The Cauchy Combination Test under Arbitrary Dependence Structures

Mingya Long, Zhengbang Li, Wei Zhang, and Qizhai Li

Abstract

Combining individual p-values to perform an overall test is often encountered in statistical applications. The Cauchy combination test (CCT) (Journal of the American Statistical Association, 2020, 115, 393–402) is a powerful and computationally efficient approach to integrate individual p-values under arbitrary dependence structures for sparse signals. We revisit this test to additionally show that (i) the tail probability of the CCT can be approximated just as well when more relaxed assumptions are imposed on individual p-values compared to those of the original test statistics; (ii) such assumptions are satisfied by six popular copula distributions; and (iii) the power of the CCT is no less than that of the minimum p-value test when the number of p-values goes to infinity under some regularity conditions. These findings are confirmed by both simulations and applications in two real datasets, thus, further broadening the theory and applications of the CCT.

1. Introduction

Combining individual p-values to perform an overall test is a long-standing problem in statistics with wide-ranging applications in, for example, genetics, genomics, and functional magnetic resonance imaging. We consider m hypothesis testing problems with a test statistic constructed for each one. Let \( p_i \) be the p-value for the ith hypothesis testing problem, \( i = 1, \ldots, m \). Here, we cite four well-known conventional approaches for combining \( p_1, \ldots, p_m \): \(-2 \sum_{i=1}^{m} \ln p_i \) (Fisher 1932), \(-\sum_{i=1}^{m} \ln(1 - p_i)\) (Pearson 1933), \(\sum_{i=1}^{m} \Phi^{-1}(1 - p_i)\) (Lipták 1958), and \(\sum_{i=1}^{m} p_i\) (Edgington 1972), where \(\Phi(\cdot)\) is the cumulative distribution function of a standard normal distribution. However, these approaches perform well only when signals are dense, that is, when most \( p_i \)'s are small.

Recent high-dimensional data collected in multiple disciplines tend to have very sparse signal, with low signal-to-noise ratio. Therefore, the application of these conventional methods could result in substantial power loss. To overcome this limitation, a number of alternative tests can be applied, including, the Tippett's minimum p-value test (MINP) (Tippett 1931), which will become important in the present work, as well as the Berk–Jones test (Berk and Jones 1979), the higher criticism test (Donoho and Jin 2004), the group-combined test (Huet al. 2017), and the generalized Berk–Jones test (Sun and Lin 2020). However, most of these tests do not provide analytic formulas for calculating p-values when individual p-values are correlated. Resampling approaches, such as permutation and bootstrap procedures, can be used to handle correlated p-values. However, such methods are computationally impractical when it comes to large-scale data, especially when the p-value of a combination test is extremely small.

Recently, Liu and Xie (2020) proposed a Cauchy combination test denoted by \(\text{CCT} = \sum_{i=1}^{m} \omega_i \tan \left( (0.5 - p_i)\pi \right) \), where the weights \(\omega_i\) are nonnegative and \(\sum_{i=1}^{m} \omega_i = 1\). They showed that the tail probability of the CCT could be well approximated by a standard Cauchy distribution under the null hypothesis, which specifies the bivariate normality and two mild assumptions in the high-dimensional setting. So far, the CCT has been used in quite a few real applications. For example, Gorfine, Schlesinger, and Hsu (2020) applied it to study right-censoring data in survival analysis, while McCaw et al. (2020) constructed a powerful test based on the CCT for quantitative trait genetic association studies. Khalid et al. (2020) also employed the CCT to perform inter-module communications for runtime hardware Trojan detection. The CCT has also been used to analyze whole-genome sequencing studies and genome-wide studies with summary statistics (Li et al. 2019; Liu et al. 2019; Bu et al. 2020; Li et al. 2020; Xu et al. 2020; Li et al. 2021).

If these assumptions in Liu and Xie (2020) are violated, the approximation of the tail distribution of the CCT with the derived Cauchy distribution might not be appropriate. Although extensive numerical simulation studies have been conducted in the literature to investigate the feasibility of such approximation, a theoretical justification is not available. In this work, we revisit the theoretical underpinnings of the CCT, showing that the approximation of the standard Cauchy distribution for...
the tail probability of the CCT is still valid under a broader range of bivariate distributions, including the six popular copula distributions. We further show that the power of the CCT is no less than that of the MINP, when the number of tests goes to infinity. These extensions broaden the theory and applications of the CCT.

This article is organized as follows. The main results are presented in Section 2. Simulation studies are conducted in Section 3 to examine the accuracy of the tail probability approximation. In Section 4, data from prostate cancer and air quality studies are analyzed to further investigate the performance of the CCT and three conventional tests. Some discussions are given in the final section, and all technical details are provided in the supplementary materials.

2. Main Results

For each individual p-value $p_i$, let $Z_i$ be the corresponding test statistic, $i = 1, \ldots, m$, $Z_i$ has zero mean and unit standard deviation under the global null hypothesis. Denote by $\rho_{ij}$ the correlation coefficient between $Z_i$ and $Z_j$, and $i, j = 1, \ldots, m$. Write $R = (\rho_{ij})_{m \times m}$ and assume that:

(C1) $(Z_i, Z_j)^\top$ follows a bivariate normal distribution for $1 \leq i < j \leq m$, where the superscript $\top$ denotes the transpose of a matrix or a vector.

(C2) $\lambda_{\text{max}}(R) < C_0$, where $\lambda_{\text{max}}(R)$ is the largest eigenvalue of $R$ and $C_0$ is a positive constant.

(C3) a constant $\rho_{\text{max}}$ exists such that $\max_{1 \leq i < j \leq m} |\rho_{ij}| \leq \rho_{\text{max}} < 1$.

Under assumption (C1) with fixed $m$ or assumptions (C1)–(C3) with $m = o(t^8)$, $\delta \in (0, 1/2)$, Liu and Xie (2020) showed that

$$\lim_{t \to \infty} P(\text{CCT} > t) = 0.5 - \arctan(t)/\pi = 1,$$

where $o(t^8)/(t^8) \to 0$ as $t \to \infty$ and the denominator is the tail probability of the standard Cauchy distribution.

It is worth pointing out that the bivariate normal distribution assumption (C1) for $(Z_i, Z_j)^\top$ can be too stringent for real applications. Instead, $Z_i$ and $Z_j$ could have an arbitrary bivariate distribution. Moreover, assumption (C2) requires the eigenvalues of $R$ be bounded by a constant, which is not always appropriate in practice. For example, the largest eigenvalue of the common spiked correlation model can go to infinity (Johnstone 2001; Lee, Zou, and Wright 2014; Zhang et al. 2020; Shi et al. 2022) (see details of the spiked correlation model in Section 3).

2.1. Tail Null Distribution of the CCT for Fixed $m$

Instead of assumption (C1), we consider the following less stringent distribution assumption on the individual $p$-values $p_1, \ldots, p_m$.

(D1) For $1 \leq i < j \leq m$, as $t \to \infty$, there exists a sequence of $\delta_t$, with $\lim_{t \to \infty} \delta_t = 0$, and the limit $\delta_t t \to \infty$, such that

$$P\left(0 < p_i < \frac{\omega_i}{\pi} \frac{m}{t}, 0 < p_j < \frac{\omega_j}{\pi} \frac{m}{\delta_t t}\right) = o\left(\frac{1}{t}\right).$$

and

$$P\left(0 < p_i < \frac{\omega_i}{\pi} \frac{m}{t}, 0 < p_j < \frac{\omega_j}{\pi} \frac{m}{\delta_t t}\right) = o\left(\frac{1}{t}\right).$$

Under assumption (D1), we can show that the standard Cauchy approximation still holds in the following theorem, the proof of which is given in the supplementary materials.

**Theorem 1.** The approximation given by (1) still holds if $p_i$ follows the uniform distribution on $[0, 1]$ and assumption (D1) is satisfied.

Assumption (D1) imposes no restriction on the type of joint distribution of $Z_i$ and $Z_j$, $i, j = 1, \ldots, m$. An arbitrary bivariate distribution (including but not limited to the bivariate normal distribution) for $(T_i, T_j)^\top$ is allowed under assumption (D1). To make this case, we give the following six bivariate copula functions which are widely used in applications including finance (Genest and MacKay 1986; Meyer 2013) and survival analysis (Geerdens, Acar, and Janssens 2018):

1. **Product Copula:**
   $$C(u_i, v_j) = u_i v_j, \quad 1 \leq i \neq j \leq m.$$

2. **Farlie–Gumbel–Morgenstern (FGM) Copula:**
   $$C(u_i, v_j) = u_i v_j [1 + \theta (1 - u_i) (1 - v_j)], \quad \theta \in [-1, 1], \quad 1 \leq i \neq j \leq m.$$

3. **Cuadras-Augé Copula:**
   $$C(u_i, v_j) = \{ \min (u_i, v_j) \}^{\theta} (u_i v_j) \{1 - \theta \}, \quad \theta \in [0, 1], \quad 1 \leq i \neq j \leq m.$$

4. **Normal Copula:**
   $$C(u_i, v_j) = \frac{1}{2\pi \sqrt{1 - \rho_{ij}}} \int_{-\infty}^{\Phi^{-1}(u_i)} \int_{-\infty}^{\Phi^{-1}(v_j)} \exp \left( -\frac{\left( x^2 + 2 \rho_{ij} x y + y^2 \right)}{2(1 - \rho_{ij}^2)} \right) dx dy, \quad 1 \leq i \neq j \leq m, \quad \max_{1 \leq i < j \leq m} |\rho_{ij}| \leq \rho_{\text{max}} < 1.$$

5. **Ali-Mikhail-Haq (AMH) Copula:**
   $$C(u_i, v_j) = \frac{u_i v_j}{1 - \theta (1 - u_i) (1 - v_j)}, \quad \theta \in [-1, 1], \quad 1 \leq i \neq j \leq m.$$

6. **Survival Copula:**
   $$C(u_i, v_j) = u_i v_j \exp \left( -\theta \ln u_i \ln v_j \right), \quad \theta \in [0, 1], \quad 1 \leq i \neq j \leq m.$$

Let the joint distribution of $p_i$ and $p_j$ be modeled by one of the six copula functions described above. We can show the following result, with the proof being given in the supplementary materials.

**Theorem 2.** Assumption (D1) is satisfied for the above six types of copula functions. Therefore, the approximation given by (1) still holds under these copula functions.
2.2. Tail Null Distribution of the CCT for Divergent m

Next, in order to relax assumptions (C1)–(C3), we establish the theory of the tail null distribution of the CCT for divergent m. The following assumption is needed.

(D2) For 1 ≤ i < j ≤ m, as t → ∞, there exists a sequence of δ_i, with \( \lim_{t \to \infty} \delta_i = 0 \), and the \( \lim_{t \to \infty} \delta_i t \to \infty \), such that

\[
\sup_{1 \leq i < j \leq m} P \left( 0 < p_i < \frac{\omega_j m}{\pi \delta_i t}, 0 < p_j < \frac{\omega_j m}{\pi \delta_i t} \right) = o \left( \frac{1}{t^{1+\gamma}} \right)
\]

and

\[
\sup_{1 \leq i < j \leq m} P \left( 0 < p_i < \frac{\omega_j m}{\pi (1 + \delta_i)t}, 1 - \frac{\omega_j m}{\pi \delta_i t} < p_j < 1 \right) = o \left( \frac{1}{t^{1+\gamma}} \right),
\]

where 0 < γ ≤ 1.

**Theorem 3.** The approximation given by (1) still holds if \( p_i \) follows the uniform distribution on [0, 1], \( m = o(t^{1/2}) \), and assumption (D2) is satisfied.

The proof of Theorem 3 is given in the supplementary materials. Similar to the case of fixed m, there is no restriction on the joint distribution of \( Z_i \) an \( Z_j \), as long as assumption (D2) is satisfied. Furthermore, we can show the following theorem.

**Theorem 4.** Assumption (D2) is satisfied for the above six types of copula functions. Therefore, the approximation given by (1) still holds under these copula functions.

The proof for this theorem is given in the supplementary materials. As noted earlier, according to Liu and Xie (2020), the conclusion in (1) may not hold for a spiked model, which violates assumption (C2), since the largest eigenvalue is not a constant. However, assumption (D2) allows the use of the spiked model. Therefore, approximate (1) is still valid for the spiked model. Simulations results shown later can confirm this.

2.3. Power Comparison between the CCT and the MINP

Previously, we stated that the power of the CCT is no less than that of the minimum p-value test when the number of p-values goes to infinity under some regularity conditions in the introduction. Accordingly, the MINP is equivalent to such maximum p-value test (hereinafter denoted as MAX), where \( \text{MAX} = \max \{ Z_1^2, \ldots, Z_m^2 \} \). In the following, we compare the performance of the CCT and MAX. By following the theoretical settings of Donoho and Jin (2004) and Liu and Xie (2020), we assume that \( (Z_1, \ldots, Z_m)^T \sim N_m(\mu, \Sigma) \), where \( \mu = (\mu_1, \ldots, \mu_m)^T \). Throughout this work, \( Z_i \) has had unit standard deviation, which means that the \( \Sigma \) is the correlation matrix. The hypothesis testing problem is \( H_0 : \mu = 0_m \) versus \( H_1 : \mu \neq 0_m \), where \( 0_m \) is the m-dimensional vector with all elements being zero. Under \( H_1 \), denote the index set of nonzero elements (signals) of \( \mu \) by \( \Omega = \{ 1 \leq i \leq m : \mu_i \neq 0 \} \), and let the total number of nonzero signals be \( ||\Omega|| = m^\nu \), where \( || \cdot || \) is the cardinality of a set and the parameter \( 0 < \nu < 1 \) measures the sparsity magnitude of signals. Similarly, denote the number of zero signals by \( ||\Omega^c|| = m^\kappa \), where \( 0 < \kappa < 1 \) and \( \Omega^c = \{ 1, \ldots, m \} \setminus \Omega \). To derive the asymptotic distribution of MAX, we first need the following assumption.

(D3) Define \( \varphi_k = \sup_{|i-j| < k} \rho_k \), \( k = 1, \ldots, m - 1 \). Then, \( \varphi_1 < 1 \) and \( \varphi_k (\log k)^{2+d} \to 0 \) for \( d > 0 \) as \( k \to \infty \).

**Corollary 1.** Under assumption (D3) and for a large enough \( t > 0 \), we have

\[
\lim_{m \to \infty} P \left( \frac{\text{MAX} - \tilde{a}_m - O((\log m)^{-1})}{\tilde{b}_m} < t \right) = \exp \left( \exp(-t) \right),
\]

where \( \tilde{a}_m = 2 \log m - \left[\log (\log m) + \log(4\pi) - \log 4 \right) + \log (\log m) + \log(4\pi) - \log 4)/2 \log m, \tilde{b}_m = \sqrt{2 \log m - 1}/(\log m)^{-1} \to C_1, C_1 \) is a constant.

Deo (1972) and Pakshirajan and Hebbar (1977) considered max \( \{ |Z_1|, \ldots, |Z_m| \} \), and showed that under assumption (D3),

\[
\lim_{m \to \infty} P \left( \frac{\max \{ |Z_1|, \ldots, |Z_m| \} - a_m - O((\log m)^{-1})}{b_m} < t \right) = \exp \left( \exp(-t) \right),
\]

where \( a_m = (2 \log m)^{1/2} - (\log (\log m) + 4\pi - 4)/(8 \log m)^{1/2} \) and \( b_m = (2 \log m)^{-1/2} \). We want to point out that Corollary 1 is similar to the conclusion by Deo (1972) and Pakshirajan and Hebbar’s (1977). However, the proof of Corollary 1 is different. Let \( \xi_1 - \alpha \) and \( \eta_1 - \alpha \) be the \( 1 - \alpha \) quantile of the standard Cauchy distribution and the standard Gumbell distribution, respectively, that is, \( \xi_1 - \alpha = \cot(\pi \alpha) \) and \( \eta_1 - \alpha = -\log(\log(\alpha))^{-1} \). The asymptotic powers of the CCT and MAX are thus given by

\[
\beta_{\text{CCT}} = P_{H_1}(\text{CCT} > \xi_1 - \alpha) \quad \text{and} \quad \beta_{\text{MAX}} = P_{H_1}(\sqrt{\text{MAX} - a_m - O((\log m)^{-1})}/b_m > \eta_1 - \alpha).
\]

We have the following result comparing the powers of the two tests.

**Theorem 5.** Assumption (D3) is satisfied and \( \min_{i=1, \ldots, m} \omega_i = O(1/m) \). Then as \( m \to \infty \),

\[
\beta_{\text{CCT}} \geq \beta_{\text{MAX}} + o(1).
\]

When \( m \) is large enough, Theorem 5 shows that the asymptotic power of the CCT is no less than that of MAX, which demonstrates the power advantage of the CCT when combining a large number of individual p-values.

3. Simulation Studies

3.1. Tail Probability Approximation

In this section, we conduct simulation studies to evaluate the accuracy of the tail probability approximation based on the standard Cauchy distribution (SCD). Two spiked models with equal and unequal correlation coefficients are considered.
- Model 1 (Unequal correlation spiked model): $\lambda_i = m/3^i$ for $i = 1, \ldots, d$ and $\lambda_i = 1$ for $i = d + 1, \ldots, m$, where $d = 4, 5$ and 6.
- Model 2 (Equal correlation spiked model): $\rho_{ij} = \rho$ for $1 \leq i \neq j \leq m$ and $\rho_{ii} = 1$ for $1 \leq i \leq m$, where $\rho = 0.2, 0.5$ and 0.8.

Model 2 has a compound symmetry correlation structure, that is, $(\rho^{\text{sgn}((i-j))})_{m \times m}$ with the largest eigenvalue being $m \rho + (1 - \rho)$ and all other eigenvalues being $1 - \rho$, where $\text{sgn}((i-j))$ is a sign function which equals 0 if $i = j$ and 1 otherwise, and $\rho \in (0, 1)$. It is clear that the largest eigenvalue goes to infinity as $m \to \infty$.

We consider $m = 10, 50$, and 500. The potential test statistics $(Z_1, \ldots, Z_m)\top$ are generated from an $m$-dimensional $t$ distribution $t(0_m, R)$ with mean vector $0_m$ and correlation matrix $R = (\rho_{ij})_{m \times m}$. The marginal distributions of $Z_1, \ldots, Z_m$ are all set to be the univariate $t$ distribution with 10 degrees of freedom.

The individual $p$-value is obtained as $p_i = 2(1 - \Psi(|Z_i|))$, where $\Psi(\cdot)$ is the cumulative distribution function of a $t$ distribution with 10 degrees of freedom. Figure 1 displays the tail probabilities of $P(\text{SCD} > t)$ (blue dotted line) and $P(\text{CCT} > t)$ (red solid line) for Model 1, where $P(\text{CCT} > t)$ is calculated based on 500,000 Monte Carlo samples. Since the exact distribution of the CCT is unknown, we use this strategy as the gold standard throughout the remainder of this work. In addition, we set the range of the horizontal axis to be the 95% and 99.97% quantiles of the standard Cauchy distribution, which are 6.314 and 1000, respectively. From Figure 1, it can be seen that the standard Cauchy distribution is a good approximation of the CCT, with both lines always coinciding with each other. The results for Model 2 are similar and details are provided in the supplementary materials.

Next, we conduct simulation studies using the AMH copula and the FGM copula mixed with the product copula. We again consider $m = 10, 50$, and 500.
- Model 3 (AMH copula mixed with product copula model): $(p_i, p_{i+1})\top \sim \mathbb{C}(u_i, v_{i+1}) = u_iv_{i+1}/(1 - \theta(1 - u_i)(1 - v_{i+1}))$ for $i = 1, 3, \ldots, 2\lfloor m/2 \rfloor - 1$, where $\lfloor m \rfloor$ is the maximum integer less than $m$, where $(p_i, p_{i+1})\top \sim \mathbb{C}(u_i, v_{i+1}) = u_iv_{i+1}$ for other $i$, and $\theta = 0.2, 0.5$, and 0.8.
- Model 4 (FGM copula mixed with product copula model): $(p_i, p_{i+1})\top \sim \mathbb{C}(u_i, v_{i+1}) = u_iv_{i+1}[1 + \theta(1 - u_i)(1 - v_{i+1})]$ for $i = 1, 3, \ldots, 2\lfloor m/2 \rfloor - 1$, where $(p_i, p_{i+1})\top \sim \mathbb{C}(u_i, v_{i+1}) = u_iv_{i+1}$ for other $i$, and $\theta = 0.2, 0.5$, and 0.8.

The $p$-values are generated based on the above two models. Figure 2 shows the tail probabilities of $P(\text{SCD} > t)$ and $P(\text{CCT} > t)$, where $P(\text{CCT} > t)$ is again calculated based on 500,000 Monte Carlo samples. The two lines in the figure are almost the same, indicating again that the tail probability...
of the CCT can be well approximated by the standard Cauchy distribution. Similar results are observed for Model 4 and are presented in the supplementary materials.

### 3.2. Power Comparison

Here, we report a simulation study comparing the power of the CCT to those of MAX, the generalized higher criticism test (GHC) (Barnett et al. 2017) and generalized Berk–Jones test (GBJ) (Sun and Lin 2020). Theorem 5 requires that \( \rho_1 < 1 \) and \( \rho_k (\log k)^{2+d} \to 0 \) for \( d > 0 \) as \( k \to \infty \), which is not satisfied by Models 1 and 2 in Section 3.1. Hence, we use two other common models to specify the correlation matrix. The potential test statistics \( (Z_1, \ldots, Z_m)^T \) are generated from an \( m \)-dimensional normal distribution \( N(\mu_m, R) \) with mean vector \( \mu_m \) and correlation matrix \( R = (\rho_{ij})_{m \times m} \). Matrix \( R \) is specified using the following two structures:

- **Structure 1** (AR(1) correlation): \( \rho_{ij} = \rho^{|i-j|} \) for \( 1 \leq i \neq j \leq m \) and \( \rho_i = 1 \) for \( 1 \leq i \leq m \), where \( \rho = 0.2, 0.5, \) and 0.8.
- **Structure 2** (Polynomial decay): \( \rho_{ij} = 1/(1 + |i-j|^a) \) for \( 1 \leq i \neq j \leq m \), where \( a = 0.5, 1.5, \) and 2.5.

We consider a sparse mean vector \( \mu_m \) by letting the proportion of its nonzero elements be \( ||\Omega||/m = 0.1, 0.2, \) and 0.3, and letting all nonzero elements be equal. Thus, the nonzero elements are equal to \( \sqrt{3 \log m/(||\Omega||)^{1/3}} \), which makes all powers comparable. The simulation results are calculated based on \( n = 2000 \) simulation replicates, when \( m \) is chosen from \( \{20, 40, 60, 80\} \), and the nominal significance level is 0.05.

Figures 3 and 4 display empirical powers of CCT, MAX, GHC, and GBJ for Structures 1 and 2, respectively. From these figures, we can see that the CCT has the power comparable to those of GHC and GBJ when individual statistics are weakly correlated. For example, in Figure 3, when \( \rho = 0.2, ||\Omega||/m = 0.1 \) and \( m = 60 \), the empirical power of the CCT is 0.4600, which is slightly smaller than that of GHC (0.4785) and GBJ (0.5005). In contrast, when the individual statistics are strongly correlated, the CCT can achieve much higher power. For example, in Figure 4, when \( a = 0.5, ||\Omega||/m = 0.3 \) and \( m = 80 \), the empirical power of the CCT is 0.7185, which is much higher than that of GHC (0.6450) and GBJ (0.6060). The figures further show that the empirical power of MAX is always lower than that of the CCT; in some cases the CCT is about 38% more powerful than MAX.

### 4. Real Data Analyses

#### 4.1. A Prostate Cancer Study

We apply MAX, GHC, GBJ, and CCT to analyze data from a prostate cancer study aimed to investigate whether gene expression levels in some pathways are different between tumor and normal prostate samples. The data consist of expression
profiles of approximately 12,600 genes from 52 tumor and 50 non-tumor prostate specimens (Singh et al. 2002). Raw data are publicly available at https://www.ncbi.nlm.nih.gov/sites/myncbi/recentactivity. Previous literatures show that gene expression levels between tumor and normal samples are different in some pathways, including the axon guidance pathway (map04360), the toll-like receptor signaling pathway (map04620), the T cell receptor signaling pathway (map04660), the caffeine metabolism pathway (map00232), the riboflavin metabolism pathway (map00740), and the neuroactive ligand-receptor interaction pathway (map04080), consisting of 163, 148, 182, 10, 12, and 263 genes, respectively (Honda and Taniguchi 2006; Barach et al. 2011; Shitivelman, Beer, and Evans 2014; Eidelman et al. 2017; Ren et al. 2018). Here, we confine our analyses to these pathways.

The Wilcoxon test is conducted for each gene to obtain individual p-values. However, for a more meaningful comparison, we use the CCT and MAX to combine the individual p-values, and use GHC and GBJ to combine the individual (Wilcoxon) test statistics. Results on these six pathways are presented in Table 1. The p-values of MAX, GHC, GBJ, and CCT are calculated based on 10,000 permutation replicates and the p-value for the CCT is obtained using the approximation formula (1). Under the nominal significance level of 0.05, Table 1 shows that all four tests, except GBJ, are able to detect significant difference between tumor and normal samples. The GBJ test fails to do so for pathway map04080, with a p-value of 0.0514.

4.2. An Air Quality Study

To further evaluate the performances of MAX, GHC, GBJ, and CCT, we apply them to analyze data from an air quality study. The dataset, which is publicly available at http://archive.ics.uci.edu/ml/datasets/Air+quality, contains samples of hourly averaged responses from an air quality chemical multi-sensor device located in an Italian city. Five air pollutants including CO, NOx, NO2, non-metanic hydrocarbons, and benzene and three air quality indicators including temperature, relative humidity, and absolute humidity, were recorded from March 2004 to February 2005 (De Vito et al. 2008). After deleting the samples with missing data, 827 remaining samples are used for analysis. Our aim is to check whether the air pollutants and air quality indicators are associated.

Individual p-values and individual test statistics are constructed from the F test of overall significance in the regression model for each air pollutant on all air quality indicators. Each pollutant could be measured two ways. Hence, $m = 10$. We calculate the p-values of MAX, GHC, and GBJ by using 100,000 permutations and p-value of the CCT using the approximation formula (1). The p-values of MAX, GHC, GBJ, and CCT are $< 10^{-6}$, $10^{-15}$, $10^{-12}$, and $< 10^{-17}$, respectively. Under the nominal significance level of 0.05, all tests could successfully detect the significant association between air pollutants and air quality indicators, consistent with the conclusion of Runge (2018).
M. LONG ET AL.

Figure 4. Empirical powers of CCTs (Squares), MAXes (Circles), GHCs (Triangles), and GBJs (Plus signs), where the test statistics $Z_i, i = 1, \ldots, m$ are generated from an $m$-dimensional normal distribution with the parameters set in Structure 2.

Table 1. $p$-values of MAX, GHC, GBJ, and CCT for six pathways of prostate cancer data.

| Pathway     | # of genes | MAX     | GHC     | GBJ     | CCT     |
|-------------|------------|---------|---------|---------|---------|
| map04360    | 163        | 0.0004  | 0.0003  | 0.0008  | 0.0004  |
| map04620    | 148        | 0.0145  | 0.0152  | 0.0118  | 0.0078  |
| map04660    | 182        | 0.0053  | 0.0149  | 0.0132  | 0.0079  |
| map00232    | 10         | 0.0220  | 0.0150  | 0.0130  | 0.0140  |
| map00740    | 12         | 0.0167  | 0.0188  | 0.0258  | 0.0146  |
| map04080    | 263        | 0.0232  | 0.0341  | 0.0514  | 0.0117  |

5. Discussion

We began by characterizing high-dimensional data signals as sparse, that is, having only a few small $p$-values. In the performance of an omnibus test, we proposed that individual $p$-values are often encountered in statistical applications. Many conventional approaches to combine $p$-values have been reported in the literatures, such as the MINP, the Berk–Jones test, and the higher criticism test. However, most of these tend to be computationally intensive, limiting their ability. The Cauchy combination test (CCT), on the other hand, is a powerful and computationally efficient approach to integrate individual $p$-values under arbitrary dependence structures for sparse signals. In this work, we revisited the CCT and put forth several assumptions to relax the original ones and, thus, further broaden the theory and applications of the CCT. Specifically, instead of the original test statistics, we first impose assumptions on individual $p$-values. Considering fixed $m$ or divergent $m$, the tail probability of the CCT can be approximated just as well under two assumptions: that (a) as $t$ goes to infinity, the joint probability of any two individual $p$-values is bounded by $1/t$, and (b) the joint probability of any two individual $p$-values is bounded by $1/(1+t^r)$. Next, in order to confirm the broader application of these two assumptions, six popular and widely used copula distributions are illustrated. Finally, following the theoretical settings of Donoho and Jin (2004) and Liu and Xie (2020), we theoretically prove that the power of the CCT is no less than that of the minimum $p$-value test when the number of $p$-values goes to infinity. These findings are confirmed by both simulation and two real datasets. R codes for the simulation study and real data analyses are given in the supplementary materials.

In the case of sparse signals, the analytic formulas for calculating the CCT make the test useful. However, as shown in Chen (2022), the CCT may be less powerful when signals are mixed, that is, when some $p$-values to be combined are small and others are large. Additionally, Liu and Lin (2019) pointed out that no uniformly powerful test exists for a multiple dimensional composite alternative hypothesis. These challenges call for a novel, robust and widely applicable combination method. In this work, we have, as noted above, broadened the theoretical basis of the CCT to meet these challenges. However, some improvements are still needed. For instance, $p$-values do not follow uniform
distribution, even though appropriate in some cases. In the case of discrete data analysis, the CCT cannot make an accurate inference when p-values do not follow uniform distribution, thus, requiring further investigation. As stated in the article, the approximate strengths of the CCT stem from the dimension \( m \), the critical threshold \( t \), and the bound of distribution probability of p-values, so another interesting direction for future study is to derive the convergence rate of the Cauchy approximation and explore different assumptions to speed it up.

**Supplementary Materials**

The Supplementary Materials contain three parts. In Part 1, we provide the technical details of theorems and Corollary in Section 2 of the main text. Part 2 includes the additional simulation results for Models 2 and 4 in Section 3. Part 3 gives the R codes for the numerical studies (Model 1 in the simulation study and real data analysis for Air Quality study).

**Acknowledgments**

We would like to thank the editor, the associate editor, and two anonymous reviewers for their insightful comments that have greatly improved the article. We also have to thank David Martin and Kai Yu for their valuable help.

**Disclosure Statement**

The authors report there are no competing interests to declare.

**Funding**

Qizhai Li was supported in part by Beijing Natural Science Foundation (Z180006).

**References**

Barach, Y. S., Lee, J. S., and Zang, X. (2011), “T Cell Coinhibition in Prostate Cancer: New Immune Evasion Pathways and Emerging Therapeutics,” *Trends in Molecular Medicine*, 17, 47–55. [139]

Barnett, I., Mukherjee, R., and Lin, X. (2017), “The Generalized Higher Criticism for Testing SNP-set Effects in Genetic Association Studies,” *Journal of the American Statistical Association*, 112, 64–76. [134,138]

Berk, R. H., and Jones, D. H. (1979), “Goodness-of-Fit Test Statistics that Explore Different Assumptions to Speed Up,” *Biometrika*, 66, 644–658. [139]

Berk, R. H., and Jones, D. H. (2017), “Goodness-of-Fit Test Statistics that Explore Different Assumptions to Speed Up,” *Biometrika*, 66, 644–658. [139]

Bu, D., Yang, Q., Meng, Z., Zhang, S., and Li, Q. (2020), “Truncated Tests for Combining Evidence of Summary Statistics,” *Biometrics*, 76, 464–474. [134]

Chen, Z. (2022), “Robust Tests for Combining p-values with Applications to Genetic Association Studies,” *Bioinformatics*, 32, 2737–2743. [134]

Christenson, L. M. (2001), “On the Distribution of the Largest Eigenvalue in Principal Components Analysis,” *The Annals of Statistics*, 29, 295–327. [135]

Chow, H., and Stoffer, D. H. (1995), “Likelihood Ratio Tests for the Selection of Covariates in Multiple Linear Regression,” *Journal of the American Statistical Association*, 90, 182–190. [135]

Cline, D. B., and Raftery, A. E. (1994), “A Framework for Detecting Noncoding Rare Variant Associations of Large-Scale Whole-Genome Sequencing Studies,” *bioRxiv*. [134]

Cline, D. B., and Raftery, A. E. (1994), “A Framework for Detecting Noncoding Rare Variant Associations of Large-Scale Whole-Genome Sequencing Studies,” *bioRxiv*. [134]

Lipták, T. (1958), “On the Combination of Independent Tests,” *Magyár Tud Akad Mat Kutato Int Kozl.*, 3, 171–197. [134]

Liu, Y., Chen, S., Li, Z., Morrison, A. C., Boerwinkle, E., and Lin, X. (2019), “A Fast and Powerful p Value Combination Method for
Rare-Variant Analysis in Sequencing Studies," *The American Journal of Human Genetics*, 104, 410–421. [134]

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. [134,135,136,140]

Liu, Z., and Lin, X. (2019), "A Geometric Perspective on the Power of Principal Component Association Tests in Multiple Phenotype Studies," *Journal of the American Statistical Association*, 114, 975–990. [140]

McCaw, Z. R., Lane, J. M., Saxena, R., Redline, S., and Lin, X. (2020), "Operating Characteristics of the Rank-based Inverse Normal Transformation for Quantitative Trait Analysis in Genome-Wide Association Studies," *Biometrics*, 76, 1262–1272. [134]

Meyer, C. (2013), "The Bivariate Normal Copula," *Communications in Statistics-Theory and Methods*, 42, 2402–2422. [135]

Pakshirajan, R., and Hebbar, H. V. (1977), "Limit Theorems for Extremes of Absolute Values of Gaussian Sequences," *Sankhyā: The Indian Journal of Statistics*, Series A, 39, 191–195. [136]

Pearson, K. (1933), "On a Method of Determining Whether a Sample of Size n Supposed to have been Drawn from a Parent Population having a Known Probability Integral has Probably Been Drawn at Random," *Biometrika*, 25, 379–410. [134]

Ren, S., Wei, G.-H., Liu, D., Wang, L., Hou, Y., Zhu, S., Peng, L., Zhang, Q., Cheng, Y., Su, H., Zhou, X., Zhang, J., Li, F., Zheng, H., Zhao, Z., Yin, C., He, Z., Gao, X., Zhou, H. E., Chu, C.-Y., Wu, J. B., Collins, C., Volik, S. V., Bell, R., Huang, J., Wu, K., Xu, D., Ye, D., Yu, Y., Zhu, L., Qiao, M., Lee, H.-M., Yang, Y., Zhu, Y., Shi, X., Chen, R., Wang, Y., Xu, W., Cheng, Y., Xu, C., Gao, X., Zhou, T., Yang, B., Hou, J., Liu, L., Zhang, Z., Zhou, Y., Qiu, C., Shao, P., Pang, J., Chung, I. W. K., Xu, J., Wu, C.-L., Zhong, W., Xu, X., Li, Y., Zhang, X., Wang, J., Yang, H., Wang, J., Huang, H., and Sun, Y. (2018), "Whole-Genome and Transcriptome Sequencing of Prostate Cancer Identify New Genetic Alterations Driving Disease Progression," *European Urology*, 73, 322–339. [139]

Runge, J. (2018), "Conditional Independence Testing based on a Nearest-Neighbor Estimator of Conditional Mutual Information," in *International Conference on Artificial Intelligence and Statistics*, pp. 938–947. [139]

Shi, Y., Zhang, W., Liu, A., and Li, Q. (2022), "Distance-based Regression Analysis for Measuring Associations," *Journal of Systems Science & Complexity*. DOI:10.1007/s11122-022-3072-0. [139]

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D’Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [139]

Sun, R., and Lin, X. (2020), "Genetic Variant Set-based Tests using the Generalized Berk–Jones Statistic with Application to a Genome-Wide Association Study of Breast Cancer," *Journal of the American Statistical Association*, 115, 1079–1091. [134,138]

Tippett, L. H. C. (1931), *The Methods of Statistics. An Introduction Mainly for Workers in the Biological Sciences*. London: Williams and Norgate, Ltd. [134]

Xu, D., Wang, C., Kiryluk, K., Buxbaum, J. D., and Ionita-Laza, I. (2020), "Co-localization between Sequence Constraint and Epigenomic Information Improves Interpretation of Whole-Genome Sequencing Data," *The American Journal of Human Genetics*, 106, 513–524. [134]

Zhang, J.-T., Guo, J., Zhou, B., and Cheng, M.-Y. (2020), "A Simple Two-Sample Test in High Dimensions Based on l 2-Norm," *Journal of the American Statistical Association*, 115, 1011–1027. [135]