Statistica Sinica Preprint No: SS-2022-0046	
Title	Heavy-Tailed Distribution for Combining Dependent P-Values With Asymptotic Robustness
Manuscript ID	SS-2022-0046
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0046
Complete List of Authors	Yusi Fang, Chung Chang, Yongseok Park and George C. Tseng
Corresponding Authors	Chung Chang
E-mails	cchang@math.nsysu.edu.tw
Notice: Accepted version subject to English editing.	

Heavy-tailed distribution for combining dependent p-values with asymptotic robustness

Yusi Fang, Chung Chang*, Yongseok Park, and George C. Tseng

University of Pittsburgh and National Sun Yat-sen University

Abstract: Combining individual p-values to aggregate multiple small effects is a long-standing statistical topic. In recent years, advances in big data analysis promoted methods to aggregate correlated, sparse, and weak signals. Under such context, we investigate a wide range of p-value combination methods, formulated as the sum of transformed p-values with the transformations by a broad family of heavy-tailed distributions, namely regularly varying distributions. Two recent methods, Cauchy and harmonic mean tests are included in the family. We explore the conditions for a method of the family to possess robustness to dependency for type I error control and optimal power in terms of detection boundary for detecting weak and sparse signals. We show that only an equivalent class of Cauchy and harmonic mean tests has sufficient robustness to dependency in a practical sense. We also propose an improved truncated Cauchy method, which belongs to the equivalent class with fast computation, to address the issue caused by the large negative penalty in the Cauchy method. In addition, we conduct comprehensive simulations to verify our theoretical insights and provide a recommendation to the real practice. Finally, we present a neuroticism GWAS application to illustrate the theoretical findings in the regularly varying distribution family and the advantages of the truncated Cauchy method.

Key words and phrases: p-value combination method, combining dependent p-values, regularly varying distribution, global hypothesis testing.

*Corresponding author: Chung Chang; National Sun Yat-sen University; Email: cchang@math.nsysu.edu.tw

1. Introduction

Combining p-values to aggregate information from multiple sources is a long-standing issue in social science and biomedical research. Classical methods mostly focus on combining multiple independent and frequent signals to increase statistical power, which can be viewed as a type of meta-analysis. Consider the combination of n independent p-values, $\boldsymbol{p} = (p_1, ..., p_n)$. Many earlier methods were developed in the form of statistics $T(\boldsymbol{p}) = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} F_U^{-1}(1 - p_i)$ to sum up transformed p-values, where the transformation $g(p)$ is the inverse CDF of a random variable U. Conventional methods in this category include Fisher's method (Fisher et al., 1934) with $T = \sum_{i=1}^{n} -2\log(p_i)$ using U as a chi-squared distribution and Stouffer's method (Stouffer et al., 1949) with $T = \sum_{i=1}^{n} -\Phi^{-1}(p_i)$ using U as a standard normal distribution, among many other choices of $g(p)$ and their corresponding U in the literature (Edgington, 1972; Pearson, 1933; Mudholkar and George, 1979). This first category of methods aims for classical meta-analysis to combine independent and relatively frequent signals and it applies light-tailed distribution (i.e., tails thinner than an exponential function) for U. Efficiency of a method is mostly considered under the asymptotic framework that the number of p-values n is fixed and sample size m to derive each p-value goes to infinity, where $p = O(e^{-m})$ in most cases. Under this setting, it has been shown that only the equivalent class of Fisher's method is asymptotically Bahadur Optimal (ABO), meaning the efficiency of the combined p-value statistics is asymptotically optimal under fixed n and $m \to \infty$ (Littell and Folks, 1971).

In the rise of big data, many scientific questions have turned to combine p-values with large n. The seminal paper by Donoho and Jin (2004) established a framework of combining p-values with weak and sparse signals and proposed the higher-criticism test with asymptotically optimal property. This second category of methods considers $n \to \infty$ and only a small number s of the n p-values ($s = n^{\beta}$ where $0 < \beta < \frac{1}{2}$) have weak signals ($p = O(n^{-r}/(\log n)^{\frac{1}{2}})$

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

All aforementioned methods were developed to combine independent p-values. Many modern large-scale data analyses have generated the need to combine a large number of dependent p-values with sparse and weak signals, which we categorize as methods of the third category. A notable application is to combine p-values of multiple correlated SNPs (can be tens to hundreds or thousands) in a SNP-set (e.g., all SNPs in a gene region or in gene regions of a pathway) in genome-wide association studies (GWAS). In this case, the neighboring SNPs often pose unknown dependency structures. Efforts have been made to extend existing tests to account for dependency using permutation or other numerical simulation approaches (e.g., Liu and Xie, 2019). However, permutation or simulation-based methods are not practical when n is large, and high precision of p-value is needed to account for multiple comparisons. The null hypothesis may also be difficult to simulate via permutation. Barnett et al. (2017) developed an analytic approximation for higher criticism incorporated with dependency structure. The method is, however, still computationally intensive and not accurate enough for small p-values needed for multiple comparisons. Motivated by these needs, Liu and Xie (2020) and Wilson (2019a) independently proposed Cauchy combination test ($T = \sum_{i=1}^{n} \tan\{(0.5 - p_i)\pi\}$) and harmonic mean combination test ($T = \sum_{i=1}^{n} \frac{1}{p_i}$) to combine p-values under unspecified dependency structure. Wilson (2019a) also provided a convenient R package called harmonicmeanp (function p.mamml)

to implement the harmonic test. A remarkable property of both methods is that the null distributions and testing procedures derived from independence assumption are robust under dependency structure in an asymptotic but practical sense to be explained in Section 3.1. In this paper, motivated by this observation, we consider a rich family of test statistics that includes the Cauchy and harmonic mean tests. More precisely, we consider the test statistics

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

where the transformation $g(p)$ corresponds to U that is from the regularly varying distribution family, a broad family of heavy-tailed distributions. We investigate the conditions such that practical robustness to dependency similar to Cauchy and harmonic mean methods can be achieved. We note that selections of U for classical meta-analysis setting (fixed n and $m \to \infty$) are all from thin-tailed distributions (e.g., chi-squared distribution for Fisher's method and Gaussian for Stouffer's method). This is reasonable since a thin-tailed distribution produces evener contributions from marginally significant p-values in the meta-analysis of frequent signals. In contrast, Cauchy and harmonic mean methods correspond to heavy-tailed distributions of U, which highly focus on small p-values and down-weigh marginally significant p-values. Figure 1 shows the transformation function of $g(p)$ in log-scale. For Fisher's method, the contributions of p-values 10^{-2} and 10^{-6} to the test statistics are 4.6 and 13.8. For heavy-tailed transformation methods, the contributions become 100 versus 10^6 for harmonic mean and 31.82052 versus 3.18×10^5 for Cauchy. With an increased focus on small p-values, the methods are more powerful for detecting sparse signals. It is worth noting that the recent work by Vovk and Wang (2020) also considered the sum of transformed p-values to combine p-values and showed an upper bound of significance level inflation under arbitrary dependence structure. We will describe the difference between our results and theirs in detail in the remark following Theorem 2. Wilson (2019b) and Wilson

(2020) and Vovk et al. (2021) also did related work on combining dependent p-values.

Throughout this paper, when we call a thin-tailed, heavy-tailed, or regularly varying method, it means that its corresponding U is a thin-tailed, heavy-tailed, or regularly varying distribution. The paper is structured as follows. We first investigate Box-Cox transformation for $g(p)$ in Section 2, which is equivalent to Pareto distribution for U. In Section 2.1, we build connections between existing methods, including $minP$, harmonic mean, Cauchy and Fisher in this framework. Particularly, we show that the Cauchy method is approximately equivalent to the harmonic mean method, which is a special case of the Box-Cox transformation. In Section 2.1, we observe that the Cauchy method can potentially suffer from the large negative penalty for p-values close to 1. To avoid the issue of the large negative penalty, we improve the Cauchy method by introducing a new test, the truncated Cauchy method, and develop a fast computing algorithm for it. In Section 3, we introduce a family of heavy-tailed distributions, namely regularly varying distribution, and investigate the conditions in the family that can provide robustness for dependency structure as in Cauchy and harmonic mean (Section 3.1-3.2). Section 3.3 studies the power of the family of methods in terms of detection boundary under the sparse and weak alternatives considered in Donoho and Jin (2004). Section 4 contains extensive simulations to demonstrate type I error control and power of different methods and numerically verify the theoretical results. Section 5 contains a GWAS application of neuroticism to compare the performance of different methods and demonstrate the improvement of the truncated Cauchy method over the Cauchy method. Section 6 provides the final conclusion and discussion.

2. Connection between minP, harmonic mean, Cauchy and Fisher

2.1 Methods by Pareto distribution to connect four existing methods

As mentioned in Section 1, we observe that many methods of the first category to combine independent and relatively frequent signals all correspond to thin-tailed distributions for U and many methods of the second and third categories for combining sparse and weak signals utilize heavy-tailed distributions. In this subsection, we consider Pareto distribution for U, which is equivalent to Box-Cox transformation for $g(p)$. Based on this transformation family, we will build the connection of 4 existing methods: $minP$, harmonic mean, Cauchy, and Fisher. Insight in Pareto distribution also provides intuition when introducing the regularly varying distribution as an extended richer family in the next section. Finally, we will prove the approximate equivalency of the harmonic mean and Cauchy combination methods. Consider the family of p-value combination methods: $T = \sum_{i=1}^{n} g(p_i)$, where $g(p) = \frac{1}{p^\eta}$ for some $\eta > 0$. We can show that $g(p) = F_U^{-1}(1-p)$ such that $U \overset{D}{\sim} Pareto(\frac{1}{\eta}, 1)$. In other words, $P(U > t) = t^{-\frac{1}{\eta}}$ for $t > 1$, which means U is a heavy-tailed distribution. A larger η corresponds to a heavier tail. Particularly, the harmonic mean method corresponds to $\eta = 1$ in Pareto distribution. We note that, by denoting $\lambda = -\eta$, we can rewrite $h(p;\lambda) = \frac{g(p;\eta)-1}{\lambda} = \frac{p^\lambda - 1}{\lambda}$, which is Box-Cox transformation. The following Proposition 1 shows that minP and Fisher are limiting cases in the Pareto distribution when $\eta \to +\infty$ and when $\eta \to 0$. Proposition 2 shows that the Cauchy combination method is approximately identical to harmonic mean for relatively small p-values.

Proposition 1. *For fixed n, $minP$ is a limiting case of methods by Pareto distribution when $\eta \to \infty$. Similarly, the Fisher's method is the limiting case of Pareto when $\eta \to 0$.*

Proof. Denote by $T_{\gamma_m} = \sum_{i=1}^{n} \frac{1}{p_i^{\gamma_m}} = \sum_{i=1}^{n} \frac{1}{p_{(i)}^{\gamma_m}}$, where $p_{(i)}$'s are ordered p-values. Note that T_{γ_m} is equivalent to $T^*_{\gamma_m} = \left(\sum_{i=1}^{n} \frac{1}{p_i^{\gamma_m}}\right)^{\frac{1}{\gamma_m}} = \frac{1}{p_{(1)}} \left(\sum_{i=1}^{n} \left(\frac{p_{(1)}}{p_{(i)}}\right)^{\gamma_m}\right)^{\frac{1}{\gamma_m}}$. As $\gamma_m \to \infty$,

$T^*_{\gamma_m} \to \frac{1}{p_{(1)}}$, which is equivalent to $minP$.

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

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

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

It is somewhat surprising that even though the forms of transformations by Cauchy and harmonic mean are very different, they are approximately equivalent when p is small, and the behaviors of both when p is small can be characterized by the index $\eta = 1$ of the Box-Cox transformation (note that the behaviors of these two transformations are fundamentally different when p is close to 1). It is natural to ask if other p-value combination methods exist in an extended rich heavy-tailed distribution family to enjoy similar finite-sample robustness property as in the Cauchy and harmonic mean methods. To answer this question, we introduce the family of regularly varying distributions and investigate the properties in Section 3.

Figure 1 shows minus log-scaled p transformation $g(p)$ versus minus log-scaled transformation $g(p)$ for the $BC_{0.5}$ (i.e., Box-Cox transformation with $\eta = 0.5$), HM (the harmonic mean method, equivalent to BC_1), CA (the Cauchy method), $BC_{1.5}$, Fisher's and Stouffer's methods. We see that as η increases, smaller p-values will be more dominant and impact of marginally significant p-values rapidly diminishes, which gives stronger power for sparse signal applications. CA and HM are approximately proportional when p is sufficiently small (roughly when $p < 10^{-2}$).

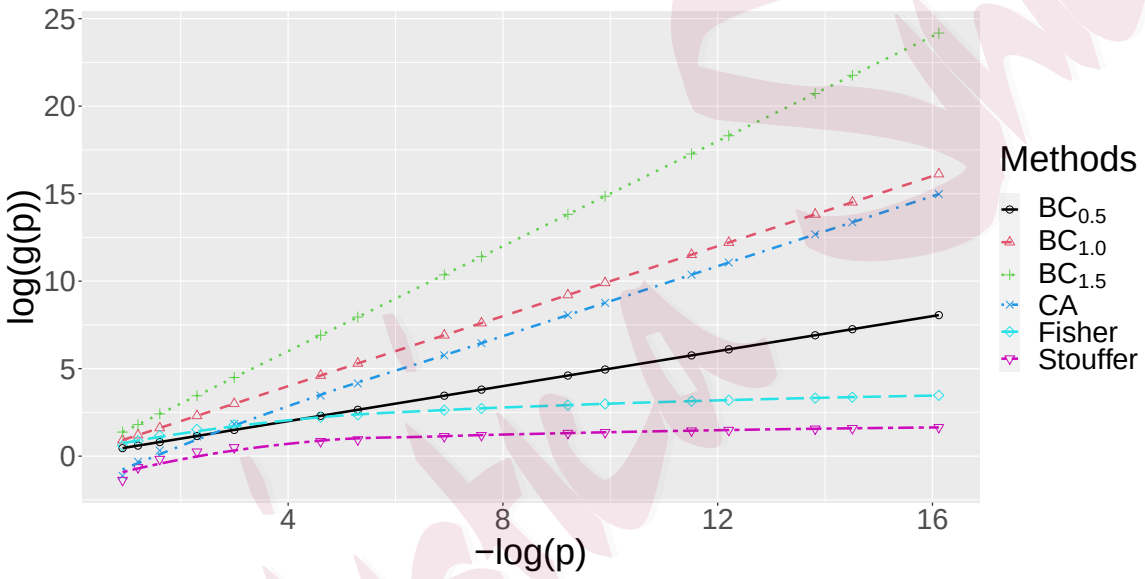

Figure 1: Comparison of transformations. We show 6 different transformations of p-values, $g(p)$, which correspond to $BC_{0.5}$, BC_1 (HM), $BC_{1.5}$, CA, Fisher and Stouffer. The x-axis is $-\log(p)$ and the y-axis shows $\log(g(p))$.

Although we have shown that HM and CA are approximately equivalent for combining relatively small p-values analytically and in simulations in Section 4.2. However, we note that when a p-value is very close to 1, the contribution in the Cauchy method is close to negative infinity, which can potentially cause numerical issues and substantial power loss; we call this phenomenon as "large negative penalty issue" in Cauchy. The situation of a p-value close to 1 can frequently happen for tests of discrete data, in which case the p-values under null hypothesis may not necessarily be $Unif(0,1)$. Two other possible situations to cause p-values close to 1 are when n is large or when the model to derive p-values is misspecified. As a simple remedy, we propose a truncated Cauchy test (CA^{tr}) that truncates any of the n p-values greater than $1-\delta$ to be $1-\delta$. For example, when $\delta = 0.01$, we have $p^{tr} = p$ if $p < 0.99$ and $p^{tr} = 0.99$ if $p \geq 0.99$. In fact, we recommend and use $\delta = 0.01$ throughout the paper. For selection of δ, conceptually δ should be large enough so that it avoids the large negative penalty issue in Cauchy. But for computational purpose, it cannot be too large so approximation by our fast-computing procedures is accurate. Detailed justification and support from simulation results for choosing $\delta = 0.01$ are given in Supplement Section S2.4. The proposed method can also be viewed in the form of sum of transformed p-values. Indeed, the statistic of CA^{tr} can be written as:

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

For more details on CA^{tr}, see Supplement Section S2.

3. Asymptotic properties of regularly varying methods for p-value combination

3.1 Regularly varying tailed distribution

Before introducing the regularly varying distributions, we first define some notations. Throughout this paper, denote by $\bar{F}$ the survival function of the distribution F (i.e., $\bar{F}(t) = 1 - F(t)$

for any t). Limits and asymptotic properties are assumed to be for $t \to \infty$ unless mentioned otherwise. For two positive functions $u(\cdot)$ and $v(\cdot)$, we write $u(t) \sim v(t)$ if $\lim_{t\to\infty} \frac{u(t)}{v(t)} = 1$. Also, if $\lim_{t\to\infty} \frac{u(t)}{v(t)} > 1$, we write $u(t) \gtrsim v(t)$; if $\lim_{t\to\infty} \frac{u(t)}{v(t)} < 1$, we write $u(t) \lesssim v(t)$. The definition of regularly varying tailed distribution is given below:

Definition 1. A distribution F is said to belong to the regularly varying tailed family with index γ (denoted by $F \in R_{-\gamma}$) if

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

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

We denote the whole family of regularly varying tailed distributions as R. It can be shown that every distribution F belonging to $R_{-\gamma}$ can be characterized by

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

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

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

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

$$P(U_1 + \ldots + U_n > t) \sim nP(U_1 > t). \tag{3.1}$$

3.2 Asymptotic tail probability approximation and robustness to dependence

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

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

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

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

Condition (A1) is mild and is also assumed in Liu and Xie (2020) when investigating the robustness of the Cauchy method under unspecified dependence structure. Throughout this paper, the term, "unspecified dependence structure", indicates unspecified Gaussian corre-

lation structure. In fact, this condition is to guarantee the tail distributions of each pair of U_i and U_j are asymptotically tailed independent; see the precise definition of asymptotically tailed independence for a pair of random variables in Supplement Section S1.

Condition (A2) includes the Box-Cox transformation (satisfying A2.1), Cauchy transformation (satisfying A2.1) and truncated Cauchy transformation (satisfying A2.2) introduced in Section 2.1. Condition (A3) is called "balance condition", which is a common condition for regularly varying tailed random variables (Goldie and Klüppelberg, 1998). For example, for the harmonic mean method, $p = 1, q = 0$; for the Cauchy method, $p = q = 1/2$, and for the truncated Cauchy method, $p = 1, q = 0$.

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

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

Here $T(\boldsymbol{X}) = \sum_{i=1}^{n} U_i$ is the sum of correlated regularly varying tailed random variables. The theorem is somewhat surprising and a general result since it is applicable to any regularly varying method and any correlation structure $\boldsymbol{\Sigma}$ with $-1 < \rho_{ij} < 1$ as long as no perfect correlation exists. This theorem is related to Theorem 3.1 in Chen and Yuen (2009), i.e., Lemma S2 in the Supplement. Roughly speaking, because of the heaviness of the tail for each U_i and the asymptotic tailed independence between each pair of U_i and U_j, asymptotically, the correlation structure has limited influence on the tail of $T(\boldsymbol{X})$. Since the approximated tail probability is independent of $\boldsymbol{\Sigma}$, an immediate application is to derive the p-value of a regularly varying method under independence assumption (i.e., $P(U_1 + \cdots + U_n > t)$ with i.i.d. $U_1, \cdots, U_n$; see Equation (3.1)). The theorem warrants its asymptotic robustness to unspecified dependence structure as similarly shown in the harmonic mean and Cauchy methods (Wilson, 2019a; Liu and Xie, 2020). Alternatively,

one may approximate the tail probability by $nP(U > t)$. We, however, note that the robustness to unspecified dependence structure is in an asymptotic sense, meaning extremely large t (corresponding to extremely small test size α) may be required for different tail heaviness in U and correlation structure to guarantee a good approximation. Throughout this paper, we approximate $P(U_1 + \cdots + U_n > t)$ under dependence structure by calculating $P(U_1 + \cdots + U_n > t)$ under independence assumption using Monte Carlo simulation.

Below we perform a simple simulation to demonstrate and investigate Theorem 1. Assume $n = 3$ and $\boldsymbol{X} = (X_1, X_2, X_3)$ is multivariate normal with unit variance and common pairwise correlation $\rho_{ij} = \rho$ $(1 \leq i < j \leq 3)$. In this simulation we set $\rho = 0$, 0.3, 0.6, 0.9 and 0.99. Here we consider 7 Box-Cox tests, $BC_{0.75}$, $BC_{0.8}$, $BC_{0.9}$, BC_1, $BC_{1.1}$, $BC_{1.25}$, and $BC_{1.5}$. From Theorem 1, we calculate $y(\alpha) = \frac{nP(U>t_\alpha)}{P(T(\boldsymbol{X})>t_\alpha)}$ from simulations, where t_α is chosen so that $P(T(\boldsymbol{X}) > t_\alpha) = \alpha$ and $\alpha = 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}$. We expect $\lim_{t_\alpha \to \infty} \log(y(\alpha)) = 0$ when $-1 < \rho < 1$. Figure 2A-2E show $\log_{10}$-scale α on the x-axis and the mean $\log(y(\alpha))$ on the y-axis for different ρ =(0, 0.3, 0.6, 0.9, 0.99). We note that, as ρ increases, smaller α will be required to observe a good approximation. Theorem 2 below further characterizes what would happen if partial of the p-values have perfect correlations $\rho_{ij} = 1$ or -1.

Theorem 2. *Suppose the conditions (A1), (A2) and (A3) in Theorem 1 hold. Define an arbitrary weight vector* $\boldsymbol{w} = (w_1, \cdots, w_n) \in R_+^n$, $T_{n,\boldsymbol{w}}(\boldsymbol{X}) = \sum_{i=1}^n w_i g(p_i)$. *Also assume* $\rho_{ij} = 1$ *or* -1 *for* $1 \leq i < j \leq m$, *and* $|\rho_{ij}| < 1$ *for* $i > m$ *or* $j > m$. *Then under the null hypothesis* $H_0 : \boldsymbol{\mu} = \boldsymbol{0}$, *we have:*

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

Note that Theorem 2 is a more general result, of which Theorem 1 is a special case. Consider a special scenario $\boldsymbol{w} = (1, \cdots, 1)$. An immediate consequence of Theorem 2 is that

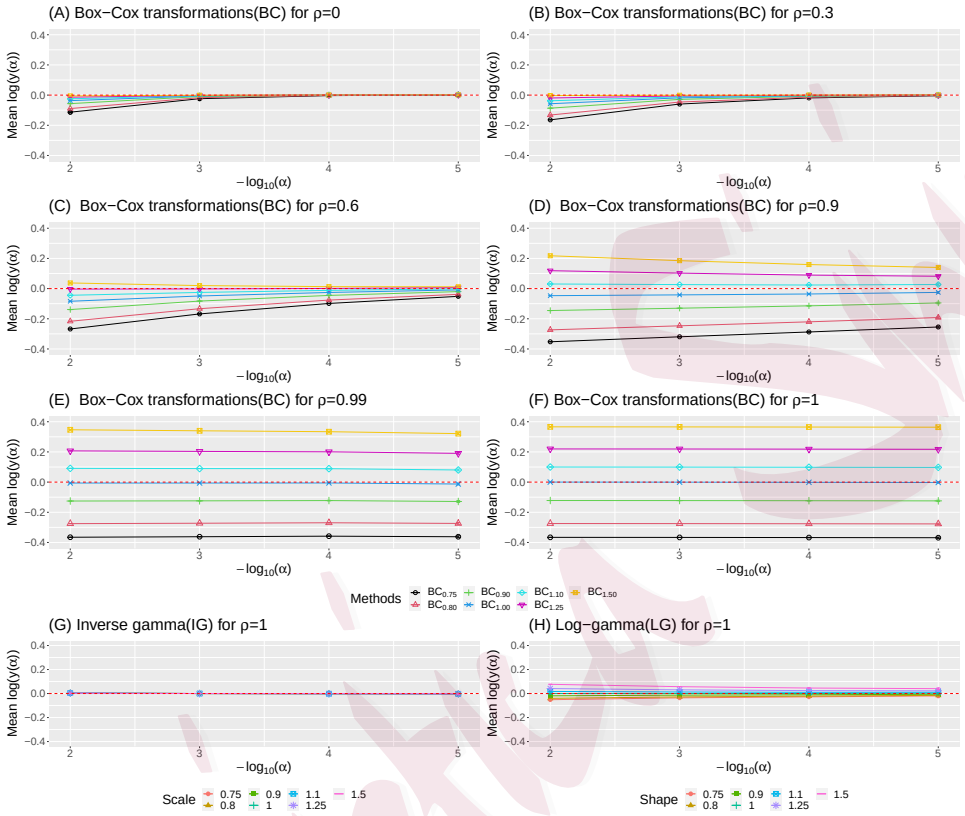

Figure 2: The mean log-scaled $y(\alpha)$ for Box-Cox transformations, inverse Gamma and log-Gamma across different significance levels α. (A)-(F) represent the results of Box-Cox transformations with different values of $\eta = 0.75, 0.8, 0.9, 1, 1.1, 1.25, 1.5$ for correlation level ρ =0, 0.3, 0.6,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

only when $\gamma = 1$ (e.g., HM or CA or CA^{tr} method) can satisfy $\{(\sum_{i=1}^{m} w_i)^{\gamma} + \sum_{i=m+1}^{n} w_i^{\gamma}\} = m^{\gamma} + (n-m) = n$, which produces the asymptotic robustness of Theorem 1. In other words, Figure 2A-2E already shows a hint that the convergence of Theorem 1 becomes more and more difficult when ρ increases to almost 1. When some of the p-values have perfect correlation, only index $\gamma = 1$ of the regularly varying distribution can still enjoy the asymptotic robustness to unspecified dependence structure. Figure 2F shows an simulation with $\rho = 1$, which satisfies the condition of Theorem 2. By assuming $w_1 = w_2 = w_3 = 1$ and $\rho = 1$, we have $P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t) \sim 3^{\gamma} P(U > t)$. Figure 2F verifies Theorem 2 that only BC_1 can reach the convergence $\lim_{t_\alpha \to \infty} \log(y(\alpha)) = 0$, showing robustness to perfect correlation. Although Figure 2E ($\rho = 0.99$) and Figure 2F ($\rho = 1$) visually look almost identical, all BC methods in Figure 2E will eventually converge to 0 as $\alpha \to 0$ by Theorem 1, although very slowly. On the other hand, in Figure 2F, only BC_1 can converge to 0 by Theorem 2.

Corollary 1. *Suppose the conditions in Theorem 2 hold and assume* $\sum_{i=1}^{n} w_i = n$*, then we have:*

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

From Corollary 1, note that, when $w_1 = \cdots = w_n = 1$ and the transformation $g(p) = 1/p^{1/\gamma}$, the test statistic $T_{n,\boldsymbol{w}}$ corresponds to the statistic $BC_\eta, \eta = 1/\gamma$. Hence, the BC tests with $\eta < 1$ (i.e., $\gamma > 1$) are anti-conservative in this situation; the higher the value of γ is, the more anti-conservative the test is. This is verified by Figure 2F for $BC_{0.9}$, $BC_{0.8}$ and $BC_{0.75}$ when $\rho = 1$. As $\eta \to 0$ (i.e., $\gamma \to \infty$), BC_η is asymptotically equivalent to the Fisher's method and is the most anti-conservative under dependence. On the other hand, for $\eta > 1$ (i.e., $\gamma < 1$), all the corresponding tests BC_η ($\eta > 1$) are conservative under this dependence structure, which is confirmed by Figure 2F for $BC_{1.1}$, $BC_{1.25}$ and $BC_{1.5}$.

In particular, when $\eta \to \infty$ ($\gamma \to 0$), BC_η becomes $minP$, which hence is expected to be very conservative. Figure 2G and 2H verifies that since inverse Gamma and log-Gamma are also regularly varying distributions with index $\gamma = 1$, they enjoy asymptotic robustness to correlation structure similar to HM (BC_1) and Cauchy even when perfect correlation exists. Another important implication for the application of this corollary is that among all the tests using transformations of regularly varying distributions, only the type I errors of those corresponding to $\gamma \leq 1$ are well preserved asymptotically (i.e., tests that are at least not anti-conservative asymptotically) under any correlation structure.

Corollary 2. *If we further assume* $-1 < \rho_{i,j} < 1, \forall 1 \leq i < j \leq n$ *(i.e.,* $m = 0$*), then we have*

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

Corollary 2 shows that the tail probability of the weighted test statistic $T_{n,\boldsymbol{w}}$ can be approximated by $\sum_{i=1}^{n} w_i^\gamma P(U > t)$. Similar to the unweighted version in Theorem 1, since the approximation in Corollary 2 is independent of the correlation structure, $P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t)$ under dependence structure can be approximated by calculating $P(w_1U_1 + \ldots + w_nU_n > t)$ under independence assumption using Monte Carlo simulation, as long as no perfect correlation among $U_i's$ exists. Also, note that this formula can be considered to be an extension of Corollary 1.3.8 in (Mikosch, 1999), in which $U_1, \ldots, U_n$ are assumed to be independent regularly varying distributed random variables.

Remark 1. Note that the robustness property of Theorems 1 and 2 is similar to (Liu and Xie, 2020; Wilson, 2019a) and only describes the asymptotic behavior of the tail probability of our proposed family. Indeed, the results of Theorem 1 and 2 only guarantee that the type I errors of the corresponding tests ($\gamma = 1$, equivalent to harmonic mean and Cauchy) can be well controlled for a small size α given fixed n and $\boldsymbol{\Sigma}$. Intuitively, as n increases, a more stringent cutoff corresponding to a small α is needed to ensure the robustness of type I error

control. An ideal robustness property for type I error should be to achieve a uniform upper tail bound in the sense of $P(T(\boldsymbol{X}) > t_\alpha) \leq c \cdot \alpha$ under any dependence structure $\boldsymbol{\Sigma}$, where t_α is the tail threshold when a nominal α is controlled under independence assumption, and c is independent of n, and $\boldsymbol{\Sigma}$ and is in a reasonable magnitude (e.g., $c = 1.5$, meaning the inflation of the type I error is at most 50% in the worst scenario). This uniform bound is, however, not achievable in general. Vovk and Wang (2020) recently provided a remarkable uniform bound for arbitrary dependency structure (note that ours is unspecified dependency structure) but dependent on n for the HM method:

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

where the adjusted factor α_n^{HM} is between $\log(n)$ and $e \cdot \log(n)$ (see Proposition 6 in the paper). This bound is, however, not practical in general applications since, considering $n = 100$ or 1000, the inflation bound $\alpha_n^{HM} \geq \log(n)$ is at least 4.6 or 6.9 folds. Furthermore, the factor α_n^{HM} is in comparison to type I error in perfect correlation situation (i.e., $\rho = 1$), instead of nominal size α under independence. On this issue, Goeman et al. (2019) pointed out an extreme case that when $n = 10^5$ and $\boldsymbol{\Sigma}$ has exchangeable correlation $\rho = 0.2$, HM has more than three folds of type I error inflation (true type I error=0.164 under nominal $\alpha = 0.05$). In Section 4.1, we will perform extensive simulations for a wide range of n and size α to investigate the limitation and develop practical guidance for applying the HM method in daily applications.

The discussion above indicates that with mild normality assumption, the upper bound for the inflation of type I errors is much smaller than that under arbitrary dependence structure. This is especially useful in the application because using a smaller upper bound of the inflation of type I errors to adjust the significance level will increase the power of the test. Also, note that based on Theorem 2 and simulations, the practical guideline to adjust significance level can easily be developed not only for HM test ($\eta = 1$) but also for

any test that is a sum of transformations by regularly varying tailed distributions, including any BC_η test. The guideline is very helpful in the application.

3.3 Detection boundary of regularly varying methods

In this subsection, we investigate the powers of regularly varying methods by deriving the detection boundary of the test $T(\boldsymbol{X})$ under sparse alternatives as $n \to \infty$ (Theorem 3), which is a popular measurement of the power performance for detecting weak and sparse signals. Below we introduce the standard setup of weak and sparse signals by Donoho and Jin (2004), which will be used for Theorem 3.

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

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

For Theorem 3 below, in addition to the setup by Donoho and Jin (2004) and the conditions (A2) and (A3), we need two additional conditions:

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

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

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

for all $t > t_1$.

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

Theorem 3. *Under conditions (A2), (A3), (C1) and (C2), for any* $0 < \gamma \leq 1$, *any significance level* $0 < \alpha < 1$, *and* τ *satisfying* $\sqrt{\tau} + \sqrt{\beta} > 1$, *then under the alternative hypothesis we have:*

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

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

Remark 2. Under the same conditions of Theorem 3, one can show that for $T_{n,\boldsymbol{w}} = \sum_{i=1}^n w_i g(p_i)$ with $\boldsymbol{w} \in R_+^n$ and $\sum_{i=1}^n w_i = n$, if $\max_i w_i \leq (\log n)^{\eta_1}$ and $\min_i w_i \geq 1/(\log n)^{\eta_2}$ for some fixed constants $\eta_1, \eta_2 > 0$, the result of Theorem 3 still holds. See Supplement Remark S6 for more details.

Theorem 3 states that the power of this test $T(\boldsymbol{X})$ converges to 1 for any significance level $\alpha > 0$ and $0 < \gamma \leq 1$, or equivalently, that the sum of Type I and II errors goes to zero given the setup. Moreover, Theorem 3 implies that the methods with $0 < \gamma \leq 1$ attain the optimal detection boundary defined in Donoho and Jin (2004) in the strong sparsity situation $0 < \beta < 1/4$. Liu and Xie (2020) showed a similar result for their proposed Cauchy's test. As described in Section 2, the Cauchy distribution has regular-varying tail

with index $\gamma = 1$. This theorem is valid for methods of regularly varying tailed distributions with index $0 < \gamma \leq 1$. Therefore, this theorem can be considered to be a generalization of Theorem 3 in Liu and Xie (2020).

4. Simulations

In this section, we perform simulations to compare the robustness performance of different p-value combination methods under varying correlation levels among p-values to verify theoretical results in Section 2 and 3. We include 7 methods discussed in Section 2, $minP$, $BC_{1.25}$, CA, CA^{tr}, $HM(BC_1)$, $BC_{0.75}$ and Fisher's method, as well as HC (Higher criticism) and BJ (Berk-Jones test). Section 4.1 firstly evaluates the type I error control of different methods under independence and varying level of correlation to verify the robustness of HM and Cauchy methods. Furthermore, since the robustness in Theorem 2 for HM and Cauchy is an asymptotic result, we further investigate the type I error control for HM under a wide range of n, ρ, and γ to ensure that the robustness of HM and Cauchy is preserved and useful in a practical sense. Section 4.2 assesses the statistical power under different dependency structures and sparsity of signals in the alternative hypothesis. In Section 4.3, we will evaluate the improvement of the truncated Cauchy method over the Cauchy method in a discrete data simulation.

4.1 Type I error control

In this subsection, we first simulate $n = 100$, $\boldsymbol{X} = (X_1, \cdots, X_n) \overset{D}{\sim} N(0, \boldsymbol{\Sigma})$, $p_i = 2(1 - \Phi(|X_i|))$ and $T = \sum_{i=1}^{n} g(p_i)$ for different aforementioned methods. We also assume that $\boldsymbol{\Sigma}$ has unit variance on the diagonal line and is exchangeable with common correlation $\rho = \text{cor}(X_i, X_j)$ for $1 \leq i \neq j \leq n$, where ρ is evaluated at 0 (independence), 0.3, 0.6, 0.9 and 0.99. Table 1 shows the type I errors of the 9 methods with different levels of

Table 1: Type I errors for 9 tests, including $Fisher$, CA, CA^{tr} (truncated Cauchy), $BC_{0.75}$, BC_1 (HM), $BC_{1.25}$, $minP$, HC and BJ, across correlation level ρ =0, 0.3, 0.6, 0.9, 0.99.

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

correlations at $\alpha = 0.001$ using 10^6 simulations under the null hypothesis. As expected all methods control type I error perfectly under independence assumption (i.e., $\rho = 0$). When correlation among p-values exists, we find that $minP$ is the most conservative in type I error control followed by $BC_{1.25}$, as expected from the theoretical result in Corollary 1. CA, CA^{tr} and HM remain with perfect type I error control in all correlation settings, showing robustness to dependency structure. Fisher and BJ are the most anti-conservative methods in the presence of correlation, followed by slight anti-conservativeness for HC and $BC_{0.75}$.

It is worth noting that according to Theorem 1 and 2 for regularly varying distribution transformation, the tail probability $P(T(\boldsymbol{X}) > t)$ under dependence can be asymptotically approximated by that under independence. However, the asymptotic result only guarantees the dependence robustness for very large t (or equivalently very small α). We also expect that larger n will require larger t (smaller α) to ensure a good approximation. Specifically, Goeman et al. (2019) has pointed out that, with $\rho = 0.2$ and $n = 10^5$, the much inflated type I error of 0.164 is obtained for size $\alpha = 0.05$. Therefore, it is of interest to explore the robustness property of $T(\boldsymbol{X})$ for dependence in HM for varying n, α and ρ to pro-

vide a practical guidance in real applications. In Table 2, we extend the simulation for HM with $n = (25, 50, 100, 500, 1000, 2000, 10000)$, $\alpha = (0.05, 0.01, 0.001, 0.0001)$, and $\rho = (0, 0.3, 0.6, 0.9, 0.99)$. Given each combination of α and n, we calculate the maximum percent of inflation (PI) across different ρ, which is defined as $\text{PI} = (\max_\rho \text{type I error} - \alpha)/(\alpha)$. The result confirms the theoretical result that larger n will generate greater type I error inflation under dependence for a fixed α and will require much smaller α to improve the type I error inflation. For example, when $\alpha = 0.01$, we have $PI = 30\%$ for $n = 25$ compared to $PI = 80\%$ for $n = 10,000$. On the other hand, when $n = 10,000$, PI decreases from 80% to 49% when α decreases from 0.01 to 0.0001. In general, the result shows robust type I error control under varying correlation levels in a practical sense when $n \leq 1,000$ and $\alpha \leq 0.05$ with the maximum $PI = 50\%$, which inflates type I error from $\alpha = 0.01$ to 0.015 at $n = 1000$ and $\rho = 0.3$. Even when n increases to 10,000, PI only minimally increases to 80%. When multiple comparisons are needed such as in the GWAS applications, small α is targeted and the robust type I error control for HM is generally achieved in a practical sense. However, if a single test is performed with a very large n, caution should be taken for the type I error inflation (e.g., type I error is 0.072 for $\alpha = 0.05$ when $n = 10,000$ and $\rho = 0.3$).

4.2 Statistical power

In this subsection, we follow the simulation setting in Section 4.1 to evaluate statistical power using different methods under different values of correlation ρ and strengths of the signal. Following the sparse and weak signal setting in Donoho and Jin (2004), we design the n signals $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)$ to contain $n - s$ with no signal ($\mu_{s+1} = \cdots = \mu_n = 0$) and the first s have non-zero signals $\mu_1 = \cdots = \mu_s = \mu_0 = \frac{\sqrt{4\log(n)}}{s^{0.1}}$, where $s/n = (5\%, 10\%, 20\%)$. We first compare the power of different methods under varying correlations, where the rejection

Table 2: Type I error control of HM evaluated at total number of p-values n =25, 50,100,500, 1000, 2000, 10000 and ρ =0, 0.3, 0.6, 0.99 for different sizes of test α =0.05, 0.01, 10^{-3} and 10^{-4}. We also calculated percent of inflation (PI) to reflect the extent of inflation of type I error under various cases given n and α. The PI is defined as PI $= \frac{\max_\rho \text{type I error} - \alpha}{\alpha}$.

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

threshold is obtained from the independence assumption and uncorrected for dependence. After that, we further demonstrate the power comparison of different methods, where the rejection threshold is corrected with precise type I error control under dependency. We note that the correction is only applicable in simulations and are generally not accessible unless extensive permutation test or simulation-based methods are applied.

Power comparison with uncorrected rejection threshold from independence assumption

In Section 4.1, BJ, HC, $BC_{0.75}$ and Fisher's method are anti-conservative when rejection threshold from independence assumption is used. In other words, the methods lose control of type I error when the dependence structure exists. As a result, we only compare HM, CA, CA^{tr}, $BC_{1.25}$, and $minP$ here to evaluate the power of different methods in varying levels of correlation ρ. Table 3 shows the power of the 5 methods. As expected, the statistical power decreases as ρ increases. HM, CA and CA^{tr} methods have almost identical power and are superior to $BC_{1.25}$. $minP$ is the least powerful method among the five. Different proportions of signals give similar patterns and conclusions.

Power comparison with corrected rejection threshold considering dependence structure

Since methods except for CA, CA^{tr} and HM are either conservative or anti-conservative in type I error control under the presence of correlation, the power comparison in the previous subsection is not completely fair. Here, we evaluate power using the rejection threshold corresponding to the accurate type I error control in each method under each correlation setting, to obtain the corrected rejection thresholds considering dependence structure, for each ρ, we simulate 10^6 Monte Carlo samples for each method by the same sampling procedure in Section 4.1 with assumed correlation, and calculate the empirical rejection threshold from the Monte Carlo samples under the null hypothesis as the critical value for each method.

Table 3: Mean uncorrected power for tests CA, CA^{tr} (truncated Cauchy), HM, $BC_{1.25}$, $minP$ across correlation $\rho = 0, 0.3, 0.6, 0.9, 0.99$ and proportion of signals $s/n = 5\%, 10\%, 20\%$. The standard error is far less than the mean power and hence not shown here.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
5%	CA	0.749	0.629	0.518	0.392	0.347
	CA^{tr}	0.749	0.629	0.518	0.393	0.347
	$BC_1(HM)$	0.749	0.629	0.518	0.393	0.347
	$BC_{1.25}$	0.735	0.617	0.505	0.374	0.321
	$minP$	0.712	0.596	0.482	0.339	0.256
10%	CA	0.870	0.690	0.533	0.371	0.319
	CA^{tr}	0.870	0.690	0.533	0.371	0.318
	$BC_1(HM)$	0.870	0.690	0.533	0.371	0.318
	$BC_{1.25}$	0.850	0.670	0.512	0.342	0.282
	$minP$	0.814	0.639	0.479	0.292	0.194
20%	CA	0.955	0.738	0.542	0.353	0.299
	CA^{tr}	0.955	0.738	0.542	0.353	0.299
	$BC_1(HM)$	0.954	0.737	0.542	0.353	0.298
	$BC_{1.25}$	0.936	0.712	0.513	0.314	0.250
	$minP$	0.895	0.670	0.469	0.249	0.145

We note that this comparison is theoretically a fairer comparison with accurate type I error control but, on the other hand, is less practical in applications unless the dependency structure is known or computationally intensive approaches are applied to precisely control the type I error.

Table 4 shows results of all 9 methods. We order the methods by the index η of Box-Cox transformation as introduced in Section 2: $minP$, $BC_{1.25}$, HM, CA, CA^{tr}, $BC_{0.75}$, Fisher, and then add HC and BJ for comparison. We first observe almost identical results of CA, CA^{tr} and HM, and decreasing power when ρ increases, as expected. We next compare the 5 methods $minP$, $CA/CA^{tr}/HM$ and Fisher with varying proportions of signals and ρ. When $\rho = 0$, Fisher is the least powerful when $s/n = 5\%$ (power= 0.640) but becomes more powerful than $CA/CA^{tr}/HM$ and $minP$ when $s/n = 10\%$ and 20%, showing its superior performance in frequent signals. $CA/CA^{tr}/HM$ consistently have good power in between $minP$ and Fisher. When ρ increases, Fisher quickly drops to almost zero power even with accurate type I error control. For each given s/n, $minP$ is slightly less powerful than $CA/CA^{tr}/HM$ at small ρ but becomes much more powerful than $CA/CA^{tr}/HM$ when ρ is large. This is reasonable since at a very high correlation (e.g., $\rho = 0.99$), all signals can almost be viewed as coming from one source, so taking the smallest p-value gives sufficiently complete information. For $BC_{0.75}$ and $BC_{1.25}$, we observe that the performance of $BC_{1.25}$ is generally intermediate in between $minP$ and $CA/CA^{tr}/HM$, and $BC_{0.75}$ is between $CA/CA^{tr}/HM$ and Fisher. We next compare HC and BJ to the other methods. Although these two methods lose control of type I error under dependency structure and are not the focus of this paper, we are curious about their power performance if correlation structure is correctly considered with Type I error control. As shown in Table 4, BJ is surprisingly powerful for all three proportions of signals when $\rho = 0$ (e.g., power= 0.91 compared to power= $0.640 - 0.778$ for the other 7 methods when $s/n = 5\%$). But similar to Fisher's

method, BJ's power quickly drops to almost 0 with the existence of dependency. The power of HC is generally similar to $CA/CA^{tr}/HM$ but becomes weaker than $CA/CA^{tr}/HM$ for larger ρ. Both HC and BJ lose much power when ρ increases. One possible explanation is that both tests compare the ordered p-values $p_{(i)}$ with the reference value i/n, which is not the correct reference under null with dependence structure (Liu and Xie, 2020).

4.3 Simulation for the large negative penalty issue in the Cauchy method

As discussed in Section 2.1, p-values close to 1 lead to large negative penalties in the Cauchy method, which can cause significant power loss. Below, we design a Fisher's exact (hypergeometric) test for a 2×2 contingency table to illustrate the issue and evaluate the improvement of the truncated Cauchy method.

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

$\delta = 0.01$ is largely improved for all different α; e.g., type I error is now 0.00077, identical to HM, when $\alpha = 0.001$.

We next evaluate power for HM and CA. Similar to Section 4.2, we simulate 10^5 Monte Carlo samples. All settings are identical to the last paragraph for type I error control except that we now generate 2×2 tables with row-column correlation. We first simulate Y from $Hypergeometric(400, 200, 200)$ under independence assumption. We then simulate $Z \overset{D}{\sim} Bin(200 - Y, p_{11})$ and take $Y + Z$ as the value for the upper-left cell. We note that $p_{11} = 0$ corresponds to the original null hypothesis and the larger effect size p_{11}, the stronger signal. We set $p_{11} = 0.2$ and 0.3 and the powers under different α are shown in Table 5. As expected, larger p_{11} generates higher power for both HM and CA. CA produces much smaller power than HM mainly due to impact from skewed p-values toward 1. CA^{tr} largely alleviates the issue and can perform almost identically to HM.

5. Application

We apply the HM, CA, CA^{tr}, and $minP$ tests to analyze a GWAS of neuroticism (Okbay et al., 2016), a personality trait characterized by easily experiencing negative emotions. The dataset contains $6,524,432$ genetic variants (SNPs) across $179,811$ individuals, and p-values are calculated for all SNPs to represent the association between the variant and neuroticism. We use genome annotations to locate the genic or intergenic region for each variant. The total number of intergenic and genic regions is $78,895$. Within each genic or intergenic region, we combine p-values of variants using the HM, CA, CA^{tr} and $minP$ methods and obtain the combined p-values. Figure 3 shows three Manhattan plots for the combined p-values using the HM, CA and $minP$ methods, respectively. As shown in Figure 3, the combined p-values using CA and HM are almost identical, and they are slightly more significant than those obtained from $minP$. The bottom right plot in Figure

Table 4: Mean corrected power for tests Fisher, $BC_{0.75}$, CA, CA^{tr}(truncated Cauchy), HM, $BC_{1.25}$, $minP$, HC and BJ across correlation $\rho = 0, 0.3, 0.6, 0.9, 0.99$ and proportion of signals $s/n = 5\%, 10\%, 20\%$. The standard errors are far less than the mean power and hence omitted.

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

Table 5: Mean proportion of rejection of CA, HM and CA^{tr}(truncated CA) across $\rho_{11} = 0$(type I error), 0.2(power), 0.3(power). The standard errors are far less than the mean proportion and hence omitted.

ρ_{11}	Methods/Cutoffs	0.05	0.01	0.005	0.001	5×10^{-4}	10^{-4}
$\rho_{11} = 0$	CA	0.00825	0.00182	0.000862	0.00016	0.0000687	0.0000100
	$BC_1(HM)$	0.0386	0.00894	0.00417	0.00077	0.000334	0.0000487
	CA^{tr}	0.0285	0.00729	0.00417	0.00077	0.0000334	0.0000487
$\rho_{11} = 0.2$	CA	0.333	0.202	0.146	0.0582	0.0408	0.0135
	$BC_1(HM)$	0.863	0.525	0.379	0.154	0.108	0.0357
	CA^{tr}	0.848	0.522	0.377	0.154	0.108	0.0361
$\rho_{11} = 0.3$	CA	0.431	0.428	0.420	0.355	0.310	0.190
	$BC_1(HM)$	1.000	0.992	0.972	0.822	0.717	0.440
	CA^{tr}	1.000	0.991	0.971	0.822	0.716	0.440

3 shows the numbers of significant genic or intergenic regions with significance thresholds determined by the Bonferroni procedure (controlling the family-wise error rate at 0.05) and the Benjamini–Hochberg FDR procedure (controlling the false discovery rate at 0.05; the significant threshold is $p_{(k)}$, where k is the largest integer such that $p_{(k)} \leq \frac{0.05k}{n}$), or p-value threshold at 10^{-4}, 10^{-5} or 10^{-6}. In all different significance thresholds, the numbers of statistically significant genes for HM and CA are almost identical, and they are generally larger than those from $minP$. Particularly, HM and CA both identify 750 regions under FDR= 5% while $minP$ only finds 476 regions.

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

functions.

We next investigate two regions, SLC2A9 and PCSK6, with small combined p-values by HM $p = 9.534 \times 10^{-4}$ for SLC2A9 and $p = 1.527 \times 10^{-3}$ for PCSK6; q-values q=0.0759 for SLC2A9 and q=0.0939 for PCSK6) but not by CA ($p = 0.9999$ and 0.9999 and q-values both equal 1). The SLC29A9 gene has been found related to Alzheimer's disease and PCSK6 is related to structural asymmetry of the brain and handedness. We suspect the difference of HM and CA comes from p-values close to 1 as described in Section 4.3. Figure S2 shows two jitter plots of p-values for SNPs in genes SLC2A9 (right) and PCSK6 (left). Both of these two genes contain multiple SNPs with very small p-values (e.g. 17 SNPs with $p < 10^{-4}$ in SLC2A9 and 8 SNPs for PCSK6) so the gene regions could potentially be significant. But since both genes also contain many SNPs with p-values close to 1 (5 SNPs with $p > 0.99$ for SLC2A9 and 9 SNPs for PCSK6), CA is impacted and produces larger combined p-values than HM, a situation similar to that described in Section 4.3. Since there are above 500 p-values to combine for both genes, by applying CA^{tr} at $\delta = 0.99$ with approximation by GCLT (Proposition S1), the p-values improve to 9.531×10^{-4} for SLC2A9 and 1.532×10^{-3} for PCSK6, which are almost identical to the p-values calculated by HM.

6. Discussion

In this paper, we investigate methods for combining dependent p-values using transformations corresponding to regularly varying distributions, which is a rich family of heavy-tailed distributions and includes Pareto distribution (Box-Cox transformation) as a special case. We first present the issue of aggregating multiple p-values in three major historical scenarios: (1) classical meta-analysis of combining independent and frequent signals (e.g., Fisher), (2) methods for aggregating independent weak and sparse signals (e.g., $minP$, higher criticism and Berk-Jones), and (3) recent methods for combining p-values with sparse signals

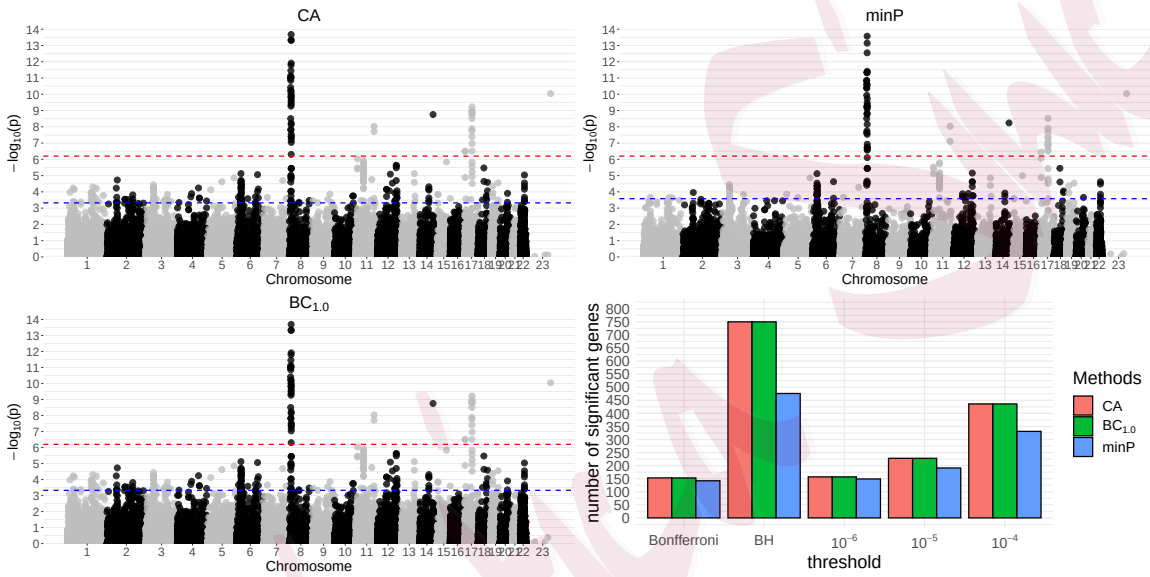

Figure 3: Mahattan plots and number of significant p-values for CA, $BC_1(HM)$ and $minP$. The red dash lines are the cutoffs of Bonferroni correction for $\alpha = 5\%$ and the blue dash lines are the cutoffs of Benjamini-Hochberg correction for FDR=5%. The significant regions (FDR=5%) detected by HM and CA are the same except two regions, DDX58 (q=0.0499 by CA and q=0.0501 by HM) and POU2F3 (q=0.0509 by CA and q=0.0492 by HM).

and unknown dependency structure (i.e., Cauchy and harmonic mean). We then examine popular methods designed for these three settings under the Pareto and regularly varying distribution to provide theoretical insight and finally present the condition of heavy-tailed transformation methods to have the robustness with dependency structure.

Our contributions are fourfold in both providing theoretical insight and practical application guidelines. Firstly, in Section 2, we use the family of Box-Cox transformations, or equivalently transformations by CDF of Pareto distributions, to provide connections among Fisher, CA, HM, and $minP$ methods that are designed to specialize in the three scenarios. We also show that the two recent methods – CA and HM – are approximately identical. Secondly, in Section 3, we focus on the dependent p-value scenario and investigate the condition for p-value combination methods with regularly varying distributions to have the robustness to dependency structure, where CA and HM are special cases. We show that only methods of the equivalent class of CA and HM (i.e., index $\gamma = 1$) in the regularly varying distributions have the robustness property. Thirdly, we demonstrate an occasional drawback of the Cauchy method when some p-values are close to 1, which contributes to the large negative penalty and causes power loss. We propose a simple yet practical solution by a truncated Cauchy method with fast and accurate computation. Finally, the simulations and a real GWAS application confirm the theoretical insights and provide a practical guideline for using the harmonic mean and Cauchy methods. Specifically, Table 2 in Section 4.1 gives guidance of the degree of possible type I error inflation of the harmonic mean method under varying n (number of combined p-values), ρ (correlation level between p-value) and α (test size).

Modern data science faces challenges from larger data dimension, increased structural complexity, and the need for models and inference to tailor for the subject domain. The three categories of p-value combination methods have motivated the development of numerous

methods in the literature and is a good example of how statistical theories can provide insight into method development and guide towards real applications. In our paper, we conclude that the condition in regularly varying distribution to have dependency structure robustness in p-value combination is those distributions with index $\gamma = 1$, which includes Cauchy and harmonic mean methods recently proposed. For future direction, it is of interest whether other methods (e.g., inverse Gamma or log-Gamma family) satisfying this condition may enjoy robustness and simultaneously obtain better statistical power in some applications of interest.

Supplementary material

Supplementary material includes the proofs of Proposition S2, Theorem 1–3 and technical lemmas, additional simulation results, as well as details of the efficient importance sampling procedure for the truncated Cauchy method.

Acknowledgements

The authors would like to thank constructive comments and suggestions from the associate editor and reviewers, which significantly improved the paper. The authors also thank Zhao Ren for multiple inspiring discussions. YF and GCT are funded by NIH R21LM012752; and CC is funded by Ministry of Science and Technology of ROC 109-2118-M-110-002.

References

Barnett, I., R. Mukherjee, and X. Lin (2017). The generalized higher criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association 112*(517), 64–76.

REFERENCES

Berk, R. H. and D. H. Jones (1979). Goodness-of-fit test statistics that dominate the Kolmogorov statistics. *Probability Theory and Related Fields 47*(1), 47–59.

Chen, Y., P. Liu, K. S. Tan, and R. Wang (2021). Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Statistica Sinica*. Also available as "`https://www3.stat.sinica.edu.tw/preprint/SS-2021-0071_Preprint.pdf`".

Chen, Y. and K. C. Yuen (2009). Sums of pairwise quasi-asymptotically independent random variables with consistent variation. *Stochastic Models 25*(1), 76–89.

Donoho, D. and J. Jin (2004). Higher criticism for detecting sparse heterogeneous mixtures. *The Annals of Statistics 32*(3), 962–994.

Edgington, E. S. (1972). An additive method for combining probability values from independent experiments. *The Journal of Psychology 80*(2), 351–363.

Fisher, R. et al. (1934). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh and London.

Goeman, J. J., J. D. Rosenblatt, and T. E. Nichols (2019). The harmonic mean p-value: Strong versus weak control, and the assumption of independence. *Proceedings of the National Academy of Sciences of the United States of America 116*(47), 23382.

Goldie, C. M. and C. Klüppelberg (1998). Subexponential distributions. *A practical Guide to Heavy Tails: Statistical Techniques and Applications*, 435–459.

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

Li, J. and D. Siegmund (2015). Higher criticism: p-values and criticism. *The Annals of Statistics 43*(3), 1323–1350.

Littell, R. C. and J. L. Folks (1971). Asymptotic optimality of Fisher's method of combining independent tests.

Journal of the American Statistical Association 66(336), 802–806.

Liu, Y. and J. Xie (2019). Accurate and efficient p-value calculation via Gaussian approximation: a novel monte-carlo method. *Journal of the American Statistical Association 114*(525), 384–392.

Liu, Y. and J. Xie (2020). Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. *Journal of the American Statistical Association 115*(529), 393–402.

Mikosch, T. (1999). *Regular variation, subexponentiality and their applications in probability theory*. Eindhoven University of Technology.

Mudholkar, G. S. and E. O. George (1979). The logit method for combining probabilities. In *Symposium on Optimizing Methods in Statistics*, pp. 345–366. Academic Press New York.

Okbay, A., B. M. Baselmans, J.-E. De Neve, P. Turley, M. G. Nivard, M. A. Fontana, S. F. W. Meddens, R. K. Linnér, C. A. Rietveld, J. Derringer, et al. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics 48*(6), 624–633.

Pearson, K. (1933). On a method of determining whether a sample of size n supposed to have been drawn from a parent population having a known probability integral has probably been drawn at random. *Biometrika*, 379–410.

Stouffer, S. A., E. A. Suchman, L. C. DeVinney, S. A. Star, and R. M. Williams Jr (1949). The American soldier: Adjustment during army life. *Studies in Social Psychology in World War II 5*.

Vovk, V., B. Wang, and R. Wang (2021). Admissible ways of merging p-values under arbitrary dependence. *Annals of Statistics*.

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

Wilson, D. J. (2019a). The harmonic mean p-value for combining dependent tests. *Proceedings of the National*

Academy of Sciences 116(4), 1195–1200.

Wilson, D. J. (2019b). Reply to Held: When is a harmonic mean p-value a Bayes factor? *Proceedings of the National Academy of Sciences 116*, 5857–5858.

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