Statistica Sinica Preprint No: SS-2025-0263	
Title	On Correlation Robustness of Supremum-based P-value Combination Tests
Manuscript ID	SS-2025-0263
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202025.0263
Complete List of Authors	Xiaohui Chen, Fangfang Wang, Hong Zhang and Zheyang Wu
Corresponding Authors	Zheyang Wu
E-mails	zheyangwu@wpi.edu
Notice: Accepted author version.	

ON CORRELATION ROBUSTNESS OF SUPREMUM-BASED p-VALUE COMBINATION TESTS

Xiaohui Chen, Fangfang Wang, Hong Zhang, and Zheyang Wu

Department of Mathematical Sciences, Worcester Polytechnic Institute

Global Biometrics & Data Management, Pfizer Inc., New York

Abstract: A hypothesis testing procedure is considered asymptotically correlation-robust (ACR) if data correlations have a diminishing impact on type I error control at higher significance levels. This property is crucial for analyzing large datasets with complex correlations, such as those in whole-genome sequencing studies. Since such data often require stringent significance thresholds, correlation-robust tests allow the use of independence approximations to reduce computational complexity while maintaining accurate type I error control. This study demonstrates that a broad range of supremum-based p-value combination tests – such as the classic minP, Simes, Higher Criticism, and some phi-divergence tests – are ACR when the data exhibit asymptotically independent tails, a condition satisfied by Gaussianity and various non-Gaussian dependencies characterized by appropriate pairwise copulas. A theoretical power analysis further demonstrates that, in the stringent-significance regime, the supremum-based p-value combination tests and the equal-weight Cauchy Combination Test possess the same first-order asymptotic power under the dependence conditions that guarantee ACR. Furthermore, we systematically investigate the non-asymptotic properties of these tests under a variety of linkage disequilibria among SNPs across the full set of genes in the human genome through extensive simulations and a gene-based SNP-set analysis of femoral neck bone mineral density.

Key words and phrases: Correlation robustness, global hypothesis testing, p-value combination, SNP-set test, whole genome sequencing study.

1. Introduction

The p-value combination test is a powerful tool for evaluating a global null hypothesis based on collective evidence from multiple input p-values. One application of this approach is in genome-wide association studies (GWAS) using whole genome sequencing (WGS) data, where the goal is to identify genetic associations between genes (or other DNA segments) and diseases. Disease-associated single nucleotide polymorphisms (SNPs) often involve rare mutations with weak association signals, but their collective contribution to disease risk is significant. Combining SNP p-values offers a promising strategy for identifying disease-associated genes in WGS studies.

A challenge in p-value combination tests is the complex correlations among the input p-values, such as those caused by linkage disequilibrium (LD) among SNPs. Complicated and computationally demanding algorithms are often required to fully account for the correlations in order to derive an accurate null distribution of the test statistic, thereby ensuring precise control of type I error. However, if a test statistic is correlation-robust – meaning that the correlations have a negligible influence on the distribution of the test statistic and, consequently, on type I error control – the computation can be greatly simplified by assuming independence among the input p-values.

In large data analyses where type I error control is stringent, it may be sufficient to consider correlation robustness in an asymptotic sense—i.e., the influence of correlation diminishes at the far-end tail of the null distribution. For example, when simultaneously testing 20,000 genes in the human genome, the genome-wide significance level is $\alpha = 0.05/20,000 = 2.5 \times 10^{-6}$. If a test is robust to correlations at this level of α and beyond, computing the test's gene p-values based on the independence assumption can simplify the computation

while maintaining genome-level type I error control.

Literature studies have shown that a group of summation-based p-value combination statistics are ACR when specific distributional and dependence conditions are satisfied. Notable examples include the Cauchy Combination Test (CCT) statistic, which sums transformed p-values using the inverse cumulative distribution function (CDF) of the Cauchy distribution (Liu and Xie, 2020), and the harmonic mean statistic, which sums the reciprocals of the input p-values (Wilson, 2019; Vovk and Wang, 2020). In general, summation-based statistics are ACR when p-values are transformed using the inverse CDF of a heavy-tailed distribution, a method referred to as the heavy-tailed transformation method (Fang et al., 2023). Specifically, Liu and Xie (2020) showed that CCT is ACR when the individual p-values are marginally uniform under the null and the corresponding test statistics satisfy a bivariate normal dependence structure; Long et al. (2023) relaxed the Gaussian dependence assumption by imposing a joint tail condition on the p-values, requiring that simultaneous extreme p-values occur with probability $o(1/t)$. The harmonic mean statistic is ACR when the individual p-values are marginally valid under the null; its robustness stems from the heavy-tailed $1/p$ transformation, which causes the null distribution of the statistic to be dominated by extreme p-values and hence largely insensitive to correlation. More generally, the summation-based statistics studied in Fang et al. (2023) are ACR under Gaussian dependence when the transformation distribution has a regularly varying tail with index equal to one.

However, there has been a lack of systematic study on another large family of p-value combination statistics, defined by the supremum over functions of the ordered p-values. This family includes many classic statistics, such as the minP test statistic, the Simes test

statistic (Simes, 1986), the Higher Criticism (HC) test statistic (Donoho and Jin, 2004), the ϕ-divergence test statistic (Jager and Wellner, 2007), and the generalized goodness-of-fit (GGOF) test statistic (Zhang et al., 2020; Zhang and Wu, 2022). Numerical studies have observed ACR behavior in the minP test statistic (Fang et al., 2023), but further theoretical research is needed. The Simes test has been shown to be conservative under positive correlation (Sarkar and Chang, 1997; Sarkar, 1998) and liberal under negative correlation (Block et al., 2008). However, its behavior under arbitrary correlation in the context of ACR remains unclear. The HC and ϕ-divergence tests are known to be optimal for detecting weak and rare signals (Donoho and Jin, 2004; Jager and Wellner, 2007), and numerical studies suggest their robustness to weak correlations (Zhang and Wu, 2022), but no theoretical results on ACR have yet been established. Studying the ACR property of supremum-based p-value combination statistics is important, as these statistics differ fundamentally from summation-based statistics in their mathematical definitions and mechanisms of correlation robustness. The heavy-tailed transformation method may not be effective for supremum-based tests; for example, applying any monotone transformation to the ordered p-values does not alter the minP test.

This paper makes two main contributions. First, it establishes the general principles and conditions for the ACR property of supremum-based p-value combination statistics under both Gaussian and copula-based dependence models. The study includes an in-depth analysis of the tail probabilities of these statistics under general correlation structures. Notably, we provide a more straightforward method for computing very small p-values in the minP test by using the upper and lower bounds of tail probability, which can be more accurate than the commonly used R function `pmvnorm`. We also present sufficient conditions for

supremum-based p-value combination statistics to be ACR, showing that the HC test and some (but not all) ϕ-divergence statistics satisfy this property. In general, the ACR property holds for a broad family of supremum-based p-value combinations when the input statistics or p-values being combined exhibit asymptotic upper-tail independence, a condition satisfied by Gaussian and many non-Gaussian distributions. Our copula-based analysis extends the ideas of Long et al. (2023), who studied input p-value copulas (asymptotic lower-tail independence) for the CCT test. In contrast, we examine both the copulas of input statistics (asymptotic upper-tail independence) and p-values, establishing general conditions for the ACR property of a different class of supremum-based p-value combination statistics.

Second, we investigate the non-asymptotic properties of relevant tests under various correlation structures, including the LD patterns within a comprehensive set of genes in the human genome. We compare their correlation robustness and statistical power through extensive simulations and a gene-based SNP-set analysis of GWAS summary data for femoral neck bone mineral density. The results show that genome-wide significant genes can be consistently identified, regardless of whether correlations are accounted for in the testing procedures of the ACR tests.

The remainder of the paper is organized as follows: Section 2 introduces the hypothesis testing problem and supremum-based p-value combination statistics. Section 3 examines the ACR properties of these statistics under the Gaussian assumption, while Section 4 addresses non-Gaussian cases. Section 5 investigates the power of tests based on supremum-based p-value combination statistics and related procedures. Numerical studies are presented in Section 6, and a real data analysis is discussed in Section 7. Finally, Section 8 concludes the paper and discusses its limitations and directions for future research. Throughout the paper,

we write $f(x) \sim g(x)$ as $x \to a$ if $\lim_{x\to a} f(x)/g(x) = 1$. The density, CDF, and survival function of the standard normal distribution are denoted by $\phi(x)$, $\Phi(x)$, and $\bar{\Phi}(x) = 1-\Phi(x)$, respectively.

2. Problem Description and Formulation

2.1 Global Hypothesis Testing

We consider the global hypothesis testing problem, in which the null hypothesis designates the joint distribution of multiple *input statistics*:

$$H_0 : \boldsymbol{X} = (X_1, ..., X_n)^\top \sim F_0, \tag{2.1}$$

where F_0 denotes the joint CDF of the statistics under H_0. The marginal distributions of individual input statistics X_i's under the null are $X_i \sim F_{0i}$, $i = 1, ..., n$. Without loss of generality, we assume that larger values of X_i give more significant evidence against H_0, from which the p-value associated with the individual input statistic X_i is

$$P_i = 1 - F_{0i}(X_i) \equiv \bar{F}_{0i}(X_i), \quad i = 1, ..., n. \tag{2.2}$$

Accordingly, $X_i = \bar{F}_{0i}^{-1}(P_i)$, $i = 1, ..., n$. For a given F_0, there is a one-to-one correspondence between the p-values $\boldsymbol{P} = (P_1, \ldots, P_n)$ and the test statistics $\boldsymbol{X}$. Therefore, the test statistic for the global hypothesis testing problem is a function of either $\boldsymbol{P}$ or $\boldsymbol{X}$. For simplicity, we will write the test statistic as $T(\boldsymbol{P})$ (or $T(\boldsymbol{X})$) when we wish to emphasize its dependence on $\boldsymbol{P}$ (or $\boldsymbol{X}$). The two notations are interchangeable, and the exact form of $T(\boldsymbol{P})$ and $T(\boldsymbol{X})$ may vary depending on the context.

P-value combination tests often offer practical advantages. When X_i's are continuously distributed, P_i follows a Uniform$(0, 1)$ distribution under the null hypothesis. Thus, p-values

provide standardized and consistent statistical evidence across different data distributions. They are useful for integrating heterogeneous data with varying F_{0i}'s using the p-value combination statistic $T(\boldsymbol{P})$. Furthermore, p-value combination tests are considered non-parametric because they do not require distributional assumptions about X_i's when testing the global H_0 (Ingster and Suslina (2000)).

2.2 Supremum-based p-value Combination

In this paper, we study a broad class of supremum-based p-value combination statistics, involving the supremum of a function of ordered p-values $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$. A notable example is the minP statistic, defined as:

$$T_{minP}(\boldsymbol{P}) = \min_{i=1,\ldots,n} P_i = P_{(1)}. \tag{2.3}$$

The Simes statistic (Simes (1986)) is another classic example:

$$T_{Simes}(\boldsymbol{P}) = \min_{i=1,\ldots,n} \left\{ P_{(i)} \frac{n}{i} \right\}. \tag{2.4}$$

Both statistics have the property that smaller values provide stronger evidence against H_0.

It is also common for a large value of the statistic to provide strong evidence against H_0; for example, the higher-criticism (HC) statistic (Donoho and Jin (2004)), which is defined as:

$$T_{HC}(\boldsymbol{P}) = \max_{i=1,\ldots,n} \frac{\sqrt{n}(\frac{i}{n} - P_{(i)})}{\sqrt{P_{(i)}\left(1 - P_{(i)}\right)}}. \tag{2.5}$$

A more general class of statistics is the ϕ-divergence family (Jager and Wellner (2007)), for which the statistics are:

$$T_{\phi}(\boldsymbol{P}; s) = \max_{i=1,\ldots,n} f_s^{\phi}(\frac{i}{n}, P_{(i)}), \text{ where} \tag{2.6}$$

$$f_s^{\phi}(\frac{i}{n},x)=\frac{1}{s(1-s)}\left(1-(\frac{i}{n})^s x^{1-s}-(1-\frac{i}{n})^s(1-x)^{1-s}\right),\ s\neq 0,1,$$
$$f_1^{\phi}(\frac{i}{n},x)=\frac{i}{n}\log\left(\frac{i/n}{x}\right)+\left(1-\frac{i}{n}\right)\log\left(\frac{1-i/n}{1-x}\right),\text{ and}$$
$$f_0^{\phi}(\frac{i}{n},x)=x\log\left(\frac{x}{i/n}\right)+(1-x)\log\left(\frac{1-x}{1-i/n}\right).$$

Here, s is a parameter that defines the statistic. For example, $s=1$ and $s=0$ correspond to the Berk-Jones (BJ) and reverse Berk-Jones statistics (Berk and Jones, 1979), respectively.

Since small p-values (i.e., $P_{(i)}<\frac{i}{n}$) indicate signals or effects, it is appropriate to consider the one-sided ϕ-divergence statistics, for example, by casting $f_s^{\phi}(\frac{i}{n},P_{(i)})$ into a strictly decreasing function (Zhang et al., 2020):

$$f_s(\frac{i}{n},P_{(i)})=\begin{cases}\sqrt{2nf_s^{\phi}(\frac{i}{n},P_{(i)})} & P_{(i)}\leq\frac{i}{n}\\ -\sqrt{2nf_s^{\phi}(\frac{i}{n},P_{(i)})} & P_{(i)}>\frac{i}{n},\end{cases}\tag{2.7}$$

which yields exactly the HC and reverse HC statistics (Donoho and Jin, 2008) at $s=2$ and $s=-1$, respectively.

To further generalize the family of statistics, we consider the supremum over arbitrary functions of ordered p-values:

$$T_f(\boldsymbol{P})=\sup_{1\leq i\leq n} f_i(P_{(i)}).\tag{2.8}$$

The function $f_i(x)$ satisfies the following condition:

Assumption 1. For any fixed i, $f_i(x)$, defined on $[0,1]$, is strictly decreasing in x, with $\lim_{x\to 0} f_i(x)=+\infty$.

Assumption 1 ensures that f_i^{-1} exists and that $\lim_{t\to\infty} f_i^{-1}(t)=0$, indicating that smaller p-values correspond to larger statistics, thereby reflecting higher statistical significance against H_0. This general form encompasses a wide range of p-value combination test statistics,

including the minP, Simes, HC, one-sided ϕ-divergence, and GGOF statistics, by specifying an appropriate f_i. For detailed discussions, see the Supplementary Material.

2.3 Asymptotic Correlation-Robust Tests

A test statistic $T(\boldsymbol{X})$ (or $T(\boldsymbol{P})$) (and thus its testing procedure) is said to be ACR if the correlations among X_i's (or P_i's) have a negligible impact on type I error control as the significance level $\alpha \to 0$. In other words, the far-end tail probability of the test statistic under the null, with arbitrary correlation structures among the data, can be well approximated by the tail probability computing under the assumption that X_i's are uncorrelated.

Specifically, consider the statistic $T_f(\boldsymbol{P})$ in (2.8), where a larger value indicates stronger evidence against H_0. It is said to be ACR if (Liu and Xie, 2020)

$$\lim_{t\to\infty} \frac{\mathbb{P}(T_f(\boldsymbol{P}) > t \mid \boldsymbol{\Sigma})}{\mathbb{P}(T_f(\boldsymbol{P}) > t \mid \mathbf{I})} = 1,$$

where $\mathbb{P}(\cdot \mid \boldsymbol{\Sigma})$ denotes the tail probability given that the input statistics $\boldsymbol{X}$ have the correlation matrix $\boldsymbol{\Sigma}$, and $\mathbb{P}(\cdot \mid \mathbf{I})$ denotes the tail probability given that $\boldsymbol{\Sigma} = \mathbf{I}$, the identity matrix. Because smaller values of the minP and Simes statistics in (2.3) and (2.4) provide stronger evidence against H_0, they are ACR if

$$\lim_{p\to 0} \frac{\mathbb{P}(T_*(\boldsymbol{P}) < p|\boldsymbol{\Sigma})}{\mathbb{P}(T_*(\boldsymbol{P}) < p|\mathbf{I})} = 1, \tag{2.9}$$

where $*$ is $minP$ or $Simes$. The asymptotic argument assumes a fixed number n of input statistics.

3. Asymptotic Correlation Robustness under Gaussianity

This section examines the asymptotic correlation robustness when the individual input statistics X_i's follow a multivariate Gaussian distribution, with the correlation matrix fully char-

acterizing the dependence structure. In real-world data analysis, individual test statistics are often approximately multivariate Gaussian, as observed in generalized linear models with large sample sizes (Zhang et al., 2023).

Definition 1 (Multivariate Gaussian model)**.**

$$H_0 : \boldsymbol{X} \sim N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{3.10}$$

where $\boldsymbol{\Sigma} = (\sigma_{ij})_{1\le i,j\le n}$ is a correlation matrix with $\sigma_{ii} = 1$ and $-1 < \sigma_{ij} < 1$, for all $i \neq j$.

In Definition 1, the input statistics X_i's are assumed to be standardized with unit variance, without loss of generality. The one-sided and two-sided input p-values are, respectively,

$$\text{One-sided: } P_i = \bar{\Phi}(X_i), \qquad \text{Two-sided: } P_i = 2\bar{\Phi}(|X_i|). \tag{3.11}$$

The two-sided p-values in (3.11) are consistent with the definition of p-value given in (2.2), noting that $P_i = \bar{F}_{\chi^2_1}(X_i^2)$, where $F_{\chi^2_1}$ denotes the CDF of a chi-squared distribution with one degree of freedom.

3.1 The minP Test

The minP statistic $T_{minP}(\boldsymbol{P})$ in (2.3) is equivalent to a supremum-based p-value combination statistic $T_f(\boldsymbol{P})$ in the form of (2.8) with a strictly decreasing function $f(x)$. For convenience, under the Gaussian model, we choose $f(x) = \bar{\Phi}^{-1}(x)$ and $f(x) = \bar{\Phi}^{-1}(x/2)$, corresponding to the one-sided and two-sided input p-values in (3.11), respectively. Accordingly, the minP statistic $T_{minP}(\boldsymbol{P})$ is equivalent to

$$\begin{aligned} T_{maxN}(\boldsymbol{X}) &\equiv \max_i X_i \quad \text{for one-sided } P_i\text{'s,} \\ \text{and } T_{maxhN}(\boldsymbol{X}) &\equiv \max_i |X_i| \quad \text{for two-sided } P_i\text{'s.} \end{aligned} \tag{3.12}$$

To study correlation robustness, we first examine the tail probabilities under the assumption of independence, where $\mathbf{\Sigma} = \mathbf{I}$.

Lemma 1. *Under the Gaussian model in Definition 1 with* $\mathbf{\Sigma} = \mathbf{I}$*, we have*

1. *For the* $T_{minP}(\boldsymbol{P})$ *statistic given in* (2.3)*, as* $p \to 0$*,*

$$\mathbb{P}(T_{minP}(\boldsymbol{P}) < p|\mathbf{I}) \sim np.$$

2. *For* $T_{maxN}(\boldsymbol{X})$ *and* $T_{maxhN}(\boldsymbol{X})$ *in* (3.12)*, as* $t \to \infty$*,*

$$\mathbb{P}(T_{maxN}(\boldsymbol{X}) > t|\mathbf{I}) \sim n\frac{\phi(t)}{t}, \quad \mathbb{P}(T_{maxhN}(\boldsymbol{X}) > t|\mathbf{I}) \sim 2n\frac{\phi(t)}{t}.$$

Theorem 1 below provides the asymptotic tail probabilities of the minP statistic under arbitrary non-perfect correlations, i.e., $|\sigma_{ij}| < 1$ for $i \neq j$.

Theorem 1. *Under the Gaussian model in Definition 1, as* $t \to \infty$*,*

$$\mathbb{P}\left(T_{maxN}(\boldsymbol{X}) > t|\mathbf{\Sigma}\right) = n\mathbb{P}(X_1 > t) + O(t^{-2}\exp(\frac{-t^2/}{1+\max_{i\neq j}\sigma_{ij}})),$$
$$\mathbb{P}\left(T_{maxhN}(\boldsymbol{X}) > t|\mathbf{\Sigma}\right) = n\mathbb{P}(|X_1| > t) + O(t^{-2}\exp(\frac{-t^2}{1+\max_{i\neq j}|\sigma_{ij}|})).$$

Moreover,

$$\mathbb{P}(T_{minP}(\boldsymbol{P}) < p|\mathbf{\Sigma}) = np + o(p), \quad p \to 0. \tag{3.13}$$

In light of the tail probabilities under both independence and non-perfect correlations, we can conclude that the minP statistic is ACR, as defined in (2.9), which is summarized in Corollary 1.

Corollary 1. *Under the Gaussian model in Definition 1, for both one-sided and two-sided* P_i*'s in* (3.11)*, the minP statistic in* (2.3) *is ACR, i.e.,*

$$\lim_{p\to 0}\frac{\mathbb{P}(T_{minP}(\boldsymbol{P}) < p|\mathbf{\Sigma})}{\mathbb{P}(T_{minP}(\boldsymbol{P}) < p|\mathbf{I})} = 1.$$

As a byproduct of Corollary 1, we also derive the non-asymptotic bounds for the tail probability of the minP test, presented in Corollary 2, which can significantly facilitate practical numerical evaluation.

Corollary 2. *Consider the Gaussian model in Definition 1. For any* $t > 0$,

$$\frac{n\phi(t)t}{1+t^2} - \sum_{1\le i<j\le n}\left\{\frac{(1+\sigma_{ij})^2\exp(-t^2/(1+\sigma_{ij}))}{2\pi t^2\sqrt{1-\sigma_{ij}^2}}\right\}$$
$$\le \mathbb{P}\left(T_{maxN}(\boldsymbol{X}) > t|\boldsymbol{\Sigma}\right) \le \frac{n\phi(t)}{t}; \tag{3.14}$$
$$\frac{2n\phi(t)t}{1+t^2} - \sum_{1\le i<j\le n}\left\{\frac{(1+\sigma_{ij})^2\exp(-t^2/(1+\sigma_{ij}))}{\pi\sqrt{1-\sigma_{ij}^2}t^2} + \frac{(1-\sigma_{ij})^2\exp(-t^2/(1-\sigma_{ij}))}{\pi\sqrt{1-\sigma_{ij}^2}t^2}\right\}$$
$$\le \mathbb{P}\left(T_{maxhN}(\boldsymbol{X}) > t|\boldsymbol{\Sigma}\right) \le \frac{2n\phi(t)}{t}. \tag{3.15}$$

It is worth mentioning that, aside from enabling fast computation, the non-asymptotic tail probability bounds given in Corollary 2 provide a more accurate approximation of the tail probability than the commonly used R function `pmvnorm` when t is large, as evidenced by the numerical results in Section 6. Therefore, these bounds can be useful for calculating small p-values in the minP test. Moreover, Equation (3.14) provides a sharper asymptotic bound than Corollary 2.1 of Li and Shao (2002); see the Supplementary Material for a detailed comparison.

3.2 Simes Test

The Simes statistic T_{Simes} in (2.4) follows the Uniform(0,1) distribution under independence, i.e., $\mathbb{P}\left(T_{Simes} < p|\mathbf{I}\right) = p$ (Simes, 1986). Theorem 2 shows that it is ACR, as outlined in (2.9).

Theorem 2. *Under the Gaussian model in Definition 1, for* $T_{Simes}(\boldsymbol{P})$ *in* (2.4)*, with* $\boldsymbol{P}$

being one-sided or two-sided p-values in (3.11)*, we have*

$$\mathbb{P}\left(T_{Simes} < p\right) = p + o(p). \tag{3.16}$$

Equations (3.13) and (3.16) suggest that the Simes test is asymptotically more powerful than the minP test, as their p-values are related by

$$\mathbb{P}(T_{minP} < p_{(1)}|\boldsymbol{\Sigma}) \sim \mathbb{P}(T_{Simes} < np_{(1)}|\boldsymbol{\Sigma}) \geq \mathbb{P}\left(T_{Simes} < \min_i \left\{p_{(i)}\frac{n}{i}\right\}|\boldsymbol{\Sigma}\right).$$

This result extends the literature on the superior power of the Simes test over the minP test (Sarkar, 1998) by considering arbitrary correlations and non-Gaussian cases, as shown in Corollaries 6 and 7 (see detailed discussions in the Supplementary Material).

Going beyond the assumption of non-perfect correlations (i.e., $\sigma_{ij} \in (-1, 1)$ for $i \neq j$), Theorem 3 establishes the correlation robustness of the Simes test in the case of perfect correlations, an assumption consistent with the literature (cf. Theorem 2 in Fang et al. (2023)).

Theorem 3. *Suppose that under the null,* $P_i \sim$ *Uniform*$(0, 1)$ *for* $i = 1, \ldots, n$*, where* m *of them are equal (perfect positive correlations) and the rest are independent of each other. For any* $1 \leq m \leq n$*, we have*

$$\mathbb{P}(T_{Simes}(\boldsymbol{P}) < p|\boldsymbol{\Sigma}) \sim p, \ \text{as } p \to 0.$$

Note that the assumption in Theorem 3 differs from the assumption that there are $n' \doteq n - m + 1$ independent X_i's. Specifically, the Simes statistic in Theorem 3, $T_{Simes}(\boldsymbol{P}) = \min_{i=1,\ldots,n}\left\{P_{(i)}\frac{n}{i}\right\}$, is indexed by n instead of by n', which would be used in the latter case.

3.3 Supremum-based p-value Combinations

For the general class of supremum-based p-value combination statistics $T_f(\boldsymbol{P})$ defined in (2.8), we provide a sufficient condition for ACR under the Gaussian model. This condition is rather weak and is satisfied by many statistics, including the HC statistic in (2.5) and the one-sided ϕ-divergence statistic in (2.7) for $s > 1$.

Theorem 4. *Consider $T_f(\boldsymbol{P})$ in (2.8), where $\boldsymbol{P}$ represents one-sided or two-sided p-values as defined in (3.11) under the Gaussian model in Definition 1. Suppose that Assumption 1 is satisfied, and that $f_1^{-1}(t) \leq f_2^{-1}(t) \leq \ldots \leq f_n^{-1}(t)$ for any t. Then,*

$$\mathbb{P}(T_f(\boldsymbol{P}) > t) = n f_1^{-1}(t) + o(f_n^{-1}(t)), \quad t \to \infty.$$

With Assumption 2, the ACR property follows directly, as formally stated in Corollary 3.

Assumption 2. The ratio $f_n^{-1}(t)/f_1^{-1}(t)$ converges to a nonzero constant as $t \to \infty$.

Corollary 3. *Suppose that the assumptions in Theorem 4 are satisfied. Under the additional Assumption 2, $T_f(\boldsymbol{P})$ is ACR.*

The HC statistic defined in (2.5) satisfies the conditions in Corollary 3, making it asymptotically robust to correlation, with the asymptotic tail probability given in Corollary 4.

Corollary 4. *Under the Gaussian model in Definition 1, for $T_{HC}(\boldsymbol{P})$ in (2.5), with $\boldsymbol{P}$ being one-sided or two-sided p-values in (3.11), we have*

$$\mathbb{P}\left(T_{HC}(\boldsymbol{P}) > t\right) = \frac{1}{t^2} + o(\frac{1}{t^2}), \text{ as } t \to \infty.$$

More generally, the family of one-sided ϕ-divergence statistics in (2.7) with $s > 1$ also satisfies the conditions in Corollary 3, with the asymptotic tail probability given in Corollary 5.

Corollary 5. *Consider the one-sided ϕ-divergence statistic in* (2.7) *with $s > 1$. Assume the input p-values $\boldsymbol{P}$ are one-sided or two-sided in* (3.11) *based on the Gaussian model in Definition 1. We have*

$$\mathbb{P}(T_{\phi s}(\boldsymbol{P}) > t) = \left(\frac{(s-1)\,s}{2}\right)^{-\frac{1}{s-1}} t^{-\frac{2}{s-1}} + o(t^{-\frac{2}{s-1}}), \ \text{as } t \to \infty.$$

As a direct application of Corollary 5, for $s = 3$, the one-sided $\phi 3$ statistic is ACR, with the tail probability given by $\mathbb{P}(T_{\phi 3}(\boldsymbol{P}) > t) = 3^{-1/2}t^{-1} + o(t^{-1})$, as $t \to \infty$. This statistic is of particular interest, in that it is powerful for detecting strong and sparse signals (Zhang and Wu, 2022), and is relatively robust to correlation in practical scenarios.

We have shown that Assumption 2 is a sufficient condition for supremum-based p-value combination tests to be ACR. It is satisfied by minP, Simes, and all one-sided ϕ-divergence statistics in (2.7) with $s > 1$, including the HC and $\phi 3$ tests, as stated in Corollaries 3–5. Although we are unable to prove that Assumption 2 is a necessary condition for supremum-based p-value combination tests to be ACR, our simulations suggest that ϕ-divergence statistics may not be robust to correlation whenever $s \leq 1$. For instance, the one-sided BJ statistic (with $s = 1$) does not satisfy Assumption 2, and our simulations indicate that it is sensitive to correlation.

4. Asymptotic Correlation Robustness under non-Gaussianity

In this section, we relax the Gaussianity assumption for the input statistics $\boldsymbol{X}$ and focus solely on their pairwise dependence using bivariate copulas. This approach allows the test statistics to have arbitrary continuous marginal distributions (Long et al., 2023).

Definition 2 (Copula model)**.** Assume that under the null hypothesis, X_i, $i = 1, ..., n$, have a marginal CDF F_{0i}, and for $1 \leq i \neq j \leq n$, the pairwise dependence between $(X_i, X_j)^\top$ is

modeled through the copula $C_X(u_i, u_j) = \mathbb{P}(X_i \leq F_{0i}^{-1}(u_i), X_j \leq F_{0j}^{-1}(u_j))$, $u_i, u_j \in (0, 1)$.

Theorem 5 presents the conditions in terms of the copulas of (X_i, X_j) (which also dictate the dependence among the input p-values P_i's), under which the statistic $T_f(\boldsymbol{P})$ is asymptotically robust to correlations.

Theorem 5. *Consider the supreme-based p-value combination statistic $T_f(\boldsymbol{P})$ defined in (2.8). Suppose that Assumption 1 is fulfilled and $f_1^{-1}(t) \leq f_2^{-1}(t) \leq \ldots \leq f_n^{-1}(t)$ for any t, and that for each $i \neq j \in \{1, ..., n\}$, the copula of (X_i, X_j) in Definition 2 satisfies the following condition*

$$C_X(u, u) = 2u - 1 + o(1 - u), \quad u \to 1. \tag{4.17}$$

Then, we have

$$\mathbb{P}(T_f(\boldsymbol{P}) > t) = n f_1^{-1}(t) + o(f_n^{-1}(t)), \quad t \to \infty. \tag{4.18}$$

Furthermore, under the additional Assumption 2, $T_f(\boldsymbol{P})$ is ACR.

Condition (4.17) essentially indicates that X_i and X_j are asymptotic upper-tail independent: $\mathbb{P}\left(X_i > F_i^{-1}(u) | X_j > F_j^{-1}(u)\right) = o(1)$ as $u \to 1$, for $i \neq j$.

For the minP statistic, to assess $\mathbb{P}(T_{minP}(\boldsymbol{P}) < p)$, we can choose $f_i(\cdot)$ as an arbitrary strictly decreasing function $f : [0, 1] \to [0, \infty)$ that satisfies $\lim_{p \to 0} f(p) = \infty$. Thus, with $t = f(p)$, we have $\mathbb{P}(T_{minP}(\boldsymbol{P}) < p) = \mathbb{P}(f(T_{minP}(\boldsymbol{P})) > f(p)) = n f^{-1}(f(p)) + o(f^{-1}(f(p))) = np + o(p)$ as $p \to 0$. This is formally stated in Corollary 6.

Corollary 6. *Consider the minP statistic defined in (2.3). Suppose that for each $i \neq j \in \{1, ..., n\}$, the copula of (X_i, X_j) defined in Definition 2 satisfies (4.17). Then, we have $\mathbb{P}(T_{minP}(\boldsymbol{P}) < p) = np + o(p)$ as $p \to 0$, and $T_{minP}(\boldsymbol{P})$ is ACR.*

For the Simes statistic, note that for $p \in (0,1)$, $\mathbb{P}\left(\min_i \left(P_{(i)}n/i\right) < p\right) = \mathbb{P}\left(\sup_i f(P_{(i)}n/i) > f(p)\right)$, where $f : [0,1] \to [0,\infty)$ is an arbitrary strictly decreasing function satisfying $\lim_{p\to 0} f(p) = \infty$. With $f_i(x) = f(nx/i)$, $x \in [0,1]$, and $t = f(p)$, we have $f^{-1}_{n-i+1}(t) = n^{-1}(n-i+1)f^{-1}(t) = n^{-1}(n-i+1)p$. We obtain a similar result for Simes test in Corollary 7.

Corollary 7. *Consider the Simes statistic defined in* (2.4)*. Suppose the copula of* (X_i, X_j) *defined in Definition 2 satisfies* (4.17) *for each* $i \neq j \in \{1, ..., n\}$*. Then,* $\mathbb{P}(T_{Simes}(\boldsymbol{P}) < p) = p + o(p)$ *as* $p \to 0$*, and* $T_{Simes}(\boldsymbol{P})$ *is ACR.*

The remark below provides examples of popular copulas that are asymptotically upper-tail independent. Therefore, if $C_X(u,u)$ corresponds to one of these copulas, (4.17) is satisfied, X_i and X_j are asymptotically upper-tail independent, enabling $T_f(\boldsymbol{P})$ to be asymptotically robust to the dependencies they define. In particular, the Gaussian copula extends the Gaussian model in Definition 1, as it also accommodates test statistics $(X_1, \ldots, X_n)^\top$ with non-Gaussian marginals.

Remark 1. *The following copulas are asymptotically upper-tail independent:*

1. *Product copula:* $C(u_i, u_j) = u_i u_j$.

2. *Farlie-Gumbel-Morgenstern (FGM) copula:*

$$C(u_i, u_j) = u_i u_j \left(1 + \theta \left(1 - u_i\right)\left(1 - u_j\right)\right), \theta \in [-1, 1].$$

3. *Ali-Mikhail-Haq (AMH) copula:* $C(u_i, u_j) = \frac{u_i u_j}{1-\theta(1-u_i)(1-u_j)}$, $\theta \in [-1, 1]$.

4. *Survival copula:* $C(u_i, u_j) = u_i u_j \exp(-\theta \ln u_i \ln u_j)$, $\theta \in [0, 1]$.

5. *Gaussian copula: With* $\max_{i\neq j} |\rho_{ij}| < 1$,

$$C(u_i, u_j) = \frac{1}{2\pi\sqrt{1-\rho_{ij}}} \int_{-\infty}^{\Phi^{-1}(u_i)} \int_{-\infty}^{\Phi^{-1}(u_j)} exp\Big(-\frac{x^2 - 2\rho_{ij}xy + y^2}{2(1-\rho_{ij}^2)}\Big)dxdy.$$

The discussion so far has concerned dependence among the input statistics. We next re-state Condition (4.17) in terms of p-values, which renders our results directly comparable with those of Long et al. (2023). By definition of the copula of (X_i, X_j),

$$C_X(u,u) = \mathbb{P}(X_i \le F_{0i}^{-1}(u), X_j \le F_{0j}^{-1}(u)) = \mathbb{P}(P_i > 1-u, P_j > 1-u).$$

Letting $p = 1-u$, an application of the inclusion - exclusion formula yields

$$\mathbb{P}(P_i \le p, P_j \le p) = 1 - \mathbb{P}(P_i > p) - \mathbb{P}(P_j > p) + \mathbb{P}(P_i > p, P_j > p) = 1 - 2u + C_X(u,u).$$

Thus, the upper-tail independence condition on (X_i, X_j) is equivalent to the lower-tail independence condition on (P_i, P_j), which is formally stated in Remark 2.

Remark 2. *Consider the bivariate copula of (P_i, P_j), defined by $C_P(u_i, u_j) = \mathbb{P}(P_i \le u_i, P_j \le u_j)$. Then, Condition* (4.17) *is equivalent to the following requirement:*

$$C_P(p,p) = o(p), \quad p \to 0. \tag{4.19}$$

If $C_P(p,p)$ belongs to one of the five copulas listed in Remark 1, then (4.19) *holds and P_i and P_j are asymptotically lower-tail independent. Condition* (4.17) *imposes asymptotic upper-tail independence on the copula of the input test statistics, whereas* (4.19) *requires asymptotic lower-tail independence for the induced p-values. This distinction is important, because some copulas exhibit different tail behaviors. For instance, the Cuadras-Augé copula, defined as $C(u_i, u_j) = (\min(u_i, u_j))^{\theta} (u_i u_j)^{1-\theta}$ with $\theta \in (0,1)$, is asymptotically lower-tail independent but upper-tail dependent. Consequently, while the Cuadras–Augé copula may be suitable for modeling dependence among p-values, it does not satisfy the upper-tail independence required for modeling the dependence of the input test statistics when ACR is desired.*

Remark 3 below provides a direct comparison with the corresponding conditions in Long et al. (2023).

Remark 3. *Due to the equivalence between Condition* (4.17) *and Condition* (4.19)*, we compare the latter with Condition (D1) in Long et al. (2023), which states that for* $1 \le i < j \le n$*, as* $t \to \infty$*, there exists a sequence* δ_t *satisfying* $\lim_{t\to\infty} \delta_t = 0$ *and* $\lim_{t\to\infty} \delta_t t = \infty$*, such that*

$$\mathbb{P}\left(0 < P_i < \frac{\omega_i n}{\pi t}, 0 < P_j < \frac{\omega_j n}{\pi \delta_t t}\right) = o\left(\frac{1}{t}\right), \tag{4.20}$$

$$\mathbb{P}\left(0 < P_i < \frac{\omega_i n}{\pi(1+\delta_t) t}, 1 - \frac{\omega_j n}{\pi \delta_t t} < P_j < 1\right) = o\left(\frac{1}{t}\right), \tag{4.21}$$

where the weights $\omega_i, \omega_j > 0$ *and* $\sum_{i=1}^n \omega_i = 1$.

The lower–lower tail condition (4.20) *alone implies Condition* (4.19)*. To see this, fix* $p \downarrow 0$ *and set* $t = \frac{\omega_i n}{\pi p}$*, so that* $t \to \infty$*. Since* $\delta_t \to 0$*, we have* $\frac{\omega_j n}{\pi \delta_t t} = \frac{\omega_j}{\omega_i} \cdot \frac{p}{\delta_t} > p$ *for all sufficiently large* t*. Hence, for sufficiently large* t*,* $\{P_i \le p, P_j \le p\} \subseteq \left\{P_i \le p, P_j \le \frac{\omega_j}{\omega_i}\frac{p}{\delta_t}\right\}$*, which implies*

$$\mathbb{P}(P_i \le p, P_j \le p) \le \mathbb{P}\left(0 < P_i < \frac{\omega_i n}{\pi t}, 0 < P_j < \frac{\omega_j n}{\pi \delta_t t}\right) = o\left(\frac{1}{t}\right) = o(p).$$

In fact, (4.20) *is strictly stronger, as it requires joint control when the two thresholds are of different asymptotic orders, whereas Condition* (4.19) *only restricts the equal-threshold case. Furthermore, Condition (D1) includes the lower–upper tail restriction* (4.21)*, which is essential for the Cauchy Combination Test because the corresponding transformed terms may cancel. Since Condition* (4.19) *only constrains lower–lower tail dependence and does not impose any lower–upper tail restriction, our requirement on the dependence among the p-values is weaker than Condition (D1) in Long et al. (2023).*

Before closing this section, we discuss how the requirements for supremum-based statistics to be ACR differ from those for summation-based statistics. To ensure that correlations are asymptotically negligible relative to their tail probabilities, supremum-based statistics

require that the joint probability of $\boldsymbol{X}$ (or the corresponding $\boldsymbol{P}$) is dominated by the marginal distributions in the far tail. That is, X_i and X_j (or P_i and P_j) are asymptotically upper-tail (or lower-tail) independent, as described in (4.17) (or (4.19)). The conditions on the statistic itself (i.e., the conditions on the functions f_i summarized in Theorem 5) are relatively weak. In contrast, for summation-based statistics, a key requirement is that the summands, corresponding to transformed p-values, have heavy tails. Specifically, a summation-based p-value combination statistic is defined as $T_g(\boldsymbol{P}) = \sum_{i=1}^{n} g(P_i) = \sum_{i=1}^{n} F_U^{-1}(1-P_i)$, where F_U^{-1} is the inverse CDF of a random variable U. For $T_g(\boldsymbol{P})$ to be ACR, U must follow a distribution with a sufficiently heavy tail (e.g., the Cauchy distribution in the CCT or a regularly varying distribution in general (Liu and Xie, 2020; Fang et al., 2023)) to ensure that the summation is asymptotically unaffected by correlations. This heavy-tail transformation does not affect supremum-based statistics. For instance, the minP test is invariant to any monotone transformation of p-values. Let $g_1(x) = F_{U_1}^{-1}(1-x)$ and $g_2(x) = F_{U_2}^{-1}(1-x)$ be transformations corresponding to U_1 and U_2 with different tail heaviness. The following events are equivalent:

$$\{\omega : P_{(1)}(\omega) < p\} = \{\omega : g_1(P_{(1)}(\omega)) > g_1(p)\} = \{\omega : g_2(P_{(1)}(\omega)) > g_2(p)\}.$$

Thus, the two transformations yield the same test, and the tail heaviness of U_1 or U_2 does not impact the minP test.

5. Power Analysis

In this section, we investigate the power of tests based on supremum-based p-value combination statistics, along with the CCT. The power of the HC test was studied in Donoho and Jin (2004), while the ϕ-divergence statistic was investigated in Jager and Wellner (2007).

Long et al. (2023) compared the power of the CCT and minP tests and showed that, when the number of p-values n is sufficiently large, the asymptotic power of the CCT is no less than that of minP. These studies were conducted under the asymptotic regime $n \to \infty$. In contrast, this paper considers fixed n and the regime $\alpha \to 0$.

For clarity, we first consider a transparent but somewhat restrictive setting to establish a baseline framework for comparing the powers of the tests under consideration: a Gaussian model with exchangeable correlation and a multi-signal alternative with equal effect sizes. We then relax these assumptions, guided by the key ideas underlying the theoretical results and their proofs, to accommodate more general dependence structures and heterogeneous signal strengths under relatively mild conditions. Proofs of all theorems and remarks are provided in the Supplementary Material.

Specifically, in this restrictive setting, suppose the test statistics satisfy the following Gaussian model:

$$X_i \sim N(\mu_i, 1), i = 1, 2, \ldots, n, \qquad \mathrm{Corr}(X_i, X_j) = \rho, i \neq j, \quad \rho \in (0, 1). \tag{5.22}$$

We test

$$H_0 : \mu_1 = \cdots = \mu_n = 0 \quad \text{vs} \quad H_1 : \mu_1 = \cdots = \mu_m = \delta > 0, \mu_{m+1} = \cdots = \mu_n = 0, \tag{5.23}$$

where $1 \leq m \leq n$. Without loss of generality, we use one-sided p-values, defined by $P_i = 1 - \Phi(X_i)$, $i = 1, \ldots, n$.

For the minP statistic $P_{(1)}$ in (2.3), its ACR property, established in Theorem 1, implies that the level-α cutoff satisfies $c_\alpha^{minP} \sim \alpha/n$ as $\alpha \to 0$. Hence, the power of the minP test satisfies

$$\beta_{minP}(\delta) = \mathbb{P}_{H_1}(P_{(1)} \leq c_\alpha^{minP}) \sim \mathbb{P}_{H_1}(P_{(1)} \leq \frac{\alpha}{n}). \tag{5.24}$$

Under the model in (5.22), $\mathbb{P}_{H_1}(P_{(1)} \le \alpha/n)$ admits a simple asymptotic approximation. Specifically, as shown in (S4.53) of the Supplementary Material,

$$\mathbb{P}_{H_1}\left(P_{(1)} \le \frac{\alpha}{n}\right) \sim m\bar{\Phi}(z_{\alpha/n} - \delta), \qquad \alpha \to 0, \tag{5.25}$$

where $z_{\alpha/n} = \Phi^{-1}(1-\alpha/n)$ is the upper α/n quantile of the $N(0,1)$ distribution. This approximation facilitates power calculations and provides a useful interpretation: under the fixed equal-signal alternative, the power is asymptotically driven by the event that one of the m signal coordinates produces an exceptionally small p-value of order α/n.

In what follows, we show that the powers of other supremum-based p-value combination tests and the equal-weight CCT are asymptotically equivalent in the stringent-significance regime, in the sense that they share the same first-order asymptotic approximation, $\mathbb{P}_{H_1}(P_{(1)} \le \alpha/n)$. In particular, for the general supremum-based p-value combination statistic $T_f(\boldsymbol{P})$ in (2.8) for testing (5.23), Theorem 6 shows that the corresponding test has the same first-order asymptotic power as the minP test, under the same conditions ensuring the ACR property as in Theorem 4.

Theorem 6. *Consider the supremum-based p-value combination statistic* $T_f(\boldsymbol{P})$ *defined in* (2.8) *for testing the hypotheses in* (5.23) *under the Gaussian model* (5.22). *Suppose that Assumptions 1 and 2 are fulfilled and that* $f_1^{-1}(t) \le f_2^{-1}(t) \le \cdots \le f_n^{-1}(t)$ *for any* t. *Let* c_α *be the level-*α *critical value, i.e.,* $\mathbb{P}_{H_0}\{T_f(\boldsymbol{P}) > c_\alpha\} = \alpha$. *Assume that* n, m, $\rho \in (0,1)$, *and* $\delta > 0$ *are fixed as* $\alpha \to 0$. *Then*

$$\beta(\delta) = \mathbb{P}_{H_1}(T_f(\boldsymbol{P}) > c_\alpha) = \mathbb{P}_{H_1}\left(P_{(1)} \le \frac{\alpha}{n}\right)(1 + o(1)), \qquad \alpha \to 0. \tag{5.26}$$

By Theorem 6, the powers of the HC statistic and the one-sided ϕ-divergence statistic with $s = 3$ are asymptotically equivalent to that of the minP test. The same conclusion also

applies to the Simes statistic $T_{Simes}(\boldsymbol{P}) = \min_{i=1,\dots,n}\left\{P_{(i)}\frac{n}{i}\right\}$ by considering an equivalent formulation. Specifically, define

$$T_f(\boldsymbol{P}) = \max_{i=1,\dots,n} \frac{i}{nP_{(i)}}.$$

Then $T_f(\boldsymbol{P}) = \frac{1}{T_{Simes}(\boldsymbol{P})}$, so the induced test is equivalent to the Simes test. Here $f_i(x) = i/(nx)$ and $f_i^{-1}(t) = i/(nt)$. Therefore, the conditions of Theorem 6 are satisfied, since $f_1^{-1}(t) \le f_2^{-1}(t) \le \cdots \le f_n^{-1}(t)$ for any t and $\frac{f_n^{-1}(t)}{f_1^{-1}(t)} = n$. The level-$\alpha$ rejection event for Simes is

$$\{T_{Simes}(\boldsymbol{P}) \le \alpha\} = \left\{T_f(\boldsymbol{P}) \ge \frac{1}{\alpha}\right\}.$$

Thus, applying Theorem 6 gives

$$\mathbb{P}_{H_1}\left(T_{Simes}(\boldsymbol{P}) \le \alpha\right) = \mathbb{P}_{H_1}\left(T_f(\boldsymbol{P}) \ge \frac{1}{\alpha}\right) = \mathbb{P}_{H_1}\left(P_{(1)} \le \frac{\alpha}{n}\right)(1+o(1)), \qquad \alpha \to 0.$$

Next, we consider a more general setting in which the dependence between the test statistics X_i and X_j is characterized by a copula. In contrast to the Gaussian framework considered earlier, the marginal distributions of the X_i's are not required to be Gaussian. Moreover, the hypotheses are formulated more generally: the null hypothesis is given in (2.1), whereas the alternative hypothesis is specified through probabilistic conditions on the resulting p-values.

Theorem 7. *Consider the supremum-based p-value combination statistic $T_f(\boldsymbol{P})$ defined in* (2.8) *for testing the null hypothesis* (2.1) *against an alternative hypothesis H_1. Suppose that under H_0, the copula of (X_i, X_j) satisfies* (4.17) *for each $i \ne j$. Suppose further that Assumptions 1 and 2 are fulfilled, and that $f_1^{-1}(t) \le f_2^{-1}(t) \le \cdots \le f_n^{-1}(t)$ for any t. Let c_α denote the level-α critical value, i.e., $\mathbb{P}_{H_0}\{T_f(\boldsymbol{P}) > c_\alpha\} = \alpha$.*

Assume further that under H_1,

$$\mathbb{P}_{H_1}(P_{(2)} \leq C\alpha) = o\left\{\mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right)\right\} \tag{5.27}$$

for every fixed $C > 0$, *and that*

$$\mathbb{P}_{H_1}\{P_{(1)} \leq f_1^{-1}(c_\alpha)\} \sim \mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right). \tag{5.28}$$

Then

$$\mathbb{P}_{H_1}\{T_f(\boldsymbol{P}) > c_\alpha\} = \mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right)(1 + o(1)), \qquad \alpha \to 0.$$

The conditions in Theorem 7 are useful for establishing power results in concrete models. In particular, they apply to copula models with Gaussian marginals. Theorem 8 below can be viewed either as an application of Theorem 7 or as a generalization of Theorem 6.

Theorem 8. *Suppose that the marginal distribution of* X_i *is* $N(\mu_i, 1)$. *Let* $\Phi_{\mu_i}(\cdot)$ *denote the CDF of* $N(\mu_i, 1)$. *Suppose that, for* $i \neq j$, *the copula of* X_i *and* X_j, $C_X(u_i, u_j) = \mathbb{P}(\Phi_{\mu_i}(X_i) \leq u_i, \Phi_{\mu_j}(X_j) \leq u_j)$, *satisfies* $C_X(u, u) = 2u - 1 + o(1 - u)$, $u \to 1$.

Consider the supremum-based p-value combination statistic $T_f(\boldsymbol{P})$ *defined in* (2.8) *for testing the hypotheses in* (5.23). *Suppose further that Assumptions 1 and 2 are fulfilled and that* $f_1^{-1}(t) \leq f_2^{-1}(t) \leq \ldots \leq f_n^{-1}(t)$ *for any* t. *Let* c_α *be the level-*α *critical value, i.e.,* $\mathbb{P}_{H_0}\{T_f(\boldsymbol{P}) > c_\alpha\} = \alpha$. *Assume that* n, m, *and* $\delta > 0$ *are fixed as* $\alpha \to 0$. *Then*

$$\beta(\delta) = \mathbb{P}_{H_1}(T_f(\boldsymbol{P}) > c_\alpha) = \mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right)(1 + o(1)), \qquad \alpha \to 0. \tag{5.29}$$

Next, for comparison purposes, we evaluate the power of the CCT statistic under the Gaussian model with hypotheses specified in (5.23).

Theorem 9. *Consider the CCT test statistic*

$$T_{CCT} = \sum_{i=1}^{n} \omega_i \tan(\pi(1/2 - P_i)) = \sum_{i=1}^{n} \omega_i \cot(\pi P_i).$$

where $\omega_i \geq 0$ *and* $\sum_{i=1}^n \omega_i = 1$. *For testing the hypotheses in* (5.23) *under the Gaussian model* (5.22), *let* c_α^{CCT} *be its level-*α *critical value, i.e.,* $\mathbb{P}_{H_0}(T_{CCT} > c_\alpha^{CCT}) = \alpha$. *Assume that* n *and* $\delta > 0$ *are fixed as* $\alpha \to 0$. *Then*

$$\beta_{CCT}(\delta) = \mathbb{P}_{H_1}(T_{CCT} > c_\alpha) = \mathbb{P}_{H_1}\left(\bigcup_{i=1,\ldots,n} \{P_i \leq \omega_i \alpha\}\right)(1 + o(1)). \tag{5.30}$$

When $\omega_i > 0$ for all i, (5.30) becomes

$$\beta_{CCT}(\delta) = \mathbb{P}_{H_1}\left(\min_{1\leq i\leq n} \frac{P_i}{\omega_i} \leq \alpha\right)(1 + o(1)). \tag{5.31}$$

Therefore, the CCT test with equal weights $\omega_i = 1/n$ has asymptotically equivalent power to the minP test, since (5.31) reduces to

$$\beta_{CCT}(\delta) \sim \mathbb{P}_{H_1}(P_{(1)} \leq \frac{\alpha}{n}). \tag{5.32}$$

In fact, the assumption of equal signal magnitudes is not essential when the signal strengths are fixed in the regime $\alpha \to 0$. We state the corresponding results for heterogeneous signal strengths in the remarks below. Proofs are provided in the Supplementary Material.

Remark 4. *Theorem 6 remains valid if the alternative hypothesis in* (5.23) *is replaced by*

$$H_1 : \mu_i = \delta_i > 0, i = 1, \ldots, m, \quad \mu_{m+1} = \cdots = \mu_n = 0, \tag{5.33}$$

where $\delta_1, \ldots, \delta_m$ *are fixed positive constants.*

Consistent with (5.25), under heterogeneous signal strengths, we can show (see (S4.63) in the Supplementary Material) that the statistical power is asymptotically equivalent to

$$\mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right) \sim \sum_{i=1}^m \bar{\Phi}\left(z_{\alpha/n} - \delta_i\right), \tag{5.34}$$

which reduces to (5.25) when $\delta_1 = \cdots = \delta_m = \delta$. Note that we require $\delta_i > 0$ for one-sided p-values. For two-sided p-values, the requirement $\delta_i > 0$ can be relaxed to $\delta_i \neq 0$. Indeed, if $P_i = 2\bar{\Phi}(|X_i|)$, then by an analogous argument, $\mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right) \sim \sum_{i=1}^{m} \bar{\Phi}\left(z_{\alpha/(2n)} - |\delta_i|\right)$ and $\mathbb{P}_{H_1}(T_f(\boldsymbol{P}) > c_\alpha) = \mathbb{P}_{H_1}\left(P_{(1)} \leq \frac{\alpha}{n}\right)(1 + o(1))$.

Heterogeneous signal strengths can also be accommodated under general dependence structures through a weaker copula-based condition, as stated in Remark 5.

Remark 5. *If the alternative hypothesis in* (5.23) *is replaced by* (5.33)*, Theorem 8 remains valid under the additional assumption that, for each* $i \neq j$*,*

$$C_X(u_i, u_j) = u_i + u_j - 1 + o\{\min(1 - u_i, 1 - u_j)\}, \qquad u_i, u_j \to 1. \tag{5.35}$$

Combining (5.24), (5.26), and (5.32), we conclude that, under the multi-signal alternative and in the stringent-significance regime $\alpha \to 0$, the first-order asymptotic powers of a broad range of supremum-based p-value combination statistics and the equal-weight CCT statistic are all governed by the same tail event $\left\{P_{(1)} \leq \frac{\alpha}{n}\right\}$. Thus, to first order, these tests derive their power from the occurrence of an extremely small minimum p-value. This theoretical conclusion is well supported by numerical studies in the Supplementary Material, which show that these statistics may have different advantages under different signal patterns but generally have similar power, especially when α is small. Nevertheless, this theoretical analysis has limitations, and further investigation of potentially meaningful power differences in higher-order terms or under fixed-α regimes would be an interesting topic for future research.

6. Numerical Studies

In this section, we present numerical studies to demonstrate the correlation robustness of the minP, Simes, HC, BJ, and $\phi 3$ tests under non-asymptotic and realistic settings of data analysis, especially in genome-wide association studies. We also include the CCT as a benchmark, given its well-known robustness to correlations. Due to space limitations, an extensive study of statistical power under the Gaussian model, linear models for binary and continuous traits, and a non-Gaussian copula is provided in the Supplementary Material. In addition, Supplementary Section S5.3.2 also includes comparisons with more summation-based p-value combination methods: the heavy-tailed Box-Cox transformation statistics (Fang et al., 2023) and the averaging combination statistics (including the harmonic mean (HM) statistic) (Vovk and Wang, 2020).

6.1 Correlation Robustness

To demonstrate correlation robustness, we simulate p-value combination statistics based on input statistics $\boldsymbol{X}$ following the Gaussian model in (3.10), with the corresponding p-values given in (3.11). We consider three types of structures for the correlation matrix $\boldsymbol{\Sigma} = (\sigma_{ij})_{1\le i,j\le n}$:

1. Equal-correlation: $\sigma_{ij} \equiv \sigma$ for $1 \le i \ne j \le n$, where σ will be specified in a given simulation setting.

2. Polynomial-decay: Positive correlations, $\sigma_{ij} = 1/|1+i-j|$, decay polynomially with the distance between components. The negative polynomial-decay structure uses a correlation matrix derived from $\boldsymbol{\Sigma}^{-1}$, where $\boldsymbol{\Sigma}$ follows the positive polynomial-decay structure.

3. Linkage disequilibrium (LD) structure: $\mathbf{\Sigma}$ represents the LD matrix of SNPs in human genes. We use SNPs from the Genetic Factors for Osteoporosis Consortium (GEFOS) 2012 study on femoral neck bone mineral density (FN-BMD) (Estrada et al., 2012). The SNPs are grouped by their physical locations into 22,648 genes. The number of SNPs per gene varies from 2 to 985, with a mean of 29.14 and a median of 12. The distribution of gene sizes is shown in Supplementary Figure S9. $\mathbf{\Sigma}$ is constructed based on the pairwise LD r values among these SNPs' genotype data obtained from the 1000 Genomes Project Phase 3 (1KG) (1000 Genomes Project Consortium, 2015).

To measure the correlation robustness of a right-sided test with statistic T at a given α level, we introduce the true-to-nominal error ratio

$$r(\alpha) = \frac{\mathbb{P}_{H_0}(T > t_\alpha^{ind}|\mathbf{\Sigma})}{\alpha}, \tag{6.36}$$

where t_α^{ind} is the α-level threshold obtained under the assumption of independent input test statistics, i.e., $\mathbb{P}_{H_0}(T > t_\alpha^{ind}|\mathbf{I}) = \alpha$, the nominal error rate. Accordingly, $\mathbb{P}_{H_0}(T > t_\alpha^{ind}|\mathbf{\Sigma})$ is the true type I error rate at the threshold t_α^{ind} when the input test statistics are actually correlated, with the correlation matrix $\mathbf{\Sigma}$. Similarly, for a left-sided test (such as the minP and Simes tests), the ratio is defined as $r(\alpha) = \alpha^{-1}\mathbb{P}_{H_0}(T < t_\alpha^{ind}|\mathbf{\Sigma})$.

If $r(\alpha)$ is close to 1, the test is not sensitive to $\mathbf{\Sigma}$ and is thus correlation robust. In this case, one can proceed as if the input test statistics were independent and use t_α^{ind} as the threshold, since the true type I error rate is close to α, the intended control level. A value of $r(\alpha) > 1$ indicates that the test assuming independence is liberal, with a type I error rate larger than α, while $r(\alpha) < 1$ suggests that the test is conservative.

By definition, t_α^{ind} is obtained based on the statistic's null distributions under independence. Specifically, for the minP test, t_α^{ind} is derived from $T_{minP} \sim \text{Beta}(1, n)$ under inde-

pendence. For the Simes test, $t_\alpha^{ind} = \alpha$, since $T_{Simes} \sim \text{Uniform}(0,1)$ under independence. For the HC, $\phi 3$, and BJ tests, we obtain t_α^{ind} using the exact distributions of T_{HC}, $T_{\phi 3}$, and T_{BJ} under independence, as given in Zhang et al. (2020). For the Cauchy combination test, $T_{cct} \sim \text{Cauchy}(0,1)$ under independence. After obtaining t_α^{ind}, we compute $r(\alpha)$ empirically based on 10^7 simulations.

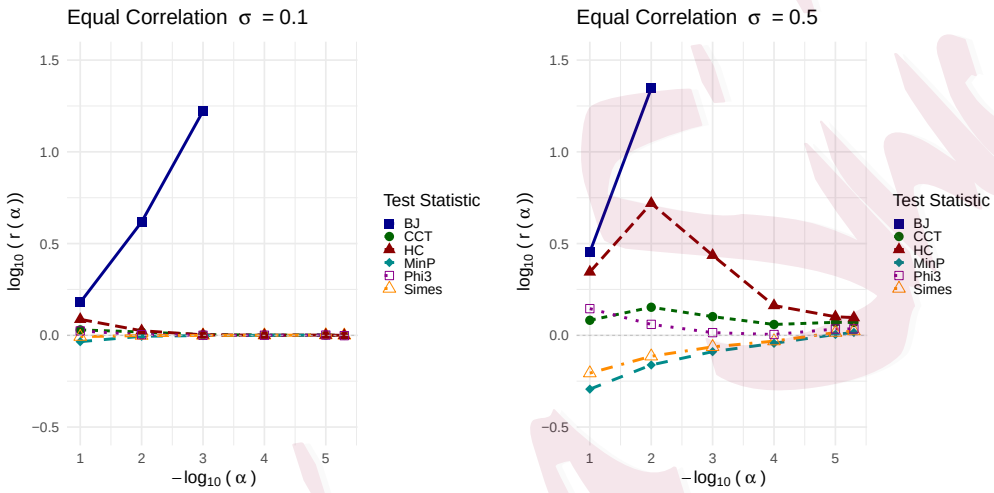

Figure 1: Plots of $\log(r(\alpha))$ against $-\log_{10}(\alpha)$ under the equal-correlation structure with $\sigma = 0.1$ and 0.5, and $n = 100$.

Figure 1 visualizes $r(\alpha)$ on a logarithmic scale for various tests under the equal-correlation structure at different correlation strengths ($\sigma = 0.1$ and 0.5). Several observations can be made. First, as expected, the $r(\alpha)$ values for these tests are close to 1 under weak correlations. Second, the minP, Simes, HC, and $\phi 3$ tests are asymptotically correlation-robust, because their $r(\alpha)$ values converge to 1 as α decreases (i.e., as $-\log_{10}(\alpha)$ increases). However, this is not the case for the BJ test, whose $r(\alpha)$ values diverge. This behavior aligns with the fact that the BJ statistic does not satisfy the condition in Theorem 4 for a test statistic to be ACR. Third, we can observe the relative correlation robustness among

these tests. A test is more robust (i.e., less sensitive) to correlation if its $r(\alpha)$ is closer to 1. From this perspective, the CCT, Simes, and $\phi 3$ tests are generally more correlation-robust than the minP and HC tests. Furthermore, if independence is assumed for testing, the HC and $\phi 3$ tests tend to be liberal (with $r(\alpha) > 1$), while the minP and Simes tests tend to be conservative. Additional results under other correlation settings, different n values (10, 50, 200), and under realistic data analysis scenarios (particularly when the input statistics are derived from genotype-phenotype association analyses), are shown in Figures S1 – S5 in the Supplementary Material. Furthermore, we provide direct empirical evidence that the asymptotic tail independence condition is reasonable in GWAS data by estimating empirical upper tail-dependence coefficients (TDCs) from genotype-based simulations across genes with varying LD strengths; see Figures S6 and S7 in the Supplementary Material. In contrast, simulations show that these tests lose their correlation-robustness when the tail independence assumption is violated. See Figure S8 for results under pairwise dependence modeled by the bivariate Cuadras-Augé copula.

Because $r(\alpha)$ consistently approaches 1 as α approaches 0, we focus on the largest α value (i.e., the least stringent control level) that allows the independence-assumed test to remain sufficiently robust. Specifically, we define a test as "sufficiently robust" if $r(\alpha) \in [0.8, 1.2]$ at a given α level. We are interested in the largest α value, denoted $\alpha^{sr} \in (0, 1]$, such that $r(\alpha)$ remains sufficiently robust for all $0 \leq \alpha \leq \alpha^{sr}$; in other words,

$$\alpha^{sr} = \max\{\alpha^* \in (0, 1] : r(\alpha) \in [0.8, 1.2] \text{ for all } \alpha \leq \alpha^*\}. \tag{6.37}$$

The value α^{sr} quantifies the test's correlation-robustness across all α levels, summarizing the curve $r(\alpha)$ for $0 \leq \alpha \leq 1$. A larger α^{sr} indicates greater robustness, as the test can tolerate higher α levels while maintaining the type I error rate near the nominal rate. Define the

nominal (i.e., independence-assumed) p-value at an observed statistic t_0 as $p \doteq \mathbb{P}_{H_0}(T > t_0|\mathbf{I})$. If $p \leq \alpha^{sr}$, then $r(p) \in [0.8, 1.2]$. This means that, when the data are correlated, the true p-value $\mathbb{P}_{H_0}(T > t_0|\mathbf{\Sigma})$ will differ from p by less than 20%.

To assess the correlation robustness of the HC, MinP, Simes, $\phi 3$, and CCT tests in the GWAS setting, we examine their α^{sr} values in SNP-set tests across 22,648 human genes. These genes vary in SNP counts (n) and LD patterns, which are reflected in the correlation matrix $\mathbf{\Sigma}$. To simplify computation, we estimate only a lower or upper bound of α^{sr}, as each gene requires extensive simulations to obtain $r(\alpha)$. This approach is reasonable because $r(\alpha)$ is nearly monotonic as α approaches 0, as shown in Figure 1. Specifically, for each gene and test, we compute $r(\alpha)$ at the following values of $\alpha : \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$. If $0.8 \leq r(10^{-1}) \leq 1.2$, then $\alpha^{sr} \geq 10^{-1}$. Otherwise, we compute $r(\alpha)$ at $\alpha = 10^{-2}$. If $0.8 \leq r(10^{-2}) \leq 1.2$, then $10^{-2} \leq \alpha^{sr} < 10^{-1}$; otherwise, we proceed to $\alpha = 10^{-3}$, and so on. If $r(\alpha)$ falls outside the range $[0.8, 1.2]$ for all five values of α, then $\alpha^{sr} < 10^{-5}$. We estimate $r(\alpha)$ in (6.36) using 10^7 Monte Carlo replicates. For each gene, the test statistic T is computed from the two-sided SNP-level p-values defined in (3.11), which are obtained from input statistics simulated under the Gaussian model (3.10) with correlation matrix $\mathbf{\Sigma}$ given by the LD matrix of the SNPs within the gene. This procedure yields a gene-specific estimate of the null tail probability under the within-gene correlation structure.

Table 1 categorizes the 22,648 genes into six groups based on their α^{sr} values: $\geq 10^{-1}$, $[10^{-2}, 10^{-1})$, $[10^{-3}, 10^{-2})$, $[10^{-4}, 10^{-3})$, $[10^{-5}, 10^{-4})$, and $< 10^{-5}$, for the five tests. Following GWAS practice, a gene is considered trait-associated if its p-value is below the genome-wide significance level, $0.05/22648 \approx 2.2 \times 10^{-6}$. Most genes have $\alpha^{sr} > 2.2 \times 10^{-6}$, meaning that if these genes are genome-wide significantly associated (i.e., their true p-values $< 2.2 \times 10^{-6} <$

Table 1: Numbers of genes out of 22,648 reaching α^{sr} levels. Proportions of conservative tests by the independence assumption (i.e., $r(\alpha) < 1$) are shown in parentheses.

α^{sr}	$\geq 10^{-1}$	$[10^{-2}, 10^{-1})$	$[10^{-3}, 10^{-2})$	$[10^{-4}, 10^{-3})$	$[10^{-5}, 10^{-4})$	$< 10^{-5}$
HC	7973 (0)	432 (0)	5667 (0)	4935 (0.003)	1985 (0.11)	1656 (0.18)
MinP	9631 (0.99)	6607 (1)	3910 (1)	1418 (0.99)	577 (0.83)	505 (0.63)
Simes	18428 (0.94)	3307 (1)	655 (1)	157 (1)	53 (0.89)	48 (0.69)
$\phi 3$	21026 (0)	1616 (0)	6 (0)	–	–	–
CCT	22648 (0)	0	–	–	–	–

α^{sr}), their p-values can be calculated under the independence assumption, as their true p-values will differ by less than 20%. Thus, for detecting trait-associated genes, complex algorithms to account for SNP dependence for these tests may generally be unnecessary due to their correlation robustness. This is further supported by the real data study in Section 7. Additionally, Table 1 shows proportions of conservative tests (in parentheses) under the independence assumption. The minP and Simes tests are often conservative, while the HC, $\phi 3$, and CCT tests are predominantly liberal, consistent with the patterns observed in Figure 1.

We further analyze how correlation strengths influence α^{sr} by examining its distribution across three groups of genes with weak, medium, and strong correlations. The results show that the tests exhibit decreasing robustness to correlation in the following order: CCT, $\phi 3$, Simes, minP, and HC. For more details, see the Supplementary Material.

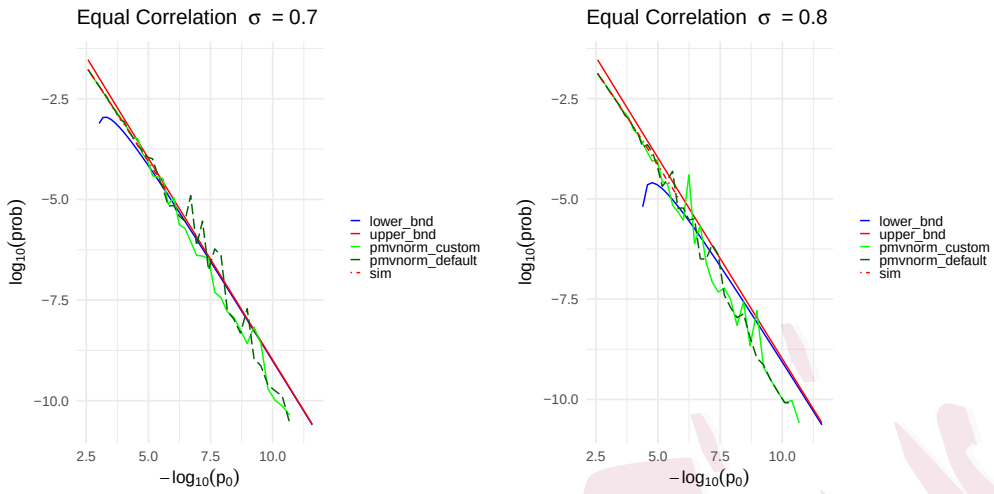

Figure 2: Plot $\log_{10}\mathbb{P}\left(T_{minP} < p_0|\mathbf{\Sigma}\right)$ against $-\log_{10}(p_0)$ under the equal-correlation structure with $\sigma = 0.7$ and 0.8 and $n = 10$, where $\mathbb{P}\left(T_{minP} < p_0|\mathbf{\Sigma}\right)$ is evaluated by simulations ("sim"), and using `pmvnorm` ("default" and "custom" with the error tolerance parameter "abseps" set to the default 10^{-6} and 10^{-15}, respectively), along with its upper and lower bounds from Corollary 2 ("lower_bnd" and "upper_bound").

6.2 On calculating p-values for the minP test

Numerical evidence shows that Corollary 2 provides tight bounds for the far-end tail probability, often outperforming the general-purpose R function `pmvnorm`. When the bounds are close, their average provides an effective way to approximate the p-value of the minP test under correlation.

Figure 2 shows the tail probability $\mathbb{P}\left(T_{minP} < p_0|\mathbf{\Sigma}\right)$ against p_0 on a logarithmic scale. The true tail probability, obtained via simulations, falls within the lower and upper bounds given in (3.15) (for two-sided p-values, $p_0 = 2\bar{\Phi}(t)$). These bounds provide an accurate approximation of the tail probability when p_0 is small. In contrast, the tail probability calculated by `pmvnorm` often exceeds these bounds and exhibits non-smooth curves due to

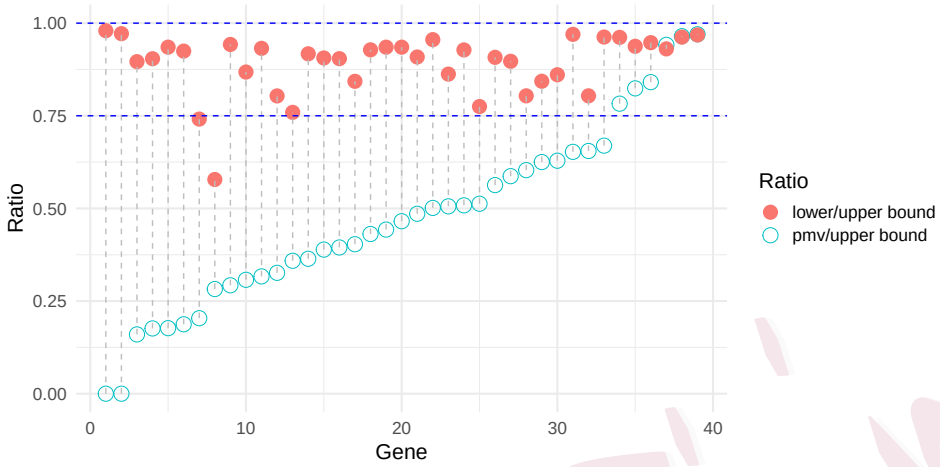

Figure 3: The lower-to-upper bound ratio (red dots) versus the `pmvnorm`-to-upper bound ratio (blue circles) for the 39 top-hit genes. The lower and upper bounds are given in (3.15).

numerical instability at the far tail. Additional simulation results, including those with LD-based correlation structures, are presented in Figures S13 and S14 in the Supplementary Material, confirming similar findings.

The lower and upper bounds in Corollary 2 are very close at the genome-wide significance level used in GWAS. Using summary statistics from the GEFOS2012-FN study, 39 genes have minP p-values below the genome-wide significance threshold of 2.2×10^{-6}. Figure 3 displays the lower-to-upper-bound ratio for these 39 genes (represented by red dots). Except for one gene (the 8th from left), all other genes have tight bounds, with a lower-to-upper-bound ratio of 0.75 or higher. Additionally, we compute the ratio of the tail probability by `pmvnorm` to the upper bound. It turns out that 36 genes have p-values computed by `pmvnorm` that fall below the lower bound (represented by blue circles, including the 8th gene from left), suggesting that `pmvnorm` tends to provide less accurate (and more liberal) p-values compared to the average of the lower and upper bounds. The remaining three genes have nearly identical values for the two ratios. Similar findings are observed for the top 43 significant genes by

the CCT test (see Figure S15 in the Supplementary Material). Moreover, `pmvnorm` does not support n values greater than 1000, and its computation time can be over 100 times slower than the bound-based calculation (see Figure S16 in the Supplementary Material).

7. Real Data Analysis

In this section, we demonstrate the performance of ACR tests by analyzing the GWAS summary dataset for GEFOS femoral neck bone mineral density (Estrada et al., 2012). Specifically, we conduct gene-based SNP-set studies using the minP, Simes, HC, $\phi 3$, and CCT tests to aggregate individual SNP p-values for each gene into a single gene-level p-value. Two methods are used for p-value calculation. The first method computes p-values for the minP, Simes, HC, and $\phi 3$ tests by applying the dominant term of the tail probability from Theorems 1 and 2, and Corollaries 4 and 5, respectively, without accounting for correlation/LD information. The second method employs the Effective Correlation Coefficient (ECC) algorithm, which explicitly accounts for correlations when calculating the tail probability for supremum-based p-value combination tests. This approach is available in the R package `SetTest` (Zhang and Wu, 2022).

Table 2 reports the number of genome-wide significant genes ("hits"), the number of literature-reported osteoporosis genes among the hits ("literature hits"), and the genomic inflation factor λ. In GWAS, $\lambda = F^{-1}_{\chi^2_1}(1-p_{median})/F^{-1}_{\chi^2_1}(0.5)$, where $F^{-1}_{\chi^2_1}$ denotes the inverse CDF of χ^2_1 distribution, and p_{median} represents the median of gene p-values (Devlin and Roeder, 1999). The λ value compares the median observed chi-squared statistics (from p-values) to the expected median under the null hypothesis, with λ close to 1 indicating accurate median-level p-value calculation. Compared to asymptotic tail probabilities, the

Table 2: The numbers of hits, literature hits, and λ values by five tests.

Tests (p-value calculation methods)	Hits	Literature hits	λ
minP (Theorem 1)	39	33	0.00
minP (ECC)	39	33	0.52
Simes (Theorem 2)	40	34	0.60
Simes (ECC)	40	34	1.04
$\phi 3$ (Corollary 5)	43	37	2.55
$\phi 3$ (ECC)	43	37	1.65
CCT	43	37	0.75
HC (Corollary 4)	47	41	3.64
HC (ECC)	47	41	1.81

ECC algorithm produces λ values closer to 1, demonstrating its greater accuracy for moderate p-values at the median level. However, both ECC and asymptotic tail probabilities identify the same sets of hits and literature hits in all relevant tests, highlighting the effectiveness of the asymptotic method in identifying top associated genes.

When comparing across tests, the Simes test identifies one additional literature hit, gene *RERE*, compared to the minP test, suggesting slightly higher power. The CCT and $\phi 3$ tests identify the same number of hits and literature hits, differing by only one osteoporosis gene: the CCT test identifies *SMG6*, while the $\phi 3$ test identifies *PKDCC*, indicating similar power for both tests. The HC test identifies the largest number of hits and literature hits, demonstrating the highest power for this dataset. The list of top-hit genes identified by these tests is provided in the supplementary file "top_gene_results.xlsx".

Beyond the literature hits, six putative novel genes were commonly identified by all five tests: *MIR6782*, *HOXC-AS2*, *LOC101927839*, *HOXC-AS1*, *MIR615*, and *SRR*. Among these, at least two genes may be relevant to bone mineral density and osteoporosis. The gene *SRR* has been shown to suppress metastasis in human osteosarcoma cells when down-regulated (Pu et al., 2016). Additionally, *MIR615* regulates growth and development during osteogenesis and angiogenesis (Godínez-Rubí and Ortuño-Sahagún, 2020) and functions as a tumor suppressor in osteosarcoma by inhibiting the enzyme Hexokinase 2 (HK2) (Sun et al., 2020).

8. Discussion

In this paper, we established the ACR property for a broad class of supremum-based p-value combination tests under Gaussian and pairwise copula models. We also conducted numerical and real-data studies to evaluate the robustness of these tests to correlation in realistic settings, particularly in genome-wide association studies. When only very small p-values are of interest (e.g., for quick screening of genome-wide significantly associated genes), the minP, Simes, $\phi 3$, and HC tests can be applied under the assumption of independence, significantly reducing computational demand.

Although this work focuses on the ACR property of supremum-based p-value combination tests, the choice among these tests should be guided by practical considerations. For computational efficiency and correlation robustness, the minP and Simes tests are preferable, as the independence approximation is highly accurate for screening purposes. These tests are also easier to compute accurately, even when accounting for correlation. The multivariate normal distribution can be used to obtain the p-value for minP. The ECC algorithm for

the Simes test is also very accurate. For higher statistical power, the HC and $\phi 3$ tests are recommended, especially when signals are relatively weak ($\phi 3$ may have a slight advantage for denser signals), as demonstrated in the literature (Donoho and Jin, 2004; Zhang and Wu, 2022). In the Supplementary Material (Section S5.3.2), we further compare the power of the supremum-based tests and the CCT against several summation-based combination methods, including the harmonic mean (HM) statistic (Wilson, 2019; Vovk and Wang, 2020), the heavy-tailed Box–Cox transformation statistics (Fang et al., 2023), and the averaging combination statistics (Vovk and Wang, 2020). The results show that the correlation-robust tests—both the supremum-based tests (minP, Simes, HC, $\phi 3$) and the summation-based tests (CCT, HM, $BC_{1.5}$, $BC_{0.5}$)—exhibit clustered power patterns, whereas BJ, GM, and AM are not correlation-robust and can lose substantial power under strong correlations.

One theoretical limitation of this study is the assumption that the number of input p-values n is fixed. Extending the ACR study to high-dimensional settings, where $n \to \infty$, is an interesting direction for future research. Another interesting area for future research is to explore the theoretical reasons why some tests are more robust to correlation than others, as the current results on the ACR property do not fully explain these differences. Such research may require more in-depth analyses that account for α levels and quantify the impact of correlations on the tail probabilities of various statistics.

Supplementary Materials

The Supplementary Materials include additional discussions, proofs of main results, and extra numerical studies. The supplementary file, *top_gene_results.xlsx*, lists the top-hit genes identified in the GEFOS2012-FN study.

Acknowledgements

We acknowledge partial financial support from NSF grants DMS-2113570 and DMS-2515791.

References

1000 Genomes Project Consortium (2015). A global reference for human genetic variation. *Nature 526*(7571), 68–74.

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.

Block, H. W., T. H. Savits, and J. Wang (2008). Negative dependence and the simes inequality. *Journal of statistical planning and inference 138*(12), 4107–4110.

Devlin, B. and K. Roeder (1999). Genomic control for association studies. *Biometrics 55*(4), 997–1004.

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

Donoho, D. L. and J. Jin (2008, Sep 30). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *Proceedings of the National Academy of Sciences of the United States of America 105*(39), 14790–14795.

Estrada, K., U. Styrkarsdottir, E. Evangelou, Y.-H. Hsu, E. L. Duncan, E. E. Ntzani, et al. (2012). Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. *Nature genetics 44*(5), 491–501.

Fang, Y., C. Chang, Y. Park, and G. C. Tseng (2023). Heavy-tailed distribution for combining dependent p-values with asymptotic robustness. *Statistica Sinica 33*, 1115–1142.

Godínez-Rubí, M. and D. Ortuño-Sahagún (2020). mir-615 fine-tunes growth and development and has a role in cancer and in neural repair. *Cells 9*(7), 1566.

Ingster, Y. I. and I. A. Suslina (2000). Minimax nonparametric hypothesis testing for ellipsoids and besov bodies. *ESAIM: Probability and Statistics 4*, 53–135.

Jager, L. and J. A. Wellner (2007). Goodness-of-fit tests via phi-divergences. *The Annals of Statistics 35*(5), 2018–2053.

Li, W. V. and Q.-M. Shao (2002). A normal comparison inequality and its applications. *Probability Theory and Related Fields 122*(4), 494–508.

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.

Long, M., Z. Li, W. Zhang, and Q. Li (2023). The cauchy combination test under arbitrary dependence structures. *The American Statistician 77*(2), 134–142.

Pu, Y., F. Zhao, W. Cai, X. Meng, Y. Li, and S. Cai (2016). Mir-193a-3p and mir-193a-5p suppress the metastasis of human osteosarcoma cells by down-regulating rab27b and srr, respectively. *Clinical & experimental metastasis 33*, 359–372.

Sarkar, S. K. (1998). Some probability inequalities for ordered mtp2 random variables: a proof of the simes conjecture. *Annals of Statistics 26*(2), 494–504.

Sarkar, S. K. and C.-K. Chang (1997). The simes method for multiple hypothesis testing with positively dependent test statistics. *Journal of the American Statistical Association 92*(440), 1601–1608.

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

Sun, L., P. Wang, Z. Zhang, K. Zhang, Z. Xu, S. Li, and J. Mao (2020). Microrna-615 functions as a tumor suppressor in osteosarcoma through the suppression of hk2. *Oncology Letters 20*(5), 1–1.

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

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

Zhang, H., J. Jin, and Z. Wu (2020). Distributions and power of optimal signal-detection statistics in finite case. *IEEE Transactions on Signal Processing 68*, 1021–1033.

Zhang, H., M. Liu, J. Jin, and Z. Wu (2023). Signal-noise ratio of genetic associations and statistical power of snp-set tests. *Annals of Applied Statistics 17*(3), 2410–2431.

Zhang, H. and Z. Wu (2022, March). The general goodness-of-fit tests for correlated data. *Computational Statistics & Data Analysis 167*, 107379.

Worcester Polytechnic Institute, Worcester, MA, USA

E-mail: (xchen8@wpi.edu)

Worcester Polytechnic Institute, Worcester, MA, USA

E-mail: (fwang4@wpi.edu)

ORCID: 0000-0002-9807-6446

Global Biometrics & Data Management, Pfizer Inc., New York, NY, USA

E-mail: (hong.zhang3@pfizer.com)

ORCID: 0000-0002-8869-8671

Worcester Polytechnic Institute, Worcester, MA, USA

E-mail: (zheyangwu@wpi.edu)

ORCID: 0000-0003-0641-0791