. In a previous attempt, a simple event detection algorithm was used to detect which molecular signal has risen earlier than the other. Depending on each adhesion's fate to mature or fail, the time order of the recruitment event was different [Han et al., 2021]. However, a functional causality has not been evaluated between the two signals.

There are two series of Talin and Vinculin in each of the 278 locations. In each location, the series has a length of at most 601. Since newly assembled IACs tend to show an increasing trend, we first detrend the data to make them stationary. Figure 3.5(a) plots the detrended time series of two variables from one randomly picked location. To find a causal influence of Talin to Vinculin or vice versa, we use a F test for the existence of Granger causalities between the two series.

As shown in equation (3.4.16) and (3.4.17), the F test compares the unrestricted model in which one series (x_t) is explained by the lagged value of both series (x_t and y_t) and the restricted model where the series is only explained by its own lagged values. For $t = 1, \ldots, T$,

$$x_t = u_0 + \sum_{i=1}^{q} u_i x_{t-i} + \sum_{i=1}^{q} v_i y_{t-i} + e_{1t}, \text{ and} \tag{3.4.16}$$

$$x_t = \tilde{u}_0 + \sum_{i=1}^{q} \tilde{u}_i x_{t-i} + e_{2t}. \tag{3.4.17}$$

Each F test has an individual null hypothesis that there is no directional causal link from y_t to x_t, that is $v_i = 0$ for all $i = 1, \ldots, q$. The lag q is selected by

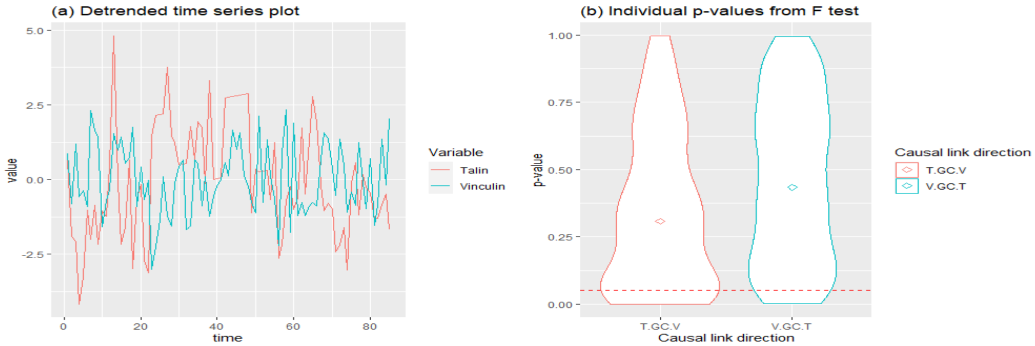

Figure 3.5: (a) The detrended time series plots of two variables. (b) The violin plots of the individual p-values from Granger causality F test of both directions.

the smallest BIC with a maximum lag order of five. The test statistic is defined as $F = \frac{(\sum_{t=1}^{T} \hat{e}_{2t}^2 - \sum_{t=1}^{T} \hat{e}_{1t}^2)/q}{\sum_{t=1}^{T} \hat{e}_{1t}^2/(T-2q-1)}$. Under the normality assumption, the F test statistic has a null distribution of $F(q, T-2q-1)$, and thus we can compute a corresponding p-value for each test statistic.

In total, there are 278 hypotheses and 278 corresponding p-values in each causal direction. Figure 3.5(b) is the violin plots of the individual p-values from the multiple Granger causality F tests of both directions. ‘T.GC.V” represents the direction from Talin to Vinculin and “V.GC.T” represents the direction from Vinculin to Talin. The dashed horizontal red line is at an intercept of 0.05. The points in the plots are the median values. From Figure 3.5(b), the p-values of a causal link from Talin to Vinculin tend to be closer to zero, whereas the other direction is generally uniformly distributed between 0 and 1.

We further fit multiple testing models to calculate a combined p-value to draw more accurate conclusions. The 278 individual p-values are fitted into the unified framework with different αs and different methods: FCT, GMP, SCT with $\beta = 0$, and SCT with $\beta = 1$. The individual p-values are truncated at 10^{-6} to avoid divergence in the calculation. Table 3.1 on the left (or right) give the combined p-values of the causal link from Talin to Vinculin (or from Vinculin to Talin) with varying α. When $\alpha = 1$, the GMP is equivalent to the HMP. SCT is $\mathrm{SCT}_{\alpha,1}$ when $\alpha \neq 1$ or

$\text{SCT}_{1,0}$, which is equivalent to CCT. The p-values are rounded to the 4th decimal place. Regarding the left part of Table 3.1, the causal link from Talin to Vinculin, all p-values with different α and different methods are very small. Therefore, we can reject the null confidently. However, on the right of Table 3.1, the conclusions drawn from the overall p-values with different αs are inconsistent. Most p-values are large, but when $\alpha > 1.6$ p-values from SCT are small. The combined p-values when $\alpha = 1.5$ are 0.13, 0.14, and 0.09 for FCT, GMP, and SCT respectively. Therefore, we reject the global null hypothesis and conclude that Vinculin does not Granger-cause Talin. The results of causal links between Talin and Vinculin are backed up by the findings in Han et al. [2021] that Talin comes a bit earlier than Vinculin. It makes sense to have Vinculin follow Talin but not the other way around.

α	FCT	GMP	SCT
0.5	0.0003	0.0003	0.0082
0.8	0.0003	0.0003	0.0010
1	0.0002	0.0002	0.0002
1.5	0.0001	0.0001	0.0009
1.8	0.0001	0.0001	0.0003

α	FCT	GMP	SCT
0.5	0.4489	0.4473	0.4473
0.8	0.4194	0.4017	0.3356
1	0.2642	0.2646	0.1672
1.5	0.1308	0.1358	0.0854
1.8	0.1201	0.1298	0.0298

Table 3.1: The table of combined p-value causal links from both directions.

3.5 Discussion

We proposed a unified p-value combination framework that embraces many additive tests in the literature. Our assumption allows a wide variety of existing methods such as the Cauchy combination test and the harmonic mean p, providing an explanation of why those methods behave somewhat similarly. Though our method belongs to the family of additive p-value combination tests and utilizes the information of all p-values, we focus more on the minimum p-value because it dominates the unified test statistic after transformation when the global null hypothesis is false. As shown in Section B.2 in Appendix B, the transformed minimum p-value goes to infinity faster than the other transformed p-values.

We proved that the right tails of this class of test statistics under the null hypothesis can be approximated by a stable random variable, ensuring good sizes for small significance levels. The power goes to one as long as the sample size is large enough under some conditions on the minimum and maximum p-values and the number of false underlying hypotheses. However, there are some drawbacks to the unified framework. First, the framework controls the family-wise error rate, which is more conservative than the false discovery rate. Second, the framework is valid based on the tail approximations of regularly varying distribution and stable distribution as well as the closure property of regularly varying random variables, and thus the significance level is required to be small enough. It remains future work how small the significance level should be to ensure the approximations are valid. Besides, the methods require the underlying p-values to be uniform under the null, which might narrow the usage. Sometimes it is of more interest to make statements of underlying hypotheses than the global hypothesis. But not all methods in the proposed framework have a valid multi-level test. Last, the performance of a method depends on the choice of transformation function ϕ and tail index α. Future studies will be needed to provide the theoretical choices of the optimal ϕ and α.

Chapter 4

A simple remedy for the multiplicity problem in rolling window Granger causality tests

Abstract

The multiplicity issue happens when the rolling window technique is utilized to capture the dynamic structure instability of Granger causality in a vector autoregression model. Bootstrap is a popular method to control the multiplicity issue, however, it is difficult to implement and costs high computations. This chapter proposes to use the additive p-value combination tests to handle the multiplicity issue caused by rolling windows. The finite sample simulations demonstrate that additive p-value combination tests are robust to dependence, have well-controlled sizes, and have comparable powers to bootstrap. Rolling window Granger causalities between the Infections Disease Equity Market Volatility Tracker (IDEMV) and S&P500 index are not founded in the full sample analysis but are discovered in rolling windows.

Keywords: Granger causality; rolling window; multiplicity; bootstrap; additive combination test.

4.1 Introduction

In economics, structural changes such as technological innovation happen occasionally. The shifts affect the data-generating process and might weaken the power of statistical analyses. In a full-sample regression, the information over the sampling period is averaged. In the presence of structural changes during the sampling period, the data-generating process was affected and thus regression estimates based on the full sample are biased, which may lead to misleading inference and may weaken the power of statistical analyses [Granger, 1996, Duchin, 1998, Boehlje, 1999, Rossi, 2005, Clark and McCracken, 2009, Zhang, 2013]. Providing a clearer picture of possible dynamic structural changes, rolling window tests have been widely used to mitigate this issue in many kinds of models, [Zivot and Wang, 2003, Diebold and Yilmaz, 2009, Mylonidis and Kollias, 2010, Diebold and Yilmaz, 2012, Diebold and Yılmaz, 2014, Guidi and Ugur, 2014, Papież and Śmiech, 2015, Chen, Mantegna, Pantelous, and Zuev, 2018, Shi, Phillips, and Hurn, 2018, Ji, Zhang, and Zhao, 2020]. In a rolling window scheme, the underlying assumption is that the parameters in a short time interval can be considered time-invariant. Statistical tests are conducted in a sub-interval with a fixed length at the beginning of the sample, moving the location of sub-intervals forward. With the dynamic tests over sub-intervals, the shift in one period does not mislead the overall picture. In the presence of structural changes, the shifts in the underlying data-generating mechanism would be reflected by the test statistics of the sub-intervals around those change points.

The rolling window technique has been broadly utilized in the Granger causality analysis [Balcilar, Ozdemir, and Arslanturk, 2010, Balcilar and Ozdemir, 2013, Lu, Hong, Wang, Lai, and Liu, 2014, Ming-Hsien and Chih-She, 2015, Nyakabawo, Miller, Balcilar, Das, and Gupta, 2015, Shi, Phillips, and Hurn, 2018]. Granger causality, first proposed by Granger [1969] is a popular concept in econometrics, involving a relation in predictions instead of the general idea of cause-and-effect. A variable X is said to *Granger-cause* a variable Y if involving past values of the former in the forecasting equation helps reduce the forecast error of the latter variable. Granger causality not only catches the causation between variables but also detects the directions. The strength and direction of causal relationships may change over time due to exogenous events, such as global financial crises and new government policies. Since these change points and effect durations are often not evident, it is necessary to apply the rolling window technique, which allows reliable assessments of the causal links regardless of the existence of the possible structural changes. If the casual linkages stay constant over time, estimates of different sub-intervals would remain approximately the same, whereas if the casual linkages are unstable, the changes could be detected.

The rolling window Granger causality method produces well-grounded evaluations, especially in the area where structural changes are suspected. Previous studies have employed the rolling window approach to examine the existence of Granger causality. For example, Aaltonen and Östermark [1997] employed a rolling F-test to assess the causal relationship between Finnish and Japanese securities markets. Balcilar et al. [2010] applied the rolling sub-samples to analyze the causal links between energy consumption and economic growth for G-7 countries. Lu et al. [2014] investigated the weighted sum of rolling sample cross-correlation between standard residuals of two series to test the time-varying causality among global crude oil markets. Ming-Hsien and Chih-She [2015] used rolling window estimation to measure the causal link between exports and GDP growth in China and Taiwan. Shi et al. [2018] examined and identified changes in causality based on the forward window, rolling window, and recursive window, concluding that both the rolling window and the recursive window work smoothly.

Although the rolling window strategy has been widely studied, the intrinsic multiplicity issue has not been thoroughly discussed yet. The number of sub-intervals in a rolling window method is often very large. A test is conducted on every sub-interval, which leads to a large number of p-values. If one uses the same threshold value as a single hypothesis test for all these p-values, the probability of false rejection would be higher than intended. However, many studies have not considered the multiplicity issue, using the same cutoff value from a single hypothesis test for all underlying p-values. For instance, Aaltonen and Östermark [1997] and Lu et al. [2014] reported the raw test statistics for each window without any adjustment, which may be associated with a higher type I error. This problem was recognized in some previous studies. For example, Cai et al. [2020] corrected the confidence interval width with respect to different rolling window lengths by deriving the asymptotic limiting distribution of rolling regression estimators. A more popular approach to maintain the overall error rate is the bootstrap-based method. The basic rule of a s-level bootstrap hypothesis testing is generating a large number of bootstrap samples under the null hypothesis, and comparing the $(1 - s)$th quantile of bootstrap test statistics with the observed test statistic. In the field of Granger causality testing, Balcilar et al. [2010] used the mean adjusted Ordinary Least Square (OLS) residual-based bootstrap p-values of observed likelihood ratio test statistic to test the null hypothesis of no causality between integrated variables. Shi et al. [2018] also employed a residual-based bootstrap to tackle the multiplicity problem in the rolling-window Granger-causality test.

Combining multiple hypotheses and controlling the multiplicity issue is of substantial interest in the genome-wide association studies field. Some ideas can be borrowed and applied to the economics field. Among all possible methods, two additive p-value combination tests, Cauchy transformation p-values combination (CCT) [Liu and Xie, 2020] introduced in Chapter 2 and Fréchet combination test (FCT) proposed in Chapter 3 are evaluated and compared with a bootstrap technique proposed by Shi et al. [2018, 2020].

In this paper, we propose an alternative approach to settle the multiplicity issue besides the bootstrap technique. We demonstrate this approach in the context of the rolling window Granger causality test but it can be easily generated in other settings as well. Our intention is to combine the large number of p-values from all sub-intervals into one combined p-value. This method is easier to implement and much faster in computation, compared to bootstrap. Our main contribution is that we not only greatly reduce the computing time but also obtain slightly better results.

The chapter is constructed as follows. Section 4.2 introduces the settings of Granger causality tests and the multiplicity issue caused by rolling windows. Section 4.3 presents and compares the simulation results of various tests in finite samples. Section 4.4 present an application of p-value combination tests on rolling window Granger causalities between the stock market and the market uncertainty due to COVID-19. Section 4.5 concludes lastly.

4.2 Rolling Window Granger Causality Test

In this section, we introduce the setting of the bivariate Granger causality F test and discuss the multiplicity issue caused by multiple windows. Then we state several additive p-value combination methods and propose to use them as an alternative to the bootstrap technique.

Granger causality, first proposed by Granger [1969], is a statistical concept that describes a prediction relation. A time series x_t is said to Granger-cause y_t if the mean square error of a forecast of the future value of y_t decreases when the past values of x_t are involved. A simple approach for the Granger causality test uses the autoregressive specification. We consider full and reduced autoregressives with lag length k, for $t = 1, \dots, T$,

$$y_t = c_1 + \sum_{i=1}^{k} a_i y_{t-i} + \sum_{i=1}^{k} b_i x_{t-i} + \varepsilon_t,$$

and

$$y_t = c_0 + \sum_{i=1}^{k} d_i y_{t-i} + e_t.$$

The null hypothesis of no Granger causality from x_t to y_t is equivalent to the zero coefficient of lagged values of x_t in the full model. That is,

$$H_0 : b_1 = b_2 = \cdots = b_k = 0.$$

Suppose ε_t is an i.i.d. Gaussian disturbance, we will conduct an F test. With OLS estimations, we calculate the test statistic

$$S = \frac{(RSS_0 - RSS_1)/k}{RSS_1/(T-2k-1)},$$

where $RSS_1 = \sum_{i=1}^{T} \hat{\varepsilon}_t^2$ is the sum of squared residuals from the full model and $RSS_0 = \sum_{i=1}^{T} \hat{e}_t^2$ is the sum of squared residuals from the reduced model. If S is greater than the $(1-s)$ critical value of an $F(k, T-2k-1)$ distribution or the p-value $\Pr(S > F_{k,T-2k-1})$ is less than s, we reject the null hypothesis and conclude that x_t does Granger cause y_t at s significance level. Reversely, the Granger causality from y_t to x_t can also be examined.

The above procedure describes the test on a full interval. In the rolling procedure, the F test is conducted on each sub-interval as the testing window moves forward. Suppose the window length is fixed at m, the ith window contains data points $(x_i, y_i), \ldots, (x_{i+m-1}, y_{i+m-1})$, where $i = 1, \ldots, T-m+1$. Compute the F test on ith window and denote the test statistic as S_i. Then the corresponding p-value is $p_i = \Pr(S_i > F_{k,m-2k-1})$. The global null hypothesis that one variate does not Granger cause another is an intersection of multiple hypotheses that the causal link does not exist in any rolling sub-intervals. However, if the raw test statistics or p-values are executed solely, the multiplicity problem occurs. The probability of falsely finding at least one significant result is inflated. In the independent case, the FWER is given by

$$\text{FWER} = 1 - (1-s)^{T-m+1},$$

which is much greater than s if the number of tests is large. In the rolling window procedure, the neighboring tests are highly positively correlated, which makes the FWER falls somewhere between s and $1-(1-s)^{T-m+1}$. To maintain the overall error rate and obtain high power simultaneously, we present and compare a bootstrap technique and additive p-value combination methods in the following.

When the analytic distribution of the test statistic is not feasible, bootstrap is a powerful replacement. White [2000] proposed and proved a bootstrap reality check strategy to approximate the p-value of multiple hypotheses. The null hypothesis is an intersection of the one-sided individual hypotheses where neither of the individual models has predictive superiority over a benchmark model. Shi et al. [2018, 2020] proposed a residual-based bootstrap algorithm to test the sup Wald statistic in the sense of forward, rolling, and recursive windows. Although the model hypotheses of White [2000] and Shi et al. [2018, 2020] are different, they share the same spirit of comparing the best of multiple models with the quantiles of the bootstrap statistics.

We use the residual-based bootstrap method proposed by Shi et al. [2018, 2020]. The test statistic for the global null hypothesis is calculated as the maximum value of the F statistics, $\max\{S_1, \ldots, S_{T-m+1}\}$. The bootstrap samples are generated from a restricted model with the null hypothesis imposed and the bootstrap residuals are randomly drawn with replacements from the estimation errors. By construction, the series of bootstrap test statistics is the emulated sampling distribution of the test statistic under the global null hypothesis. The global null hypothesis is rejected if the test statistic is greater than the 95% percentile of the bootstrap statistics series at the 0.05 significance level.

Instead of involving test statistics, a bunch of multiple hypotheses tests address the multiplicity issue by checking the p-values. For example, the most simple and widely used approach is Bonferroni corrections. Many other methods are proposed aiming to improve the power. However, many multiple hypotheses methods cannot be utilized to handle rolling windows. For example, Stouffer et al. [1949], Simes [1986], and Van der Sluis et al. [2013] do not work because of the strong dependencies. Robustness to strong dependence is essential because nearby windows are always highly correlated. Some recently developed additive p-value combination methods are robust to strong dependence structure among the p-values, which makes them possible to tackle the multiplicity issue for rolling windows.

Additive p-value combination tests are a set of methods that transform the underlying p-values into new random variables and then further linearly combine them. The p-values are usually "combined" following a simple recipe, so it is easy to implement for a practitioner. These methods include the HMP [Wilson, 2019], GMP [Wilson, 2020], CCT [Liu and Xie, 2020], SCT introduced in Chapter 2, and UCT introduced in Chapter 3. The aforementioned methods all require the validity of the p-value in each sub-interval, that is, the p-values follow a uniform distribution. SCT ensures the in-distribution relation under some mixing conditions, which regulates long-range and short-range dependencies. In the case of the rolling window, there is a natural order among the individual tests, and thus the conditions are easily satisfied [Ling and Rho, 2022]. Besides, SCT has good performance from finite sample simulation when the correlation between tests is decaying exponentially. Rolling windows also have correlations decay fast, so SCT will be a good choice for the rolling window multiple tests. Within the framework of the UCT, a new method called Fréchet combination test (FCT) was proposed in Chapter 3. It is close to GMP since the transformation functions $[-\log(1-p)]^{-1/\alpha}$ is close to $p^{-1/\alpha}$ when p is small. FCT also has good finite sample simulation performance when the correlation between tests is decaying exponentially.

Among all of these methods, we utilize the CCT and FCT to test the Granger causality in rolling samples because of their robustness to dependence and high statistical powers. The FCT involves a user-chosen tail parameter. We set it to be one because of its simplicity and robustness to arbitrary dependence structure. Let T_C and T_F denote the combined test statistics of CCT and FCT, respectively. It is reasonable to set equal weights for each rolling window, so the test statistics are calculated as the averages of transformed individual p-values $p_1, \ldots, p_{T-m+1}$,

$$T_C = \frac{1}{T-m+1} \sum_{i=1}^{T-m+1} \tan\{(0.5 - p_i)\pi\},$$

and

$$T_F = \frac{1}{T-m+1} \sum_{i=1}^{T-m+1} \frac{-1}{\log(1-p_i)} - \left[\log(T-m+1) + 1 - 2c_E\right],$$

where $c_E \approx 0.57721$ is the Euler's constant. Under the null hypothesis of no Granger causality, T_C either has the same upper tail probability under the assumption of [Liu and Xie, 2020] or has a standard Cauchy distribution under the conditions in Chapter 2. Although the assumptions and asymptotic results of Liu and Xie [2020] and Chapter 2 are different, the implementation is the same. Under the null, the tail probability of T_F is the same as stable distribution $\boldsymbol{S}(1, 1, \pi/2, 0)$. Given the tail probability or distribution of the combined test statistic under the null hypothesis, it's easy to obtain the combined p-value from the cumulative distribution function.

4.3 Simulations

This section presents the finite sample simulations in a bivariate VAR model under similar settings to Shi et al. [2018]. The simulated performance of bootstrap and additive p-value combination methods are compared under different sample sizes, different rolling window lengths, and different values of parameters.

Suppose the data-generating procedure is a restricted bivariate VAR model with lag one and no intercept. Only the causal link from x_t to y_t is considered and the other link is set to 0 for simplicity. The model is

$$\begin{bmatrix} y_t \\ x_t \end{bmatrix} = \begin{bmatrix} \phi_{11} & \phi_{12}(t) \\ 0 & \phi_{22} \end{bmatrix} \begin{bmatrix} y_{t-1} \\ x_{t-1} \end{bmatrix} + \begin{bmatrix} u_{1,t} \\ u_{2,t} \end{bmatrix}, \quad t = 1, \ldots, T$$

where $u_{1,t}$ and $u_{2,t}$ are i.i.d. $N(0, 1)$, ensuring the F tests are exact. The existence of Granger causality from x_t to y_t is determined by $\phi_{12}(t)$, which is a function of t. Under the null hypothesis, $\phi_{12}(t) = 0$, indicating there is no Granger causality at any t. Under the alternative hypothesis, $\phi_{12}(t) = \phi_s I_D$, where ϕ_s is a constant taking a value between 0 to 1, corresponding to different causality strengths, and I_D is an indicator function showing when the causality takes place or disappears. As a result, the Granger causality does not always exist but switches on in part D under the alternative hypothesis. In the simulation, the duration D is taken to be

$[0.5T, 0.7T]$. We consider three pairs of the coefficients (ϕ_{11}, ϕ_{22}) = (0.5, 0.5), (-0.5, 0.8), and (0.5, -0.8), two different sample sizes $T = 100$ and 300, and two casual link strengths $\phi_s = 0.5, 0.8$. Let $f_0 = m/T$ be the proportion of rolling window length. We consider six values of $f_0 = 0.18, 0.24, 0.30, 0.36, 0.42$, and 0.48. The number of replication is 1000 for each circumstance.

In the ith window, the F test statistic S_i follows a $F(1, m-2)$ distribution under the null hypothesis. The rolling algorithm of a residual-based bootstrap technique in Shi et al. [2018, 2020] is conducted for purpose of comparison. The number of bootstraps is 500. It's worth pointing out that Shi et al. [2018] set the bootstrap sample size to be $T + m - 1$ in order to control the size over the entire sample period. However, our unreported simulation results suggest that it deflates the error rate. As a result, we set the bootstrap sample size as T with the purpose to preserve the same number of rolling windows. Since the disturbances are generated from a normal distribution, the F tests are exact, and thus the corresponding p-values are uniformly distributed under the null hypothesis, which satisfies the requirement of the additive combination tests we utilized.

The proportion of rejections out of 1000 replications under the null ($\phi_{12}(t) = 0$) and two alternatives ($\phi_s = 0.5, 0.8$) are reported in Tables 4.1 and 4.2. The FCT and HMP have almost the same results, so only the FCT is presented. The CCT and FCT have similar results with well-controlled sizes and comparable powers to bootstrap in most cases. Under the setting (ϕ_{11}, ϕ_{22}) = (-0.5, 0.8), bootstrap and FCT have slightly conservative sizes when $T = 100$, and the under-rejections are relieved when T increased to 300. Although the size of CCT also increases slightly when T increases, the sizes of CCT under this setting are close to the target value. When $T = 100$, the powers when $\phi_s = 0.5$ and 0.8 of FCT beat CCT in all rolling window lengths despite comparable error rates. CCT even has higher powers than FCT. When $T = 300$, FCT and bootstrap still have close error rates. When $\phi_s = 0.5$, FCT has the highest power with small windows (f_0 = 0.18, 0.24, 0.36), and CCT has

	Bootstrap			CCT			FCT		
T=100, (ϕ_{11}, ϕ_{22}) = (-0.5,0.8)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.034	0.483	0.811	0.041	0.549	0.86	0.036	0.548	0.849
0.24	0.044	0.546	0.854	0.044	0.596	0.872	0.042	0.58	0.866
0.30	0.042	0.521	0.807	0.052	0.571	0.836	0.041	0.552	0.826
0.36	0.042	0.504	0.774	0.047	0.565	0.818	0.048	0.55	0.799
0.42	0.043	0.505	0.753	0.052	0.569	0.793	0.041	0.535	0.77
0.48	0.042	0.49	0.72	0.056	0.541	0.773	0.039	0.513	0.75
T=100, (ϕ_{11}, ϕ_{22}) = (0.5,0.5)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.053	0.259	0.622	0.074	0.365	0.743	0.067	0.354	0.734
0.24	0.048	0.316	0.692	0.065	0.393	0.748	0.059	0.371	0.742
0.30	0.044	0.29	0.659	0.063	0.381	0.724	0.051	0.351	0.707
0.36	0.041	0.319	0.638	0.061	0.37	0.702	0.05	0.356	0.682
0.42	0.05	0.32	0.614	0.062	0.382	0.673	0.052	0.349	0.648
0.48	0.049	0.313	0.611	0.06	0.366	0.652	0.05	0.335	0.614
T=100, (ϕ_{11}, ϕ_{22}) = (-0.5,-0.8)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.043	0.507	0.839	0.047	0.522	0.862	0.044	0.513	0.859
0.24	0.05	0.571	0.868	0.052	0.584	0.87	0.044	0.566	0.866
0.30	0.058	0.559	0.841	0.051	0.574	0.847	0.045	0.552	0.844
0.36	0.056	0.547	0.798	0.057	0.579	0.809	0.049	0.543	0.793
0.42	0.048	0.536	0.766	0.054	0.556	0.79	0.047	0.53	0.772
0.48	0.047	0.515	0.76	0.05	0.544	0.778	0.041	0.502	0.752

Table 4.1: Size and power of bootstrap, CCT, and FCT when sample size $T = 100$.

	Bootstrap			CCT			FCT		
T=300, (ϕ_{11}, ϕ_{22}) = (-0.5,0.8)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.046	0.983	1	0.051	0.983	1	0.048	0.985	1
0.24	0.047	0.976	1	0.051	0.981	1	0.042	0.982	1
0.30	0.04	0.957	1	0.049	0.956	1	0.039	0.958	1
0.36	0.043	0.938	0.997	0.049	0.946	0.996	0.043	0.941	0.997
0.42	0.046	0.925	0.995	0.055	0.935	0.996	0.043	0.925	0.994
0.48	0.056	0.913	0.992	0.058	0.921	0.993	0.047	0.913	0.991
T=300, (ϕ_{11}, ϕ_{22}) = (0.5, 0.5)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.061	0.858	0.997	0.067	0.876	0.998	0.065	0.883	0.997
0.24	0.056	0.871	0.998	0.065	0.88	0.998	0.054	0.876	0.998
0.30	0.051	0.828	0.995	0.062	0.845	0.995	0.053	0.837	0.997
0.36	0.056	0.788	0.985	0.066	0.821	0.991	0.053	0.797	0.986
0.42	0.061	0.77	0.977	0.06	0.809	0.984	0.05	0.78	0.979
0.48	0.061	0.751	0.97	0.056	0.782	0.978	0.048	0.754	0.974
T=300, (ϕ_{11}, ϕ_{22}) = (0.5,-0.8)									
f_0	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$	$\phi_{12} = 0$	$\phi_s = 0.5$	$\phi_s = 0.8$
0.18	0.043	0.98	1	0.042	0.982	1	0.039	0.984	1
0.24	0.046	0.985	1	0.042	0.982	1	0.036	0.979	1
0.30	0.045	0.971	1	0.042	0.969	1	0.035	0.969	0.999
0.36	0.047	0.954	0.996	0.043	0.951	0.998	0.035	0.951	0.997
0.42	0.045	0.93	0.993	0.05	0.941	0.996	0.039	0.927	0.995
0.48	0.047	0.914	0.988	0.051	0.923	0.99	0.044	0.907	0.99

Table 4.2: Size and power of bootstrap, CCT, and FCT when sample size $T = 300$.

the highest powers with larger windows (f_0 = 0.36, 0.42, 0.48). When $\phi_s = 0.8$, the powers are all close to 1.

Under the setting (ϕ_{11}, ϕ_{22}) = (0.5, 0.5), the error rates tend to be inflated but power is deflated compared with other parameter settings. when $T = 100$ and 300, CCT has inflated sizes for all rolling sample lengths f_0, and FCT has inflated size only when f_0 is small. Bootstrap controls the size well at the target value when $T = 100$ but suffers from over-rejections when $T = 300$. The over-rejections of CCT and FCT tend to be relieved as the window lengths get larger. Such a trend is not observed for bootstrap when $T = 300$. When $T = 100$, CCT has the highest powers and bootstrap has the lowest powers, which is reasonable because CCT has the highest error rates and bootstrap has the lowest error rates. However, when $T = 300$ this trend still exists while bootstrap does not have the smallest error rates anymore. For example, when $f_0 = 0.24, 0.36, 0.42, 0.48$, FCT has lower error rates and higher powers than bootstrap.

Under the setting (ϕ_{11}, ϕ_{22}) = (0.5, -0.8), bootstrap and CCT control the sizes well when $T = 100$ and 300. FCT controls the size well when $T = 100$ but under-rejects when $T = 300$. When $T = 100$, CCT has higher power than bootstrap in all windows even though the error rate of CCT is less than bootstrap when $f_0 = 0.3$. Bootstrap has higher powers than FCT in all cases when $\phi_s = 0.5$ except $f_0 = 0.18$. In some cases of $\phi_s = 0.8$, the powers of bootstrap are lower than FCT ($f_0 = 0.18, 0.3, 0.42$). When $T = 300$ and $\phi_s = 0.5$, there is no method that works best in all cases. When $f_0 = 0.18$, FCT has the smallest error rate with the highest power. When $f_0 = 0.24$, 0.3, 0.36, bootstrap has the highest powers. When $f_0 = 0.42, 0.48$, CCT has the highest powers. In the case of $\phi_s = 0.8$, all of the bootstrap, CCT, and FCT have power close to 1, while CCT is slightly higher than the other two methods when the window length is large.

Furthermore, we can observe that a smaller fraction of window lengths behave better from Table 4.1 and 4.2. For example, when $T = 100$, (ϕ_{11}, ϕ_{22}) = (0.5, -0.8) and (-0.5, 0.8), $f_0 = 0.24$ produces the best performance in all methods. When (ϕ_{11}, ϕ_{22}) = (0.5, 0.5), $f_0 = 0.24$ has the highest power in all methods except that $\phi_s = 0.5$ of bootstrap. When $T = 300$ and (ϕ_{11}, ϕ_{22}) = (-0.5, 0.8), $f_0 = 0.18$ results in best error rates and powers in all methods. when (ϕ_{11}, ϕ_{22}) = (0.5, 0.5), $f_0 = 0.24$ produces the best power for bootstrap and $f_0 = 0.24$ produces the best power for both CCT and FCT. When (ϕ_{11}, ϕ_{22}) = (0.5, -0.8), bootstrap achieves its best power when $f_0 = 0.24$, CCT achieves the same highest power when $f_0 = 0.18$ and 0.24, and FCT achieves its highest power when $f_0 = 0.18$. In practice, the choice of window lengths affects the size and power. The optimal choice usually depends on the periodicity of the data. Longer rolling window sizes tend to yield smoother estimates than shorter sizes. When the individual tests are not exact, a longer size is required to ensure that the p-values are asymptotically uniform.

4.4 Application

In this section, we conduct a sample application on the dynamic Granger causality analysis between the stock market and the market uncertainty related to COVID-19. The proxy for the US stock market is the daily return of the S&P 500 index downloaded from Google Finance.

We use the Infections Disease Equity Market Volatility Tracker (IDEMV) proposed by [Baker et al., 2020] as the proxy for the market consensus of financial uncertainties due to COVID-19. The IDEMV index is constructed by counting the number of articles including keywords related to the pandemic, economy, stock market, and uncertainty across approximately 3000 US Newspapers [1]. Then, the raw counts are scaled by the total number of all articles on the same day. After this, the authors multiplicatively rescaled the resulting series to match the level of the Chicago Board

[1] The list of keywords can be founded and data can be downloaded at `https://www.policyuncertainty.com/infectious_EMV.html`.

Options Exchange's Volatility Index, by using the overall EMV index. In the final step, the series is rescaled to reflect the ratio of the IDEMV articles to the total EMV articles. Hence, this index naturally reflects society's uncertainty about the impact of COVID-19 on the stock market. The higher value, the higher uncertainty.

There are many studies investing the impact of IDEMV on financial variables. For example, Sosa-Castro [2022] analyzed the effect of the 42 category-specific EMV trackers (including IDEMV) on nine S&P 500 sectors indexes using a Neural Network approach given there was a negative and symmetric long-run relationship between the EMV and section indexes. Surprisingly the Information and Communication Technologies was the only sector that was driven by IDEMV, whereas the IDEMV was not one of the main drivers of the Health Care sector. Bouri et al. [2021] ran quantile regressions between IDEMV and the total connectedness index (an index that represents the connectedness across various assets). They found that the connectedness among gold, crude oil, world equities, currencies, and bonds is positively related to the IDEMV. Their evidence also suggested that the higher IDEMV is, the higher the market risk across the five assets. They found that the increased COVID-19 uncertainty in the US has increased the market risk across. Li et al. [2020] found that the IDEMV contained useful information in predicting the realized volatility for the France and UK stock markets but was ineffective for the German stock market during the global pandemic period.

Among the many studies of IDEMV impact, there are several papers examining the effect of IDEMV on the stock markets in the Granger causality sense. For example, Coronado et al. [2021] found the existence of the causality from IDEMV to the volatility of five of the most important Latin American stock and exchange markets. They used both the classical Granger causality test and a time-varying Granger multivariate causality test of Ajmi et al. [2015]. They also highlighted the importance of having IDEMV as a new indicator for financial agents. Cepni et al. [2022] examined both a classical constant parameter Granger causality test and a time-varying robust Granger causality test of Rossi and Wang [2019]. They investigated the adverse impact of IDEMV on the overall economy and the more significant impact on the tourism and hospitality industry. Most of the above papers used time-varying models where parameters were treated as a function of time. However, the rolling window approach has the advantage of tremendous simplicity and coherence as claimed by Diebold and Yılmaz [2014]. In addition, we consider bi-directions between IDEMV and stock market return, while all of the above papers studied only the impact of IDEMV on financial variables. The analysis of the direction from the stock market to IDEMV makes sense because the fluctuation of stock markets naturally impacts society's expectations of volatility associated with COVID-19.

The sample period is from July 15th, 2020 to April 08th, 2022, in a total of 438 business days. In order to fit in a stationary VAR model, the unit root test suggests taking the first difference of log prices, which is known as the daily return. The daily S&P 500 closing price and return are plotted in Figure 4.1(a) and (b). We take the log transformation of the IDEMV index and then subtract its average value to get the centering log data. The unit root test suggests this series is stationary. The original IDEMV index and its demean log series are plotted in Figure 4.1(c) and (d), respectively.

In the full sample analysis, the selected lag is 2 with the smallest BIC -10.41. In the direction from IDEMV to the stock return, the F test statistic and the corresponding p-value are 0.0276 and 0.9727, respectively. The critical value derived from F distribution is 3.006 and the value derived from bootstrap is 3.0256. In the other direction from the stock return to IDEMV, the F test stat and the corresponding p-value are 1.3347 and 0.2638 respectively. The critical value derived from F distribution is still 3.006 and the value derived from bootstrap is 3.5910. In full sample analysis, there's no significant evidence of any causal relations.

The dynamic Granger causality analyses are conducted with rolling window sizes m = 10, 22, 66, 132, 264, and 438 corresponding to two weeks, one month, three

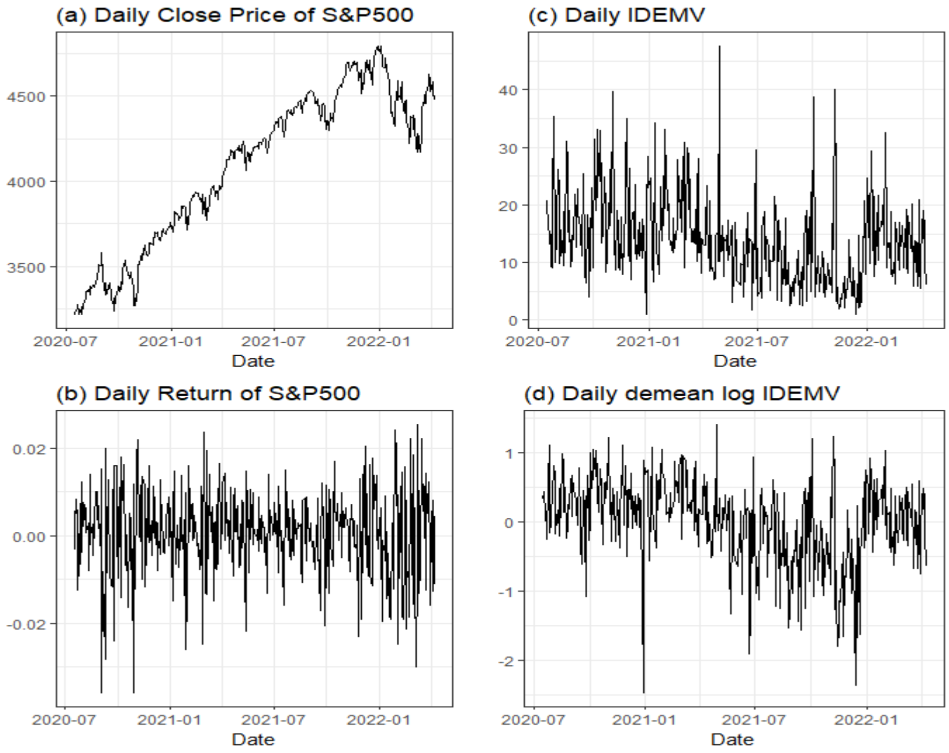

Figure 4.1: Time series plots of (a) daily S&P 500 closing price, (b) daily return of S&P 500, (c) IDEMV index, and (d) IDEMV index after log transformation and subtracting the mean value.

	IDEMV G.C. Return			Return G.C. IDEMV		
m	FCT	CCT	Bootstrap TS(CV)	FCT	CCT	Bootstrap TS(CV)
10	**0.0000**	**0.0000**	**557.8971(182.7373)**	0.2252	0.2327	18.2102(169.3577)
22	**0.0048**	**0.0047**	13.9862(26.7586)	**0.0080**	**0.0076**	14.5486(28.8266)
66	0.5192	0.1994	6.7535(11.0782)	0.0743	0.8413	7.8076(12.2330)
132	0.0797	0.0582	**9.8925(9.1393)**	**0.0329**	**0.0301**	**11.4755(9.3835)**
264	1.0000	0.7398	1.5640(6.8791)	0.1907	0.0878	5.7752(7.0138)

Table 4.3: Results of directional Granger causality bootstrap and additive combination tests with different rolling window sizes.

months, a half year, one year, and the entire sample period. We set up the above choices because we don't have theoretical guidance to choose the optimal window length. In each rolling sample, the daily return and demean log IDEMV index are fitted into a stationary bivariate VAR model without intercept. The lag order is selected by the smallest BIC with a maximum lag length of 5 when the window size is greater than 10. The maximum lag length is set to 2 when the window size is 10. After selecting the lag order, a Granger causality F test as described in Section 4.2 is conducted and leads to an F test statistic and a corresponding p-value. In the next step, we handle the multiplicity issue with bootstrap and additive p-value combination as discussed in Section 4.2.

Table 4.3 presents the results of directional causal links between the daily return of S&P 500 and demean log IDEMV index under different rolling window sizes m. The left three columns under the name "IDEMV G.C. Return" show the outcomes of the global null hypothesis that IDEMV doesn't Granger cause the stock market return in any sub-samples. The right three columns with the name "Return G.C. IDEMV" display the outcomes for the null hypothesis that the stock market return doesn't Granger cause IDEMV. The "FCT" and "CCT" stand for combined additive p-values. "TS" and "CV" are short for combined test statistics and critical values with a 0.05 significance level from the bootstrap approach. The significant entries are in bold with a 5% level.

Bootstrap and additive p-value combinations give different results in some cases, although they have the same peaks in unreported dynamic plots. The results of FCT and CCT are always consistent. In the direction from IDEMV to Return, additive combinations deliver significant results when window sizes are small (10, 22) and insignificant results when window sizes are larger. But this trend is not observed in the other direction or in the bootstrap results. Different rolling window sizes have different significant results. This may be caused by sampling errors. When the window size is 10, the effect of IDEMV on return is significant while the other direction is not. When the window size is 22, additive p-value combination approaches suggest that both directions are significant, whereas the critical values calculated from bootstrap are not significant in either direction. When the window size is 66 or 264, no significant result is detected by either method. When the window size is 66 in the direction from the stock return to IDEMV, although FCT and CCT have consistently significant results, the two combined p-values differ a lot. This is the consequence that the maximum p-value 0.99987 is too close to 1. CCT is more sensitive to large p-values than FCT. When the window size is 132, bootstrap suggests that both directions are significant. However, p-value combinations suggest that only the direction from the stock return to IDEMV is significant.

In the rolling window analysis, most lags are selected to be 1 or 2. When the window length is 10, 62% sub-intervals have lag order 1 and the remaining have 2. When window sizes are 66 and 132, over 90% of sub-intervals have lag order 1. When the window size is 264, 70% of lags are 1 and 30% of lags are 2. However, there is a problem when the window size is 22. More specifically, 75% (312 out of 417) of sub-intervals have lag order 1, 12% (51 out of 417) of sub-intervals have lag order 2, 1% (5 out of 417) have lag order 3, 4% (15 out of 417) have lag 4, and the remaining 8% (34 out of 417) have larger lag order 5. Because larger lag orders are selected, the results are not robust to the manually set maximum lag order. If the maximum lag is set to be 4 or 5, both FCT and CCT suggest the rolling tests are significant. But if the maximum lag is set to less than 3, both FCT and CCT are not significant. The empirical test of Granger causality is sensitive to the choice of lag length Hamilton [2020], Chuen [2015].

After uncovering the existence of causalities, we want to locate the significant subsamples. Bootstrap naturally locates the significant subsample by comparing the elementary F test statistic series $\{S_1, \dots, S_{T-m+1}\}$ with the critical value computed by bootstrap samples. If there exists any S_i that exceeds the critical value, then the global null is rejected and the ith sample significantly observes Granger causality. Figures 4.2 present the dynamic F test statistic for different window lengths and directions. The y-axis is the F test statistic verse the x-axis is the ending date of a rolling subsample. Each point represents the F test statistic estimated from a rolling subsample. The dashed horizontal lines indicate the critical values derived from the bootstrap sample. If a point exceeds the critical line, the x-coordinate is the ending date of a subsample where significance is identified by bootstrap. The left column of Figure 4.2 shows the test statistics and their corresponding critical values for the Granger causalities direction from IDEMV to stock return. Except for (a2), the highest points in the direction of IDEMV causes return all happen in November of 2021 in (a1), (a3), (a4), and (a5). The right column of Figure 4.2 shows the test statistics and their corresponding critical values in the direction from the stock return to IDEMV. Most graphs except (b5) have a peak in July 2021. In particular, (b4) is significant at the end of July 2021 and the start of August 2021.

In the context of additive p-value combination, the date-stamp requires the method to be able to control the FWER in the strong sense for any configuration of true and false elementary null hypotheses. As discussed in Remark 3.3 in Chapter 3, FCT is valid for a multilevel test, whereas CCT is not for a computationally feasible shortcut.

The elementary p-value p_i from one subsample is significant if for $i = 1, \ldots, T-m+1$,

$$\frac{-1}{\log(1-p_i)} > (T-m+1)s_{0.95} + (T-m+1)[\log(T-m+1) + 1 - 2c_E],$$

where $s_{0.95}$ is the 0.95 quantile of $\boldsymbol{S}(1, 1, \pi/2, 0)$. Similarly as bootstrap, we present Figure 4.3 to show the dynamic transformed values of elementary p-values. The y-axis represents $\frac{-1}{\log(1-p_i)}$ and the x-axis is the ending date of a rolling subsample as well. The dashed horizontal lines are the threshold values over which the elementary p-value is significant in the strong sense. If any point exceeds the threshold, causality exists significantly in this subsample. The exact ending dates of significant subsamples for both bootstrap and FCT are displayed in Table 4.4 for clear comparison. The em-dashes in cells indicate that no significant subsample is located.

Overall, FCT suggests more significant subsamples in shorter rolling windows and bootstrap has more in longer windows. As reported in Table 4.4, when the window size is 10, both FCT and bootstrap recognize 2021-11-29 and 2021-11-30 are significant ending dates, which means that the IDEMV Granger causes the S&P 500 index return during the last half month of November 2021. The show-up of causality from IDEMV to return might result from the approval of booster and the occurrence of Omicron. FDA approved COVID-19 boosters for all adults on Nov 19, which might affect the causality structure. The World Health Organization classified a new variant on Nov 26 2021 and named it Omicron on Nov 30, 2021. The first confirmed case of the Omicron variant was detected in the US on Dec 1, 2021. There were two more causal links during the first half month of both February 2021 and November 2021 in the direction from IDEMV to return suggested by FCT but not noticed bootstrap. Pfizer asked for booster emergency use authorization for all adults on Nov 9. This news might affect the causality structure by releasing more confidence in the stock market and investors. The coefficient of the lag value of IDEMV on return is -0.00535 during that window. This suggests that a high past value of market uncertainty has a negative impact on the return. In contrast, the coefficient of the lagged value of the return is 0.8655. On Feb 1 2021 it was announced that more Americans were

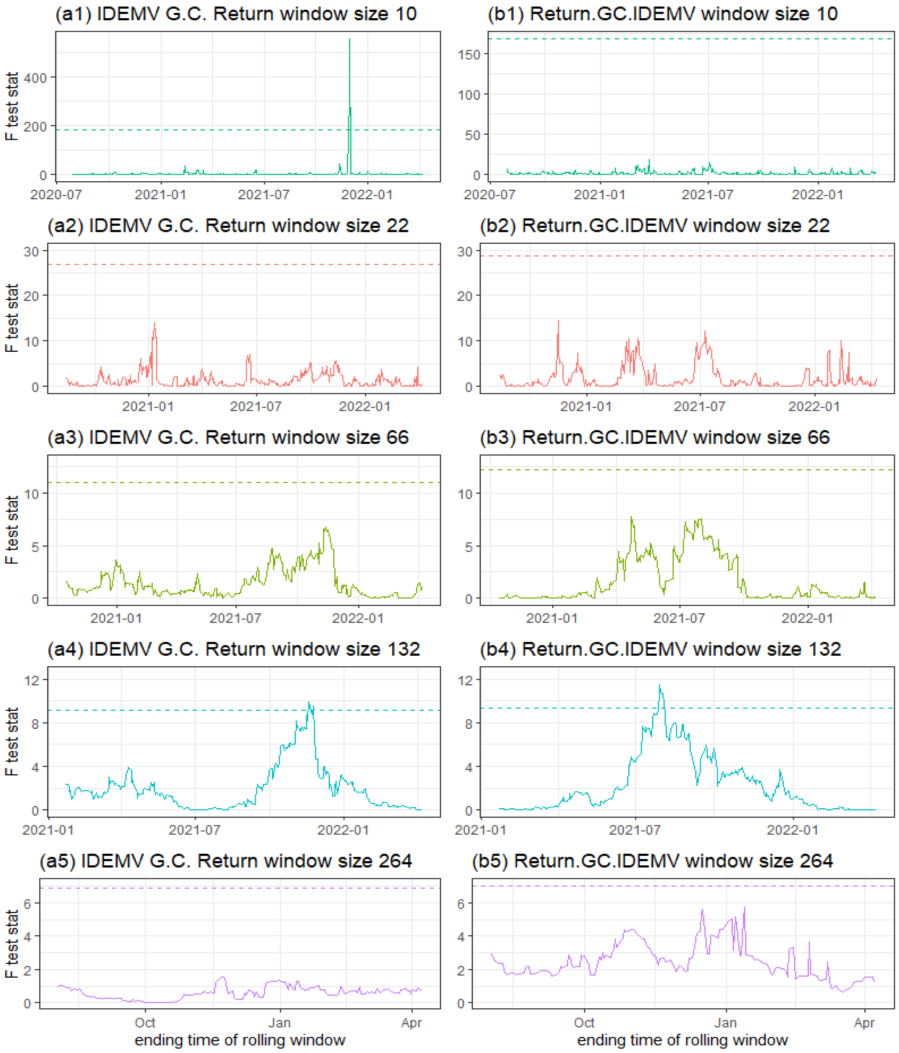

Figure 4.2: Dynamic plot of the F test statistics and critical values for different rolling window lengths.

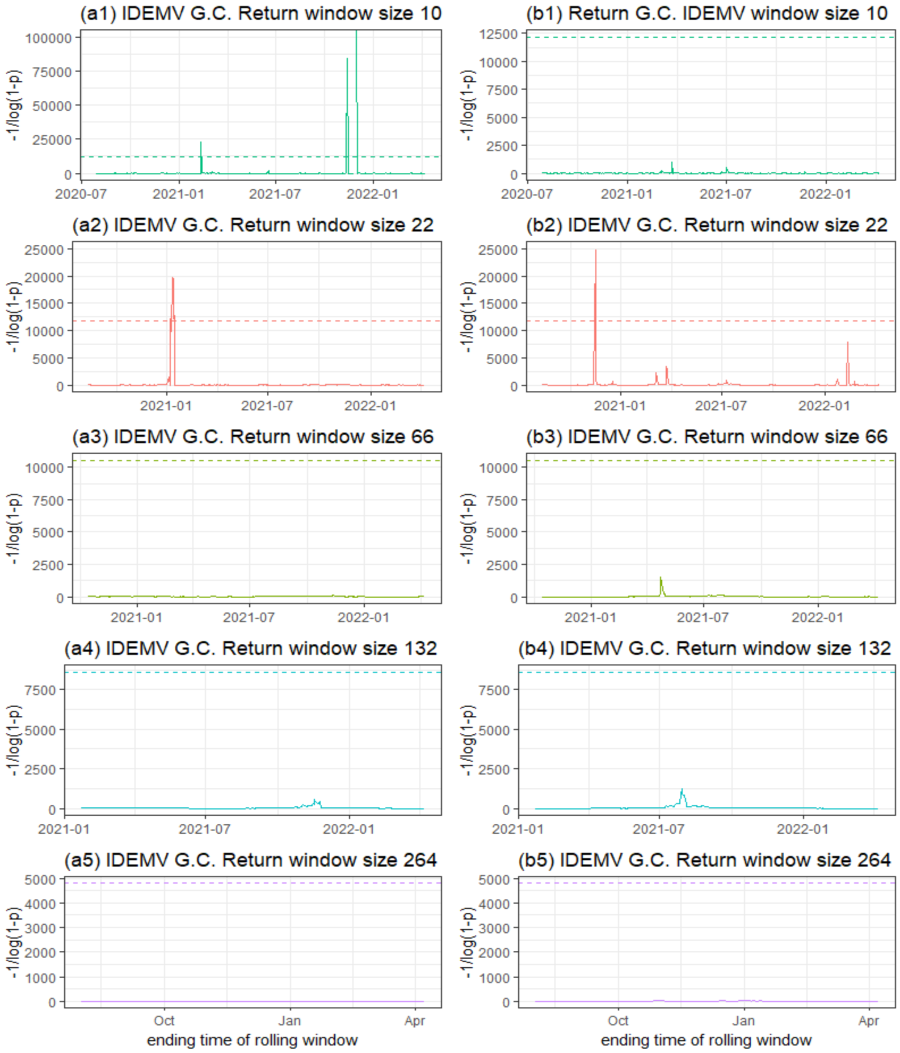

Figure 4.3: Dynamic plot of transformed p-values in each rolling windows.

	IDEMV G.C. Return		Return G.C. IDEMV	
m	FCT	Bootstrap	FCT	Bootstrap
10	2021-02-11 2021-11-12 2021-11-29 2021-11-30	2021-11-29 2021-11-30	—	—
22	2021-01-07 2021-01-11 2021-01-12 2021-01-13 2021-01-14	—	2020-11-16	—
66	—	—	—	—
132	—	2021-11-18 2021-11-19 2021-11-24	—	2021-07-28 2021-07-29 2021-07-30 2021-08-02 2021-08-03
264	—	—	—	—

Table 4.4: The ending dates of significant rolling subsamples identified by FCT and bootstrap. The em-dashes in cells indicate that no significant subsample is located.

vaccinated than infected. The coefficient of return's lag value is positive at 0.018737.

When the window size is 22, FCT suggests that the causality from IDEMV to stock return exists from mid-December 2020 to mid-January 2021. The appearance of causality might be affected by FDA's emergency use authorization for the Pfizer-BioNTech vaccine on Dec 11 and for Moderna on Dec 18. The other causal direction is also advised by FCT from mid-October to mid-November 2020. In this window, the lag is selected as 5 and the coefficients of lag values of return are much larger in magnitude than the coefficients of lag values of IDEMV itself. For comparison, the coefficients of return from 1st lag to 5th lag are -6.0917, -15.4741, -23.9347, 14.6330, and 0.7438, while that of IDEMV are 0.2832, -0.2437, 0.3018, 0.05829, and 0.7654. This suggests that the return negatively impacts the market uncertainty. The 59th US presidential election held on Nov 3rd, 2020 may incur some changes in the stock market expectation and the pandemic policies. Although bootstrap has peaks at these significant dates, the threshold is not exceeded for either direction.

When the window size is 66 and 264, no elementary hypothesis is found to be significant. When the window size is 132, bootstrap indicates some significant subsamples while FCT does not. Bootstrap finds the causality from IDEMV to return exists in the period from July 2021 to November 2021. Causality exists in the other direction from March 2021 to August 2021. The combined p-value(0.0797) from the IDEMV to return is not significant, nor is not any elementary p-value. In the other direction from the return to IDEMV, although the combined FCT p-value(0.0329) is significant, no rolling window is found to be significant. Instead, a collection of 29 elementary p-values is found to be significant. The ending dates of these subsample collections range from 2021-07-09 to 2021-08-18, which is about one-month larger than the set given by bootstrap.

4.5 Discussion

We borrow the idea of additive p-values combination test from the field of multiple hypotheses testing into the dynamic structural examination in the field of econometrics. Under the context of Granger causality tests in rolling windows, each rolling window produces one hypothesis test and thus one corresponding p-value. Bootstrap is a popular technique to deal with the multiplicity issue in econometrics, despite its cost in computation. We propose to use additive p-value combination tests that are robust to dependencies to handle the multiplicity issue as an efficient replacement for bootstrap. It could be used in field of engineering under the context of analysing the relationship among several time series.

We compare the performance in finite sample simulations of additive p-value combination tests and bootstrap. It is found that CCT, HMP, and FCT are robust to dependencies and produce well-controlled sizes and comparable powers to bootstrap. Moreover, the additive p-value combination test is easier to be implemented and consumes less time than bootstrap. We conduct a case study examining the bidirectional Granger causalities between stock market return and market uncertainty due to COVID-19. Additive p-value combination methods reject more than bootstrap overall. Similar to bootstrap, FCT also has the property of identifying the significant subsample but CCT does not. However, in the application of rolling window techniques in real data, one of the mysteries is the choice of window size. Different window sizes sometimes produce different results.

However, there are some drawbacks to applying the additive p-value combination tests on the Granger causality test. First, not all p-value combination methods can make statements of individual hypotheses. For example, SCT with $\alpha \geq 1$ or $\beta \neq 1$ only controls the error rate at the weak sense when all underlying hypotheses are true. Even though some methods like FCT and HMP can examine each underlying hypothesis, it works differently from the bootstrap. In bootstrap, if and only if the global null is significant, there is at least one significant underlying hypothesis. In the context of unified additive p-value combination tests, if there is at least one significant underlying hypothesis, the global null is also significant. But the other direction is not generally true. Another disadvantage of the combination test is that the elementary p-values are required to be uniform under the null. We assume the data is normal to take the advantage of exact F test of Granger causality. However, this is a strong assumption in real data. When the data is non-normal, a Wald form of the OLS χ^2 test should be utilized. This makes the p-values not uniform anymore and violates the assumption of the additive combination tests. In this case, we require that both the sample size and window size are large enough and thus the p-values would be uniform asymptotically.

References

J. Aaltonen and R. Östermark. A rolling test of granger causality between the finnish and japanese security markets. *Omega*, 25(6):635–642, 1997.

R. P. Abelson. *Statistics as principled argument.* Psychology Press, 2012.

A. N. Ajmi, S. Hammoudeh, D. K. Nguyen, and J. R. Sato. On the relationships between co2 emissions, energy consumption and income: the importance of time variation. *Energy Economics*, 49:629–638, 2015.

S. R. Baker, N. Bloom, S. J. Davis, K. Kost, M. Sammon, and T. Viratyosin. The unprecedented stock market reaction to covid-19. *The review of asset pricing studies*, 10(4):742–758, 2020.

M. Balcilar and Z. A. Ozdemir. The export-output growth nexus in japan: a bootstrap rolling window approach. *Empirical Economics*, 44(2):639–660, 2013.

M. Balcilar, Z. A. Ozdemir, and Y. Arslanturk. Economic growth and energy consumption causal nexus viewed through a bootstrap rolling window. *Energy Economics*, 32(6):1398–1410, 2010.

J. Beirlant, Y. Goegebeur, J. Segers, and J. L. Teugels. *Statistics of extremes: theory and applications.* John Wiley & Sons, 2006.

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

M. Boehlje. Structural changes in the agricultural industries: how do we measure, analyze and understand them? *American Journal of Agricultural Economics*, 81 (5):1028–1041, 1999.

E. Bouri, O. Cepni, D. Gabauer, and R. Gupta. Return connectedness across asset classes around the covid-19 outbreak. *International review of financial analysis*, 73:101646, 2021.

T. T. Cai, W. Liu, and Y. Xia. Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 76(2):349–372, 2014.

Z. Cai, T. Juhl, et al. The distribution of rolling regression estimators. Technical report, University of Kansas, Department of Economics, 2020.

O. Cepni, T. Dogru, and O. Ozdemir. The contagion effect of covid-19-induced uncertainty on us tourism sector: Evidence from time-varying granger causality test. *Tourism Economics*, page 13548166221077633, 2022.

Y. Chen, R. N. Mantegna, A. A. Pantelous, and K. M. Zuev. A dynamic analysis of s&p 500, ftse 100 and euro stoxx 50 indices under different exchange rates. *PloS one*, 13(3):e0194067, 2018.

D. L. K. Chuen. *Handbook of digital currency: Bitcoin, innovation, financial instruments, and big data.* Academic Press, 2015.

T. E. Clark and M. W. McCracken. Improving forecast accuracy by combining recursive and rolling forecasts. *International Economic Review*, 50(2):363–395, 2009.

S. Coronado, J. N. Martinez, and R. Romero-Meza. Time-varying multivariate causality among infectious disease pandemic and emerging financial markets: the case of the latin american stock and exchange markets. *Applied Economics*, pages 1–9, 2021.

R. A. Davis and S. I. Resnick. Limit theory for bilinear processes with heavy-tailed noise. *The Annals of Applied Probability*, 6(4):1191–1210, 1996.

F. X. Diebold and K. Yilmaz. Measuring financial asset return and volatility spillovers, with application to global equity markets. *The Economic Journal*, 119(534):158–171, 2009.

F. X. Diebold and K. Yilmaz. Better to give than to receive: Predictive directional measurement of volatility spillovers. *International Journal of forecasting*, 28(1): 57–66, 2012.

F. X. Diebold and K. Yılmaz. On the network topology of variance decompositions: Measuring the connectedness of financial firms. *Journal of econometrics*, 182(1): 119–134, 2014.

A. Dmitrienko, A. C. Tamhane, and F. Bretz. *Multiple testing problems in pharmaceutical statistics*. CRC press, 2009.

F. Duchin. *Structural economics: measuring change in technology, lifestyles, and the environment*. Island Press, 1998.

W. Feller. An introduction to probability theory and its applications, vol. 1, 3rd edna wiley. *New York*, 1968.

R. A. Fisher. Statistical methods for research workers. In *Breakthroughs in statistics*, pages 66–70. Springer, 1992.

B. Geiger, J. P. Spatz, and A. D. Bershadsky. Environmental sensing through focal adhesions. *Nature reviews Molecular cell biology*, 10(1):21–33, 2009.

J. Geluk and Q. Tang. Asymptotic tail probabilities of sums of dependent subexponential random variables. *Journal of Theoretical Probability*, 22(4):871–882, 2009.

C. W. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, pages 424–438, 1969.

C. W. Granger. Can we improve the perceived quality of economic forecasts? *Journal of Applied Econometrics*, 11(5):455–473, 1996.

F. Guidi and M. Ugur. An analysis of south-eastern european stock markets: Evidence on cointegration and portfolio diversification benefits. *Journal of International Financial Markets, Institutions and Money*, 30:119–136, 2014.

J. D. Hamilton. *Time series analysis*. Princeton university press, 2020.

S. J. Han, E. V. Azarova, A. J. Whitewood, A. Bachir, E. Guttierrez, A. Groisman, A. R. Horwitz, B. T. Goult, K. M. Dean, and G. Danuser. Pre-complexation of talin and vinculin without tension is required for efficient nascent adhesion maturation. *Elife*, 10:e66151, 2021.

K. S. Henning and P. H. Westfall. Closed testing in pharmaceutical research: Historical and recent developments. *Statistics in biopharmaceutical research*, 7(2):126–147, 2015.

Y. Hochberg. A sharper bonferroni procedure for multiple tests of significance. *Biometrika*, 75(4):800–802, 1988.

J. D. Humphries, P. Wang, C. Streuli, B. Geiger, M. J. Humphries, and C. Ballestrem. Vinculin controls focal adhesion formation by direct interactions with talin and actin. *Journal of Cell Biology*, 179(5):1043–1057, 2007.

A. Jakubowski and M. Kobus. α-stable limit theorems for sums of dependent random vectors. *Journal of multivariate analysis*, 29(2):219–251, 1989.

S. Janson. Stable distributions. *arXiv preprint arXiv:1112.0220*, 2011.

H. A. Jessen and T. Mikosch. Regularly varying functions. *Publications de L'institut Mathematique*, 80(94):171–192, 2006.

Q. Ji, D. Zhang, and Y. Zhao. Searching for safe-haven assets during the covid-19 pandemic. *International Review of Financial Analysis*, 71:101526, 2020.

J. Karamata. Sur un mode de croissance régulière. théorèmes fondamentaux. *Bulletin de la Société Mathématique de France*, 61:55–62, 1933.

S. C. Kim, S. J. Lee, W. J. Lee, Y. N. Yum, J. H. Kim, S. Sohn, J. H. Park, J. Lee, J. Lim, and S. W. Kwon. Stouffer's test in a large scale simultaneous hypothesis testing. *Plos one*, 8(5):e63290, 2013.

M.-X. Li, H.-S. Gui, J. S. Kwan, and P. C. Sham. Gates: a rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293, 2011.

Y. Li, C. Liang, F. Ma, and J. Wang. The role of the idemv in predicting european stock market volatility during the covid-19 pandemic. *Finance research letters*, 36: 101749, 2020.

X. Ling and Y. Rho. Stable combination tests. *Statistica Sinica*, 32:641–644, 2022.

Y. Liu and J. Xie. 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, 2020.

F.-b. Lu, Y.-m. Hong, S.-y. Wang, K.-k. Lai, and J. Liu. Time-varying granger causality tests for applications in global crude oil markets. *Energy Economics*, 42: 289–298, 2014.

L. Mattner. Combining individually valid and conditionally iid p-variables. *arXiv preprint arXiv:1008.5143*, 2010.

N. Meinshausen, M. H. Maathuis, P. Bühlmann, et al. Asymptotic optimality of the westfall–young permutation procedure for multiple testing under dependence. *The Annals of Statistics*, 39(6):3369–3391, 2011.

Y. Ming-Hsien and W. Chih-She. Revisit export and gdp nexus in china and taiwan: A rolling window granger causality test. *Theoretical & Applied Economics*, 22(3), 2015.

M. D. Moran. Arguments for rejecting the sequential bonferroni in ecological studies. *Oikos*, 100(2):403–405, 2003.

F. Mosteller and R. R. Bush. *Selected quantitative techniques.* Addison-Wesley, 1954.

N. Mylonidis and C. Kollias. Dynamic european stock market convergence: Evidence from rolling cointegration analysis in the first euro-decade. *Journal of Banking & Finance*, 34(9):2056–2064, 2010.

J. P. Nolan. *Univariate stable distributions.* Springer, 2020.

W. Nyakabawo, S. M. Miller, M. Balcilar, S. Das, and R. Gupta. Temporal causality between house prices and output in the us: A bootstrap rolling-window approach. *The North American Journal of Economics and Finance*, 33:55–73, 2015.

P. C. O'Brien. Procedures for comparing samples with multiple endpoints. *Biometrics*, 40(4):1079–1087, 1984.

M. Papież and S. Śmiech. Dynamic steam coal market integration: Evidence from rolling cointegration analysis. *Energy Economics*, 51:510–520, 2015.

N. S. Pillai and X.-L. Meng. An unexpected encounter with cauchy and lévy. *The Annals of Statistics*, 44(5):2089–2097, 2016.

G. A. Rempala and Y. Yang. On permutation procedures for strong control in multiple testing with gene expression data. *Statistics and its interface*, 6(1), 2013.

G. Robinson. *FMStable: Finite Moment Stable Distributions*, 2012. URL `https://CRAN.R-project.org/package=FMStable`. R package version 0.1-2.

J. P. Romano, A. M. Shaikh, and M. Wolf. Formalized data snooping based on generalized error rates. *Econometric Theory*, pages 404–447, 2008.

B. Rossi. Optimal tests for nested model selection with underlying parameter instability. *Econometric theory*, 21(5):962–990, 2005.

B. Rossi and Y. Wang. Vector autoregressive-based granger causality test in the presence of instabilities. *The Stata Journal*, 19(4):883–899, 2019.

L. Rüschendorf. Random variables with maximum sums. *Advances in Applied Probability*, 14(3):623–632, 1982.

S. Shi, P. C. Phillips, and S. Hurn. Change detection and the causal impact of the yield curve. *Journal of Time Series Analysis*, 39(6):966–987, 2018.

S. Shi, S. Hurn, and P. C. Phillips. Causal change detection in possibly integrated systems: Revisiting the money–income relationship. *Journal of Financial Econometrics*, 18(1):158–180, 2020.

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

M. Sosa-Castro. Equity market volatility impact on s&p 500 sector indexes, 1989-2021. *Applied Econometrics and International Development*, 22(1):39–60, 2022.

S. A. Stouffer, E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr. *The american soldier: Adjustment during army life.(studies in social psychology in world war ii), vol. 1.* Princeton Univ. Press, 1949.

V. V. Uchaikin and V. M. Zolotarev. *Chance and stability: stable distributions and their applications*. Walter de Gruyter, 2011.

S. Van der Sluis, D. Posthuma, and C. V. Dolan. Tates: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet*, 9(1):e1003235, 2013.

V. Ventura, C. J. Paciorek, and J. S. Risbey. Controlling the proportion of falsely rejected hypotheses when conducting multiple tests with climatological data. *Journal of Climate*, 17(22):4343–4356, 2004.

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

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

H. White. A reality check for data snooping. *Econometrica*, 68(5):1097–1126, 2000.

S. S. Wilks. The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The annals of mathematical statistics*, 9(1):60–62, 1938.

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

D. J. Wilson. Generalized mean p-values for combining dependent tests: comparison of generalized central limit theorem and robust risk analysis. *Wellcome Open Research*, 5, 2020.

D. Wuertz, M. Maechler, and R. core team members. *stabledist: Stable Distribution Functions*, 2016. URL `https://CRAN.R-project.org/package=stabledist`. R package version 0.7-1.

I. Zaliapin, Y. Y. Kagan, and F. P. Schoenberg. Approximating the distribution of pareto sums. *Pure and Applied geophysics*, 162(6-7):1187–1228, 2005.

S. Zhang, H.-S. Chen, and R. M. Pfeiffer. A combined p-value test for multiple hypothesis testing. *Journal of Statistical Planning and Inference*, 143(4):764–770, 2013.

W.-B. Zhang. *Synergetic economics: time and change in nonlinear economics*, volume 53. Springer Science & Business Media, 2013.

E. Zivot and J. Wang. Rolling analysis of time series. In *Modeling Financial Time Series with S-Plus®*, pages 299–346. Springer, 2003.

Appendix A

Technical Work From Chapter 2

A.1 Lemmas

This section presents lemmas for the proof of Theorem 2.2. Recall that $p(x) = 2 - 2\Phi(|x|)$ for all $x \in \mathbb{R}$, as defined in the beginning of Section 2.2.2. Lemmas A.1 and A.2 help find the lower bound of $F^{-1}(1 - \min_{i \in S} p_i)$ and $F^{-1}(1 - \max_{i \in S} p_i)$, respectively. Lemma A.3 presents a lower bound for a_n.

Lemma A.1. *Define* $g(x) = c_{\alpha,\beta} x^{1/\alpha} e^{\frac{x^2}{2\alpha}}$ *with constant* $c_{\alpha,\beta} = \left[\frac{1+\beta}{\sqrt{2\pi}}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)\right]^{1/\alpha}$, *where* $0 < \alpha < 2$ *and* $-1 < \beta \leq 1$*. For* $x \to \infty$*,*

$$F^{-1}[1 - p(x)|\alpha, \beta] > g(x) = c_{\alpha,\beta} x^{1/\alpha} e^{\frac{x^2}{2\alpha}}.$$

Proof of Lemma A.1. When $x \to \infty$, $g(x) \to \infty$. Therefore, we can apply the right tail approximation of a stable distribution in Theorem 1.2 form Nolan [2020]. When $0 < \alpha < 2$ and $-1 < \beta \leq 1$,

$$\begin{aligned} 1 - F[g(x)|\alpha, \beta] &= \Pr[W_0 > g(x)] \\ &\sim \frac{1+\beta}{\pi}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)[g(x)]^{-\alpha} \\ &= \sqrt{\frac{2}{\pi}} x^{-1} e^{-x^2/2}. \end{aligned}$$

From Mill's ratio inequality that $1-\Phi(x) \leq \frac{\phi(x)}{x}$ for any $x>0$, where $\Phi(\cdot)$ and $\phi(\cdot)$ represent the distribution function and probability density function of a standard normal random variable respectively, we have

$$\begin{aligned} p(x) &= 2[1-\Phi(x)] \\ &\leq 2\frac{\phi(x)}{x} \\ &= \sqrt{\frac{2}{\pi}} x^{-1} e^{-x^2/2}. \end{aligned}$$

Therefore, $p(x) \leq 1-F[g(x)|\alpha,\beta]$ for $x \to \infty$. Since F^{-1} is increasing, $F^{-1}[1-p(x)] > g(x)$ for large enough x.

■

Lemma A.2. *Define* $\tilde{g}(x) = -\tilde{c}_{\alpha,\beta} x^{-1/\alpha} e^{\frac{x^2}{2\alpha}}$ *where* $\tilde{c}_{\alpha,\beta} = \left[\frac{1-\beta}{\sqrt{2\pi}}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)\right]^{1/\alpha}$ *is a constant depends on* $\alpha \in (0,2)$ *and* $\beta \in [-1,1]$. *When* $x \to 0^+$,

$$F^{-1}[1-p(x)|\alpha,\beta] > \tilde{g}(x) = -\tilde{c}_{\alpha,\beta} x^{-1/\alpha} e^{\frac{x^2}{2\alpha}}.$$

Proof of Lemma A.2. We first consider the case where $-1 \leq \beta < 1$. Similarly to the proof of Lemma A.1, when $x \to 0^+$, $\tilde{g}(x) \to -\infty$, and thus we can apply the left tail approximation from Theorem 1.2 of Nolan [2020] when $0<\alpha<2$ and $-1 \leq \beta < 1$:

$$\begin{aligned} F[\tilde{g}(x)|\alpha,\beta] &= \Pr[W_0 < \tilde{g}(x)] \\ &\sim \frac{1-\beta}{\pi}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)[-\tilde{g}(x)]^{-\alpha} \\ &= \sqrt{\frac{2}{\pi}} x e^{-x^2/2}. \end{aligned}$$

The standard normal distribution function, $\Phi(x)$, can be rewritten with integration by parts,

$$\begin{aligned} 1-p(x) &= 2\Phi(x)-1 \\ &= \sqrt{\frac{2}{\pi}} x e^{-x^2/2} + Q(x), \end{aligned}$$

where $Q(x) = \sqrt{\frac{2}{\pi}} e^{-x^2/2}(\frac{x^3}{3}+\frac{x^5}{3*5}+\cdots) > 0$ if $x>0$. Therefore, $1-p(x) > F[\tilde{g}(x)|\alpha,\beta]$ for $x \to 0^+$.

When $\beta = 1$, the distribution is totally skewed to the right, and the left tail probability does not follow a power law. Instead, we know that the left tail probability of $W_{0;\alpha,1}$ is smaller than that of $W_{0;\alpha,\beta}$ with $-1 \leq \beta < 1$, where $W_{0;\alpha,\beta}$ is the stable random variable with distribution $\boldsymbol{S}(\alpha, \beta)$. That is,

$$F[\tilde{g}(x)|\alpha, 1] = \Pr[W_{0;\alpha,1} < \tilde{g}(x)] \leq \Pr[W_{0;\alpha,\beta} < \tilde{g}(x)].$$

Therefore, $1 - p(x) > F[\tilde{g}(x)|\alpha, \beta] > F[\tilde{g}(x)|\alpha, 1]$ for all $-1 \leq \beta \leq 1$. Since F^{-1} is increasing, we have $F^{-1}[1 - p(x)] > \tilde{g}(x)$, which completes the proof. ■

Lemma A.3. *Let* $w_i \in (0, 1)$ *be nonnegative weights such that* $\sum_{i=1}^n w_i = 1$. *The normalizing constant* $a_n = (\sum_{i=1}^n w_i^\alpha)^{-1/\alpha} \geq \min\{n^{1-1/\alpha}, 1\}$.

Proof of Lemma A.3. The lower bound of a_n is considered in three separate cases. First, when $\alpha = 1$, $a_n = 1$. The second case is when $0 < \alpha < 1$. By Hölder's inequality,

$$\sum_{i=1}^n w_i^\alpha \leq \left[\sum_{i=1}^n (w_i^\alpha)^{1/\alpha}\right]^\alpha n^{1-\alpha} = n^{1-\alpha},$$

which is equivalent to $a_n \geq n^{1-1/\alpha}$. The last case is when $1 < \alpha < 2$. From the fact that l_α norm is smaller l_1 norm,

$$\left(\sum_{i=1}^n |w_i|^\alpha\right)^{1/\alpha} \leq \sum_{i=1}^n |w_i| = 1,$$

and therefore, $a_n \geq 1$. Combining the above three cases, $a_n = (\sum_{i=1}^n w_i^\alpha)^{-1/\alpha} \geq \min\{n^{1-1/\alpha}, 1\}$.

■

A.2 Proof of Theorem 2.2

Proof of Theorem 2.2. Recall that the test statistic is defined as $T_n(\boldsymbol{p}) = T_n(\boldsymbol{X}) = a_n \sum_{i=1}^n w_i F^{-1}[1 - p(X_i)|\alpha, \beta]$, where $a_n = \left(\sum_{j=1}^n w_j^\alpha\right)^{-1/\alpha}$. Under Assumption 2.4, the test statistic $T_n(\boldsymbol{X})$ can be decomposed into two parts:

$$T_n(\boldsymbol{X}) = a_n \sum_{i\in S} w_i F^{-1}[1-p(X_i)|\alpha,\beta] + a_n \sum_{i\in S^c} w_i F^{-1}[1-p(X_i)|\alpha,\beta] := A_n + B_n.$$

In order to show $T_n(\boldsymbol{X}) \to \infty$ as $n \to \infty$, we will show that $A_n \to \infty$ with probability 1 and that B_n cannot be arbitrary large negative.

Part A_n can be further decomposed as follows:

$$A_n \geq a_n c_0 n^{-1} \max_{i\in S} F^{-1}[1-p(X_i)|\alpha,\beta] + a_n(\sum_{j\in S} w_j - c_0 n^{-1}) \min_{i\in S} F^{-1}[1-p(X_i)|\alpha,\beta]$$

$$:= A_{n,1} + A_{n,2}$$

In the following arguments, we will prove that $A_n \to \infty$ with probability 1 by showing that $A_{n,1}$ can be arbitrarily large whereas $A_{n,2} > o_p(1)$ as $n \to \infty$.

Giving that $F^{-1}[1-p(x)|\alpha,\beta]$ is increasing in x, $A_{n,1}$ can be rewritten as $A_{n,1} = a_n c_0 n^{-1} F^{-1}[1-p(\max_{i\in S}|X_i|)|\alpha,\beta]$. Recall that the set of positive signals (S_+) is assumed to have cardinality no less than $|S|/2$. From Lemma 6 of Cai et al. [2014] and using the same argument as in the proof of Theorem 3 of Liu and Xie [2020], $\max_{i\in S}|X_i| \geq \mu_0 + \sqrt{2\log|S_+|} + o_p(1)$. Given the assumptions $\mu_0 = \sqrt{2r\log n}$ and $\sqrt{2\log|S_+|} \geq \sqrt{2(k_0\log n - \log 2)}$, we have $\max_{i\in S}(|X_i|) \to \infty$ with probability 1. Lemma A.1 implies that, as $n \to \infty$,

$$\Pr\left\{F^{-1}\left[1-p\left(\max_{i\in S}|X_i|\right)|\alpha,\beta\right] > g\left(\max_{i\in S}|X_i|\right)\right\} \to 1,$$

which is equivalent to

$$\Pr\left\{A_{n,1} \geq a_n c_0 n^{-1} c_{\alpha,\beta}\left(\max_{i\in S}|X_i|\right)^{1/\alpha} \exp\left[\frac{(\max_{i\in S}|X_i|)^2}{2\alpha}\right]\right\} \to 1.$$

Noting that $\max_{i\in S}|X_i| \geq \sqrt{2r\log n} + \sqrt{2\log|S_+|} + o_p(1)$, $\Pr\{\max_{i\in S} X_i > 1\} \to 1$ and $\sqrt{\log|S_+|} \geq \sqrt{2(k_0\log n - \log 2)} \approx \sqrt{2k_0\log n}$, we have

$$\Pr\left\{A_{n,1} \geq a_n c_0 n^{-1} c_{\alpha,\beta}\left[n^{(\sqrt{k_0}+\sqrt{r})^2}\right]^{1/\alpha}\right\} \to 1.$$

From Lemma A.3, $a_n = \left(\sum_{j=1}^n w_j^\alpha\right)^{-1/\alpha} \geq \min\{n^{(\alpha-1)/\alpha}, 1\}$, therefore, as $n \to \infty$,

$$\Pr\left\{A_{n,1} \geq c_0 c_{\alpha,\beta}\left[n^{(\sqrt{k_0}+\sqrt{r})^2/\alpha-1+\min\{1-1/\alpha,0\}}\right]\right\} \to 1.$$

By Pert 3 of Assumption 2.4, $\sqrt{r}+\sqrt{k_0} > \max\{\sqrt{\alpha}, 1\}$, we have $n^{(\sqrt{k_0}+\sqrt{r})^2/\alpha-1/\alpha-1} \to \infty$ as $n \to \infty$. Therefore, we obtain that $A_{n,1} \to \infty$ with probability tending to 1 as $n \to \infty$.

Next consider the part $A_{n,2} = a_n\left(\sum_{j\in S} w_j - c_0 n^{-1}\right)\min_{i\in S} F^{-1}[1-p(X_i)]$. Suppose $\mu_1 = \mu_0$ without loss of generality, thus $X_1 = \mu_0 + Z_1$, where $Z_1 \sim N(0,1)$. Let $\epsilon_n = n^{\alpha\gamma_0-1}$ with $k_0 < \gamma_0 < \frac{1-k_0}{\alpha}$. Similarly to the proof of Liu and Xie [2020], $\min_{i\in S}|X_i|$ is greater than any ϵ_n with probability 1 as $n \to \infty$ because

$$\begin{aligned}\Pr\left(\min_{i\in S}|X_i| < \epsilon_n\right) &\leq \sum_{i\in S}\Pr\left(|X_i| < \epsilon_n\right)\\ &= n^{k_0}\Pr\left(|X_i| < \epsilon_n\right)\\ &= n^{k_0}\left[\Phi(\mu_0+\epsilon_n) - \Phi(\mu_0-\epsilon_n)\right]\\ &< n^{k_0}\left[2\epsilon_n\phi(\mu_0-\varepsilon_n)\right]\\ &< n^{k_0}\epsilon_n = n^{k_0+\alpha\gamma_0-1} = o(1).\end{aligned} \tag{A.2.1}$$

Apply the increasing function $F^{-1}[1-p(x)|\alpha,\beta]$ on both $\min_{i\in S}|X_i|$ and ϵ_n, equation (A.2.1) is then equivalent to the statement that $F^{-1}[1-p(\min_{i\in S}|X_i|)|\alpha,\beta]$ is greater than any $F^{-1}[1-p(\epsilon_n)|\alpha,\beta]$ with probability 1 as $n \to \infty$. Since $\epsilon_n \to 0$ as $n \to \infty$, we can apply Lemma A.2 to find the lower bound of $F^{-1}[1-p(\epsilon_n)|\alpha,\beta]$, which implies the bound of $A_{n,2}$ as follows:

$$\Pr\left\{|A_{n,2}| > a_n(n^{k_0-1} - c_0 n^{-1})|\tilde{g}(\epsilon_n)|\right\} \to 1.$$

With the assumption that there is a constant c_0 such that $\min_{i=1}^n w_i \geq c_0/n$, we have $a_n < n^{1-1/\alpha}c_0^{-1}$. As $n \to \infty$, $e^{\epsilon_n^2/(2\alpha)} \to 1$, and thus

$$\begin{aligned}a_n(n^{k_0-1} - c_0 n^{-1})|\tilde{g}(\varepsilon_n)| &< \tilde{c}_{\alpha,\beta}a_n n^{k_0-1}\varepsilon_n^{-1/\alpha}e^{\epsilon_n^2/(2\alpha)}\\ &\leq \tilde{c}_{\alpha,\beta}c_0^{-1}n^{k_0-\gamma_0}\\ &= o(1).\end{aligned}$$

Therefore, $A_{n,2} > o_p(1)$, which completes the proof of the statement that $A_n \to \infty$ with probability 1 as $n \to \infty$.

Next, we show B_n cannot be arbitrary large negative. Under Part 1 of Assumption 2.4, Theorem 2.1 implies that as $n \to \infty$,

$$\left(\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{k\in S^c} w_k^\alpha}\right)^{1/\alpha} B_n = \left(\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{k\in S^c} w_k^\alpha}\right)^{1/\alpha} a_n \sum_{i\in S^c} w_i F^{-1}\left(1 - p_i|\alpha,\beta\right) \xrightarrow{d} W_0,$$

where W_0 follows $\boldsymbol{S}(\alpha,\beta)$.

Let $\delta_{\epsilon_n} = \left[\frac{1-\beta}{\pi\epsilon_n}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)\left(\frac{\sum_{k\in S^c} w_k^\alpha}{\sum_{j=1}^n w_j^\alpha}\right)\right]^{1/\alpha}$ with $\epsilon_n = n^{\alpha\gamma_0 - 1}$ and $k_0 < \gamma_0 < \frac{1-k_0}{\alpha}$. Notice that as $n \to \infty$, $\epsilon_n \to 0$, and $\delta_{\epsilon_n}\left(\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{k\in S^c} w_k^\alpha}\right)^{1/\alpha} = \left[\frac{1-\beta}{\pi\epsilon_n}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)\right]^{1/\alpha} \to \infty$. We first show the case when $-1 < \beta < 1$. According to the tail approximation of Theorem 1.2 of Nolan [2020], when $0 < \alpha < 2$ and $-1 \le \beta < 1$,

$$\begin{aligned}\Pr\left(B_n < -\delta_{\epsilon_n}\right) &\sim \Pr\left[W_0 < -\delta_{\epsilon_n}\left(\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{k\in S^c} w_k^\alpha}\right)^{1/\alpha}\right] \\ &\sim \frac{1-\beta}{\pi}\Gamma(\alpha)\sin\left(\frac{\pi\alpha}{2}\right)\delta_{\epsilon_n}^{-\alpha}\left(\frac{\sum_{j=1}^n w_j^\alpha}{\sum_{k\in S^c} w_k^\alpha}\right)^{-1} \qquad \text{(A.2.2)} \\ &= \varepsilon_n,\end{aligned}$$

for large enough n. Equation (A.2.2) implies that for any $\epsilon_n \to 0$, there exist an $\delta > \delta_{\epsilon_n}$ such that $\Pr\left(B_n < -\delta\right) < \epsilon_n$ as $n \to \infty$.

When $\beta = 1$, the distribution is totally skewed to the right, and consequently, for all $i \in S^c$, $\Pr\left(W_{i;\alpha,1} < -\delta_{\epsilon_n}\right) < \Pr\left(W_{i;\alpha,\beta} < -\delta_{\epsilon_n}\right)$ for any $\beta < 1$, where $W_{i;\alpha,\beta}$ is the transformed p-value with parameters α and β. Therefore, equation (A.2.2) holds for $\beta = 1$ as well. That is, B_n cannot be arbitrary large negative for all $0 < \alpha < 2$ and $-1 < \beta \le 1$, which finishes the proof. ■

Appendix B

Technical Work From Chapter 3

B.1 Proof of Theorem 3.1

Proof. Under Assumption 3.2, for any $i \neq j$

$$\lim_{x\to\infty} \frac{\Pr\left[w_i\phi(p_i) > x, w_j\phi(p_j) > x\right]}{\Pr[w_1\phi(p_1) > x]} \to 0.$$

For any fixed c,

$$\lim_{x\to\infty} \frac{\Pr\left\{w_i[\phi(p_i) - c] > x, w_j[\phi(p_j) - c] > x\right\}}{\Pr[w_1\phi(p_1) > x]} \to 0.$$

When $0 < \alpha < 2$, $a_n = (\sum_{i=1}^n w_i^\alpha)^{-1/\alpha}$ and $b_n = n\mathrm{E}\sin[\phi(p_1)/n]I(\alpha = 1) + a_n\mathrm{E}[\phi(p_1)]I(1 < \alpha < 2)$, where I is an indicator function. For any fixed n, $\frac{t}{a_n} \to \infty$ as $t \to \infty$. Lemma 3.1 from Jessen and Mikosch [2006] stated that the closure property of regularly varying random variables holds under the asymptotic tail independence condition. Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \frac{\Pr\left[T_n(\boldsymbol{p}) > t\right]}{\Pr\left[a_n w_1 \phi(p_1) - w_1 b_n > t\right]} &= \lim_{t\to\infty} \frac{\Pr\left\{\sum_{i=1}^{n} w_i \left[\phi(p_i) - \frac{b_n}{a_n}\right] > \frac{t}{a_n}\right]}{\Pr\left\{w_1 \left[\phi(p_1) - \frac{b_n}{a_n}\right] > \frac{t}{a_n}\right\}} \\
&= \sum_{i=1}^{n} \left\{ \lim_{t\to\infty} \frac{\Pr\left\{w_i \left[\phi(p_i) - \frac{b_n}{a_n}\right\} > \frac{t}{a_n}\right]}{\Pr\left\{w_1 \left[\phi(p_1) - \frac{b_n}{a_n}\right] > \frac{t}{a_n}\right\}} \right\} \\
&= \sum_{i=1}^{n} \left\{ \lim_{t\to\infty} \frac{\Pr\left[\phi(p_i) - \frac{b_n}{a_n} > \frac{t}{w_i a_n}\right]}{\Pr\left[\phi(p_1) - \frac{b_n}{a_n} > \frac{t}{w_1 a_n}\right]} \right\} \\
&= \sum_{i=1}^{n} \left\{ \lim_{t\to\infty} \frac{\Pr\left[\phi(p_i) > \frac{t}{w_i a_n}\right]}{\Pr\left[\phi(p_1) > \frac{t}{w_1 a_n}\right]} \right\}.
\end{aligned}$$

The last equality is true because both $\frac{b_n}{a_n}$ and $\frac{b_n}{w_1 a_n}$ are fixed constants for any given n.

Recall from Assumption 3.1 we represent $\Pr\left[\phi(p) > x\right]$ by $q_1 x^{-\alpha} L(x)$ as $x \to \infty$, where $L(x)$ is a slowly varying function with limit l_ϕ. For any fixed n, we have $\lim_{t\to\infty} L\left(\frac{t}{w_i a_n}\right) = \lim_{t\to\infty} L(t) = l_\phi$. Therefore, the above equation equals

$$\sum_{i=1}^{n} \left\{ \lim_{t\to\infty} \frac{q_1 \left(\frac{t}{w_i a_n}\right)^{-\alpha} L\left(\frac{t}{w_i a_n}\right)}{q_1 \left(\frac{t}{w_1 a_n}\right)^{-\alpha} L\left(\frac{t}{w_1 a_n}\right)} \right\} = \frac{\sum_{i=1}^{n} w_i^\alpha}{w_1^\alpha}.$$

According to Theorem 1.2 of Nolan [2020], the right tail probability of a random variable $W_0 \sim S(\alpha, \beta, \gamma, 0)$ with $0 < \alpha < 2$ and $-1 < \beta \leq 1$ is

$$\lim_{t\to\infty} \Pr(W_0 > t) = \lim_{t\to\infty} t^{-\alpha} \gamma^\alpha \sin(\pi\alpha/2)\Gamma(\alpha)(1+\beta)/\pi. \tag{B.1.1}$$

Therefore,

$$\begin{aligned}
\lim_{t\to\infty} \frac{\Pr(W_0 > t)}{\Pr\left[a_n w_1 \phi(p_1) - w_1 b_n > t\right]} &= \lim_{t\to\infty} \frac{t^{-\alpha} \gamma^\alpha \sin(\pi\alpha/2)\Gamma(\alpha)(1+\beta)/\pi}{q_1 w_1^\alpha t^{-\alpha} a_n^\alpha l_\phi} \\
&= \lim_{t\to\infty} \frac{t^{-\alpha} \gamma^\alpha \sin(\pi\alpha/2)\Gamma(\alpha)(1+\beta)/\pi}{q_1 w_1^\alpha t^{-\alpha} a_n^\alpha l_\phi} \\
&= \left(\frac{\sum_{i=1}^{n} w_i^\alpha}{w_1^\alpha}\right) \left[\frac{\gamma^\alpha \sin(\pi\alpha/2)\Gamma(\alpha)(1+\beta)/\pi}{q_1 l_\phi}\right].
\end{aligned}$$

Let $\beta = q_1 - q_2 = \lim_{t\to\infty} \frac{\Pr[\phi(p_1)>t]-\Pr[\phi(p_1)<-t]}{\Pr[|\phi(p_1)|>t]}$ and $\gamma = \left[\frac{\pi l_\phi q_1}{\sin(\pi\alpha/2)\Gamma(\alpha)(1+\beta)}\right]^{1/\alpha} = \left[\frac{\pi l_\phi}{2\sin(\pi\alpha/2)\Gamma(\alpha)}\right]^{1/\alpha}$ so that the last term in the above equation is 1. Therefore, we have proved that when $0 < \alpha < 2$

$$\lim_{t\to\infty} \frac{\Pr[T_n(\boldsymbol{p}) > t]}{\Pr(W_0 > t)} = 1.$$

■

B.2 Proof of Theorem 3.2

Before proving Theorem 3.2, the following lemma helps to control the magnitude of the test statistic.

Lemma B.1. *Under Assumption 1.1 and Assumption 3.2,* $\lim_{t\to\infty} \Pr[T_n(\boldsymbol{p}) < -t] = 0$ *for any* n *when* $0 < \alpha < 2$.

Proof of Lemma B.1. When $q_2 = 0$, the statement is true. When $q_2 \neq 0$, we will show $\Pr[T_n(\boldsymbol{p}) < -t] \sim q_2 t^{-\alpha} L(t) \to 0$ as $t \to \infty$. Similarly to the right tail argument in the proof of Theorem 3.1, for any fixed n we have the following argument for the left tail,

$$\lim_{t\to\infty} \frac{\Pr[T_n(\boldsymbol{p}) < -t]}{\Pr\left[a_n w_1 \phi(p_1) - w_1 b_n < -t\right]} = \sum_{i=1}^{n} \frac{w_i^\alpha}{w_1^\alpha},$$

based on Lemma 3.1 of Jessen and Mikosch [2006] and equation (3.2.3). Therefore, as $t \to \infty$,

$$\begin{aligned}\Pr[T_n(\boldsymbol{p}) < -t] &\sim \sum_{i=1}^{n} \left(\frac{w_i^\alpha}{w_1^\alpha}\right) \Pr\left[a_n w_1 \phi(p_1) - w_1 b_n < -t\right] \\ &\sim \left(t - \frac{b_n}{a_n}\right)^{-\alpha} q_2 l_\phi \sim t^{-\alpha} q_2 l_\phi.\end{aligned}$$

■

Proof of Theorem 3.2. In this proof, we show that $T_n(\boldsymbol{p}) \to \infty$ with probability 1 as $n \to \infty$. Let $T_{|S^c|}(\boldsymbol{p})$ be the test statistic for the true null hypothesis $\cap_{i\in S^c} H_i$, where

$|S^c|$ is the cardinality of set S^c. The test statistic $T_n(\boldsymbol{p})$ is decomposed into three parts:

$$\begin{aligned} T_n(\boldsymbol{p}) &= a_n \sum_{i \in S} w_i \phi(p_i) + a_n \sum_{i \in S^c} w_i \phi(p_i) - b_n \\ &= a_n \sum_{i \in S} w_i \phi(p_i) + \frac{a_n}{a_{|S^c|}} T_{|S^c|}(\boldsymbol{p}) + \left(\frac{a_n}{a_{|S^c|}} b_{|S^c|} - b_n \right) \\ &:= A_n + B_n + C_n. \end{aligned}$$

It is sufficient to show $A_n \to \infty$ with probability 1 and neither B_n or C_n can be arbitrary large negative.

Using a similar decomposition strategy as in the proof of Theorem 3 of Liu and Xie [2020], the lower bound of A_n can be decomposed as follows:

$$\begin{aligned} A_n &\geq a_n \min_{1 \leq i \leq n} w_i \max_{i \in S} \phi(p_i) + a_n \left(\sum_{j \in S} w_j - \min_{1 \leq i \leq n} w_i \right) \min_{i \in S} \phi(p_i) \\ &:= A_{n,1} + A_{n,2}. \end{aligned}$$

In the following arguments, we will prove that A_n is arbitrary large by showing that $A_{n,1}$ can be arbitrarily large whereas $A_{n,2} > o_p(1)$ as $n \to \infty$.

We first prove that for any $\varepsilon_1 > 0$,

$$\Pr\left(A_{n,1} > M_1 n^{\varepsilon_1}\right) \to 1 \tag{B.2.2}$$

as $n \to \infty$, where M_1 is a positive constant as defined in Assumption 3.3. Equation (B.2.2) implies that $A_{n,1}$ is arbitrarily large; i.e., $A_{n,1} \to \infty$ as $n \to \infty$ with probability 1.

The following proves equation (B.2.2). Since ϕ is nonincreasing, $\max_{i \in S} \phi(p_i) = \phi(\min_{i \in S} p_i)$. From Lemma A.3 of Ling and Rho [2022], the lower bound of a_n is $\min\left(n^{1-1/\alpha}, 1\right)$. In addition, the assumption that $\min w_i \geq c_0/n$ suggests that $A_{n,1} \geq \min\left(n^{-1/\alpha}, n^{-1}\right) c_0 \phi(\min_{i \in S} p_i)$ given that $\phi(\min_{i \in S} p_i)$ is positive. Therefore, let $m_1 = \varepsilon_1 + \max(1/\alpha, 1)$, we have

$$\begin{aligned}\Pr\left[\frac{A_{n,1}}{n^{\varepsilon_1}} > M_1\right] &\geq \Pr\left[\phi(\min_{i\in S} p_i) \geq M_1 c_0^{-1} n^{m_1}\right] \cdot \Pr\left[\phi(\min_{i\in S} p_i) > 0\right] \\ &= \Pr\left[\min_{i\in S} p_i \leq \phi^{-1}\left(M_1 c_0^{-1} n^{m_1}\right)\right] \cdot \Pr\left[\phi(\min_{i\in S} p_i) > 0\right],\end{aligned}$$

where ϕ^{-1} is the inverse function of ϕ. Note that the first inequality of the above equation is true because

$$\Pr\left[\frac{A_{n,1}}{n^{\varepsilon_1}} > M_1 \middle| \phi(\min_{i\in S} p_i) \leq 0\right] = 0.$$

Since ϕ is nonincreasing, $\phi^{-1} : (-\infty, \infty) \to (0, 1)$ is also nonincreasing, implying $\phi^{-1}(t) \leq \phi^{-1}(0)$ for any $t > 0$. As $n \to \infty$,

$$\begin{aligned}\Pr\left[\phi\left(\min_{i\in S} p_i\right) > 0\right] &= \Pr\left[\min_{i\in S} p_i \leq \phi^{-1}(0)\right] \\ &\geq \Pr\left[\min_{i\in S} p_i \leq \phi^{-1}\left(M_1 c_0^{-1} n^{m_1}\right)\right] \to 1.\end{aligned}$$

Therefore, $\phi(\min_{i\in S} p_i)$ is positive with probability tending to 1 as $n \to \infty$. Therefore, the condition (3.2.9) in Assumption 3.3 implies that

$$\Pr\left[A_{n,1} > M_1 n^{\varepsilon_1}\right] \geq \Pr\left[\min_{i\in S} p_i \leq \phi^{-1}(M_1 c_0^{-1} n^{m_1})\right] \cdot \Pr\left[\phi(\min_{i\in S} p_i) > 0\right] \to 1,$$

proving equation (B.2.2).

Now we show that $A_{n,2} > o_p(1)$. When $\phi(\max_{i\in S} p_i)$ is nonnegative, it is trivial that $A_{n,2}$ is also nonnegative and A_n can be arbitrary large. The nontrivial case is when $\phi(\max_{i\in S} p_i)$ is negative, where we have the following proof to show $A_{n,2} > o_p(1)$.

The assumption $\min_{1\leq i\leq n} w_i \geq c_0/n$ implies a_n has an upper bound at $c_0^{-1} n^{1-1/\alpha}$. Recall that the sum of the weights of false individual null hypotheses is assumed to be $\sum_{j\in S} w_j = n^{k_0-1}$ with $0 < k_0 < 0.5$, and thus, $a_n(\sum_{j\in S} w_j - \min_j w_j) < a_n \sum_{j\in S} w_j < c_0^{-1} n^{k_0-1/\alpha}$. Given $\phi(\max_{i\in S} p_i)$ is negative, $A_{n,2} > c_0^{-1} n^{k_0-1/\alpha} \phi(\max_{i\in S} p_i)$, and the condition in equation (3.2.10) implies that, as $n \to \infty$,

$$\begin{aligned}\Pr\left(A_{n,2} > M_2 n^{-\varepsilon_2}\right) &> \Pr\left[c_0^{-1} n^{k_0-1/\alpha} \phi(\max_{i\in S} p_i) > M_2 n^{-\varepsilon_2}\right] \\ &= \Pr\left[\phi(\max_{i\in S} p_i) > M_2 c_0 n^{m_2}\right] \\ &= \Pr\left[\max p_i < \phi^{-1}(M_2 c_0 n^{m_2})\right] \to 1\end{aligned}$$

where $m_2 = 1/\alpha - \varepsilon_2 - k_0$. Since $m_2 > 0$ and $M_2 < 0$, $M_2 c_0 n^{m_2} \to -\infty$ as $n \to \infty$, and thus $\phi^{-1}(M_2 c_0 n^{m_2}) \to 1$. Recall that $\varepsilon_2 > 0$, which leads to $M_2 n^{-\varepsilon_2} = o(1)$. Therefore, $A_{n,2} > o_p(1)$ with probability tending to 1 as $n \to \infty$.

Next we show that B_n cannot be arbitrarily large negative. Let $0 < \varepsilon_3 < \varepsilon_1$, as $n \to \infty$,

$$\begin{aligned}\Pr(B_n < -n^{\varepsilon_3}) &= \Pr\left[T_{|S^c|}(\boldsymbol{p}) < (-n^{\varepsilon_3})\left(\frac{\sum_{j=1}^{n} w_j^{\alpha}}{\sum_{i \in S^c} w_i^{\alpha}}\right)^{1/\alpha}\right] \\ &\leq \Pr\left[T_{|S^c|}(\boldsymbol{p}) < -n^{\varepsilon_3}\right].\end{aligned} \tag{B.2.3}$$

According to Lemma B.1,

$$\lim_{n \to \infty} \Pr\left[T_{|S^c|}(\boldsymbol{p}) < -n^{\varepsilon_3}\right] = 0.$$

Therefore, B_n cannot be arbitrary large negative and $B_n / n^{\varepsilon_3} > o_p(1)$.

When $\alpha \neq 1$, $C_n = 0$ for any n. When $\alpha = 1$, $a_n = 1$ and $a_{|S^c|} = \sum_{j \in S^c} w_j \geq 1 - c_0 n^{k_0 - 1}$ due to the assumption on the minimum weight in Assumption 3.3.2. The number of true individual hypothesis $|S^c|$ is $n - n^{k_0}$ from Assumption 3.3.2 for $0 < k_0 < 0.5$. Noting that $\lim_{x \to 0} \frac{\mathrm{E}\sin[\phi(p)x]}{x} = \mathrm{E}[\phi(p)]$, $C_n \leq \frac{|S^c|}{1 - c_0 n^{k_0 - 1}} \mathrm{E}\sin[\phi(p)/|S^c|] - n\mathrm{E}\sin[\phi(p)/n] \sim \mathrm{E}[\phi(p)]\left(\frac{1}{1 - c_o n^{k_0 - 1}} - 1\right) \to 0$ with uniform p. Therefore, $C_n = o(1)$.

Recall that $A_{n,1}$ goes to infinite with order $\varepsilon_1 > \varepsilon_3$, and $A_{n,2} > o_p(1)$. Therefore, the sum of $A_{n,1}$, $A_{n,2}$, B_n, and C_n is dominated by $A_{n,1}$ and can be arbitrary large positive. As a result, $T_n(\boldsymbol{p}) \to \infty$ as $n \to \infty$, and thus for any fixed significant level s,

$$\lim_{n \to \infty} \Pr\left[T_n(\boldsymbol{p}) > t_s\right] = 1,$$

where the t_s is the cutoff value defined under the global null hypothesis. ■

B.3 Comparisons of Assumptions

Theorem 3 in Liu and Xie [2020] assumes some conditions of the number and strength of nonzero signals of Z-score test statistics to study the asymptotic power of CCT. In this section we show that their assumption and our assumption 3.3 in Chapter 3 are comparable.

CCT's z-score test has assumptions $\mu_0 = \sqrt{2r \log n}$, $|S| = n^{k_0}$, and $\sqrt{r} + \sqrt{k_0} > 1$. These conditions can be changed into assumptions on minimum and maximum p-values. Since $\max_{i \in S} |X_i| \geq \mu_0 + \sqrt{2 \log |S_+|} + o_p(1)$ and $p(x) = 2[1 - \Phi(|x|)]$, $\min_{i \in S} p_i \leq 2[1 - \Phi(\mu_0 + \sqrt{2 \log |S_+|})] + o_p(1)$. By Mill's ratio inequality that $1 - \Phi(x) \leq x^{-1}\phi(x)$, where $x > 0$ and $\phi(x) = 1/\sqrt{2\pi} e^{-x^2/2}$, thus we have,

$$\begin{aligned}\min_{i \in S} p_i = p(\max_{i \in S} |X_i|) &\leq 2\left(\mu_0 + \sqrt{2 \log |S_+|}\right)^{-1} \phi\left(\mu_0 + \sqrt{2 \log |S_+|}\right) + o_p(1) \\ &\approx (\sqrt{\pi})^{-1} \left(\sqrt{r} + \sqrt{k_0}\right)^{-1} \left(\sqrt{\log n}\right)^{-1} n^{-\left(\sqrt{r}+\sqrt{k_0}\right)^2} + o_p(1),\end{aligned}$$

where $\sqrt{r} + \sqrt{k_0} > 1$. When $\alpha = 1$, recall from equation (3.2.11), our framework requires that

$$\min_{i \in S} p_i < \pi^{-1} \left(M_1 c_0^{-1}\right)^{-\alpha} n^{-\varepsilon_1 - 1} + o_p(1)$$

for any $\varepsilon_1 > 0$. Our condition is comparable to CCT's condition by selecting an ε_1 such that $n^{\varepsilon_1} = \sqrt{\log n}$.

Next let's consider the condition on $\max p_i$. Ling and Rho [2022] showed in Appendix B equation (A.1) that $\min_{i \in S} |X_i|$ is greater than any ϵ_n with probability tending 1 as $n \to \infty$ if $\epsilon_n n^{k_0} = o(1)$. Therefore, $1 - \max p_i = 1 - p(\min |X_i|) > 1 - p(\epsilon_n) + o_p(1) = 2\Phi(\epsilon_n) - 1 + o_p(1)$. By Taylor series expansions of $\Phi(\epsilon_n)$, we have $2\Phi(\epsilon_n) - 1 = O(\epsilon_n)$. Let $\epsilon_n = n^{\varepsilon_2 - 1}$, where $0 < \varepsilon_2 < 1 - k_0$ is defined in Assumption 3.3. Therefore, the condition on $\max p_i$ of CCT satisfies

$$\Pr\left[\max_{i \in S} p_i < 1 - C n^{\varepsilon_2 - 1}\right] \to 1,$$

where C is some positive constant. Recall from equation (3.2.12) the assumption of our framework when $\alpha = 1$ is

$$\Pr\left[\max_{i \in S} p_i < 1 - \pi^{-1}(-M_2 c_0)^{-1} n^{\varepsilon_2 + k_0 - 1}\right] \to 1.$$

Since $n^{\varepsilon_2 - 1} < n^{\varepsilon_2 + k_0 - 1}$, the condition of CCT is weaker than ours.

Appendix C

More Simulation Results From Chapter 3

Figures C.1-C.4 are the plots of finite sample sizes, raw powers, and size-adjusted powers of the FCT under the significance levels 0.01 and 0.005 and different correlation strengths. The sample size $n = 40, 100, 300$. The x-axis indicates the tail index $\alpha = 0.1, 0.3, 0.5, 0.7, 1.0, 1.3, 1.5, 1.7, 1.9$. The y-axis represents the percentage of rejections. The black horizontal line is the nominal significance level. The varying correlation coefficients $\rho = 0, 0.2, 0.4, 0.6, 0.8$ in Figures C.1 and C.2. The correlation coefficients are weaker at $\rho = 0, 0.05, 0.1, 0.15, 0.2$ in Figures C.3 and C.4. The numbers of Monte Carol replication are $5 * 10^4$ and 10^5 when the nominal significance level is 0.01 and 0.005 respectively. The size distortion of FCT when $\alpha > 1$ and ρ close to 1 is relived compared with Figures 3.1-3.3 in the main body of the paper where the significance level is 0.05.

Tables C.1-C.3 reports more detailed information of Figure 3.4 when sample size $n = 40, 100$, and 300. The nominal significance levels in all tables is 0.05. The highlighted cells have the best cells or powers. The size, raw power and size-adjusted power of six methods with $\rho = 0, 0.1, \ldots, 1$ and sample size $n = 40, 100, 300$ are reported. The same as in Figure 3.4, “FCT1” refers to FCT with α =1, “FCT1.5”

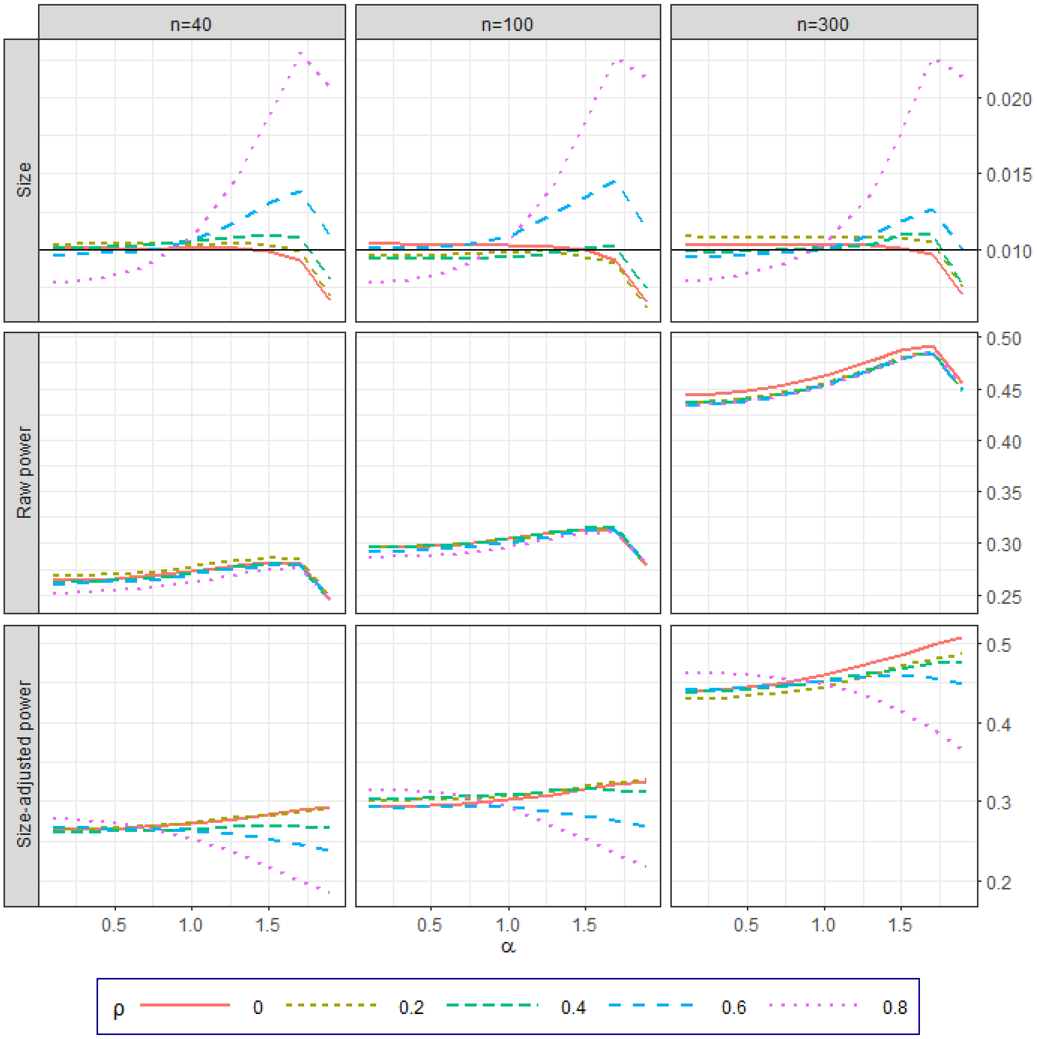

Figure C.1: Size, raw power, and size-adjusted powers of the FCT when the significance level is 0.01.

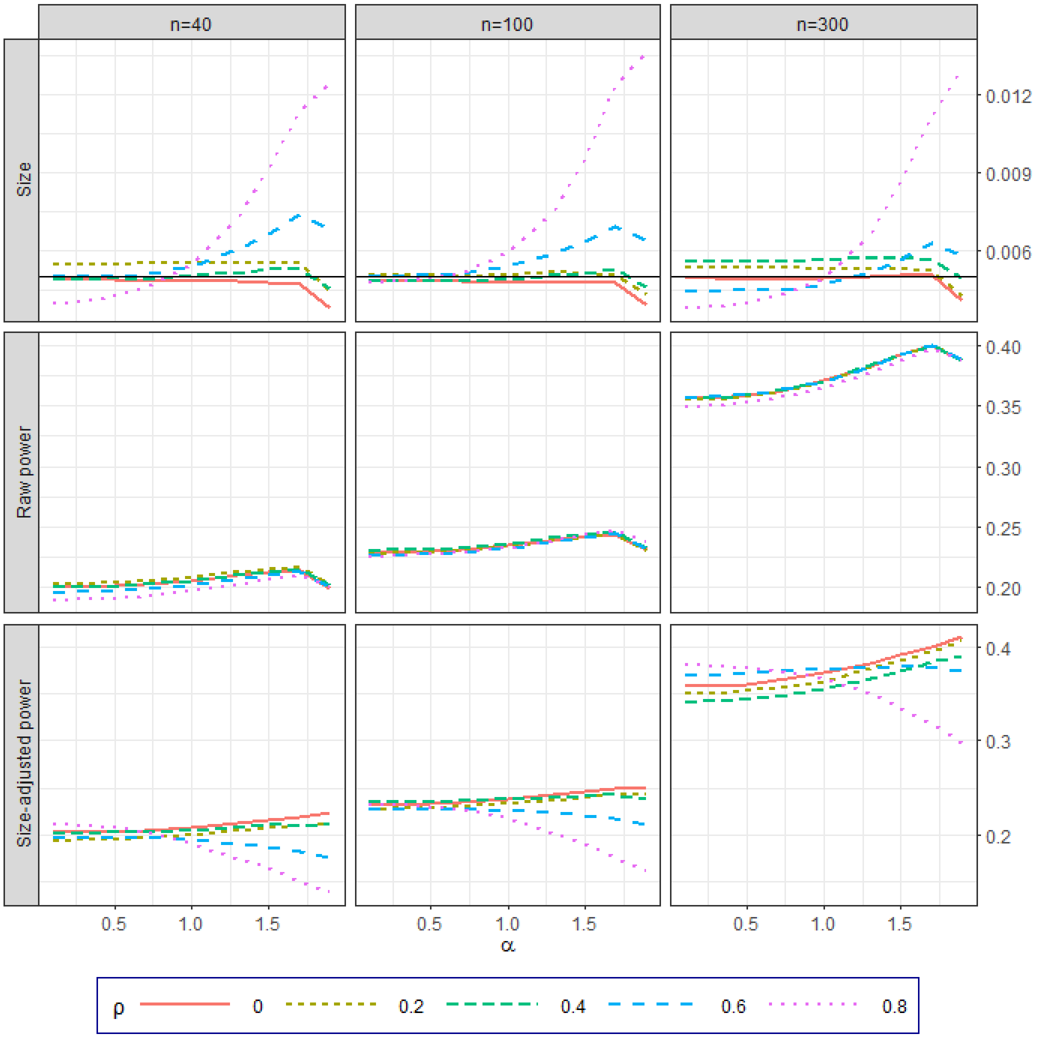

Figure C.2: Size, raw power, and size-adjusted powers of the FCT when the significance level is 0.005.

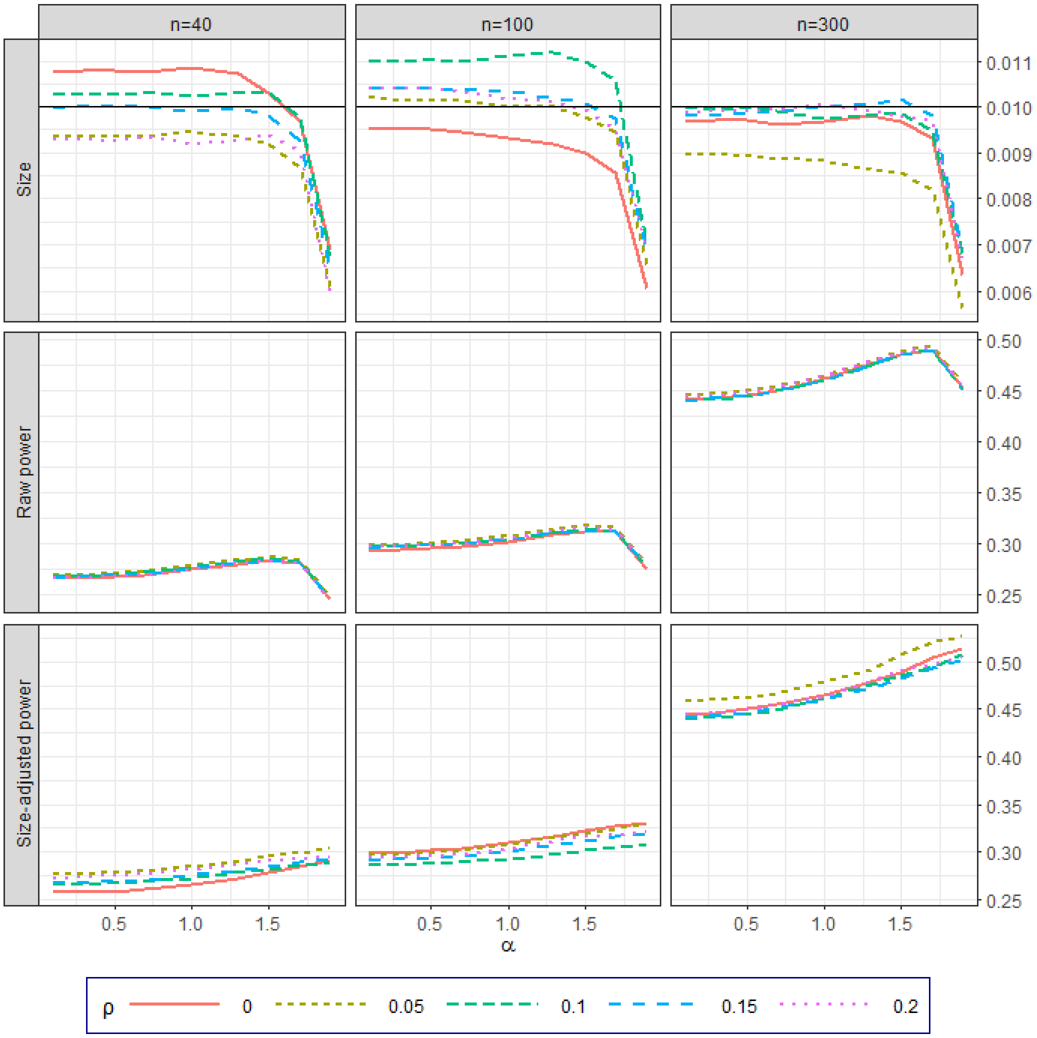

Figure C.3: Size, raw power, and size-adjusted powers of the FCT when the significance level is 0.01 with varying weak correlation coefficients.

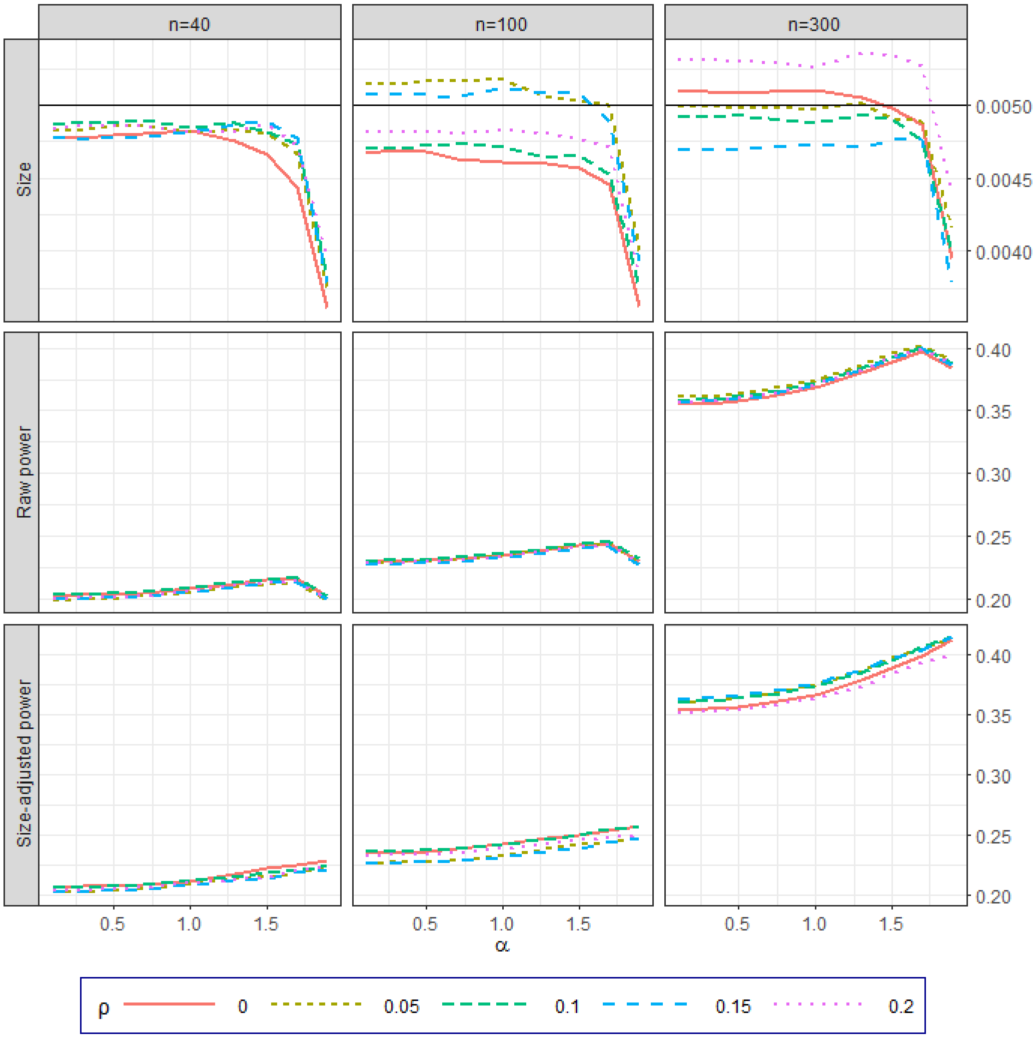

Figure C.4: Size, raw power, and size-adjusted powers of the FCT when the significance level is 0.01 with varying weak correlation coefficients

refers to FCT with $\alpha = 1.5$, "GMP1" refers GMP with $\alpha = 1$, which is the same as the HMP, "GMP1.5" refers GMP with $\alpha = 1.5$, "SCT1" refers to SCT with $(\alpha, \beta) = (1, 0)$, which is the same as CCT, and "SCT1.5" represents SCT with $(\alpha, \beta) = (1.5, 1)$.

Sample size $n = 40$

Size

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.0479	0.0495	**0.0498**	0.0518	0.0523	**0.0504**	**0.0533**	**0.0557**	**0.0539**	**0.0500**	0.0392
GMP1	0.0479	0.0495	**0.0498**	**0.0519**	0.0523	**0.0504**	**0.0533**	**0.0557**	**0.0539**	0.0499	0.0392
SCT1	**0.0481**	**0.0499**	0.0516	0.0528	0.0518	0.0530	0.0554	0.0578	0.0573	0.0566	**0.0513**
FCT1.5	0.0444	0.0447	0.0464	0.0480	0.0508	0.0538	0.0606	0.0662	0.0736	0.0819	0.0930
GMP1.5	0.0438	0.0444	0.0463	0.0473	**0.0499**	0.0525	0.0598	0.0653	0.0725	0.0800	0.0888
SCT1.5	0.0407	0.0498	0.0509	0.0533	0.0595	0.0646	0.0756	0.0834	0.0962	0.1148	0.1774

Raw power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.4919	0.4963	0.4996	0.4943	0.4849	0.4901	0.4810	0.4841	0.4576	0.4396	0.3249
GMP1	0.4919	0.4962	0.4996	0.4943	0.4848	0.4901	0.4807	0.4841	0.4576	0.4394	0.3249
SCT1	0.4882	0.4945	0.4952	0.4922	0.4827	0.4862	0.4764	0.4803	0.4560	0.4383	0.3466
FCT1.5	0.4883	0.494	0.4959	0.4917	0.4828	0.4870	0.4797	0.4796	0.4568	0.4327	0.3273
GMP1.5	0.4866	0.4921	0.4938	0.4911	0.4811	0.4863	0.4791	0.4785	0.4556	0.4319	0.3253
SCT1.5	**0.5046**	**0.5113**	**0.5150**	**0.5095**	**0.4991**	**0.5061**	**0.4959**	**0.4986**	**0.4738**	**0.4492**	**0.3655**

Size-adjusted power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.4998	0.4977	0.4999	0.4903	0.4769	**0.4870**	**0.4713**	0.4650	**0.4454**	0.4397	0.3570
GMP1	0.4998	0.4976	0.4998	0.4903	0.4769	0.4869	**0.4713**	**0.4651**	**0.4454**	**0.4398**	**0.3574**
SCT1	0.4949	0.4946	0.4903	0.4862	0.4752	0.4776	0.4629	0.4568	0.4319	0.4179	0.3429
FCT1.5	0.5128	**0.5143**	0.5118	**0.5009**	0.4803	0.4771	0.4480	0.4305	0.3846	0.3540	0.2367
GMP1.5	0.5131	0.5135	0.5117	0.5006	**0.4821**	0.4796	0.4484	0.4295	0.3868	0.3555	0.2402
SCT1.5	**0.5141**	0.5117	**0.5119**	0.4980	0.4704	0.4654	0.4350	0.4128	0.3603	0.3269	0.2084

Table C.1: Size, raw power, and size-adjusted power of six methods when sample size $n = 40$.

Sample size n=100

Size

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.0479	0.0485	0.0450	0.0542	0.0489	**0.0519**	0.0517	**0.0523**	**0.0512**	**0.0496**	0.0345
GMP1	0.0479	0.0485	0.0450	0.0542	0.0489	**0.0519**	0.0517	**0.0523**	**0.0512**	**0.0496**	0.0343
SCT1	0.0484	**0.0492**	0.0453	0.0544	0.0491	0.0540	**0.0514**	0.0532	0.0536	0.0542	**0.0483**
FCT1.5	0.0440	0.0444	0.0424	0.0521	**0.0501**	0.0578	0.0604	0.0663	0.0769	0.0946	0.1101
GMP1.5	0.0434	0.0439	0.0418	**0.0518**	0.0493	0.0577	0.0597	0.0656	0.0755	0.0922	0.1042
SCT1.5	**0.0495**	0.0489	**0.0460**	0.0561	0.0567	0.0669	0.0713	0.0800	0.0945	0.1235	0.2080

Raw power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.5227	0.5175	0.5198	0.5249	0.5254	0.5116	0.5139	0.5122	0.5018	0.4842	0.3275
GMP1	0.5227	0.5174	0.5198	0.5247	0.5254	0.5117	0.5139	0.5122	0.5018	0.4842	0.3275
SCT1	0.5152	0.5136	0.5188	0.5185	0.5226	0.5076	0.5091	0.5062	0.4964	0.4796	0.3518
FCT1.5	0.5251	0.5184	0.5237	0.5265	0.5267	0.5164	0.5163	0.5144	0.5055	0.4854	0.3268
GMP1.5	0.5237	0.5173	0.5224	0.5257	0.5253	0.5151	0.5156	0.5136	0.5048	0.4846	0.3235
SCT1.5	**0.5340**	**0.5278**	**0.5367**	**0.5356**	**0.5371**	**0.5271**	**0.527**	**0.5274**	**0.5182**	**0.4986**	**0.3739**

Size-adjusted power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.5285	0.5212	0.5410	0.5090	**0.5275**	**0.505**	**0.5111**	0.5050	0.4956	0.4869	0.3720
GMP1	0.5285	0.5212	0.5411	0.5089	**0.5275**	**0.5050**	**0.5111**	**0.5051**	**0.4960**	**0.4871**	**0.3727**
SCT1	0.5225	0.5158	0.5348	0.5057	0.5256	0.4983	0.504	0.4967	0.4897	0.4688	0.3554
FCT1.5	0.5428	0.5343	0.5485	0.5214	0.5267	0.4911	0.4849	0.4702	0.4425	0.3940	0.2272
GMP1.5	**0.5432**	**0.5345**	0.5486	**0.5230**	0.5272	0.4921	0.4853	0.4706	0.4439	0.3977	0.2306
SCT1.5	0.5352	0.5302	**0.5495**	0.5164	0.5200	0.4832	0.4750	0.4569	0.4231	0.3554	0.2029

Table C.2: Size, raw power, and size-adjusted power of six methods when sample size $n = 100$.

Sample size n=300

Size

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	**0.0504**	0.0468	**0.0500**	**0.0487**	0.0536	0.0483	**0.0517**	**0.0527**	0.0596	**0.0516**	0.0349
GMP1	**0.0504**	0.0467	**0.0500**	**0.0487**	0.0536	0.0483	**0.0517**	**0.0527**	0.0596	**0.0516**	0.0348
SCT1	0.0505	**0.0492**	0.0507	0.0482	**0.0530**	**0.0490**	0.0523	0.0545	**0.0590**	0.0542	**0.0520**
FCT1.5	0.0489	0.0447	0.0489	0.0476	0.0547	0.0525	0.0615	0.0682	0.0880	0.1073	0.1270
GMP1.5	0.0489	0.0446	0.0485	0.0474	0.0547	0.0520	0.0613	0.0678	0.0873	0.1048	0.1205
SCT1.5	0.0469	0.0434	0.0477	0.0475	0.0552	0.0551	0.0662	0.0772	0.1011	0.132	0.2302

Raw power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.6854	0.6866	0.6910	0.6876	0.6929	0.6811	0.6849	0.6856	0.6764	0.6580	0.3707
GMP1	0.6854	0.6866	0.6910	0.6876	0.6929	0.6811	0.6848	0.6855	0.6764	0.6581	0.3707
SCT1	0.6803	0.6798	0.6849	0.6807	0.6864	0.6757	0.6797	0.6819	0.6714	0.6483	0.3930
FCT1.5	**0.7010**	**0.7031**	**0.7080**	**0.7004**	**0.7104**	**0.6978**	0.7021	**0.7013**	**0.6868**	**0.6666**	0.3726
GMP1.5	0.7006	0.7023	0.7075	0.7001	0.7098	0.6972	**0.7022**	0.7006	0.6863	0.6654	0.3700
SCT1.5	0.6944	0.6928	0.6962	0.6884	0.7023	0.6812	0.6771	0.6711	0.6474	0.6006	**0.4154**

Size-adjusted power

ρ	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
FCT1	0.6837	0.6935	0.6925	0.6932	0.6847	0.6865	0.6807	0.6784	**0.6536**	**0.6500**	0.4069
GMP1	0.6837	0.6935	0.6923	0.6931	0.6847	0.6865	0.6807	**0.6785**	**0.6536**	**0.6500**	**0.4076**
SCT1	0.6786	0.6819	0.6815	0.6863	0.6797	0.6776	0.6723	0.6707	0.6502	0.6384	0.3902
FCT1.5	0.7038	0.7182	**0.7114**	**0.7056**	**0.7006**	0.6921	0.6753	0.6509	0.6035	0.5550	0.2437
GMP1.5	**0.7050**	**0.7183**	0.7111	0.7055	0.7001	**0.6925**	**0.6773**	0.6518	0.6048	0.5575	0.2457
SCT1.5	0.7024	0.7158	0.7037	0.6968	0.6875	0.6576	0.6176	0.5573	0.4388	0.3111	0.1020

Table C.3: Size, raw power, and size-adjusted power of six methods when sample size $n = 300$.