Statistica Sinica **35** (2025), 2137-2158
doi:https://doi.org/10.5705/ss.202023.0117

COMBINING *P*-VALUES USING HEAVY-TAILED DISTRIBUTIONS AND THEIR ASYMPTOTIC RESULTS WITH APPLICATIONS TO GENOMIC DATA

Junsik Kim and Junyong Park*

Duksung Women's University and Seoul National University

Abstract: Combining individual p-values to handle large scale inferences or to aggregate results of different studies is one of major interest in meta-analysis which has been traditionally based on independent p-values. In contrast to combining methods that are constructed when p-values are independent, recently proposed combinations of p-values transformed into heavy-tailed distribution are known to be robust to the dependence structure of p-values. In this paper, we investigate theoretical properties of combining p-value methods for different heaviness of transformation under a wider class of correlation structures compared to existing studies from the viewpoint of controlling Type I error and obtaining powers. We also investigate relationships between harmonic mean type combination methods and combining methods that use transformation of p-values into stable distribution including Cauchy and Lévy combination methods. We provide extensive numerical studies supporting theoretical results. We also apply these p-value combining methods to real example of Crohn's disease data and present some idea on how to validate these methods.

Key words and phrases: Cauchy transformation, combining p-values, GWAS, Lévy transformation, meta-analysis.

1. Introduction

Meta-analysis has been used in various fields as a statistical technique to draw more reliable conclusions by combining the results of different studies or experiments. Most of research has focused on how to aggregate p-values generated from the assumption that all experiments or studies are independent. Among the methods for combining such independent p-values, well known methods including Fisher's method in Fisher (1934) and Stouffer's method in Stouffer et al. (1949) are based on aggregating some transformed p-values leading to chi-square distribution and normal distribution, respectively. In particular, the distribution of the sum of independent p-values or other forms of statistics needs to be derived relatively easily for practical use. For example, if the p-value is transformed into an infinitely divisible random variable, the distribution of

*Corresponding author. E-mail: junyongpark@snu.ac.kr

the sum of those transformed p-values is the same type of distribution as the transformed distribution as in the case of Fisher's method and Stouffer's method.

On the other hand, there are limited studies on combining dependent p-values. When p-values are dependent, Kost and McDermott (2002) and Hartung (1999) aggregated the dependent p-values by estimating the correlation coefficients. Recently, Liu and Xie (2020) showed that combining dependent p-values transformed into the Cauchy distribution is robust to the overall dependency when a significance level is fairly small. Numerical studies in Wilson (2021) demonstrated the robustness of Lévy combination test for dependent p-values. Fang, Tseng and Chang (2023) showed that combinations of dependent p-values have the same tail behavior as that from independent p-values, however their result is restricted to the case when the number of p-values is fixed. Liang and Rho (2022) used stable combination tests under long range or short range dependence which is somewhat restricted dependence in meta analysis. Wilson (2019) proposed to use a harmonic mean of p-values to combine dependent p-values.

In this paper, we consider a broader class of dependence structure induced from Gaussian copula used in Liu and Xie (2020) and investigate the asymptotic performance of existing test statistics from the viewpoint of controlling a given Type I error and obtaining power. Liu and Xie (2020) assumed that eigenvalues of correlation matrix are bounded which cannot include highly correlated p-values. The dependence structure in Liang and Rho (2022) is also somewhat restricted in practice. We focus on figuring out the scopes of dependence structure in which existing methods control a given Type I error and obtain powers.

Among these two criteria in hypothesis testing, the first requirement is controlling Type I error while obtaining power is pursued thereafter. In order to obtain the robustness to dependence of p-values, it is advantageous to have test statistics dominated by a small number of small p-values. However this property is undesirable to obtain testing power under non-null hypothesis especially when there are many p-values generated from non-null hypothesis. Test statistics depending on a relatively small number of p-values may lose testing power since transformation of fairly small values of p-values tends to be extremely large, so they dominate transformation of some other p-values which are also generated from non-null hypotheses.

With this motivation, for dependent p-values, we investigate the properties of methods of combining p-values in controlling Type I error and obtaining power for different tail behaviors from different transformations. In particular, it is highlighted that all test statistics considered in this paper are shown to have a trade-off relationship between size and power. In other words, if the size is controlled stably at strongly dependent p-values, the power will be low, whereas it becomes difficult to control the size of a test statistic with an advantage in power at p-values with strong dependence. We demonstrate the trade-off theoretically

and numerically via presenting asymptotic results, simulations and real data examples.

Test statistics considered in this paper are classified by the heaviness of the tails of the transformed p-values. Depending on the heaviness of the tail, we consider heavier tailed distributions such as the Cauchy distribution and the Lévy distribution which have been studied in Liu and Xie (2020), Chen et al. (2023) and Wilson (2021), in addition to well known methods such as Stouffer's and Fisher's methods. We provide various results on how different heaviness of tails affects the performance of test statistics for dependent p-values.

We also investigate asymptotic relations between p-values combining methods for heavy-tailed distribution and for a type of harmonic mean of p-values. Combining methods using the Cauchy and Lévy distribution and harmonic mean are designed to rely on a small subset of transformed p-values which dominates all other transformed p-values due to the heavy tail of Cauchy or Lévy distributions and inverse of p-values in harmonic mean. This property results in diminishing the effect of dependence of p-values since a few p-values dominating all other p-values avoid the accumulation of dependence. Chen et al. (2023) showed that Cauchy combination method and harmonic mean method are asymptotically equivalent when the number of p-values is fixed and the smallest of p-value goes to zero. On the other hand, Fisher's method and Stouffer's method are not explained by a few number of p-values, since these two methods use chi-square and normal distributions which do not have heavy tails. As an extreme case, the Tippet's method (Tippett, 1931) which uses only the smallest p-value is also considered as the heaviest transformation.

This paper is organized as follows. Section 2 introduces the test statistics for different transformations. Section 3 presents theoretical results from the viewpoint of asymptotic size and power of the test statistics. In Section 4, relations between p-values combining methods for heavy-tailed distribution are investigated. Numerical studies and real data examples are provided in Sections 5 and 6, respectively. Section 7 presents concluding remarks.

2. Test Statistics for Combining p-Values

We consider the following global hypothesis testing:

$$H_0 : H_{0i} \text{ are all true for } 1 \le i \le d \quad vs. \quad H_1 : \text{at least one } H_{0i} \text{ is false}, \tag{2.1}$$

where $H_{0i} : \mu_i = 0$ for two-sided test, and $H_{0i} : \mu_i \le (\ge) 0$ for one-sided test, including right-tailed and left-tailed tests. For each hypothesis H_{0i}, we obtain p-values, $p_1, \ldots, p_d$ by marginally testing each hypothesis. In meta analysis, it may not be known about the marginal data as well as the joint distribution of those test statistics, but Liu and Xie (2020) assumed that the original data or the test statistic $\mathbf{X} = (X_1, \ldots, X_d)^T$ follows the multivariate normal distribution,

$\mathbf{X} \sim N_d(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_d)^T$ and Σ is unknown. The p-values are represented by $p_i = 1 - \Phi(X_i)$ or $\Phi(X_i)$ depending on right-tailed or left-tailed test and $p_i = 2\{1 - \Phi(|X_i|)\}$ for two-sided test. Σ characterizes dependence structure of p-values; however, since it is unknown, it is not feasible to identify such dependence.

Various methods of combining p-values in meta-analysis are constructed through an appropriate transformation of p-values such as $T = \sum_{i=1}^{d} h(p_i)$ where $h(\cdot)$ is a decreasing function. Typical examples are Stouffer's method, $T_{\text{Stouffer}} = \sum_{i=1}^{d} \bar{\Phi}^{-1}(p_i)/\sqrt{d}$, and Fisher's method, $T_{\text{Fisher}} = -2\sum_{i=1}^{d} \log(p_i)$. When p-values are independent, T_{Stouffer} and T_{Fisher} have the standard normal distribution and chi-square distribution with $2d$ degrees of freedom under H_0, respectively. We reject H_0 if $T_{\text{Stouffer}} > z_\alpha$ and if $T_{\text{Fisher}} > \chi^2_\alpha(2d)$, where z_α and $\chi^2_\alpha(2d)$ are the upper α quantiles of a standard normal distribution and chi-square distribution with degrees of freedom $2d$, respectively. However, when p-values are dependent, distributions of T_{Stouffer} and T_{Fisher} are not guaranteed to be the standard normal and chi-square distributions, respectively, so such decision rules are not valid any more. If there exists relatively strong dependency among p-values, T_{Stouffer} and T_{Fisher} with the critical values z_α and $\chi^2_\alpha(2d)$ are not able to control Type I error.

Recently, methods for combining dependent p-values using heavy-tailed distributions have been studied intensively since combination of dependent heavy-tailed distributions may have similar behavior to the combination of those independent distributions. For example, the Cauchy transformation, $T_{Cauchy} = \sum_{i=1}^{d} \tan\{\pi(1/2 - p_i)\}/d$, in Liu and Xie (2020), follows Cauchy distribution under H_0 and independent p-values. Liu and Xie (2020) showed that, under dependent p-values with some conditions, the tail behavior of T_{Cauchy} is asymptotically equivalent to the standard Cauchy distribution. Then H_0 is rejected when $T_{Cauchy} > c_\alpha$ where c_α is the upper α quantile of Cauchy distribution. With a similar motivation, we can also consider a heavier tailed distribution than Cauchy distribution such as Lévy distribution (Wilson, 2021). We define $T_{\text{Lévy}} = \sum_{i=1}^{d} [\Phi^{-1}\{(1 + p_i)/2\}]^{-2}/d^2$ and reject H_0 if $T_{\text{Lévy}} > l_\alpha$ where l_α is the upper α quantile of Lévy distribution.

Note that tail behaviors of $T_{\text{Stouffer}}, T_{\text{Fisher}}, T_{\text{Cauchy}}$ and $T_{\text{Lévy}}$ are different in that the thickness of the tail part has an order of $T_{\text{Stouffer}} < T_{\text{Fisher}} < T_{\text{Cauchy}} < T_{\text{Lévy}}$. Lastly, we also consider the Tippet's method which uses the minimum of p-values, $T_{\text{minP}} = \max_i \Phi^{-1}(1 - p_i)$, or $T'_{\text{minP}} = \min_i p_i$, equivalently. As a critical value which is based on independence, we reject H_0 if $T_{\text{minP}} \geq \Phi^{-1}\{(1 - \alpha)^{1/d}\}$ or equivalently $T'_{\text{minP}} = \min_i p_i \leq 1 - (1 - \alpha)^{1/d}$, for a given significant level α.

Throughout this paper, we evaluate different properties of T_{Stouffer}, T_{Fisher}, T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} by their different distributions with different tail behaviors. Each of test statistics uses critical values derived under independent p-values such as z_α, $\chi^2_\alpha(2d)$, c_α, l_α and $\Phi^{-1}\{(1 - \alpha)^{1/d}\}$, respectively. These critical values

may departure from the true critical values of $T_{\text{Stouffer}}, T_{\text{Fisher}}, T_{\text{Cauchy}}, T_{\text{Lévy}}$ and T_{minP} under dependent p-values. In practice, since we do not know the dependent structure induced from unknown Σ, we have no choice but to use them derived under independent p-values. We highlight that, as the tail of the transformed distribution $h(\cdot)$ is heavier, a smaller number of p-values dominate the variation of the test statistics which avoids the accumulation of dependence of p-values. We also address the issue of testing power of these test statistics which have reverse order of performances of robustness in controlling Type I error.

In the following sections, we provide theoretical studies on the properties of $T_{\text{Stouffer}}, T_{\text{Fisher}}, T_{\text{Cauchy}}, T_{\text{Lévy}}$ and T_{minP} and present numerical studies including simulations and real data examples. All proofs of theoretical studies are presented in the Supplementary Material.

3. Type I Error and Testing Power

In this section, we present theoretical results on controlling a given Type I error and obtaining power of T_{Stouffer}, T_{Fisher}, T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP}. Throughout this section, we see that the heavier the distribution of the transformed p-values, the more advantageous it is to control the Type I error. However, we also see that such heavy-tailed distributions can be detrimental to the achievement of testing power, which is $1 -$ Type II error. As an overall performance, we investigate the conditions for each of test statistics under which we have the following properties: Let α be a given level of Type I error. Then, as $d \to \infty$ and $\alpha \to 0$,

$$\text{Type I error} \le \alpha \to 0 \quad \text{and} \quad \text{Type II error} \ \to 0. \tag{3.1}$$

We compare conditions satisfying (3.1) for different types of testing procedures and observe there exists trade-off relationship to obtain these two criteria. To implement dependent p-values, we use the Gaussian copula as in Liu and Xie (2020) rather than considering arbitrary dependence in Vovk and Wang (2020) and Chen et al. (2023). As seen in Vovk and Wang (2020), a modification of some test statistics for arbitrary dependent p-values become impractical due to their fairly low powers.

We use the Gaussian copula to model dependent p-values which are not that restrictive, and we derive conditions for realistic powers of some existing test statistics. More specifically, the raw data or statistics X_i's for $1 \le i \le d$ are normally distributed with a correlation matrix $\Sigma = (\rho_{ij})_{1 \le i,j \le d}$ which is an element in a collection of correlation matrices, $\mathcal{F}_{d,\rho} = \{\Sigma : 0 \le \rho_{ij} \le \rho < 1 \ \text{ for } 1 \le i \ne j \le d\}$.

Although our dependence is implemented via the Gaussian copula, $\mathcal{F}_{d,\rho}$ covers various dependent structures including a correlation matrix in Liu and Xie (2020). In particular, as d increases, the correlation matrices in Liu and Xie (2020) are limited in that they represent weak dependence among all p-

values due to the condition $\lambda_{\max}(\Sigma) \leq C$ for some constant C where $\lambda_{\max}(\Sigma)$ is the maximum eigenvalue of Σ. For example, they exclude a case of equally correlated data. Specifically, since it is known that $1 + \sum_{i \neq j} \rho_{ij}/d \leq \lambda_{\max}(\Sigma)$, under a high dimensional setting or a case where the dimension increases, an assumption of bounded eigenvalues requires each correlation coefficient becomes very small or diminishes to zero order of reciprocal of dimension. In practice, especially for applications to genomic data, there are many cases that the assumption of bounded eigenvalues is not realistic due to dependence among SNPs in genes in high dimensionality. On the other hand, $\mathcal{F}_{d,\rho}$ covers highly correlated data including equally correlated case since $\mathcal{F}_{d,\rho}$ allows $\lambda_{\max}(\Sigma)$ to diverge as d increases. $\mathcal{F}_{d,\rho}$ includes correlation matrices since it allows a case where $\lambda_{\max}(\Sigma)$ diverges as d increases. Similarly, Liang and Rho (2022) used the weakly dependent structures such as serially correlated cases also included in $\mathcal{F}_{d,\rho}$. And in the real data, since true correlation coefficients are unknown, the condition in this paper can include the worst case scenario. From this point, we deal with a broader class of correlation matrices, $\mathcal{F}_{d,\rho}$.

We introduce some notations used in this paper. Denote $f(x) \sim g(x)$ as $x \to c$ if $\lim_{x \to c} f(x)/g(x) = 1$. The statement $a \lesssim b$ means that $a \leq \gamma \cdot b$, where $\gamma > 0$ is a fixed constant independent of dimension d. Let $\phi(x)$, $\Phi(x)$ and $\bar{\Phi}(x)$ be the density function, the cumulative distribution function and the survival function of standard normal distribution, respectively. For $f(x) > 0$ and $g(x) > 0$, we define $f(x) \ll g(x)$ if $\limsup_{x \to \infty} f(x)/g(x) = C$ for a constant $C \in (1, \infty)$. The notation $[x]$ means the integer part of a positive value x.

3.1. Type I error control

We present theorems showing that each of test statistics, T_{Stouffer}, T_{Fisher}, T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} controls Type I error under different scenarios such as ρ_{ij} or α is fixed while other parameters depends on d. We first present a theorem implying that for an exchangeable case, both T_{Stouffer} and T_{Fisher} fail to control a given Type I error when the correlation coefficient is fixed. This shows the weakness of T_{Stouffer} and T_{Fisher} which do not use the heavy-tailed distribution. On the other hand, Type I errors of T_{Stouffer} and T_{Fisher} are approximately a given level α if the correlation coefficient diminishes to zero under some conditions. In the following theorem, we define P_{H_0} as the probability measure under H_0.

Theorem 1.

1. *Suppose that* $0 < \rho = \mathsf{Corr}(p_i, p_j)$ *for* $1 \leq i \neq j \leq d$ *is fixed. Then for any* $\alpha \in (0, 1)$*, under* H_0 *in* (2.1)*, we have* $\lim_{d \to \infty} P_{H_0}(T_{Stouffer} \geq z_\alpha) = 1/2$. *However, if* $\rho = o(1/d)$*, then* $\lim_{d \to \infty} P_{H_0}(T_{Stouffer} \geq z_\alpha) = \alpha$.

2. *Suppose that* $0 < \xi = \mathsf{Corr}(-2\log p_i, -2\log p_j)$ *for* $1 \leq i \neq j \leq d$ *is fixed. Then for any* $\alpha \in (0, 1)$*, under* H_0 *in* (2.1)*, we have*

$\lim_{d\to\infty} P_{H_0}\left(T_{Fisher} \geq \chi^2_\alpha(2d)\right) = 1/e$. *However, if* $\xi = o(1/d)$ *and* $d\sqrt{\xi} \to c \in [0, +\infty]$, *then* $\lim_{d\to\infty} P_{H_0}\left(T_{Fisher} \geq \chi^2_\alpha(2d)\right) = \alpha$.

We note that T_{Fisher} can be expressed as a special combining method of generalized averaging of p-values. Let $p_1, \ldots, p_d$ be any p-values following the uniform distribution marginally and $M_r = d^{-1/r}(p_1^r + \cdots p_d^r)^{1/r}$ be a combining function. From Vovk and Wang (2020), the combining function is precise if for each $\epsilon \in (0, 1)$, $\sup P(M_r \leq \epsilon) = \epsilon$, where the supremum encompasses arbitrary dependence structure between $p_1, \ldots, p_d$. Vovk and Wang (2020) showed that, if a constant a_d depending on d and $a_d \to e$ as $d \to \infty$, the supremum of adjusted statistic $a_d M_0$ for $M_0 := (\prod_{i=1,\ldots,d} p_i)^{1/d}$ achieves precise probability, that is, $\sup P(a_d M_0 \leq \epsilon) = \epsilon$. Therefore, T_{Fisher} also can be considered to be precise. However, in the perspective of hypothesis testing procedure for Fisher's method, the adjusted method of Vovk and Wang (2020) becomes too conservative. Indeed, from the precise probability, for a given significance level $\alpha \in (0, 1)$, $\alpha = \sup P(a_d M_0 \leq \alpha) = \sup P\{T_{\text{Fisher}} \geq 2d \log(a_d/\alpha)\}$. However, for large d, since $\log(a_d/\alpha) > 1$, the rejection region, $[2d \log(a_d/\alpha), +\infty)$ is narrower than that of Fisher's method. Hence, the adjusted statistic becomes conservative which is a cost for validity under arbitrary dependence structure. Numerical studies in Chen et al. (2023) showed that, under arbitrary dependence structure, the powers of methods based on the generalized averaging are lower than those of the Bonferroni method, so that the adjusted statistics are hard to be used in practice. Therefore, in this paper, we adapt the quantile constructed under independence assumption that is called VI (Valid for Independent) method for a practical balance between power and size. In the following theorem, we define $P^\Sigma(\cdot)$ as the probability under a given correlation matrix $\Sigma \in \mathcal{F}_{d,\rho}$.

Theorem 2.

1. *For any* $\Sigma \in \mathcal{F}_{d,\rho}$ *and any* $\alpha \in (0, 1)$, *under* H_0 *in* (2.1), *we have*

$$\inf_{\Sigma\in\mathcal{F}_{d,\rho}} P^\Sigma_{H_0}(T_{Stouffer} \geq z_\alpha) = P^I_{H_0}(T_{Stouffer} \geq z_\alpha) = \alpha. \tag{3.2}$$

Furthermore, for any given $\alpha \in (0, 1)$ *and if* $\sum_{i\neq j} \rho_{ij} = o(d)$, *under* H_0 *in* (2.1), *we have*

$$\lim_{d\to\infty} P^\Sigma_{H_0}(T_{Stouffer} \geq z_\alpha) = \alpha. \tag{3.3}$$

2. *For any* $\Sigma \in \mathcal{F}_{d,\rho}$ *and any* $\alpha \in (0, 1)$, *under* H_0 *in* (2.1), *we have*

$$\inf_{\Sigma\in\mathcal{F}_{d,\rho}} P^\Sigma_{H_0}\{T_{Fisher} \geq \chi^2_\alpha(2d)\} = P^I_{H_0}\{T_{Fisher} \geq \chi^2_\alpha(2d)\} = \alpha. \tag{3.4}$$

Furthermore, let $\xi_{ij} = \mathsf{Corr}(-2\log p_i, -2\log p_j)$ *for* $i \neq j = 1, \ldots, d$. *If* $\sum_{i\neq j} \xi_{ij} = o(1/d)$, *then*

$$\lim_{d\to\infty} P_{H_0}^{\Sigma}\{T_{Fisher} \geq \chi_{\alpha}^2(2d)\} = \alpha. \tag{3.5}$$

3. *For any* $\Sigma \in \mathcal{F}_{d,\rho}$ *and any* $\alpha \in (0,1)$*, under* H_0 *in* (2.1)*, we have*

$$\sup_{\Sigma\in\mathcal{F}_{d,\rho}} P_{H_0}^{\Sigma}(T_{minP} \geq x_{\alpha}) = P_{H_0}^{I}(T_{minP} \geq x_{\alpha}) = \alpha \tag{3.6}$$

where $x_{\alpha} = \Phi^{-1}\{(1-\alpha)^{1/d}\}$*. Furthermore, for any* Σ *with* $\rho_{ij} \geq \epsilon > 0$*, under* H_0 *in* (2.1)*, we have*

$$\lim_{d\to\infty} P_{H_0}^{\Sigma}(T_{minP} \geq x_{\alpha}) = 0. \tag{3.7}$$

Note that (3.2), (3.4) and (3.6) are non-asymptotic results while (3.3), (3.5) and (3.7) are asymptotic results for varying correlation coefficients depending on d. Theorems 1 and 2 show the weakness of T_{Stouffer}, T_{Fisher} and T_{minP} in the sense that T_{Stouffer} and T_{Fisher} may fail in controlling a given Type I error and T_{minP} is too conservative when there exist some dependence among p-values. Liu and Xie (2020) and Wilson (2019) presented numerical studies showing that Stouffer's test and Fisher's test fail in controlling Type I error when p-values are correlated. However, the results of Theorems 1 and 2 have the novelty in that they provide theoretical reasons for the failures of Stouffer's and Fisher's tests in controlling Type I error. These results are related to how many components in T_{Stouffer}, T_{Fisher} and T_{minP} contribute to explain the whole variation of those three tests, respectively. In other words, T_{Stouffer} and T_{Fisher} are aggregating all transformed p-values while T_{minP} relies on only one term, maximum of X_is. We leave the following remarks as implications of Theorem 2.

Remark 1.

1. For any given Σ, T_{Stouffer} and T_{Fisher} control Type I error non-asymptotically only when all p-values are independent, i.e., $\Sigma = I$. Otherwise, T_{Stouffer} and T_{Fisher} cannot control Type I error for dependent p-values. However, when the number of p-values increases, T_{Stouffer} and T_{Fisher} may control Type I error asymptotically when the effect of overall correlations is not large enough such as $\sum_{i\neq j}\rho_{ij} = o(d)$ and $\sum_{i\neq j}\xi_{ij} = o(1/d)$, respectively.

2. For exchangeable p-values under H_0, if, for all $1 \leq i \neq j \leq d$, $\mathsf{Corr}(p_i, p_j) = \rho$ is positive and fixed, $P_{H_0}(T_{\text{Stouffer}} > z_{\alpha}) \to 1/2$ under the null, as $d \to \infty$, whereas the Type I error of T_{Fisher} converges to e^{-1}, if $0 < \xi = \mathsf{Corr}(-2\log p_i, -2\log p_j)$ is fixed. Since $1/e \approx 0.37$, Type I error of T_{Stouffer} is larger than that of T_{Fisher} so that T_{Stouffer} obtains more seriously inflated Type I error than T_{Fisher}. These results can be confirmed by Figure 1 in Section 5.

3. $T_{\text{minP}} = \max_i X_i$ or $\min_i p_i$ can control Type I error for any $\Sigma \in \mathcal{F}_{d,\rho}$ and exact α is obtained only when all p-values are independent. In particular, when there exist serious correlations such that $\rho_{ij} \geq \epsilon > 0$ for some ϵ, T_{minP} is fairly conservative in the sense that Type I error converges to 0 as d increases.

Compared to the weaknesses of T_{Stouffer}, T_{Fisher} and T_{minP} which are either too liberal or conservative for dependent p-values, the motivation of T_{Cauchy} and $T_{\text{Lévy}}$ is developing test statistics which are intermediate positions between T_{Stouffer} (or T_{Fisher}) and T_{minP} in the sense that some small subset of transformed p-values can explain most of variations.

Before we present the asymptotic properties of T_{Cauchy} and $T_{\text{Lévy}}$, we provide the following lemma. Roughly speaking, Lemma 1 shows that the lower bound of $\max_{1\leq i\leq d} X_i$ is decreasing in ρ which leads the corresponding p-value $\min_{1\leq i\leq d} p_i$ to be larger since $\min_{1\leq i\leq d} p_i$ is derived under the critical value with $\rho = 0$. In fact, this result is also related to the third result in Theorem 2 since Type I error of $\max_{1\leq i\leq d} X_i$ converges to 0 when $\rho_{ij} \geq \epsilon > 0$ for some ϵ.

Lemma 1. *Let* $\mathbf{X} \sim N_d(\mathbf{0}, \Sigma)$ *for* $\Sigma \in \mathcal{F}_{d,\rho}$, *then* $\max_{1\leq i\leq d} X_i \geq \sqrt{1-\rho}\sqrt{2\log d} + o_p(1)$.

Now we present asymptotic properties of T_{Cauchy} and $T_{\text{Lévy}}$ under varying quantities such as the nominal level of Type I error α and the dimension d. Our asymptotic results of T_{Cauchy} and $T_{\text{Lévy}}$ in the following theorem are different from Liu and Xie (2020) and Liang and Rho (2022) in that they dealt with the case of weakly dependent cases, such as covariance matrix with bounded eigenvalues and serially correlated cases while we discuss the cases of $\Sigma \in \mathcal{F}_{d,\rho}$ which also includes exchangeable matrix as an example of strong dependence. Furthermore, the following theorem shows an answer to the robustness of T_{Cauchy} and $T_{\text{Lévy}}$ for $\Sigma \in \mathcal{F}_{d,\rho}$ as d diverges while Fang, Tseng and Chang (2023) considered the case of given Σ with fixed d.

Theorem 3. *Suppose* $X \sim N_d(0, \Sigma)$ *where* $\Sigma \in \mathcal{F}_{d,\rho}$. *Let* c_α *and* l_α *be upper* α *quantiles of standard Cauchy and Lévy distributions, respectively.*

1. *For* $0 < a_C < (1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2$ *and* $d = [(c_\alpha)^{a_C}]$, *we have,*

$$\lim_{\alpha\to 0} \sup_{\Sigma\in\mathcal{F}_{d,\rho}} \frac{P^{\Sigma}_{H_0}(T_{Cauchy} \geq c_\alpha)}{P^{I}_{H_0}(T_{Cauchy} \geq c_\alpha)} = \lim_{\alpha\to 0} \sup_{\Sigma\in\mathcal{F}_{d,\rho}} \frac{P^{\Sigma}_{H_0}(T_{Cauchy} \geq c_\alpha)}{\alpha} = 1.$$

2. *For* $0 < a_L < (1-\rho^2)/(5\rho^2 + 4\sqrt{3}\rho + 2)$ *and* $d = [(\ell_\alpha)^{a_L}]$, *we have,*

$$\lim_{\alpha\to 0} \sup_{\Sigma\in\mathcal{F}_{d,\rho}} \frac{P^{\Sigma}_{H_0}(T_{Lévy} \geq l_\alpha)}{P^{I}_{H_0}(T_{Lévy} \geq l_\alpha)} = \lim_{\alpha\to 0} \sup_{\Sigma\in\mathcal{F}_{d,\rho}} \frac{P^{\Sigma}_{H_0}(T_{Lévy} \geq l_\alpha)}{\alpha} = 1.$$

3.2. Comparison of T_{Cauchy} and $T_{\text{Lévy}}$ in controlling Type I error

Theorem 3 shows that T_{Cauchy} and $T_{\text{Lévy}}$ are robust to dependence under given dimension $d_C = [(c_{\alpha_C})^{a_C}]$ and $d_L = [(\ell_{\alpha_L})^{a_L}]$ for some constraints on a_C and a_L. Such constraints on d can be considered as the scope of the robustness to the accumulation of dependence among p-values which is due to increasing d. From the relation between d and α, the robustness in controlling Type I error under the dependence consists of conditions on d and α. If $d \to \infty$, to handle the accumulation of dependence, the robustness of T_{Cauchy} and $T_{\text{Lévy}}$ in controlling Type I error is attained when $\alpha \to 0$. From this motivation, we compare T_{Cauchy} and $T_{\text{Lévy}}$ from the following viewpoints: for Type I error and dimension (α_C, d_C) and (α_L, d_L) corresponding to T_{Cauchy} and $T_{\text{Lévy}}$, (i) under $\alpha_C = \alpha_L \to 0$, we compare divergence rates of d_C and d_L which are required to control Type I error, and (ii) under $d_C = d_L \to \infty$, we compare the rates of α_C and α_L converging to 0 which are required to control Type I error. Based on these two criteria, we now show that $T_{\text{Lévy}}$ is more robust than T_{Cauchy} in controlling Type I error for $\Sigma \in \mathcal{F}_{d,\rho}$. We first present the following lemma which is used in comparison of T_{Cauchy} and $T_{\text{Lévy}}$.

Lemma 2. *For $\alpha \in (0,1)$, we have $2\pi c_\alpha^2 \sim \ell_\alpha$ as $\alpha \to 0$.*

In Theorem 3, the dimension d is expressed based on the a_C and a_L such that the polynomial orders of the upper α quantiles are $d_C = [(c_\alpha)^{a_C}]$ and $d_L = [(l_\alpha)^{a_L}]$. In the proof of Lemma 2, approximations of tail probabilities of Cauchy and Lévy distributions are obtained by $\alpha \sim 1/\pi c_\alpha$ and $\alpha \sim \sqrt{2/\pi \ell_\alpha}$. Then, three parameters (d, α, a) for T_{Cauchy} and $T_{\text{Lévy}}$ have the following relationship:

$$\alpha_L \sim \sqrt{\frac{2}{\pi}} \left(\frac{1}{d_L}\right)^{1/2a_L}, \quad \alpha_C \sim \frac{1}{\pi}\left(\frac{1}{d_C}\right)^{1/a_C} \tag{3.8}$$

where (d_C, α_C, a_C) and (d_L, α_L, a_L) are three parameters corresponding to T_{Cauchy} and $T_{\text{Lévy}}$.

Theorem 3 shows that T_{Cauchy} and $T_{\text{Lévy}}$ control Type I error α asymptotically under $0 < a_C < (1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2$ and $0 < a_L < (1-\rho^2)/(5\rho^2 + 4\sqrt{3}\rho + 2)$, respectively. It is obvious that $\{(2\alpha_L, d_L)\}$ satisfying (3.8) includes $\{(\alpha_C, d_C)\}$ since $0 < 2a_L < 2(1-\rho^2)/(5\rho^2 + 4\sqrt{3}\rho + 2)$ includes $0 < a_C < (1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2$ due to $(1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2 < 2(1-\rho^2)/(5\rho^2 + 4\sqrt{3}\rho + 2)$. Figure S3 in the Supplementary Material shows the upper bounds of $2a_L$ and a_C for the values of $0 \le \rho < 1$, and it is shown that the upper bound of $2a_L$ is larger than that of a_C.

Based on these, we define ranges of a_C and $2a_L$ which are

$$\mathcal{I}_C = \left\{ a_C : 0 < a_C < \frac{1-\rho^2}{(\sqrt{3}\rho + \sqrt{2})^2} \right\},$$

$$\mathcal{I}_L = \left\{ 2a_L : 0 < 2a_L < \frac{2\,(1-\rho^2)}{5\rho^2 + 4\sqrt{3}\rho + 2} \right\}$$

and regions of (α, d) for T_{Cauchy} and $T_{\text{Lévy}}$

$$\mathcal{G}_C = \left\{ (\alpha_C, d_C) : \alpha_C \sim \frac{1}{\pi}\left(\frac{1}{d_C}\right)^{1/a_C}, a_C \in \mathcal{I}_C \right\},$$
$$\mathcal{G}_L = \left\{ (\alpha_L, d_L) : \alpha_L \sim \sqrt{\frac{2}{\pi}}\left(\frac{1}{d_L}\right)^{1/2a_L}, a_L \in \mathcal{I}_L \right\}$$

where $\mathcal{G}_C$ is the collection of sequences (α, d) satisfying $\alpha \sim c \cdot (1/d)^{1/a}$ for $a \in \mathcal{I}_C$ where c is a constant and $\mathcal{G}_L$ is similarly interpreted.

We note that, to control Type I error, Liu and Xie (2020) presents the condition of dimension d in T_{Cauchy} corresponding to $0 < a_C < 1/2$ which is wider than $0 < a_C < (1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2$ in Theorem 3. Indeed, $0 < a_C < (1-\rho^2)/(\sqrt{3}\rho + \sqrt{2})^2$ equals to $0 < a_C < 1/2$ when $\rho = 0$. Hence, the conditions in Theorem 3 can be considered as a cost to encompass arbitrary structures of correlations matrix.

$\mathcal{G}_C$ and $\mathcal{G}_L$ provide the scopes of α and d under which T_{Cauchy} and $T_{\text{Lévy}}$ control Type I error for dependent p-values. We have the following corollary derived from Theorem 3 which states some advantage of $T_{\text{Lévy}}$ over T_{Cauchy} in controlling Type I errors.

Corollary 1. *Under the conditions in Theorem* 3*, we have the following results for any* $0 < \rho < 1$*:*

(i) $\mathcal{G}_C \subset \mathcal{G}_L$.

(ii) *For* $a_L \in \mathcal{I}_L - \mathcal{I}_C$*, if* $\alpha_L \sim \alpha_C$*, we have* $d_C/d_L \to 0$ *for any* $a_C \in \mathcal{I}_C$ *and the corresponding* d_C*. If* $d_L \sim d_C$*, we have* $\alpha_C/\alpha_L \to 0$ *for any* $a_C \in \mathcal{I}_C$ *and the corresponding* α_C*.*

The first result in Corollary 1 shows that $T_{\text{Lévy}}$ controls Type I error α in a wider class of (α, d) than T_{Cauchy}. The second result shows the behavior of α and d of $T_{\text{Lévy}}$ in $\mathcal{G}_L$ but outside of $\mathcal{G}_C$. From these results, $T_{\text{Lévy}}$ can control Type I error asymptotically in additional region of (α, d) such as larger values of dimension d and α compared to those of T_{Cauchy}.

3.3. Power studies

To find sufficient conditions under which combination test statistics have asymptotically power converging to 1, we consider Type II error of each test which is $1 - \textit{power}$. If Type II error of a test goes to zero asymptotically, it can be concluded that the test is asymptotically powerful. Denote $\mathcal{N}$ and $\mathcal{S}$ as index sets of null and signal part under non-null hypothesis, respectively, i.e., for $i \in \mathcal{N}$,

$X_i \overset{d}{=} Z_i$ and for $i \in \mathcal{S}$, $X_i = Z_i + \mu_i$ where $Z_i \sim N(0,1)$ and $\mu_i \equiv \mu = \sqrt{2\tau \log d}$ for $\tau > 0$. Assume that $|\mathcal{S}| = d^\beta$, $0 < \beta < 1$. Donoho and Jin (2008) called the case of $0 < \beta < 1/2$ a strongly sparse case and the case of $1/2 < \beta < 1$ a moderately sparse case.

The following theorem states that neither of T_{Cauchy} nor $T_{\text{Lévy}}$ dominates the other in terms of controlling Type I error and obtaining asymptotic power 1.

Theorem 4. *Assume that* $|\mathcal{S}| = d^\beta$ *for* $0 < \beta < 1$.

(i) *If* $1/(\beta + \tau - 1) < a_C$, *Type II error of* T_{Cauchy} *converges to 0, as* $d \to \infty$.

(ii) *If* $1/(\beta + 2\tau - 2) < a_L$, *Type II error of* $T_{Lévy}$ *converges to 0, as* $d \to \infty$.

From Theorems 3 and 4, we have sufficient conditions to achieve two criteria in (3.1). The following corollary presents conditions for T_{Cauchy} and $T_{\text{Lévy}}$ to achieve (3.1) as $\alpha \to 0$ and $d \to \infty$.

Corollary 2. *Under the settings of Theorem* 3 *and* 4*:*

(i) T_{Cauchy} *has Type I error + Type II error* $\to 0$ *under* $d^{1-\beta-\tau} \ll \alpha \ll d^{-(\sqrt{3}\rho+\sqrt{2})^2/(1-\rho^2)}$ *if* $\beta + \tau > (\sqrt{2}\rho + \sqrt{3})^2/(1-\rho^2)$.

(ii) $T_{Lévy}$ *has Type I error + Type II error* $\to 0$ *under* $d^{1-\beta/2-\tau} \ll \alpha \ll d^{-(5\rho^2+4\sqrt{3}\rho+2)/2(1-\rho^2)}$ *if* $\beta + 2\tau > (\sqrt{3}\rho + 2)^2/(1-\rho^2)$.

Corollary 2 presents the sufficient asymptotic conditions of α to attain the robustness and obtain the power for T_{Cauchy} and $T_{\text{Lévy}}$. For a given ρ, the upper bounds of α implies that α of T_{Cauchy} converges to 0 faster than $T_{\text{Lévy}}$ to attain the robustness. On the other hand, for given moderately large values of β and τ, the lower bounds of α implies that $T_{\text{Lévy}}$ requires d to increase faster than T_{Cauchy} to obtain the power. Therefore, the relationship between α and d with β, τ and ρ shows the trade-off between robustness and power for T_{Cauchy} and $T_{\text{Lévy}}$.

4. Asymptotic Equivalence between Heavy-Tailed Stable Distribution and Harmonic Mean in Combining *p*-Values

In this section, we discuss that hypothesis testing procedures based on p-value combining methods using heavy-tailed stable distribution are asymptotically equivalent to those using a type of harmonic mean. Wilson (2019) showed that the harmonic mean of p-values can be considered as a global p-value of hypothesis testing since it approximately tends to the Landau distribution which is a specific form of heavy-tailed stable distribution. Fang, Tseng and Chang (2023) and Chen et al. (2023) showed that Cauchy combination is asymptotically equivalent to the harmonic mean. Also, Wilson (2021) discussed the equivalence of Lévy combination and squared type of harmonic mean.

We investigate more general cases than those in Fang, Tseng and Chang (2023) and Wilson (2021). Fang, Tseng and Chang (2023) used the Pareto distribution which has different tail behaviors to represent heavy-tailed distribution for different parameters such as $P(X > t) = C(\eta)/t^{\eta}$ for some constant $C(\eta)$. In particular, the Pareto distributions with $\eta = 1$ and $\eta = 1/2$ have the same asymptotic tail behavior as Cauchy and Lévy distributions, respectively except some constants. Indeed, the Cauchy and Lévy distributions are special cases of a family of stable distribution, referred to as the Lévy alpha-stable distribution of which the tail probability is given by $P(X > t) \sim C^*(\eta, \psi, \gamma)/t^{\eta}$, as $t \to \infty$ where C^* is a constant depending on parameters η, ψ and γ. Here, η and ψ are parameters reflecting the heaviness of tail and the skewness, respectively, and γ is a scale parameter. With these parameters, the Pareto and the Lévy alpha-stable distributions reflect the properties of the heaviness of the tail probability.

Although Fang, Tseng and Chang (2023) and Wilson (2021) analyzed tail behaviors of heavy-tailed distributions and their properties, in this section, we investigate the connection between combining method based on transforming into the Lévy alpha-stable distribution and the harmonic mean type method in detail. Concretely, we highlight the result of the equivalence of combining method of heavy-tailed transformation and generalized harmonic mean method when dimension d goes to infinity while Fang, Tseng and Chang (2023), Wilson (2021) and Chen et al. (2023) used the assumption of fixed d.

First, we define the generalized harmonic mean of p-values by $p_{global}^{gHMP(\eta)}$: for $p_1, \ldots, p_d$ and $\eta > 0$, $p_{global}^{gHMP(\eta)} = d/(\sum_{j=1}^{d} p_i^{-1/\eta})^{\eta}$. The generalized harmonic mean includes typical cases, for example, when $\eta = 1$, $p_{global}^{gHMP(1)}$ is a harmonic mean of p-values. Note that the generalized harmonic mean is also used in Vovk and Wang (2020), Chen et al. (2023) and Wilson (2021) with the notation $M_{r,K}$.

Using the theory of regularly varying functions as in Wilson (2019), we can derive the following result: for $\eta > 0$ and weights $w_i > 0$, $i = 1, \ldots, d$, as $\epsilon \to 0$

$$P\left\{\left(\sum_{j=1}^{d} w_i p_i^{-1/\eta}\right)^{\eta} \geq \frac{1}{\epsilon}\right\} \sim \left(\sum_{j=1}^{d} w_i\right) P\left(p_i^{-1/\eta} \geq \epsilon^{-1/\eta}\right) = \epsilon,$$

where p_i's are allowed to be arbitrary dependent and d is fixed. That is, $p_{global}^{gHMP(\eta)}$ can be considered as a global p-value approximately.

In the perspective of testing procedure, we define a test statistic $T_{gHMP(\eta)}$ based on the generalized harmonic mean of p-values as a global p-value. Let $\alpha > 0$ and $\eta > 0$ be given. Define $T_{gHMP(\eta)} = I(p_{global}^{gHMP(\eta)} < \alpha)$, where $I(\cdot)$ is an indicator function. If $T_{gHMP(\eta)} = 1$, the null hypothesis is rejected and if $T_{gHMP(\eta)} = 0$, the null hypothesis cannot be rejected.

The following Theorems 5 and 6 present the robustness of Type I error and asymptotic power for $T_{gHMP(\eta)}$, respectively. In the theorem below, let $\mathsf{E}_{H_0}^{\Sigma}$ be the expectation under H_0 with a given correlation matrix $\Sigma \in \mathcal{F}_{d,\rho}$.

Theorem 5. *Let* $X \sim N_d(0, \Sigma)$ *where* $\Sigma \in \mathcal{F}_{d,\rho}$. *Let* $d = [\alpha^{-a_G}]$. *Then for* $0 < a_G < (1-\rho^2)/\{(\eta+2)\rho^2 + 2\eta + 2\rho\sqrt{(\eta+1)(2\eta+1)}\}$, *under* H_0, *we have* $\lim_{d\to\infty} \sup_{\Sigma\in\mathcal{F}_{d,\rho}} \mathsf{E}^{\Sigma}_{H_0} T_{gHMP(\eta)}/\alpha = 1$, *as* $\alpha \to 0$.

Theorem 6. *Assume that* $|\mathcal{S}| = d^{\beta}$ *for* $0 < \beta < 1$. *If* $1/(\eta\beta + \tau - 1) < a_G$, *Type II error of* $T_{gHMP(\eta)}$ *converges to 0 as* $d \to \infty$.

From Theorems 5 and 6, the following corollary presents the condition of α and d under which both Type I and Type II errors of $T_{gHMP(\eta)}$ converge to 0.

Corollary 3. *Under settings of Theorems* 5 *and* 6, $T_{gHMP(\eta)}$ *has Type I error* + *Type II error* $\to 0$ *under* $d^{1-\eta\beta-\tau} \ll \alpha \ll d^{-\{(\eta+2)\rho^2+2\eta+2\rho\sqrt{(\eta+1)(2\eta+1)}\}/(1-\rho^2)}$, *if* $\tau + \eta\beta > \{(\eta+1)\rho^2 + 2\eta + 1 + 2\rho\sqrt{(\eta+1)(2\eta+1)}\}/(1-\rho^2)$.

Similar to $p^{gHMP(\eta)}_{global}$, we also generalize the transformation of p-values to heavy-tailed stable distribution. For this, let $h_{\eta,\psi}(p_i)$ be a transformation function of p-value to a random variable of heavy-tailed stable distribution with $0 < \eta \leq 1$ and $-1 \leq \psi \leq 1$. $h_{\eta,\psi}(p_i)$ also can be considered as a form of inverse cumulative distribution function of heavy-tailed stable distribution. If $\eta = 1$ and $\psi = 0$, then $h_{1,0}$ is a Cauchy transformation and if $\eta = 1/2$ and $\psi = 1$, then $h_{1/2,1}$ is a Lévy transformation. We denote a global p-value of transformation of stable distribution by $g^{S(\eta,\psi)}_{global} = P_{H_0}\{T_{\eta,\psi} > \sum_{i=1}^{d} d^{-1/\eta} h_{\eta,\psi}(p_i)\}$, where $T_{\eta,\psi}$ is a combination test statistic of heavy-tailed transformation. It follows from Nolan (2020) that, for observed p-values, $p_1, \ldots, p_d$,

$$p^{S(\eta,\psi)}_{global} = P_{H_0}\left(T_{\eta,\psi} > \sum_{i=1}^{d} d^{-1/\eta} h_{\eta,\psi}(p_i)\right) \sim d \cdot c_{\eta,\psi} \cdot \left(\sum_{i=1}^{d} h_{\eta,\psi}(p_i)\right)^{-\eta},$$

where $c_{\eta,\psi}$ denotes a generic constant depending on η and ψ.

We now present the following Theorem 7 which investigates the asymptotic equivalence between $p^{gHMP(\eta)}_{global}$ and $p^{S(\eta,\psi)}_{global}$ under some conditions when $d \to \infty$.

Theorem 7. *For* $\Sigma \in \mathcal{F}_{d,\rho}$, $0 < \eta \leq 1$ *and* $-1 \leq \psi \leq 1$, *if one of the following conditions:* (i) $\max_j \sum_{i=1}^{d} \rho_{ij}^2 \leq C_0$ *for some constant* $C_0 > 0$, *and* (ii) *for* $0 < \eta < 1$, $0 < \rho < 1 - \eta$ *and for* $\eta = 1$, $0 \leq \rho_{ij} \leq \rho = o(\log\log d/\log d)$ *is satisfied, then we have* $p^{S(\eta,\psi)}_{global}/p^{gHMP(\eta)}_{global} \xrightarrow{p} 1$ *as* $d \to \infty$.

We note that Fang, Tseng and Chang (2023) proved the equivalence between transformation functions of Cauchy combination and harmonic mean for one p-value. On the other hand, Vovk and Wang (2020) showed the asymptotic equivalence of Cauchy combination and harmonic mean when the minimum of p-values goes to zero for fixed dimension. However, Theorem 7 presents the equivalence of combining method of heavy-tailed transformation and generalized harmonic mean method when $d \to \infty$ which is more practical assumption than that of Fang, Tseng and Chang (2023) and Vovk and Wang (2020).

Remark 2.

1. We request some conditions on Σ in Theorem 7, such as (i) $\max_j \sum_{i=1}^{d} \rho_{ij}^2 \leq C_0$ for some constant $C_0 > 0$, or (ii) $\rho = o(\log\log d/\log d)$. Neither (i) nor (ii) implies the other. The first one allows sparse ρ_{ij}^2 such that a small number of ρ_{ij}^2 can have values not diminishing to 0 and all others are fairly negligible. This does not satisfy the condition (ii) since there exist some ρ_{ij} which does not diminish to zero. On the other hand, (ii) can include the case of dense such that most of ρ_{ij} diminishing to zero, but decreasing order is faster than $\log\log d/\log d$. This does not satisfy (i) since $\sum_{j=1}^{d} \rho_{ij}^2 = o(d\log\log d/\log d)$ may diverge.

2. The asymptotic equivalence between $p_{global}^{S(\eta,\psi)}$ and $p_{global}^{gHMP(\eta)}$ shown in Theorem 7 implies that the condition of robustness of two methods are also equivalent. In Theorems 5 and 6, when $\eta = 1/2$, the upper and lower bounds of a_G can be expressed by 2 times of the upper and lower bounds of a_L in Theorems 3 and 4, respectively. On the other hand, when $\eta = 1$, the upper and lower bounds of a_G equal to the upper and lower bounds of a_C in Theorems 3 and 4, respectively.

5. Numerical Studies

In this section, we evaluate Type I error and power for various dependency structures to compare p-values combining methods according to heaviness of transformation functions and simulate the relationships between p-values combining methods. To investigate properties of p-values combining methods, we construct p-values from the multivariate normal distribution with mean $\boldsymbol{\mu}$ and correlation matrix Σ in $\mathbf{X} \sim N_d(\mu, \Sigma)$ and then derive one-sided right-tailed p-values. The correlation matrix $\Sigma = (\rho_{ij})_{i,j=1,\ldots,d}$ is defined by as follows:

1. Exchangeable: For $0 < \rho < 1$, $\rho_{ij} = \rho$ if $i \neq j$, and 1 if $i = j$.

2. Polynomial decay: For $r > 0$, $\rho_{ij} = 1/(1 + |i - j|^r)$ for $i \neq j$ and 1 if $i = j$.

3. Exponential decay, AR(1): For $0 < \rho < 1$, $\rho_{ij} = \rho^{|i-j|}$.

5.1. Asymptotic equivalence between heavy-tailed combination test and generalized harmonic mean method

In this subsection, we present simulation studies for three types of correlation matrices which demonstrate the similarity of $p_{\text{global}}^{S(\eta,\psi)}$ and $p_{\text{global}}^{gHMP(\eta)}$ for finite d. We consider $(\eta, \psi) = (1, 0)$ and $(1/2, 1)$ in $p_{\text{global}}^{S(\eta,\psi)}$ which are corresponding to the Cauchy method and Lévy method, respectively.

Figures S4 and S5 in the Supplementary Material provide plots of $p_{\text{global}}^{S(1,0)}$ *vs.* $p_{\text{global}}^{gHMP(1)}$ and $p_{\text{global}}^{S(1/2,1)}$ *vs.* $p_{\text{global}}^{gHMP(1/2)}$. As mentioned, we observe that the

Cauchy method and the harmonic mean method are quite close to each other for the cases of polynomial decay with large coefficient (bottom panel) and exponentially decay (middle panel) since the scatter plots are almost located around the diagonal line. However, under exchangeable dependency structures indicating strong dependence, the top panels in Figure S4 show that these two methods tend to be different. All plots in Figure S5 show that the Lévy method $p_{\text{global}}^{S(1/2,1)}$ looks fairly close to $p_{\text{global}}^{gHMP(1/2)}$ for all three correlation matrices. From Figure S4 and S5, we see that $p_{\text{global}}^{S(1/2,1)}$ and $p_{\text{global}}^{gHMP(1/2)}$ are close to each other compared to $p_{\text{global}}^{S(1,0)}$ and $p_{\text{global}}^{gHMP(1)}$. This is due to the fact that $p_{\text{global}}^{S(1/2,1)}$ and $p_{\text{global}}^{gHMP(1/2)}$ are dominated by a smaller number of p-values which are affected less severely by correlation structures compared to $p_{\text{global}}^{S(1,0)}$ and $p_{\text{global}}^{gHMP(1)}$.

5.2. Type I error

We present Figure 1 showing Type I error of each method at significant level 0.05 under exchangeable cases where ρ varies from 0 to 0.9 and d is fixed at 200. As expected from Theorem 1, T_{Stouffer} and T_{Fisher} fail in controlling a given Type I error 0.5 except $\rho = 0$, the case of independent p-values. We also see that as ρ increases, Type I errors of T_{Stouffer} and T_{Fisher} converge to $1/2$ and $1/e$, respectively, shown in Theorem 1. In Figure 1, patterns of Type I errors of $T_{\text{Lévy}}$ and T_{minP} are similar, since two methods are influenced by a small number of dominating p-values. Additionally, as ρ increases, both $T_{\text{Lévy}}$ and T_{minP} are getting more conservative. On the other hand, T_{Cauchy} fails to control Type I error at the mid-level dependency structures while Type I error is well controlled for independent or extremely correlated p-values. Since T_{Stouffer} and T_{Fisher} tend to have inflated Type I errors, we focus on the rest of tests such as T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP}.

Figure 2 shows Type I errors of T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} under exchangeable, polynomially and exponentially decaying dependency structures at significance levels 0.05, respectively. As seen in Figure 1, Figure 2 shows almost similar patterns of Type I errors under three types of correlation matrices. In addition, to evaluate the effects of dimension and significance level, we present the Type I errors of each method with $d = 2000$ and 3000 in Figure S7, and at significance level 0.01 in Figure S6 in the Supplementary Material also showing similar patterns.

In order to compare the finite-sample performances of $p_{\text{global}}^{gHMP(\eta)}$ and $p_{\text{global}}^{S(\eta)}$, Figure S8 in the Supplementary Material represents Type I error of T_{Cauchy}, $T_{\text{gHMP}(1)}$, $T_{\text{Lévy}}$ and $T_{\text{gHMP}(1/2)}$ under the three types of correlation. As mentioned in Section 5.1 and as shown in Figures S4 and S5, $p_{\text{global}}^{S(1/2,1)}$ and $p_{\text{global}}^{gHMP(1/2)}$ are close to each other compared to $p_{\text{global}}^{S(1,0)}$ and $p_{\text{global}}^{gHMP(1)}$ in the finite samples.

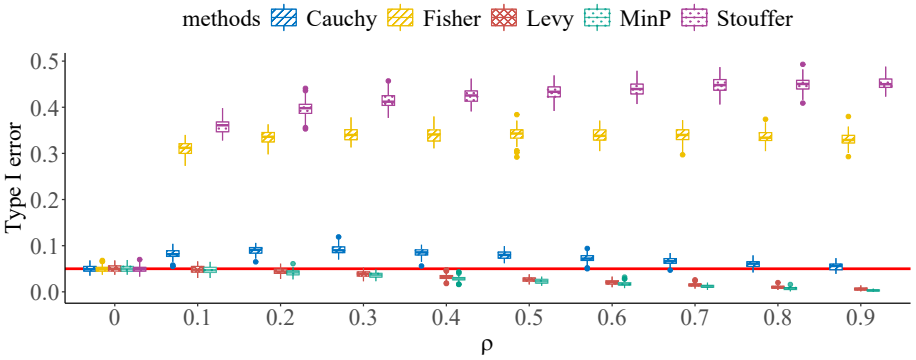


Figure 1. Box plots of Type I errors of T_{Stouffer}, T_{Fisher}, T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} under exchangeable case. Horizontal line indicates the significance level 0.05.

5.3. Power

We compare powers of T_{Stouffer}, T_{Fisher}, T_{Cauchy}, $T_{\text{gHMP}(1)}$, $T_{\text{Lévy}}$ and T_{minP} under the exchangeable case in Figure 3. We also provide Figures S9 and S10 in the Supplementary Material for polynomially and exponentially decaying dependency structures at significance level $\alpha = 0.05$. Simulation settings are similar to Liu and Xie (2020) under the model $\mathbf{X} \sim N_d(\mu, \Sigma)$, where $\mu = (\mu_i)_{1 \le i \le d}$ and Σ is a correlation matrix. All signals for non-null are defined to have the same strength and to account for effects of sparsity of signals such that $\mu_i := \sqrt{2\log(d)}/s^{1/3}$ for $i \in \mathcal{S}$, where $\mathcal{S}$ denotes the index set of signals and s is the cardinality of $\mathcal{S}$. The range of d is 20 to 500.

In Figure 3, powers of T_{Stouffer} and T_{Fisher} are unreliable since Type I errors are inflated under $\rho > 0$ shown in Figure 1. On the other hand, powers of $T_{\text{Lévy}}$ and T_{minP} are reliable regardless of $\rho > 0$. Powers of T_{Cauchy} depend on ρ since its Type I errors are inflated except the cases where correlations are close to 0 or 1. Figures S9 and S10 show similar patterns to Figure 3.

Compared to T_{Cauchy} and $T_{\text{Lévy}}$, T_{minP} has the lowest power which is related to the conservativeness represented in numerical studies of Type I errors. There are additional patterns in the powers of each method. For example, in exchangeable case in Figure 3, T_{Cauchy} has more power than $T_{\text{Lévy}}$ and T_{minP}. In addition, for more weak dependency structures such that correlations are polynomially and exponentially decaying , $T_{\text{Lévy}}$ achieves comparable power compared to T_{Cauchy} (Figs. S9 and S10). As signals become dense such that the cardinality of $\mathcal{S}$ increases, the power of $T_{\text{Lévy}}$ is getting lower than that of T_{Cauchy}. This is from the fact that $T_{\text{Lévy}}$ is dominated by a smaller number of components than T_{Cauchy}. This indicates the trade-off relationship between Type I error and power discussed in Section 3 that $T_{\text{Lévy}}$ has advantage in controlling Type I error under dependent p-values at the cost of losing powers under dense signals.

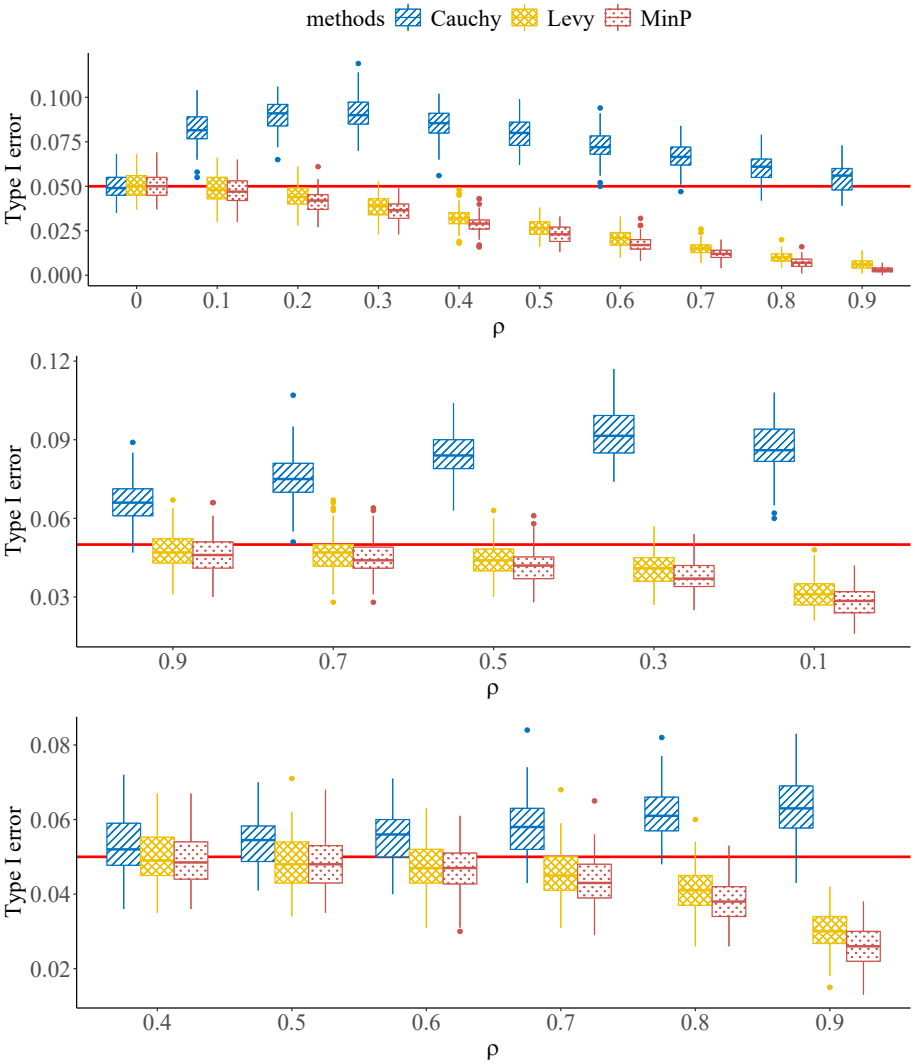


Figure 2. Box plots of Type I error of T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} at 0.05 with exchangeable, polynomially and exponentially decaying dependency structures.

We also compare size-adjusted powers of each method in finite samples under exchangeable, polynomially and exponentially decaying cases in Figures S11, S12 and S13, respectively, in the Supplementary Material.

6. Real Data Example

For a real-data analysis, we apply the combination methods to the Crohn's disease GWAS (Duerr, Taylor and Brant, 2006) which is also used in Liu and Xie (2020). For this data, the goal is identifying genes that are associated with the disease.

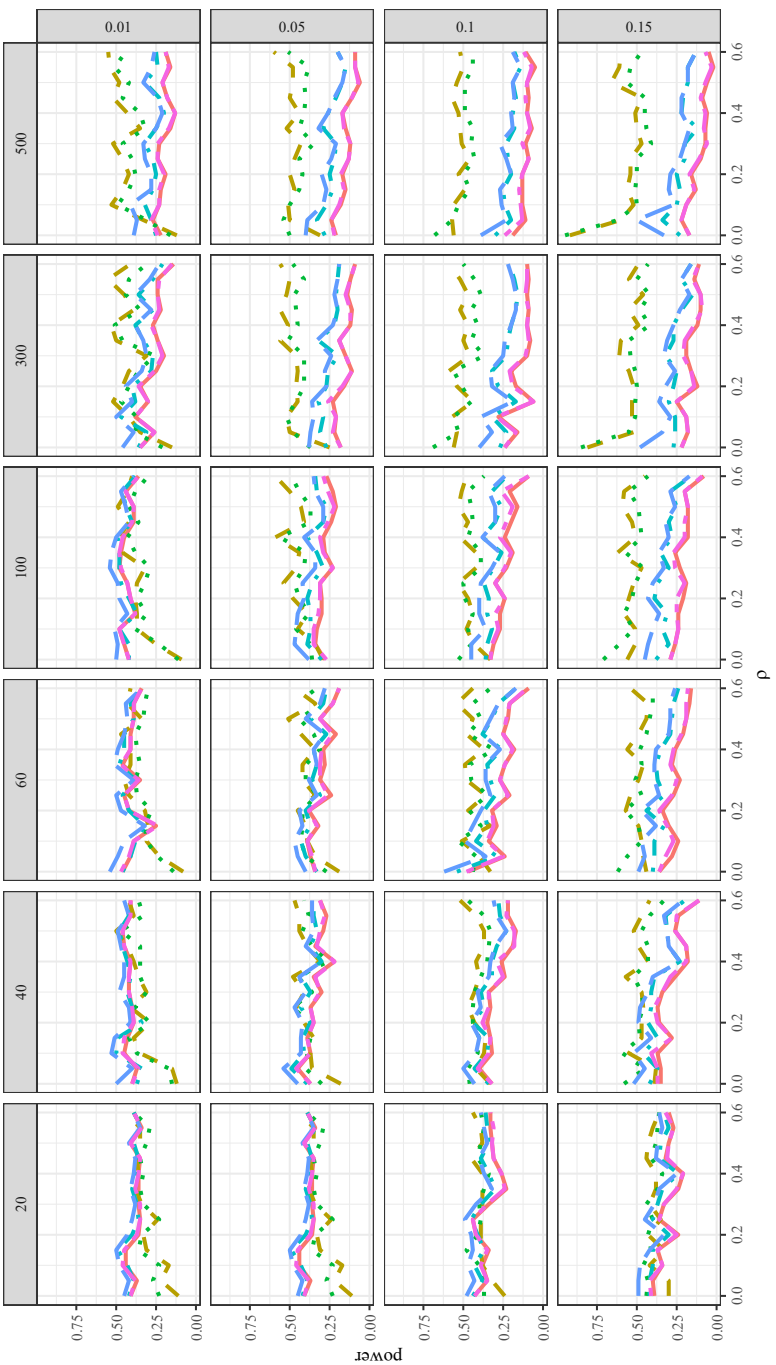


Figure 3. Power comparison for exchangeable case. The columns from left to right correspond to the dimension $d = 20, 40, 60, 100, 300, 500$. The rows from top to bottom correspond to the signal proportion $1\%, 5\%, 10\%, 15\%$.

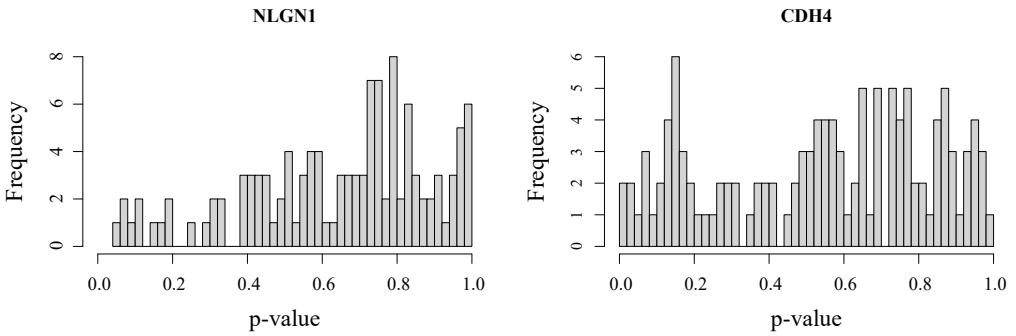


Figure 4. Histogram of p-values of SNPs for the "NLGN1" and "CDH4" genes.

The dataset is based on 968 cases and 995 controls using the Illumina HumanHap300 Genotyping BeadChip. The cases were selected to have Crohn's disease, and the controls were matched to the cases based on sex and year of birth. Subjects were drawn from two cohorts: (1) persons with non-Jewish, European ancestry (561 cases and 563 controls), and (2) persons with Jewish ancestry (407 cases and 432 controls). To analyze the association between genes and the disease, we grouped SNPs to genes via Genome Browser in a Box. Each SNP can be contained in multiple genes. As a result, all 242,535 SNPs are grouped into 19,769 genes according to the Genome Reference Consortium Human Build 38. The number of SNPs in each gene ranges from 1 to 676. Among all genes, only 4,969 genes having more than 10 SNPs are used. The p-values are constructed by using a Cochran-Mantel Haenszel chi-square test separately for each SNP. The dataset is downloaded from the database of Genotypes and Phenotypes.

Figure 4 shows histograms of p-values of SNPs in two genes, "NLGN1" and "CDH4". We present these two genes since they show some strong deviation from the uniform distribution which may be expected for highly correlated p-values. There are 118 and 123 SNPs in each gene, respectively. The histogram of p-values in each gene deviate from the uniform distribution since there are more null p-values, which is p-values from the null distribution than those of ideal uniform distribution. There may be possible cases such that either the p-values are highly correlated or there are many weak signals which are hard to be detected. However, it is more reasonable to assume that p-values in a specific gene are highly correlated based on biological properties of GWAS. From this point of view, the objective of the study is to test if there are any significant signals among the highly correlated SNPs in each gene and to compare combination methods. Since the dataset consists of only p-values of individual SNPs, a dependency structure between p-values is unknown and can not be estimated.

Tables 1 and 2 in the Supplementary Material represent p-values generated from T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP}. Table 1 contains 15 genes rejected by T_{Cauchy} and

$T_{\text{Lévy}}$ simultaneously. Table 2 contains 15 genes rejected by T_{Cauchy} but cannot be rejected by $T_{\text{Lévy}}$ and T_{minP}. Tables 1 and 2 show that p-values from T_{Cauchy}, $T_{\text{Lévy}}$ and T_{minP} tend to be ordered for each gene which matches to our theoretical and numerical results.

On the other hand, T_{Stouffer} and T_{Fisher} produce unstable p-values since their p-values have large variability compared to the other three methods. Although it is difficult to validate this for real data examples, we conjecture that T_{Stouffer} and T_{Fisher} are not robust to dependent p-values and all the other three methods have robustness to dependent p-values in controlling Type I error. We also discuss Type I error of each method using simulated null p-values which is discussed in detail in the Supplementary Material. In addition, we present Quantile-Quantile plots (Fig. S14) of real dataset to check Type I errors of each methods.

7. Concluding Remarks

In this paper, we analyze properties of different p-value combining methods according to their heaviness of transformation functions of individual p-values, when p-values are correlated and the dependence structure is unknown. We investigated Type I error and power of several methods under a wide class of dependence structures both theoretically and numerically. We also provide the theoretical study on the asymptotic equivalence between two types of recently proposed p-value combining methods, stable combination and harmonic mean.

Throughout this paper, we provide intensive results on characteristic of different p-value transformations from the viewpoint of the thickness of the tail which affects the robustness to dependence structure. Our contributions are more extensive compared to existing studies in the following sense : Firstly, we introduce a class of dependence structure based on Gaussian copula with increasing dimension while Chen et al. (2023) used arbitrary dependence under fixed dimension and Liu and Xie (2020) and Liang and Rho (2022) used a correlation matrix with bounded eigenvalues and serially correlated p-values. Secondly, we present theoretical studies on controlling Type I error for different dependent structures and provide insight on the reason for such failures for different transformations. Thirdly, we provide theoretical results showing the trade-off relationship in controlling Type I error and obtaining powers depending on different heaviness of the tail. Lastly, we investigate relationships between harmonic mean type methods and combining methods with stable distribution.

Supplementary Material

All proofs of theoretical studies and additional numerical results are contained in the online Supplementary Material.

Acknowledgments

Research of J. Park was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1A2C1A01100526).

References

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.

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

Duerr, R. H., Taylor, K. D. and Brant, S. R. (2006). A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* **314**, 1461–1463.

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

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

Hartung, J. (1999). A note on combining dependent tests of significance. *Biometrical Journal* **45**, 849–855.

Kost, J. T. and McDermott, M. P. (2002). Combining dependent p-values. *Statistics and Probability Letters* **60**, 183–190.

Liang, X. and Rho, Y. (2022). Stable combination tests. *Statistica Sinica* **32**, 641–644.

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.

Nolan, J. P. (2020). *Univariate Stable Distributions: Models for Heavy Tailed Data*. Springer, Switzerland.

Stouffer, S. A., Suchman, E. A., Dinney, L. C., Star, S. A. and Williams, R. M., Jr. (1949). *The American Soldier: Adjustment During Army Life*. Princeton University Press, New Jersey.

Tippett, L. H. C. (1931). *The Methods of Statistics*. Williams and Norgate, London.

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

Wilson, D. J. (2019). The harmonic mean p-value for combining dependent tests. *Proceedings of the National Academy of Sciences* **1116**, 1195–1200.

Wilson, D. J. (2021). The Lévy combination test. *arXiv:2105.01501v1*.

(Received April 2023; accepted October 2023)