Statistica Sinica **33** (2023), 1115-1142
doi:https://doi.org/10.5705/ss.202022.0046

HEAVY-TAILED DISTRIBUTION FOR COMBINING DEPENDENT p-VALUES WITH ASYMPTOTIC ROBUSTNESS

Yusi Fang[1], Chung Chang[2], Yongseok Park[1] and George C. Tseng[1]

[1]*University of Pittsburgh and* [2]*National Sun Yat-sen University*

Abstract: In statistics, researchers sometimes combine individual p-values to aggregate multiple small effects. Recent advances in big data analysis have led to methods that aggregate correlated, sparse, and weak signals. In this context, we investigate a wide range of p-value combination methods, formulated as the sum of p-values that are transformed using a broad family of heavy-tailed distributions, namely, regularly varying distributions. Here, we also include the Cauchy and harmonic mean tests. We explore the conditions under which a method of the family is robust to dependency for type-I error control and possesses optimal power in terms of the boundary used to detect 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 that belongs to the equivalent class with fast computation to address the problem caused by the large negative penalty in the Cauchy method. We use comprehensive simulations to verify our theoretical insights and provide practical recommendations. Finally, we apply the truncated Cauchy method to data from a neuroticism genome-wide association study to illustrate our theoretical findings in the regularly varying distribution family and the advantages of the method.

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

1. Introduction

Combining p-values to aggregate information from multiple sources is popular in the social sciences and biomedical research. Classical methods focus on combining multiple independent and frequent signals to increase the statistical power, which can be viewed as a type of meta-analysis. Consider the combination of n independent p-values, $\boldsymbol{p} = (p_1, \ldots, p_n)$. Early methods used $T(\boldsymbol{p}) = \sum_{i=1}^{n} g(p_i) = \sum_{i=1}^{n} F_U^{-1}(1 - p_i)$ to sum transformed p-values, where the transformation $g(p)$ is the inverse cumulative distribution function (CDF) of a random variable U. Conventional methods in this category include Fisher's

Corresponding author: Chung Chang, Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804201, Taiwan. Email: cchang@math.nsysu.edu.tw.

method (Fisher (1934)) where $T = \sum_{i=1}^{n} -2\log(p_i)$ and U is a chi-squared distribution, and Stouffer's method (Stouffer et al. (1949)) where $T = \sum_{i=1}^{n} -\Phi^{-1}(p_i)$ and U is a standard normal distribution, among others (Edgington (1972); Pearson (1933); Mudholkar and George (1979)). These methods use a classical meta-analysis to combine independent and relatively frequent signals, and apply a light-tailed distribution (i.e., tails thinner than an exponential function) for U. The efficiency of such methods is mostly considered under the asymptotic framework that the number of p-values n is fixed and sample size m used to derive each p-value goes to infinity, where $p = O(e^{-m})$ in most cases. Under this setting, it has been shown that only the equivalent class of Fisher's method is asymptotically Bahadur optimal (ABO), meaning that the efficiency of the combined p-value statistics is asymptotically optimal under fixed n and $m \to \infty$ (Littell and Folks (1971)).

With the advent of big data, many studies now combine p-values with large n. The seminal paper by Donoho and Jin (2004) established a framework for combining p-values with weak and sparse signals, and proposed the higher-criticism test with the asymptotically optimal property. This second category of methods considers $n \to \infty$, and only a small number s of the n p-values ($s = n^{\beta}$ where $0 < \beta < 1/2$) have weak signals ($p = O(n^{-r}/(\log n)^{1/2})$ 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 ($minP$) $T = min_{1 \le i \le n} p_i$ is asymptotically optimal in terms of the detection boundary only for $0 < \beta < 1/4$, whereas higher criticism attains an optimal detection boundary for all possible $0 < \beta < 1/2$. Several methods, including the Berk-Jones test (Berk and Jones (1979); Li and Siegmund (2015)), have subsequently been proposed to improve the finite-sample power of higher criticism, while maintaining an optimal detection boundary.

All of the aforementioned methods were developed to combine independent p-values. However, many modern large-scale data analyses need to combine a large number of dependent p-values that have sparse and weak signals, which we categorize as methods of the third category. A notable application is to combine p-values of multiple correlated SNPs (there may be tens to hundreds or thousands) in an SNP set (e.g., all SNPs in a gene region or in gene regions of a pathway) in a genome-wide association study (GWAS). In this case, neighboring SNPs often have unknown dependency structures, prompting efforts to extend existing tests to account for dependency using permutations or other numerical simulation approaches (e.g., Liu and Xie (2019)). However, permutation or simulation-based methods are not practical when n is large, and a precise p-value is needed for multiple comparisons. The null hypothesis may also be difficult to simulate using a permutation. Barnett, Mukherjee and Lin (2017) developed an analytic approximation for higher criticism that incorporated a dependency structure. However, the method is still computationally intensive and not sufficiently accurate for the small p-values needed for multiple comparisons. Motivated by these needs, Liu and Xie (2020) and Wilson (2019a) independently proposed the Cauchy combination test ($T = \sum_{i=1}^{n} \tan\{(0.5 - p_i)\pi\}$) and the harmonic mean combination test ($T = \sum_{i=1}^{n} 1/p_i$), respectively, to combine p-values under an unspecified dependency structure. Wilson (2019a) also provided a convenient R package called harmonicmeanp (function p.mamml) to implement the harmonic test. A remarkable property of both methods is that the null distributions and testing procedures derived from the independence assumption are robust under a dependency structure in an asymptotic, but practical sense; see Section 3.1. Motivated by this observation, we consider a rich family of test statistics that includes the Cauchy and harmonic mean tests. More precisely, we consider the test statistic

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

where the transformation $g(p)$ corresponds to U from a regularly varying distribution family, which is a broad family of heavy-tailed distributions. We investigate the conditions required to achieve the practical robustness to dependency of the Cauchy and harmonic mean methods. Note that selections of U in classical meta-analysis settings (fixed n and $m \to \infty$) are all from thin-tailed distributions (e.g., chi-squared distribution for Fisher's methods and the Gaussian distribution for Stouffer's method). This is reasonable, because a thin-tailed distribution produces contributions that are more even from marginally significant p-values in meta-analyses of frequent signals. In contrast, the Cauchy and harmonic mean methods correspond to heavy-tailed distributions of U, which focus on small p-values and down-weigh marginally significant p-values. Figure 1 shows the transformation function of $g(p)$ in log-scale. For Fisher's method, the contributions of the p-values 10^{-2} and 10^{-6} to the test statistics are 4.6 and 13.8, respectively. For heavy-tailed transformation methods, the contributions become 100 and 10^6 for the harmonic mean, and 31.82052 and 3.18×10^5 for the Cauchy method. With an increased focus on small p-values, the methods are more powerful for detecting sparse signals. Note that Vovk and Wang (2020) also considered the sum of transformed p-values to combine p-values, and showed an upper bound of the significance level inflation under an arbitrary dependence structure. The comparison of our results with theirs is provided in the remark following Theorem

2. Wilson (2019b), Wilson (2020), and Vovk, Wang and Wang (2022) also did related work involving combining dependent p-values.

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

2. Connection between minP, Harmonic Mean, Cauchy, and Fisher

2.1. Using a Pareto distribution to connect four existing methods

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

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

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

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

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

Proof. By Taylor's expansion, $g^{(CA)}(p) = \tan\{(0.5 - p)\pi\} \approx 1/\pi p - \pi p/3 - (\pi p)^3/45 + \cdots$. The result follows immediately. Chen et al. (2023) also showed a similar result.

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

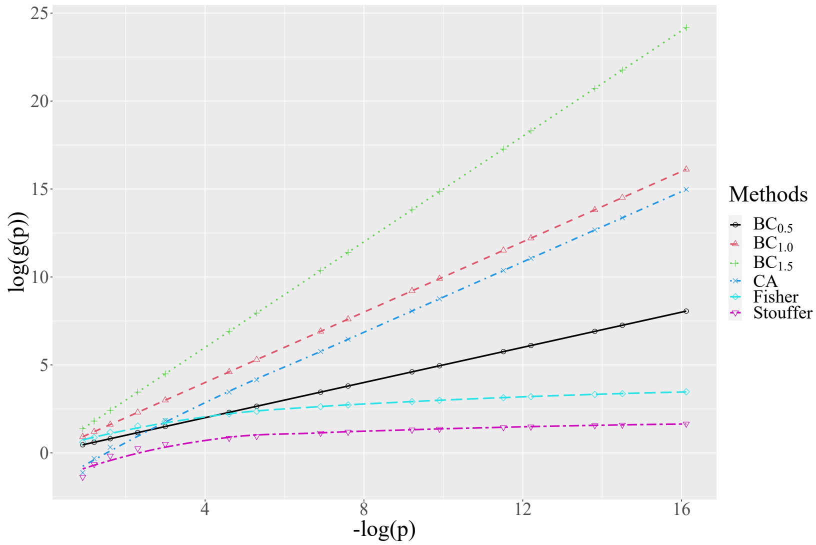

Figure 1. Comparison of transformations. We show six transformations of p-values, $g(p)$, i.e., $BC_{0.5}$, BC_1 (HM), $BC_{1.5}$, CA, Fisher, and Stouffer. The x-axis is $-\log(p)$, and the y-axis shows $\log(g(p))$.

to that of the Cauchy and harmonic mean methods. To answer this question, we introduce a family of regularly varying distributions, and investigate its properties in Section 3.

Figure 1 shows a minus log-scaled p transformation $g(p)$ versus a minus log-scaled transformation $g(p)$ for $BC_{0.5}$ (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 method and Stouffer's method. We find that as η increases, smaller p-values become more dominant and the effect of marginally significant p-values rapidly diminishes, yielding stronger power for sparse signal applications. CA and HM are approximately proportional when p is sufficiently small (roughly when $p < 10^{-2}$).

Although HM and CA are approximately equivalent when combining relatively small p-values, when a p-value is very close to one, the contribution in the Cauchy method is close to negative infinity, which can cause numerical issues and substantial power loss; we refer to this as the "large negative penalty issue" in relation to the Cauchy method. A p-value close to one happens often in tests of discrete data, in which case, the p-values under the null hypothesis may not necessarily be $Unif(0, 1)$. The p-values may also be close to one when n is large or when the model used to derive the p-values is misspecified. As a simple remedy, we propose a truncated Cauchy test (CA^{tr}) that truncates any of the n p-values greater than $1-\delta$ to be $1-\delta$. For example, when $\delta = 0.01$, we have $p^{tr} = p$ if $p < 0.99$, and $p^{tr} = 0.99$ if $p \geq 0.99$. We recommend using $\delta = 0.01$. Conceptually, δ should be sufficiently large so that it avoids the large negative penalty issue in Cauchy. However, for computational purposes, it cannot be too large, or the approximation by our fast-computing procedures may not be accurate. A detailed justification for choosing $\delta = 0.01$, with support from simulation results, is given in the Supplementary Material, Section S2.4. The proposed method can also be viewed as a sum of transformed p-values. Indeed, the CA^{tr} statistic can be written as

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

For more details on CA^{tr}, see the Supplementary Material, Section S2.

3. Asymptotic Properties of Regularly Varying Methods for p-value Combination

3.1. Disbributions with regularly varying tails

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

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

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

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

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

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

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

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

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

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

3.2. Asymptotic tail probability approximation and robustness to dependence

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

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

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

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

Condition (A1) is mild and is also assumed in Liu and Xie (2020) when investigating the robustness of the Cauchy method under an unspecified dependence structure. Throughout this paper, the term "unspecified dependence structure" indicates an unspecified Gaussian correlation structure. This condition guarantees that the tail distributions of each pair of U_i and U_j are asymptotically tailed independent; see the precise definition of asymptotically tailed independence for a pair of random variables in the Supplementary Material, 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 the "balance condition", and is a common condition for random variables with regularly varying tails (Goldie and Klüppelberg (1998)). For example, for the harmonic mean method, $p = 1$ and $q = 0$, for the Cauchy method, $p = q = 1/2$, and for the truncated Cauchy method, $p = 1$ and $q = 0$.

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

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

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

Below, we perform a simple simulation to demonstrate and investigate Theorem 1. Assume $n = 3$, and $\boldsymbol{X} = (X_1, X_2, X_3)$ is multivariate normal with unit variance and common pairwise correlation $\rho_{ij} = \rho$ $(1 \le i < j \le 3)$. In this simulation, we set $\rho = 0$, 0.3, 0.6, 0.9, and 0.99. Here, we consider seven 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 $y(\alpha) = nP(U > t_\alpha)/P(T(\boldsymbol{X}) > t_\alpha)$ 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 $-1 < \rho < 1$. Figures 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). Note that as ρ increases, a smaller α will be required for a good approximation. Theorem 2 further characterizes what would happen if some of the p-values have perfect correlations $\rho_{ij} = 1$ or -1.

Theorem 2. *Suppose conditions* (A1), (A2), *and* (A3) *in Theorem* 1 *hold. Define an arbitrary weight vector* $\boldsymbol{w} = (w_1, \ldots, w_n) \in R_+^n$, $T_{n,\boldsymbol{w}}(\boldsymbol{X}) = \sum_{i=1}^n w_i g(p_i)$. *Furthermore, assume* $\rho_{ij} = 1$ *or* -1 *for* $1 \le i < j \le m$, *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 $\boldsymbol{w} = (1, \ldots, 1)$. An immediate consequence of Theorem 2 is that only when $\gamma = 1$ (e.g., the HM, 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, Figures 2A-2E already show a hint that the convergence of Theorem 1 becomes increasingly difficult when ρ increases to almost one. When some of the p-values have perfect correlation, only index $\gamma = 1$ of the regularly varying distribution is asymptotically robust to an unspecified dependence structure. Figure 2F shows a 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

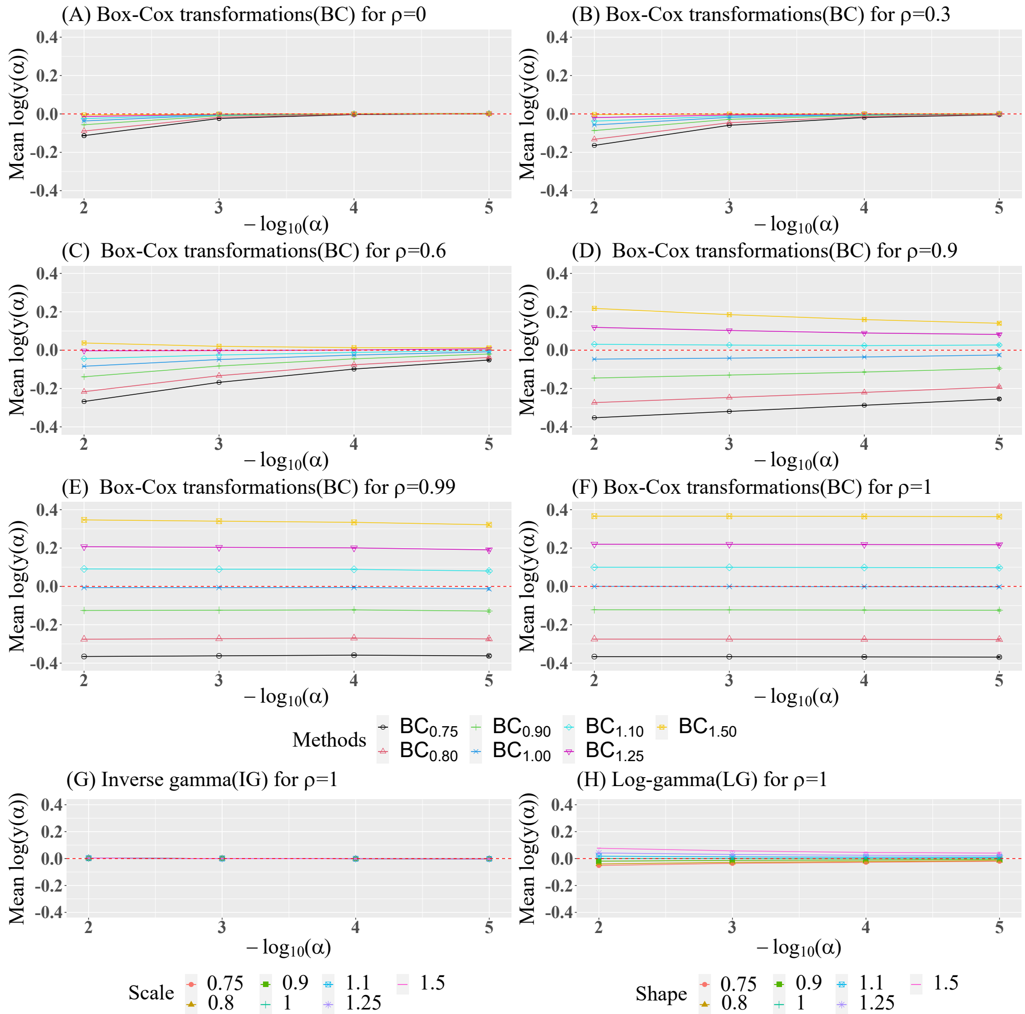

Figure 2. The mean log-scaled $y(\alpha)$ for Box-Cox transformations, inverse gamma and log-gamma across different significance levels α. (A)-(F) represent the results of Box-Cox transformations with values of $\eta = 0.75$, 0.8, 0.9, 1, 1.1, 1.25, 1.5 for correlation level ρ =0, 0.3, 0.6, 0.9, 0.99,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

robustness to perfect correlation. Although Figure 2E ($\rho = 0.99$) and Figure 2F ($\rho = 1$) are visually similar, all BC methods in Figure 2E eventually converge to zero as $\alpha \to 0$, by Theorem 1, although very slowly. On the other hand, in Figure 2F, only BC_1 converges to zero, by Theorem 2.

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

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

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, 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 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. Figures 2G and 2H verify that because the inverse gamma and log-gamma are also regularly varying distributions with index $\gamma = 1$, they enjoy an asymptotic robustness to the correlation structure, similar to that of HM (BC_1) and Cauchy, even when perfect correlation exists. Another important implication of this corollary is that among all the tests that use transformations of regularly varying distributions, only the type-I errors of those corresponding to $\gamma \leq 1$ are well preserved asymptotically (i.e., tests that are at least not anti-conservative asymptotically) under any correlation structure.

Corollary 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 $\sum_{i=1}^n w_i^\gamma P(U > t)$. Similarly to the unweighted version in Theorem 1, because the approximation in Corollary 2 is independent of the correlation structure, $P(T_{n,\boldsymbol{w}}(\boldsymbol{X}) > t)$ under the dependence structure can be approximated by calculating $P(w_1 U_1 + \cdots + w_n U_n > t)$ under the independence assumption using a Monte Carlo simulation, as long as there are no perfect correlations between U_i. Furthermore, note that this formula can be considered an extension of Corollary 1.3.8 in (Mikosch (1999)), in which $U_1, \ldots, U_n$ are assumed to be independent regularly varying distributed random variables.

Remark 1. Note that the robustness property of Theorems 1 and 2 is similar to (Liu and Xie (2020); Wilson (2019a)) and describes only the asymptotic behavior of the tail probability of our proposed family. Indeed, the results of Theorems 1 and 2 guarantee only that the type-I errors of the corresponding tests ($\gamma = 1$, equivalent to the harmonic mean and Cauchy) can be well controlled for a small size α, given fixed n and $\mathbf{\Sigma}$. Intuitively, as n increases, a more stringent cutoff corresponding to a small α is needed to ensure the robustness of type I error control. An ideal robustness property for the type-I error should achieve a uniform upper tail bound in the sense of $P(T(\boldsymbol{X}) > t_\alpha) \leq c \cdot \alpha$ under any dependence structure $\mathbf{\Sigma}$, where t_α is the tail threshold when a nominal α is controlled under the independence assumption, c is independent of n, and $\mathbf{\Sigma}$ is in a reasonable magnitude (e.g., $c = 1.5$, meaning the inflation of the type-I error is at most 50%, in the worst scenario). However, this uniform bound is not achievable, in general. Vovk and Wang (2020) recently provided a remarkable uniform bound for arbitrary dependency structure (note that ours is an unspecified dependency structure), but dependent on n for the HM method:

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

where the adjusted factor α_n^{HM} is between $\log(n)$ and $e \cdot \log(n)$ (see Proposition 6 in Vovk and Wang (2020)). However, this bound is not practical in general applications because, considering $n = 100$ or 1000, the inflation bound $\alpha_n^{HM} \geq \log(n)$ is at least 4.6- or 6.9-fold greater. Furthermore, the factor α_n^{HM} is comparable with the type-I error in the case of a perfect correlation (i.e., $\rho = 1$), instead of the nominal size α under independence. On this issue, Goeman, Rosenblatt and Nichols (2019) pointed out an extreme case that when $n = 10^5$ and $\mathbf{\Sigma}$ has exchangeable correlation $\rho = 0.2$, HM has a more than threefold type-I error inflation (true type-I error $= 0.164$ under nominal $\alpha = 0.05$). In Section 4.1, we perform extensive simulations for a wide range of n and size α to investigate the limitation and develop practical guidance for applying the HM method.

The discussion above indicates that with a mild normality assumption, the upper bound for the inflation of type-I errors is much smaller than that under an arbitrary dependence structure. This is especially useful, because using a smaller upper bound of the inflation of type I errors to adjust the significance level increases the power of the test. Furthermore, based on Theorem 2 and simulations, we can develop a practical guideline to adjust the significance level for the HM test ($\eta = 1$), and for any test that is a sum of transformations by a distribution with a regularly varying tail, including any BC_η test.

3.3. Detection boundary of regularly varying methods

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

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

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

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

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

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

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

for all $t > t_1$.

Condition (C2) is for the tail probability of U_i and is a mild condition because $\bar{F}(t) = P(U_i > t) = L(t)/t^\gamma$ ($L(t)$ is a slowly varying function). This condition holds for all commonly used distributions with regularly varying tails with index γ. In the Supplementary Material, 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 the Supplementary Material, Remark S6 for more details.

Theorem 3 states that the power of this test $T(\boldsymbol{X})$ converges to one for any significance level $\alpha > 0$ and $0 < \gamma \leq 1$, or equivalently, that the sum of the Type-I and Type-II errors goes to zero, given the setup. Moreover, Theorem 3 implies that the methods with $0 < \gamma \leq 1$ attain the optimal detection boundary defined in Donoho and Jin (2004) in the strong sparsity situation $0 < \beta < 1/4$. Liu and Xie (2020) showed a similar result for their proposed Cauchy test. As discussed in Section 2, the Cauchy distribution has a regularly varying tail with index $\gamma = 1$. This theorem is valid for methods of distributions with regularly varying tails with index $0 < \gamma \leq 1$. Therefore, this theorem can be considered a generalization of Theorem 3 in Liu and Xie (2020).

4. Simulations

In this section, we perform simulations to compare the robustness of different p-value combination methods under varying correlation levels between p-values in order to verify the theoretical results presented in Sections 2 and 3. We include the seven 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 first evaluates the type-I error control of the methods under independence and varying levels of correlation to verify the robustness of the HM and Cauchy methods. Furthermore, because the robustness in Theorem 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 evaluate

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

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

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

Note that according to Theorems 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 guarantees only the dependence robustness for very large t (or equivalently very small α). We also expect that larger n will require a larger

t (smaller α) to ensure a good approximation. Specifically, Goeman, Rosenblatt and Nichols (2019) noted that with $\rho = 0.2$ and $n = 10^5$, the much inflated type-I error of 0.164 is obtained for size $\alpha = 0.05$. Therefore, it is of interest to explore the robustness property of $T(\boldsymbol{X})$ for dependence in HM for varying n, α, and ρ in order to provide practical guidance in real applications. In Table 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 $\mathrm{PI} = (\max_\rho \text{type-I error} - \alpha)/(\alpha)$. The result confirms the theoretical result that a larger n generates greater type-I error inflation under dependence for a fixed α, and requires a much smaller α to improve the type-I error inflation. For example, when $\alpha = 0.01$, we have $PI = 30\%$ for $n = 25$, compared with $PI = 80\%$ for $n = 10000$. On the other hand, when $n = 10000$, PI decreases from 80% to 49% when α decreases from 0.01 to 0.0001. In general, this result shows robust type-I error control under varying correlation levels, in a practical sense, when $n \leq 1{,}000$ and $\alpha \leq 0.05$ with the maximum $PI = 50\%$, which inflates type I error from $\alpha = 0.01$ to 0.015 at $n = 1000$ and $\rho = 0.3$. Even when n increases to 10,000, PI only minimally increases to 80%. When multiple comparisons are needed, such as in GWAS applications, a small α is targeted, and HM achieves robust type-I error control, in general, in a practical sense. However, if a single test is performed with a very large n, we need to be careful with the type-I error inflation (e.g., type-I error is 0.072 for $\alpha = 0.05$ when $n = 10000$ and $\rho = 0.3$).

4.2. Statistical power

In this subsection, we follow the simulation setting in Section 4.1 to evaluate the statistical power of the methods under different values of correlation ρ and strengths of the signal. Following the sparse and weak signal setting in Donoho and Jin (2004), we design the n signals $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ to contain $n - s$ with no signal ($\mu_{s+1} = \cdots = \mu_n = 0$), and the first s to have nonzero signals $\mu_1 = \cdots = \mu_s = \mu_0 = \sqrt{4\log(n)}/s^{0.1}$, where $s/n = (5\%, 10\%, 20\%)$. We first compare the power of the methods under varying correlations, where the rejection threshold is obtained from the independence assumption, and is uncorrected for dependence. Furthermore, we compare the power of the methods, where the rejection threshold is corrected with precise type-I error control under dependency. Note that the correction applies only in simulations, and is not accessible, in general, without applying extensive permutation tests or simulation-based methods.

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

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}
1,000	$\rho = 0.6$	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}
2,000	$\rho = 0.6$	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}
10,000	$\rho = 0.6$	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%

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

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

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

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

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

Table 4 shows the results for all nine methods. We order the methods by the index η of the 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 for CA, CA^{tr}, and HM, and decreasing power when ρ increases, as expected. We next compare the five methods $minP$, $CA/CA^{tr}/HM$, and Fisher with varying proportions of signals and ρ. When $\rho = 0$, Fisher is the least powerful when $s/n = 5\%$ (power $= 0.640$), but becomes more powerful than $CA/CA^{tr}/HM$ and $minP$ when $s/n = 10\%$ and 20%, showing its superior performance in frequent signals. $CA/CA^{tr}/HM$ consistently have good power between that of $minP$ and Fisher. When ρ increases, Fisher quickly drops to almost zero power, even with accurate type-I error control. For each given s/n, $minP$ is slightly less powerful than $CA/CA^{tr}/HM$ at small ρ, but becomes much more powerful than $CA/CA^{tr}/HM$ when ρ is large. This is reasonable because at a very high correlation (e.g., $\rho = 0.99$), all signals can be viewed as coming from one source, so taking the smallest p-value gives sufficiently complete information. For $BC_{0.75}$ and $BC_{1.25}$, we observe that, in general, the performance of $BC_{1.25}$ lies between that of $minP$ and $CA/CA^{tr}/HM$, and that of $BC_{0.75}$ lies between that of $CA/CA^{tr}/HM$ and Fisher. We next compare HC and BJ with the other methods. Although these two methods lose control of the type-I error under a dependency structure, and are not the focus of this study, we are curious about their power performance if the correlation structure is correctly considered with type-I error control. As shown in Table 4, BJ is surprisingly powerful for all three proportions of signals when $\rho = 0$ (e.g., power$= 0.91$ compared with power $= 0.640 - 0.778$ for the other seven methods when $s/n = 5\%$). However, similarly to Fisher's method, the power of BJ drops quickly to almost zero with the existence of dependency. The power of HC is, in general, similar to that of $CA/CA^{tr}/HM$, 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 the null with a 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 one lead to large negative penalties in the Cauchy method, which can cause significant power loss. Below, we design a Fisher's exact (hypergeometric) test for a 2×2 contingency table to illustrate the issue and evaluate the improvement offered by the truncated Cauchy method.

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

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

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 are omitted.

s/n	Methods	$\rho = 0$	$\rho = 0.3$	$\rho = 0.6$	$\rho = 0.9$	$\rho = 0.99$
5%	$Fisher$	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
	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
10%	$Fisher$	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
	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
20%	$Fisher$	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
	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

are skewed toward one. CA^{tr} largely alleviates the issue and performs 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 data set contains 6,524,432 genetic variants (SNPs) across 179,811 individuals, and the p-values are calculated for all SNPs to represent the association between the variant and neuroticism. We use genome annotations to locate the genic or intergenic region for each variant. The total

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 are omitted.

ρ_{11}	Methods/Cutoffs	0.05	0.01	0.005	0.001	5×10^{-4}	10^{-4}
	CA	0.00825	0.00182	0.000862	0.00016	0.0000687	0.0000100
$\rho_{11} = 0$	$BC_1(HM)$	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
	CA	0.333	0.202	0.146	0.0582	0.0408	0.0135
$\rho_{11} = 0.2$	$BC_1(HM)$	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
	CA	0.431	0.428	0.420	0.355	0.310	0.190
$\rho_{11} = 0.3$	$BC_1(HM)$	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

number of intergenic and genic regions is 78,895. Within each genic or intergenic region, we combine the p-values of the variants using the HM, CA, CA^{tr}, and $minP$ methods. 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 are slightly more significant than those obtained from $minP$. The bottom-right plot in Figure 3 shows the numbers of significant genic or intergenic regions, with the significance thresholds determined using the Bonferroni procedure (controlling the family-wise error rate at 0.05) and the Benjamini–Hochberg FDR procedure (controlling the false discovery rate at 0.05; the significant threshold is $p_{(k)}$, where k is the largest integer such that $p_{(k)} \leq 0.05k/n$), or p-value threshold at 10^{-4}, 10^{-5}, or 10^{-6}. For all significance thresholds, the numbers of statistically significant genes for HM and CA are almost identical, and are larger, in general, than those from $minP$. In particular, HM and CA both identify 750 regions under FDR= 5%, whereas $minP$ finds only 476 regions.

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

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

6. Discussion

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

Our results contribute to the literature in four ways. First, in Section 2, we use the family of Box-Cox transformations, or equivalently, transformations by the CDF of Pareto distributions, to connect the Fisher, CA, HM, and $minP$ methods, which are designed to specialize in the three scenarios. We also show

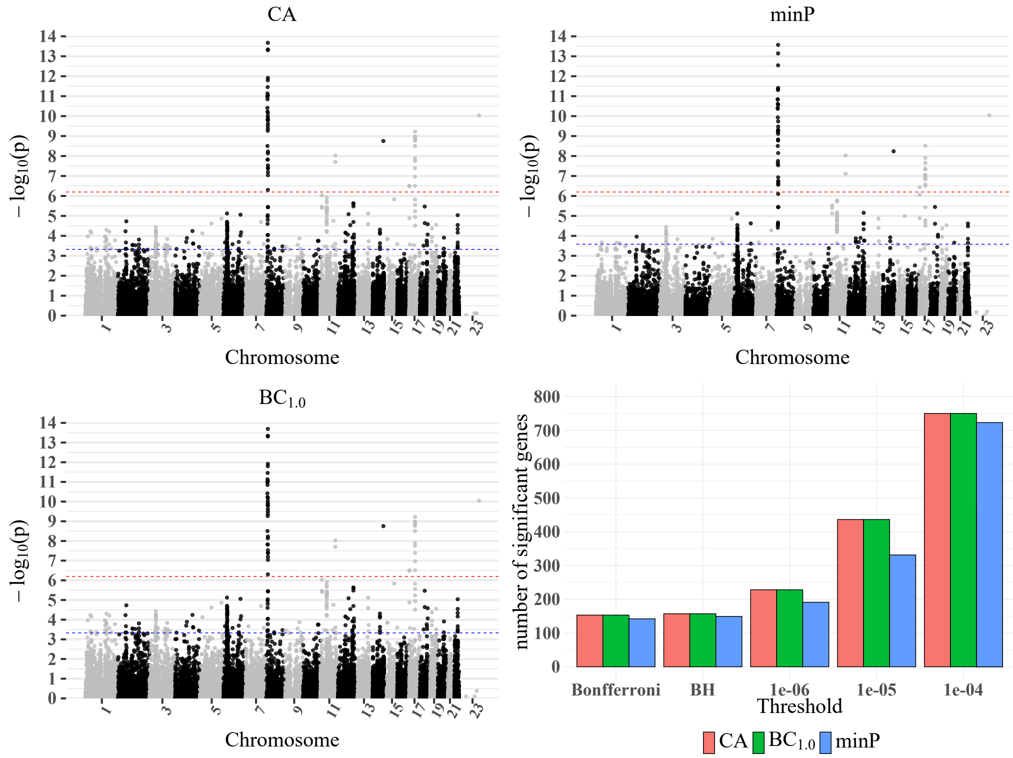

Figure 3. Mahattan plots and number of significant p-values for CA, $BC_1(HM)$, and $minP$. The red dash lines are the cutoffs of the Bonferroni correction for $\alpha = 5\%$, and the blue dash lines are the cutoffs of the Benjamini-Hochberg correction for FDR $= 5\%$. The significant regions (FDR $=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

that two recent methods, CA and HM, are approximately identical. Second, in Section 3, we focus on the dependent p-value scenario, and investigate the condition that p-value combination methods with regularly varying distributions be robust to the dependency structure, where CA and HM are special cases. We show that only methods of the equivalent class of CA and HM (i.e., index $\gamma = 1$) in the regularly varying distributions have the robustness property. Third, we demonstrate an occasional drawback of the Cauchy method when some p-values are close to one, which contributes to the large negative penalty and causes a loss of power. We propose a simple, yet practical solution using a truncated Cauchy method with fast and accurate computation. Finally, the simulations and a real GWAS application confirm our theoretical insights, and provide a practical guideline for using the harmonic mean and Cauchy methods. Specifically, Table

2 in Section 4.1 shows the degree of possible type-I error inflation of the harmonic mean method under varying n (number of combined p-values), ρ (correlation level between p-value), and α (test size).

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

Supplementary Material

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

Acknowledgments

The authors would like to thank the associate editor and reviewers for their constructive comments and suggestions. The authors also thank Zhao Ren for multiple inspiring discussions. YF and GCT were funded by NIH R21LM012752; and CC was funded by the Ministry of Science and Technology of ROC 109-2118-M-110-002.

References

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

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

Chen, Y., Liu, P., Tan, K. S. and Wang, R. (2023). Trade-off between validity and efficiency of merging p-values under arbitrary dependence. *Statistica Sinica* **33**, 851–872.

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

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

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

Fisher, R. (1934). *Statistical Methods for Research Workers.* Oliver and Boyd, Edinburgh and London.

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

Goldie, C. M. and Klüppelberg, C. (1998). Subexponential distributions. In *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, 435–459. Birkhauser Boston Inc., Cambridge.

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

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

Littell, R. C. and Folks, J. L. (1971). Asymptotic optimality of Fisher's method of combining independent tests. *Journal of the American Statistical Association* **66**, 802–806.

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

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

Mikosch, T. (1999). *Regular Variation, Subexponentiality and their Applications in Probability Theory.* Eindhoven University of Technology, Eindhoven.

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

Okbay, A., Baselmans, B. M., De Neve, J.-E., Turley, P., Nivard, M. G., Fontana, M. A. et al. (2016). Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nature Genetics* **48**, 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* **25**, 379–410.

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams Jr, R. M. (1949). The American soldier: Adjustment during army life. In *Studies in Social Psychology in World War II* **1**. Princeton University Press, Princeton.

Vovk, V., Wang, B. and Wang, R. (2022). Admissible ways of merging p-values under arbitrary dependence. *The Annals of Statistics* **50**, 351–375.

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

Wilson, D. J. (2019a). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **116**, 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. In *Wellcome Open Research* **5**. Web: `https://wellcomeopenresearch.org/articles/5-55`.

Yusi Fang

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.

E-mail: yuf31@pitt.edu

Chung Chang

Department of Applied Mathematics, National Sun Yat-sen University, Kaohsiung 804201, Taiwan.

E-mail: cchang@math.nsysu.edu.tw

Yongseok Park

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.

E-mail: yongpark@pitt.edu

George C. Tseng

Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA 15260, USA.

E-mail: ctseng@pitt.edu

(Received January 2022; accepted October 2022)