

# Optimal detection of weak positive latent dependence between two sequences of multiple tests

Sihai Dave Zhao<sup>a,\*</sup>, T. Tony Cai<sup>b</sup>, Hongzhe Li<sup>c</sup>

<sup>a</sup> Department of Statistics, University of Illinois at Urbana–Champaign, IL, United States

<sup>b</sup> Department of Statistics, The Wharton School, University of Pennsylvania, Philadelphia, PA, United States

<sup>c</sup> Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

## ARTICLE INFO

### Article history:

Received 8 September 2016

Available online 14 July 2017

### Keywords:

Detection boundary

Higher criticism

Independence testing

Optimal adaptivity

Sparsity

## ABSTRACT

It is frequently of interest to jointly analyze two paired sequences of multiple tests. This paper studies the problem of detecting whether there are more pairs of tests that are significant in both sequences than would be expected by chance. The asymptotic detection boundary is derived in terms of parameters such as the sparsity of non-null cases in each sequence, the effect sizes of the signals, and the magnitude of the dependence between the two sequences. A new test for detecting weak dependence is also proposed, shown to be asymptotically adaptively optimal, studied in simulations, and applied to study genetic pleiotropy in 10 pediatric autoimmune diseases.

© 2017 Elsevier Inc. All rights reserved.

## 1. Introduction

### 1.1. Overview

Joint analysis of two paired sequences of multiple tests, each arising from a separate independent study, arises in many applications. It has been particularly motivated by genomics research, where it is natural to investigate similarities in how genomic features, such as genes or genetic markers, behave across studies. For example, recent interest has focused on features that may be significant in both of two sequences of multiple tests. In differential gene expression experiments, enrichment analysis [32] is often used to test whether two experiments share more significantly differentially expressed genes than would be expected by chance. In the integration of an expression quantitative trait loci study and a genome-wide association study, the goal is frequently to detect and identify genetic variants that are associated with both gene expression and disease [26,44]. Replicability analysis [27,29,30] aims to discover significant findings that have been replicated across genomic studies. Finally, studies of genetic pleiotropy investigate whether the same genetic variants may be simultaneously associated with different traits [8,11–13,22,38].

These examples broadly fall into two categories of questions: the detection of whether there exist features that are significant in both of two studies, and the identification of those simultaneously significant features. This paper focuses on the detection problem; the identification problem is studied elsewhere [11,30,45,55]. Specifically, let  $I_{kj}$  be unobserved latent indicators of whether the  $j$ th test,  $j \in \{1, \dots, p\}$ , is truly non-null in the  $k$ th study,  $k \in \{1, 2\}$ . Let  $T_{kj}$  be the corresponding test statistic such that, for  $k \in \{1, 2\}$ ,

$$T_{kj} \mid I_{kj} = 0 \sim F_k^0, \quad T_{kj} \mid I_{kj} = 1 \sim F_k^1, \quad I_{kj} \sim \text{Ber}(\pi_k), \quad (1)$$

\* Corresponding author.

E-mail addresses: [sdzhao@illinois.edu](mailto:sdzhao@illinois.edu) (S.D. Zhao), [tcai@wharton.upenn.edu](mailto:tcai@wharton.upenn.edu) (T.T. Cai), [hongzhe@upenn.edu](mailto:hongzhe@upenn.edu) (H. Li).

where the  $\pi_k$  quantify the proportion of non-null tests in each study. The  $F_k^0$  and  $F_k^1$  can be viewed as mixtures of possibly different null and non-null distributions for different  $j$ . For each  $k$ , model (1) corresponds to a two-group mixture model for  $T_{kj}$ , which is common in the literature [15,16,21,49,50]. It will be assumed that the  $T_{kj}$  are two-tailed test statistics and are thus stochastically larger when  $I_{kj} = 1$ . Because the two sequences of tests arise from different studies, which typically are conducted on independent samples, it is assumed that  $T_{1j}$  and  $T_{2j}$  are independent conditional on the latent indicators  $I_{1j}$  and  $I_{2j}$ .

The goal of this paper is to test whether there are more features  $j$  that are significant in both studies than would be expected by chance. Formally, if  $\Pr(I_{1j} = 1, I_{2j} = 1) = \epsilon$ , the goal is to test

$$\mathcal{H}_0 : \epsilon = \pi_1\pi_2 \quad \text{vs.} \quad \mathcal{H}_A : \epsilon > \pi_1\pi_2. \quad (2)$$

This is motivated by a study of genetic pleiotropy in 10 pediatric autoimmune diseases conducted by Hakonarson and colleagues at the Children's Hospital of Pennsylvania [41,42]. More details about the data can be found in Section 4.6. Testing (2) using genome-wide association study summary statistics from a pair of diseases can assess whether the two conditions have some degree of shared genetic architecture, which can lead to a better understanding of their etiologies.

Several features make testing (2) difficult for existing methods. First, the  $I_{kj}$  are not directly observed. Second, in genomics applications, non-null features are typically rare and have weak effect sizes. For example, only a relatively small proportion of the human genome is expected to be associated with a given phenotype, and then only weakly so. Finally, positive dependence between  $I_{1j}$  and  $I_{2j}$  can be very weak when it exists, because cross-study heterogeneity makes it unlikely that more than a handful of features will be simultaneously non-null in both of two independently conducted genomics studies, even if the studies are closely related.

This paper proposes a new test for (2) under these challenging conditions. The proposed test statistic is shown to be asymptotically adaptively optimal, so that it performs as well as the optimal likelihood ratio test statistic but without needing to specify parameter values under  $\mathcal{H}_0$  and  $\mathcal{H}_A$ . In fact the proposed test is entirely nonparametric, so neither  $F_k^0$  nor  $F_k^1$  needs to be known. It is also computationally efficient to implement and can be computed for 10 million pairs of tests in under one minute. It is available in the R package *ssa*.

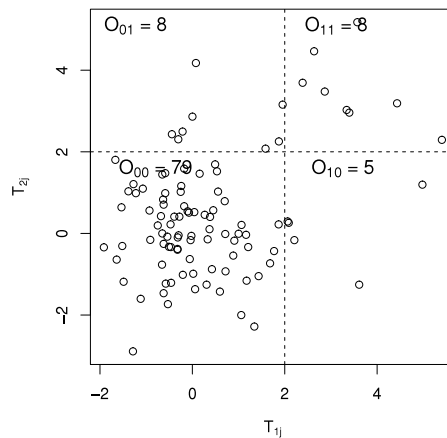
## 1.2. Related work

Because model (1) assumes that  $T_{1j}$  and  $T_{2j}$  are independent conditional on  $I_{1j}$  and  $I_{2j}$ , testing (2) is equivalent to testing for independence between  $T_{1j}$  and  $T_{2j}$ . Classical methods are based on goodness-of-fit tests comparing the empirical bivariate distribution of  $(T_{1j}, T_{2j})$  to the product of the marginal empirical distributions. Variations include Cramér–von Mises, Anderson–Darling, and Kolmogorov–Smirnov type tests [20,31,48,52]. A number of methods for detecting positive quadrant dependence have also been studied in the actuarial sciences [37]. Independence testing has seen renewed interest in the statistical literature, where the focus is on detecting arbitrary types of dependence [19,46,51]; see in particular Heller et al. [28]. In contrast, this paper is concerned with detecting a particular form of dependence between  $T_{1j}$  and  $T_{2j}$ , induced by the weak positive latent dependence between  $I_{1j}$  and  $I_{2j}$ . It appears that this type of dependence has not yet been specifically considered, and existing methods may be suboptimal. Furthermore, the fundamental limits of detection have not been studied.

Testing (2) can also be seen as an extension of the single-sequence signal detection problem. There, given test statistics  $T_{kj}$  from a single study  $k$ , the goal is to determine whether there are any non-null signals:  $\mathcal{H}_0 : \Pr(I_{kj} = 1) = 0$  vs.  $\mathcal{H}_A : \Pr(I_{kj} = 1) > 0$ . The fundamental limits of detection for this problem have been derived, and asymptotically adaptively optimal tests have also been developed [2,9,10,14,15,33–36]. Special attention has been paid to the setting where  $\pi_k$  is very close to zero and  $F_k^1$  is not too different from  $F_k^0$ . As previously noted, this rare and weak signal setting is also the focus of this paper. However, results for the single sequence problem do not apply to testing (2).

Several additional methods for testing (2) have been developed in the genomics literature. A popular approach is to estimate the  $I_{kj}$ , by thresholding the  $T_{kj}$ , and then to test for dependence using the estimated  $I_{kj}$  [32,47]. However, it is unclear how the thresholds on  $T_{kj}$  should be chosen. Alternatively, the GPA method [11] fits the  $(T_{1j}, T_{2j})$  to a four-group mixture model, each group corresponding to one of the four possible values of the tuple  $(I_{1j}, I_{2j})$ , and uses a generalized likelihood ratio test for (2). However, GPA imposes parametric assumptions on  $F_k^0$  and  $F_k^1$ . In addition, theoretical results from the single-sequence detection problem suggests that generalized likelihood ratio tests will have poor asymptotic properties when non-null  $T_{kj}$  are rare and weak [6,25]. Recently, Zhao et al. [56] proposed a simple test for (2) and studied its asymptotic properties. However, their theoretical results require distributional assumptions on the  $T_{kj}$ , and their test is only asymptotically optimal under specialized conditions.

The rest of the paper is organized as follows. Section 2 introduces the proposed test statistic and Section 3 studies its asymptotic adaptive optimality. Section 4 presents the results of simulation studies and the pediatric autoimmune disease analysis. The paper concludes with a discussion in Section 5. Additional simulation and data analysis results, and all proofs, can be found in the Supplementary Material; see Appendix A.



**Fig. 1.** The  $2 \times 2$  table induced in  $(T_{1j}, T_{2j}), j = 1, \dots, 100$ , generated according to (4), by the tuple  $(t_1, t_2) = (2, 2)$ . The cell counts are denoted by  $O_{lm}$ ,  $l, m = 0, 1$ .

## 2. Proposed method

### 2.1. Test statistic

Because testing (2) is equivalent to detecting dependence between  $T_{1j}$  and  $T_{2j}$ , let  $\hat{S}_{12}(t_1, t_2)$  and  $\hat{S}_k(t_k)$  denote the empirical bivariate and marginal survival functions, respectively:

$$\hat{S}_{12}(t_1, t_2) = \frac{1}{p} \sum_{j=1}^p \mathbf{1}(T_{1j} \geq t_1, T_{2j} \geq t_2), \quad \hat{S}_k(t_k) = \frac{1}{p} \sum_{j=1}^p \mathbf{1}(T_{kj} \geq t_k),$$

where  $k \in \{1, 2\}$ . The proposed test statistic is

$$\hat{\mathcal{D}} = \sup_{(t_1, t_2) \in \mathcal{S}} p^{1/2} \frac{|\hat{S}_{12}(t_1, t_2) - \hat{S}_1(t_1)\hat{S}_2(t_2)|}{\{\hat{S}_1(t_1)\hat{S}_2(t_2) - \hat{S}_1^2(t_1)\hat{S}_2^2(t_2)\}^{1/2}}, \quad (3)$$

where the set  $\mathcal{S}$  is defined as

$$\mathcal{S} = [T_{1(1)}, T_{1(p)}] \times [T_{2(1)}, T_{2(p)}] \setminus \{(T_{1(1)}, T_{2(1)})\},$$

and  $T_{k(j)}$  is the  $j$ th order statistics of the  $T_{kj}$ . This is the supremum version of an Anderson–Darling type goodness-of-fit test for independence between  $T_{1j}$  and  $T_{2j}$ , and is motivated by the higher criticism statistic of Donoho and Jin [15] for signal detection in a single sequence of multiple tests. Properties of an oracle version of the statistic  $\hat{\mathcal{D}}$ , where the  $\hat{S}_k$  are replaced by the true marginal survival functions, have been previously studied [17,18], but not in the present context of weak latent dependency detection. One advantage of  $\hat{\mathcal{D}}$  is that it makes no assumptions about the distributions  $F_k^0$  and  $F_k^1$ .

To better understand its properties, first consider the numerator  $\hat{S}_{12} - \hat{S}_1\hat{S}_2$ . This is a natural way to test for dependence between  $T_{1j}$  and  $T_{2j}$  and thus (2), but there is a useful alternative interpretation. Fig. 1 is a scatterplot of 100 realizations from the following data-generating mechanism:

$$\begin{aligned} T_{kj} | I_{kj} = 0 &\sim \mathcal{N}(0, 1), \quad T_{kj} | I_{kj} = 1 \sim \mathcal{N}(3, 1), \quad k \in \{1, 2\}, \\ \Pr(I_{1j} = 1, I_{2j} = 1) &= 0.1, \quad \Pr(I_{1j} = 1, I_{2j} = 0) = 0.05, \quad \Pr(I_{1j} = 0, I_{2j} = 1) = 0.05, \\ \Pr(I_{1j} = 0, I_{2j} = 0) &= 0.8. \end{aligned} \quad (4)$$

The figure illustrates that any tuple  $(t_1, t_2)$  divides the observed data into a  $2 \times 2$  contingency table. Blum et al. [7] recognized that the numerator is closely related to testing for independence using the cell counts of the  $2 \times 2$  table induced by  $(t_1, t_2)$ . Later, Thas and Ottoy [52] and most recently Heller et al. [28] extended this idea to  $m \times m$  tables for  $m \geq 2$ , which Heller et al. [28] showed can have greater power.

Next consider the supremum in  $\hat{\mathcal{D}}$ . It is difficult to know *a priori* which tuple  $(t_1, t_2)$  will induce the  $2 \times 2$  table that gives the largest test statistic. The optimal  $(t_1, t_2)$  depends on the distributions  $F_k^0$  and  $F_k^1$ , the proportions  $\pi_k$ , and the degree of dependence  $\epsilon$ . Thus  $\hat{\mathcal{D}}$  takes the supremum over all possible  $(t_1, t_2)$ , allowing it to adapt to any combination of these unknown parameters. Instead of the supremum, Thas and Ottoy [52] proposed a statistic that integrates over all tuples; their statistic turns out to be closely related to summing the Pearson chi-square test statistics calculated from each  $2 \times 2$  table induced by each of the observed tuples  $(T_{1j}, T_{2j})$ . Heller et al. [28] proposed several procedures that either sum or take the maximum over statistics arising from all possible  $m \times m$  tables, then combines these statistics across multiple choices for  $m$ .

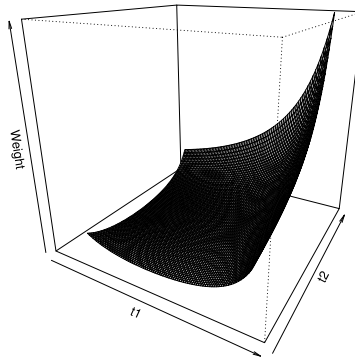


Fig. 2. Plot of the weight function  $\{S_1(t_1)S_2(t_2) - S_1^2(t_1)S_2^2(t_2)\}^{-1/2}$  for  $S_k(x) = 1 - x$ .

Finally, consider the denominator of  $\widehat{\mathcal{D}}$ . It is a natural standardizing weight in that it is the variance of  $\widehat{S}_{12}$  under the independence null hypothesis of (2). Furthermore, it is the reason why  $\widehat{\mathcal{D}}$  can have power for detecting even weak dependence. Fig. 2 plots the inverse of the denominator when the marginal survival functions are known and equal to  $S_k(x) = 1 - x$ . It is largest for large  $t_1$  and  $t_2$ , which corresponds to the upper right-hand quadrant of Fig. 1. This implies that  $\widehat{\mathcal{D}}$  can be large even when only a few points are observed in this quadrant, which will be the case when the  $T_{kj}$  are stochastically larger when  $I_{kj} = 1$  but only weakly dependent. Other denominators are also possible but may not be optimal for detecting weak positive latent dependence; see Section 4.4.

## 2.2. Inference

When the test statistics  $T_{kj}$  are independent across  $j$ , Einmahl [17] showed that the oracle statistic

$$\mathcal{D} = \sup_{-\infty < t_1, t_2 < \infty} p^{1/2} \frac{|\widehat{S}_{12}(t_1, t_2) - S_1(t_1)S_2(t_2)|}{\{S_1(t_1)S_2(t_2) - S_1^2(t_1)S_2^2(t_2)\}^{1/2}},$$

where the marginal survival functions are known, satisfies

$$\Pr_{\mathcal{H}_0}\{(\ln p)^{-1/2}\mathcal{D} > x\} \rightarrow 1 - \exp(-x^2)$$

under the null hypothesis  $\mathcal{H}_0$  of independence between  $I_{1j}$  and  $I_{2j}$ . However, this oracle result may not be applicable to the proposed  $\widehat{\mathcal{D}}$  (3). Furthermore, the convergence rates of these types of extreme values statistics are usually too slow to be useful [4,9,15].

Instead, this paper considers a simple permutation procedure to provide  $p$ -values. Fixing the indices of  $T_{2j}$ , randomly permute the indices of  $T_{1j}$  to induce independence between the two sequences of tests. Let  $\widehat{\mathcal{D}}^{(\ell)}$  be the proposed statistic (3) calculated after the  $\ell$ th permutation. Then the  $p$ -value after  $B$  permutations is  $\{1 + \sum_{\ell=1}^B \mathbf{1}(\widehat{\mathcal{D}}^{(\ell)} \geq \widehat{\mathcal{D}})\}/(B+1)$  [39]. Even large numbers of permutations are feasible because  $\widehat{\mathcal{D}}$  can be computed very quickly, as described below in Section 2.3.

In many genomic applications, the  $T_{kj}$  are likely to be dependent across  $j$ . For example, if each  $T_{kj}$  is the test statistic for association between genetic variant  $j$  and phenotype  $k$ , the  $T_{kj}$  will be correlated across  $j$  due to linkage disequilibrium. Interestingly, simulations with real genotype data in Section 4.3 indicate that using the random permutation  $p$ -value is still able to maintain type I error.

## 2.3. Implementation

A simple algorithm for calculating  $\widehat{\mathcal{D}}$  requires  $O(p^2)$  operations: the  $T_{1j}$  and  $T_{2j}$  are first sorted using quicksort, which on average requires  $O(p \ln p)$  operations and at most requires  $O(p^2)$ . Next, the algorithm iterates from the largest to the smallest order statistics  $T_{1(l)}$ , where for each  $l$  it iterates from the largest to the smallest  $T_{2(m)}$  in order to calculate

$$D_{\ell m} = p^{1/2} \frac{|\widehat{S}_{12}(T_{1(\ell)}, T_{2(m)}) - \widehat{S}_1(T_{1(\ell)})\widehat{S}_2(T_{2(m)})|}{\{\widehat{S}_1(T_{1(\ell)})\widehat{S}_2(T_{2(m)}) - \widehat{S}_1(T_{1(\ell)})^2\widehat{S}_2(T_{2(m)})^2\}^{1/2}}$$

for all  $\ell, m \in \{1, \dots, p\}$ . Finally,  $\widehat{\mathcal{D}} = \max_{\ell m} D_{\ell m}$ . This algorithm has been implemented in C in the R package *ssa*.

An additional computational shortcut can be implemented. Because the  $T_{kj}$  are stochastically larger when  $I_{kj} = 1$ , the largest  $D_{\ell m}$  is likely to be found when  $\ell$  and  $m$  are large. Therefore the algorithm only needs to iterate over  $T_{1(p-m_1+1)}, \dots, T_{1(p)}$  and  $T_{2(p-m_2+1)}, \dots, T_{2(p)}$ , where  $m_1$  and  $m_2$  can be close to  $p$ . Even if the true maximum  $D_{\ell m}$  is not attained for these test statistics, the largest  $D_{ij}$  in this restricted region may still be large enough to reject the null hypothesis. This truncated

calculation should at worst provide a conservative test, and  $m_1$  and  $m_2$  can be set as large as computationally feasible. As an example, this algorithm can calculate  $\widehat{\mathcal{D}}$  for  $p = 10^7$  and  $m_1 = m_2 = 10^4$  in 29 s on a laptop with a 2.5 GHz Intel Core i5 processor with 8 GB RAM.

### 3. Theoretical justification

#### 3.1. Assumptions

As introduced in Section 1, for any feature  $j$  the observed  $T_{1j}$  and  $T_{2j}$  are assumed to follow model (1). Because they are derived from two different studies, they will be independent conditional on the  $I_{kj}$ . They are also assumed to be two-tailed test statistics and thus stochastically larger when  $I_{kj} = 1$  than when  $I_{kj} = 0$ .

**Assumption 1.** For  $k = 1, 2$ ,  $F_k^1(t) \leq F_k^0(t)$  for all  $t$ .

The dependency detection problem (2) and the proposed test statistic  $\widehat{\mathcal{D}}$  (3) will be studied under the asymptotic testing framework [39], where the asymptotics apply to the total number of tests  $p$ . This is meaningful because in practice  $p$  can be very large, such as in applications to genome-wide association studies. If the parameters  $\epsilon$ ,  $\pi_k$ ,  $F_k^0$ , and  $F_k^1$ ,  $k \in \{1, 2\}$  were fixed with  $p$ , any reasonable test would be able to distinguish  $\mathcal{H}_0$  from  $\mathcal{H}_A$ . Instead, the parameters will be calibrated to vary with  $p$ . This allows for a more meaningful comparison between possible testing procedures, and in addition formalizes the setting of weak positive latent dependence and rare and weak signals, described in Section 1.

Specifically,  $\epsilon$  and  $\pi_k$  will be calibrated to approach 0, which models weak dependence and rare signals:

$$\begin{aligned} \pi_k &= p^{-\beta_k}, & 1/2 \leq \beta_k \leq 1, & \quad k \in \{1, 2\}, \\ \epsilon &= \pi_1 \pi_2 + p^{-\beta}, & 1/2 < \beta < 1, & \quad (\beta_1 \vee \beta_2) \leq \beta. \end{aligned} \quad (5)$$

In genomics problems, typically very few of the  $T_{kj}$  are non-null, which is reflected in the regime  $1/2 \leq \beta_k$  [9,10,15]. Analogously, this paper models weak dependence by letting  $\beta > 1/2$ . The additional restriction  $\beta \geq \beta_k$  ensures that  $\epsilon \leq (\pi_1 \wedge \pi_2)$ .

Given (5),  $F_k^1$  must be calibrated to separate from  $F_k^0$ , otherwise testing (2) would be very difficult. This divergence will be expressed in terms of the likelihood ratio between the two distributions. Because no parametric assumptions are made on  $F_k^1$  and  $F_k^0$ , the exact form of this calibration is fairly abstract. Let  $f_k^1$  and  $f_k^0$  be the corresponding density functions and let  $x \vee y$  denote  $\max(x, y)$ .

**Assumption 2.** There exist measurable functions  $\alpha_k^-, \alpha_k^+ : \mathbb{R} \rightarrow \mathbb{R}$  such that  $\alpha_k(a) = \alpha_k^-(a) \vee \alpha_k^+(a) > 0$  on a set of positive Lebesgue measure and that for  $k \in \{1, 2\}$ , the log-likelihood ratios  $\ell_k = \ln(f_k^1/f_k^0)$  satisfy

$$\lim_{p \rightarrow \infty} \frac{\ell_k \{(F_k^0)^{-1}(p^{-a})\}}{\ln(p)} = \alpha_k^-(a), \quad \lim_{p \rightarrow \infty} \frac{\ell_k \{(F_k^0)^{-1}(1 - p^{-a})\}}{\ln(p)} = \alpha_k^+(a),$$

uniformly in  $a \geq \log_p 2$ .

Assumption 2 guarantees the existence of limiting functions  $\alpha_k^-$  and  $\alpha_k^+$  that characterize the likelihood ratios at small and large values, specifically  $p^{-a}$  and  $1 - p^{-a}$ . The assumption essentially calibrates the likelihood ratios to grow only polynomially in  $p$ , which models weak signals. Restricting  $a \geq \log_p 2$  is necessary because otherwise the  $\alpha$  functions would simply be reparametrizations of each other. Since  $p^{-\log_p 2} = 1 - p^{-\log_p 2} = 0.5$ ,  $p^{-a}$  and  $1 - p^{-a}$  correspond to numbers smaller and larger than the median of  $F_k^0$ , respectively. The value of separately characterizing the likelihood ratios on the left and right sides of the null median will become clear in the theoretical results in Section 3.2.

Assumption 2 was used in Cai and Wu [10] in their study of the single-sequence detection problem and generalizes similar assumptions made in previous work. For example, suppose  $F_k^0 \equiv \mathcal{N}(0, 1)$ . Then  $\Phi\{-(2a \ln p)^{1/2}\} \approx p^{-a}$  as long as  $2a \ln p$  is sufficiently large, which is guaranteed by the condition  $a \geq \log_p 2$ . Therefore the  $(p^{-a})$ th quantile of  $F_k^0$  is  $-(2a \ln p)^{1/2}$ , and by similar reasoning the  $(1 - p^{-a})$ th quantile is  $(2a \ln p)^{1/2}$ . The setting of  $F_k^1 \equiv \mathcal{N}\{(2r_k \ln p)^{1/2}, 1\}$ , a popular model for weak signals [9,15,33–35], can be shown to correspond to

$$\alpha_k^-(a) = -2(ar_k)^{1/2} - r_k, \quad \alpha_k^+(a) = 2(ar_k)^{1/2} - r_k \quad (6)$$

in the notation of Assumption 2.

Finally, for the purpose of studying the asymptotic properties of  $\widehat{\mathcal{D}}$ , it will be assumed that in each sequence of tests, the test statistics are mutually independent. This is a simplification, but for dependent tests the asymptotic theory of these types of detection problems is still under development for arbitrary correlation structures [1,3,24,43]. In contrast, the theoretical properties when tests are independent are well understood, at least for single-sequence problems [9,10,15]. To facilitate comparison with these established results, this paper assumes that  $T_{kj}$  and  $T_{kj'}$  are independent for  $j \neq j'$ , and leaves consideration of dependent tests for future work.

### 3.2. Asymptotic properties

For the proposed  $\widehat{\mathcal{D}}$  (3), consider the test

$$\text{reject } \mathcal{H}_0 \text{ of (2) if } \widehat{\mathcal{D}} > \ln p(\ln \ln p)^2 + 3(\ln \ln p)^2. \quad (7)$$

The critical value  $\ln p(\ln \ln p)^2 + 3(\ln \ln p)^2$  is chosen such that test (7) can achieve type I and type II errors that sum to zero as  $p \rightarrow \infty$ ; this will be shown below. Furthermore, it will also be shown that test (7) is in a certain sense asymptotically optimal among all possible tests. These results support the use of the proposed  $\widehat{\mathcal{D}}$  for detecting weak positive latent dependence.

**Theorem 1** characterizes a region of the parameter space where test (7) will be successful. This region can be expressed in terms of  $\beta_k$  and  $\beta$  from calibration (5) and  $\alpha_k^-$  and  $\alpha_k^+$  from Assumption 2.

**Theorem 1.** Suppose  $F_k^0 \neq F_k^1$ ,  $k = 1, 2$  and define

$$v_k^-(x) = \text{ess sup}_{a \geq x} \{\alpha_k^-(a) - a\}, \quad v_k^+(x) = \text{ess sup}_{a \geq x} \{\alpha_k^+(a) - a\}.$$

Under calibration (5) and Assumption 2, the sum of the type I and II errors of (7) goes to 0 if one of the following is true:

$$\begin{aligned} & \sup_{\substack{x_1, x_2 > 0, \\ x_1 + x_2 < 1}} \left( \frac{1}{2} - \beta + \sum_{k=1}^2 \left\{ (-x_k) \vee v_k^+(x_k) + \frac{x_k \wedge \{\beta_k - v_k^+(x_k)\}}{2} \right\} \right) > 0, \text{ or} \\ & \sup_{\substack{x_1, x_2 > 0, \\ x_2 < 1}} \left[ \frac{1}{2} - \beta + (-x_1) \vee v_1^-(x_1) + (-x_2) \vee v_2^+(x_2) + \frac{x_2 \wedge \{\beta_2 - v_2^+(x_2)\}}{2} \right] > 0, \text{ or} \\ & \sup_{\substack{x_1, x_2 > 0, \\ x_1 < 1}} \left[ \frac{1}{2} - \beta + (-x_1) \vee v_1^+(x_1) + (-x_2) \vee v_2^-(x_2) + \frac{x_1 \wedge \{\beta_1 - v_1^+(x_1)\}}{2} \right] > 0, \text{ or} \\ & \sup_{x_1, x_2 > 0} \left\{ \frac{1}{2} - \beta + (-x_1) \vee v_1^- + (-x_2) \vee v_2^- + \frac{x_1 \wedge \beta_1 \wedge x_2 \wedge \beta_2}{2} \right\} > 0. \end{aligned} \quad (8)$$

It is also possible to derive the fundamental limits of detecting weak positive latent dependence (2). Theorem 2 characterizes a region of the parameter space where successful detection is impossible, in the sense that the sum of the type I and II errors of any hypothesis test of (2) goes to at least 1 as  $p \rightarrow \infty$ . It involves the essential supremum, which for a measurable function  $f$  and a measure  $\mu$  is defined as

$$\text{ess sup}_x f(x) = \inf[a \in \mathbb{R} : \mu\{f(x) > a\} = 0].$$

Here, essential suprema are taken with respect to the Lebesgue measure.

**Theorem 2.** Suppose  $F_k^0 \neq F_k^1$ ,  $k = 1, 2$ . Under calibration (5) and Assumption 2, the sum of the type I and II errors of any test goes to at least 1 if each of the following holds:

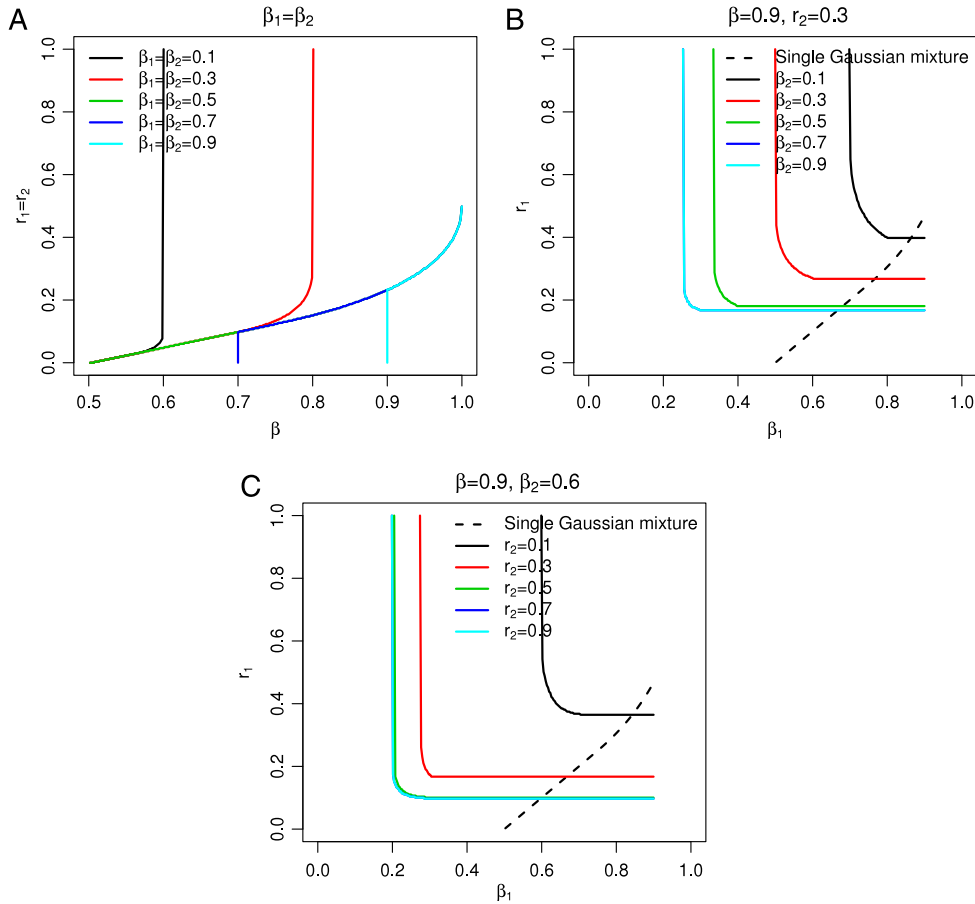
$$\begin{aligned} & 1 - 2\beta + \text{ess sup}_{a > 0} \{\alpha_k(a) + \alpha_k(a) \wedge \beta_k - a\} < 0, \quad k \in \{1, 2\}, \text{ and} \\ & 1 + \text{ess sup}_{a_1, a_2 > 0} \{-\beta + \alpha_1(a_1) + \alpha_2(a_2)\} \wedge \\ & \{-2\beta + \alpha_1(a_1) + \alpha_2(a_2) + \alpha_1(a_1) \wedge \beta_1 + \alpha_2(a_2) \wedge \beta_2\} - a_1 - a_2 < 0, \end{aligned}$$

where  $\alpha_k(a) = \alpha_k^-(a) \vee \alpha_k^+(a)$  as defined in Assumption 2.

When the  $T_{kj}$  are stochastically ordered according to Assumption 1, it turns out that the union of the two regions defined in Theorems 1 and 2, and the boundary that separates them, constitutes the entire parameter space. In other words, this boundary, called the detection boundary, partitions the parameter space into two regions. In the undetectable region, successful detection is impossible for any test, while in the detectable region, there exists a test, namely (7), that can perfectly separate  $\mathcal{H}_0$  and  $\mathcal{H}_A$ .

**Theorem 3.** Under Assumption 1, the region described by Theorem 2 is the interior of the complement of the region described by Theorem 1. In particular, the detectable region is entirely described by inequality (8).

The asymptotic optimality of the proposed  $\widehat{\mathcal{D}}$  is encapsulated in Theorem 3. It implies that whenever detection of weak positive latent dependence is possible, (7) already achieves asymptotically zero error. In other words, it can perform as well as the optimal likelihood ratio test, but has the added benefit that it is entirely data-driven and automatically adapts to the unknown values of  $\beta$ ,  $\beta_k$ ,  $F_k^0$ , and  $F_k^1$  under both  $\mathcal{H}_0$  and  $\mathcal{H}_A$ .



**Fig. 3.** Detection boundary for normally distributed  $T_{kj}$  following (9). Dotted line corresponds to the detection boundary for the single sequence of test statistics  $T_{ij}$  when  $\beta_1 > 1/2$ . Panel A fixes  $r_1 = r_2$  and  $\beta_1 = \beta_2$ . Panel B shows how the boundary varies with  $\beta_2$ , and panel C shows how it varies with  $r_2$ . The region below each colored line is the undetectable region.

For a concrete example of the detection boundary, suppose that

$$F_k^0 \sim \mathcal{N}(0, 1), \quad F_k^1 \sim \mathcal{N}((2r_k \ln p)^{1/2}, 1) \quad (9)$$

for some positive constants  $r_k, k \in \{1, 2\}$ , which satisfies Assumptions 1 and 2. The corresponding  $\alpha$  functions, which appear in the inequalities from Theorem 2, were presented above in (6). Then the detection boundary can be illustrated by plotting the boundary of the undetectable region. This is shown in Fig. 3 for various values of  $\beta, \beta_k$ , and  $r_k$ . It is interesting to compare these results to the boundary for detecting sparse mixtures in a single sequence of tests, e.g., testing  $\mathcal{H}_0 : \pi_1 = 0$ , which was computed under (9) by Donoho and Jin [15] and is plotted in Fig. 3.

### 3.3. Implications

Theorems 1–3 and Fig. 3 reveal a number of interesting features that decide the difficulty of testing weak positive latent dependence (2). Most obviously, detection is easier for smaller  $\beta$ , corresponding to stronger dependence. It is also in general easier for larger  $\alpha_k(a)$  and  $v_k(x)$ , which correspond to larger differences between the null and alternative distributions. To illustrate this, for normally distributed signals (9) it can be shown that  $v_k^+(x) = -(x^{1/2} - r_k^{1/2})_+^2$ . For large signal strengths  $r_k \geq 1$ ,  $v_k^+(x) = 0$  on  $x \in (0, 1)$ , so by Theorem 3 and inequality (8) the detectable region is

$$0 < \sup_{\substack{x_1, x_2 > 0, \\ x_1 + x_2 < 1}} \left( \frac{1}{2} - \beta + \sum_{k=1}^2 \frac{x_k \wedge \beta_k}{2} \right) = \frac{1}{2} - \beta + \frac{1 \wedge (\beta_1 + \beta_2)}{2}.$$

This implies that for strong signals, and when the individual latent indicator sequences  $I_{kj}$  are sufficiently sparse such that  $\beta_1 + \beta_2 > 1$ , any  $\epsilon = \pi_1 \pi_2 + p^{-\beta}$  is detectable. In this setting it would be more interesting to calibrate  $\epsilon$  to approach  $\pi_1 \pi_2$  at faster than a polynomial rate.



Another implication is that for fixed  $\beta$  and  $\alpha_k(a)$ , dependency detection is more difficult for smaller  $\beta_k$ . Even when  $r_k \geq 1$ , the previous inequality shows that dependence may be undetectable if  $\beta > (1 + \beta_1 + \beta_2)/2$ . When there are many non-null signals in the two sequences of test statistics, many features with both  $I_{1j} = 1$  and  $I_{2j} = 1$  are necessary to provide significant evidence for dependence, even if the  $I_{kj}$  were directly observed.

Finally, Fig. 3 reveals an interesting connection to the single-sequence sparse mixture detection problem. First, since signals must exist in both sequences of test statistics for there to exist dependence, a test for weak dependence such as (7) can also be used as a method to detect sparse mixtures in a single sequence of test statistics. Second, panels B and C of Fig. 3 show that a portion of the undetectable region of the single-sequence problem lies within the detectable region of dependency detection. This means that the proposed test (7) using  $T_{1j}$  and  $T_{2j}$  can actually detect signal in one of the sequences even when detection is theoretically impossible using that sequence alone. Intuitively, this can occur when the non-null signals of one sequence, say the  $T_{2j}$ , are strong enough to be easily identified. Then dependency could be detected simply by checking only the  $T_{1j}$  paired with the non-null  $T_{2j}$  to see if they are also non-null. This greatly reduces the dimensionality of the problem, and so could succeed even if the non-null signals in the  $T_{1j}$  are so weak that they cannot be detected by single-sequence methods.

## 4. Numerical results

### 4.1. Methods studied

The proposed statistic  $\widehat{D}$  (3) was compared to several other existing procedures for testing (2). Spearman's correlation is the most straightforward naive approach. Brownian distance covariance [51] is a recently developed nonparametric method designed for omnibus power. The GPA method [11] was specifically developed for test statistics following model (1), though it was designed for strong rather than weak dependence and makes parametric assumptions on the  $T_{kj}$ , namely  $F_k^0 \sim \mathcal{U}(0, 1)$  and  $F_k^1 \sim \mathcal{B}(\alpha_1, \alpha_2)$ . The  $M_{m \times m}^{DDP}$  test of Heller et al. [28] generalize several classical tests for independence. It calculates the Pearson chi-square test statistics for independence across all possible  $m \times m$  contingency tables induced by the observed  $(T_{1j}, T_{2j})$ , as illustrated in Fig. 1, and aggregates them by taking their maximum. It then combines this max statistic across all  $m \in \{2, \dots, M\}$ . For computational reasons, in these simulations  $M$  was set to equal 3. Finally, the method of Zhao et al. [56], referred to here as the max test, tests (2) using  $\max_j \{\min(T_{1j}, T_{2j})\}$  and provides a closed-form expression for the permutation  $p$ -value. Two hundred permutations were used to calculate  $p$ -values for the  $\widehat{D}$ , Brownian distance covariance, and  $M_{m \times m}^{DDP}$  tests.

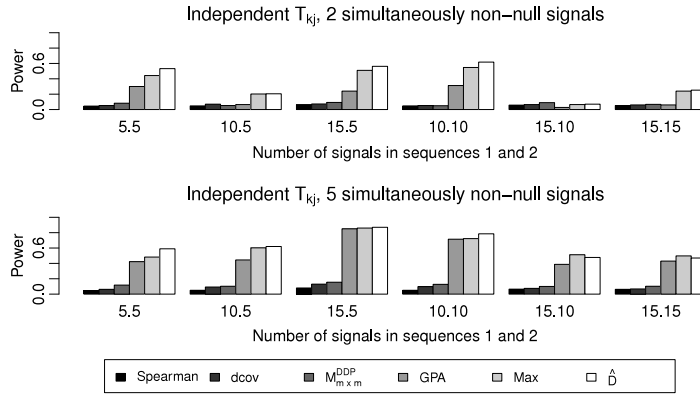
All simulations were conducted under a “fixed-effect” sampling scheme, where the non-null indicators  $I_{kj}$  were generated once and then fixed across replications. This was done because in many applications, for example in statistical genomics, whether or not a genomic feature exhibits a non-null effect does not change across repeated sampling. To generate the  $I_{kj}$ , under  $\mathcal{H}_0$ ,  $p\pi_k$  of the  $I_{kj}$  were randomly set to 1, independently for  $k = 1$  and  $k = 2$ . Under  $\mathcal{H}_A$ ,  $p\epsilon$  of the features were randomly chosen to be simultaneously non-null in both sequences, with  $I_{1j} = I_{2j} = 1$ , while maintaining a total of  $p\pi_k$  non-null signals in each sequence. Finally, conditional on the  $I_{kj}$ ,  $T_{kj}$  were generated according to the mixture model (1). All simulations were conducted under the rare and weak signal setting, as described in Section 1, where the number of non-null signals, as well as their effect sizes, are small.

### 4.2. Independent tests

These simulations consider test statistics  $T_{kj}$  that are independent across  $j$ . Null and non-null signals were generated according to  $T_{kj} \sim |\mathcal{N}(0, 1)|$  and  $T_{kj} \sim |\mathcal{N}(\mu_{kj}, \sigma_{kj}^2)|$ , respectively. To set the parameters of the non-null distribution, the  $\mu_{kj}$  were generated from  $\mathcal{N}(2.5, 1)$  and the  $\sigma_{kj}^2$  were generated from a Gamma distribution with shape equal to 2 and scale equal to 1. These parameters, like the  $I_{kj}$ , were generated once and then fixed across all replications. The total number of features,  $p = 10^3$ , was relatively small in order to accommodate the computationally intensive nature of the distance covariance and  $M_{m \times m}^{DDP}$  methods. The proposed statistic  $\widehat{D}$  (3) could therefore be calculated without using the truncated version described in Section 2.3. To implement GPA, which requires  $p$ -values as input, the  $T_{kj}$  were transformed according to  $2\Phi(-T_{kj})$ .

Table 1 reports the empirical type I errors for simulation settings with different numbers of non-null tests in each sequence of test statistics. The proposed method was able to control the type I error at the nominal  $\alpha = 0.05$  level. Fig. 4 reports the empirical powers under various simulation settings. Detecting dependence was easier for all methods when there were more simultaneous signals, corresponding to smaller  $\beta$  from calibration (5). The proposed  $\widehat{D}$  had the highest power in almost all settings. Figure A1 in the Supplementary Material in Appendix A plots the powers versus the number of simultaneous signals when there were 15 non-null signals in each sequence. GPA had the highest power under strong dependence, when there were many simultaneous signals, but  $\widehat{D}$  was the most powerful method under weak dependence. The proposed method was closely matched by the max test of Zhao et al. [56] under weak dependence but outperformed the max test when there were more than 10 simultaneous signals.





**Fig. 4.** Empirical powers for  $p = 10^3$  independent tests at nominal significance level  $\alpha = 0.05$  over 400 replications. dcov = Brownian distance covariance;  $M_{m \times m}^{DDP}$  = max aggregation method of Heller et al. [28]; GPA = method of Chung et al. [11]; Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

**Table 1**

Empirical type I errors for  $p = 10^3$  independent tests at nominal significance level  $\alpha = 0.05$  over 400 replications. dcov = Brownian distance covariance;  $M_{m \times m}^{DDP}$  = max aggregation method of Heller et al. [28]; GPA = method of Chung et al. [11]; Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

	Number of signals in sequences 1 and 2					
	(5,5)	(10,5)	(15,5)	(10,10)	(15,10)	(15,15)
Spearman	0.04	0.04	0.06	0.05	0.06	0.06
dcov	0.05	0.05	0.07	0.05	0.06	0.05
$M_{m \times m}^{DDP}$	0.06	0.05	0.09	0.05	0.09	0.07
GPA	0.02	0.01	0.01	0.01	0.01	0.01
Max	0.07	0.04	0.02	0.04	0.03	0.03
$\hat{D}$	0.04	0.03	0.04	0.03	0.02	0.02

**Table 2**

Empirical type I errors for  $p = 10^3$  dependent tests at nominal significance level  $\alpha = 0.05$  over 400 replications. dcov = Brownian distance covariance;  $M_{m \times m}^{DDP}$  = max aggregation method of Heller et al. [28]; GPA = method of Chung et al. [11]; Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

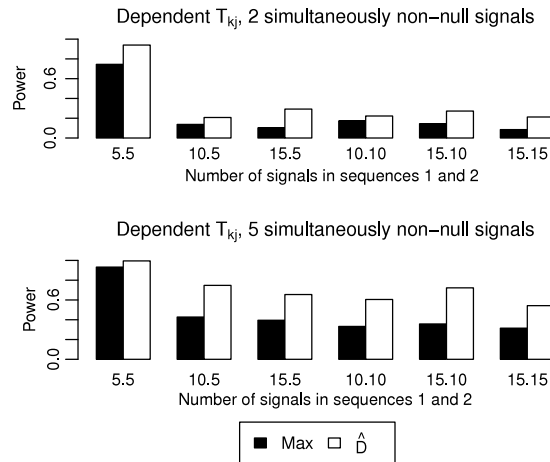
	Number of signals in sequences 1 and 2					
	(5,5)	(10,5)	(15,5)	(10,10)	(15,10)	(15,15)
Spearman	0.23	0.26	0.21	0.24	0.22	0.26
dcov	0.35	0.41	0.46	0.45	0.41	0.49
$M_{m \times m}^{DDP}$	0.46	0.51	0.60	0.56	0.54	0.64
GPA	0.06	0.21	0.24	0.33	0.25	0.34
Max	0.01	0.01	0.00	0.00	0.01	0.00
$\hat{D}$	0.04	0.05	0.03	0.01	0.03	0.04

#### 4.3. Dependent tests

These simulations generate  $T_{kj}$  that are dependent across  $j$ . The total number of features was again  $p = 10^3$ . Realistic correlation structures were generated using real genotype data from a randomly chosen set of  $p$  adjacent genetic variants on human chromosome 1, obtained from the pediatric autoimmune disease data discussed in Section 4.6.

In each replication,  $n = 200$  subjects from these data were selected at random to serve as data from hypothetical study  $k = 1$ , and another  $n = 200$  were independently selected to serve as data from hypothetical study  $k = 2$ . To generate test statistics  $T_{kj}$  from these studies, simulated outcomes  $Y_k$  were first generated according to linear models  $Y_k = S_k \theta_k + \varepsilon_k$ , where the  $S_k$  were  $n \times p$  matrices of additively coded genotypes of all variants, the  $\theta_k = (\theta_{k1}, \dots, \theta_{kp})^\top$  were  $p \times 1$  coefficient vectors, and the  $\varepsilon_k$  were  $n \times 1$  vectors of independent standard normal errors. The  $\theta_{kj}$  corresponding to variants with  $I_{kj} = 0$  were set to zero. The remaining non-zero  $\theta_{kj}$ , corresponding to variants with  $I_{kj} = 1$ , were generated from  $\mathcal{N}(0.5, 0.2)$  and then randomly multiplied by either 1 or  $-1$ . All  $\theta_{kj}$  were generated once and then fixed across all replications. Finally, the  $T_{kj}$  were taken to be the absolute values of the Z-statistics for the marginal regressions of  $Y_k$  on the  $j$ th variant.

Table 2 reports the empirical type I errors under different simulation settings for dependent test statistics. It is interesting that the proposed  $\hat{D}$ , which uses the simple permutation procedure described in Section 2.2, was still able to control the



**Fig. 5.** Empirical powers for  $p = 10^3$  dependent tests at nominal significance level  $\alpha = 0.05$  over 400 replications. Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

**Table 3**

Empirical type I errors for  $p = 10^3$  independent tests at nominal significance level  $\alpha = 0.05$  over 400 replications for variations of the procedures.  $M_{m \times m}^{DDP}$  = max aggregation method of Heller et al. [28];  $S_{m \times m}^{DDP}$  = sum aggregation method of Heller et al. [28];  $\hat{D}$  = statistic (10);  $\hat{D}_x$  = truncated version of the proposed method with  $m_1 = m_2 = x$ ;  $\hat{D}$  = proposed method without truncation.

	Number of signals in sequences 1 and 2					
	(5,5)	(10,5)	(15,5)	(10,10)	(15,10)	(15,15)
$M_{m \times m}^{DDP}$	0.06	0.05	0.09	0.05	0.09	0.07
$S_{m \times m}^{DDP}$	0.05	0.04	0.05	0.04	0.05	0.04
$\hat{D}$	0.03	0.05	0.04	0.04	0.02	0.05
$\hat{D}_{10}$	0.04	0.05	0.03	0.03	0.01	0.03
$\hat{D}_{100}$	0.04	0.03	0.04	0.03	0.02	0.02
$\hat{D}$	0.04	0.03	0.04	0.03	0.02	0.02

type I error in this setting. The only other method able to achieve this was the max test of Zhao et al. [56]. Fig. 5 reports the empirical powers and power curves of only those methods with proper type I error control. The proposed  $\hat{D}$  was consistently more powerful than the max test. Figure A2 in the Supplementary Material in Appendix A plots the power curves as a function of the number of simultaneous signals, and  $\hat{D}$  was the most powerful at all levels of dependence.

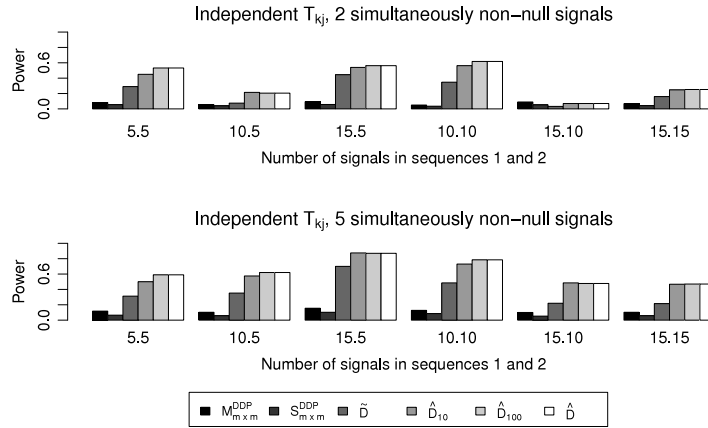
#### 4.4. Alternative dependency detection procedures

Several variants of the compared dependency detection procedures were also explored. First, truncated versions of the proposed  $\hat{D}$ , described in Section 2.3, can be calculated with different truncation parameters  $m_1$  and  $m_2$ . Next, instead of taking the maximum of the Pearson test statistics from all induced  $m \times m$  tables, Heller et al. [28] also proposed the sum aggregation test  $S_{m \times m}^{DDP}$ , which adds them. Finally, define the test statistic

$$\hat{D} = \sup_{(t_1, t_2) \in \mathcal{S}} p^{1/2} \frac{|\hat{S}_{12}(t_1, t_2) - \hat{S}_1(t_1)\hat{S}_2(t_2)|}{[\hat{S}_1(t_1)\{1 - \hat{S}_1(t_2)\}\hat{S}_2(t_2)\{1 - \hat{S}_2(t_2)\}]^{1/2}}. \quad (10)$$

Unlike the denominator  $\hat{D}$ , which as discussed in Section 2.1 favors tuples  $(t_1, t_2)$  where both  $t_1$  and  $t_2$  are large, the denominator of  $\hat{D}$  gives higher weights whenever both  $t_1$  and  $t_2$  are both extreme, regardless of whether they are extremely large or extremely small. This denominator also makes  $\hat{D}$  closely related to the maximum of the square roots of Pearson chi-square test statistics [52]. Two hundred permutations were used to calculate  $p$ -values for each of these methods.

These variations were applied to the independent test statistic simulations from Section 4.2. Table 3 indicates that most were able to maintain the nominal type I error rate. From Fig. 6, and the power curves in Figure A3 in the Supplementary Material in Appendix A,  $\hat{D}$  was the best performer and had the same power as the truncated  $\hat{D}_{100}$ , which had truncation parameters  $m_1 = m_2 = 100$ . The more heavily truncated  $\hat{D}_{10}$ , with  $m_1 = m_2 = 10$ , was slightly less powerful, especially



**Fig. 6.** Empirical powers for  $p = 10^3$  independent tests at nominal significance level  $\alpha = 0.05$  over 400 replications for variations of the procedures.  $M_{m \times m}^{DDP}$  = max aggregation method of Heller et al. [28];  $S_{m \times m}^{DDP}$  = sum aggregation method of Heller et al. [28];  $\tilde{D}$  = statistic (10);  $\hat{D}_x$  = truncated version of the proposed method with  $m_1 = m_2 = x$ ;  $\hat{D}$  = proposed method without truncation.

in the presence of a large number of simultaneous signals, but was still the best of the remaining procedures. This suggests that significant computational speedup can be achieved without sacrificing much power. The modified statistic  $\tilde{D}$  was the next best performer. Max aggregation  $M_{m \times m}^{DDP}$  always outperformed sum aggregation  $S_{m \times m}^{DDP}$  and had more power than  $\tilde{D}$  under strong dependence.

#### 4.5. Detection of single-sequence sparse mixture

As discussed in Section 3.3 and illustrated in Fig. 3, one implication of the detection boundary results is that dependency detection is sometimes possible when single-sequence signal detection is not. This section studies this phenomenon in simulations using  $p = 10^5$  pairs of test statistics.

The number of non-null signals in the first sequence of test statistics was either 282, 100 or 32. This corresponds to  $\beta_1$  from (5) equal to either 0.51, 0.6, or 0.7. The  $T_{1j}$  were generated following

$$T_{1j} \mid I_{1j} = 0 \sim |\mathcal{N}(0, 1)|, \quad T_{1j} \mid I_{1j} = 1 \sim |\mathcal{N}[(2\beta_1 - 1) \ln p]^{1/2}, 1|.$$

Existing results for the single-sequence detection problem imply that for these  $T_{1j}$ , it is impossible to detect the presence of non-null signals using single-sequence detection methods [9,10,15,33]. The second sequence of test statistics was generated with 316 non-null signals, corresponding to  $\beta_2 = 0.5$ . The  $T_{2j}$  followed

$$T_{2j} \mid I_{2j} = 0 \sim |\mathcal{N}(0, 1)|, \quad T_{2j} \mid I_{2j} = 1 \sim |\mathcal{N}[(2 \ln p)^{1/2}, 1]|,$$

so that the non-null signals were very strong. Finally, under  $\mathcal{H}_A$ , the dependency parameter  $\beta$  was set to  $\beta_1 \vee \beta_2 + 0.01$ , corresponding to either 251, 89, or 28 signals that were non-null in both sequences.

The distance covariance method of Székely et al. [51] and the  $M_{m \times m}^{DDP}$  test of Heller et al. [28] were not implemented for computational reasons, as  $p$  is quite large in these simulations. The remaining dependency detection procedures were applied for the purpose of testing  $\mathcal{H}_0 : \pi_1 = 0$ . For comparison, the higher criticism method was also applied to the  $T_{1j}$ . Donoho and Jin [15] showed that for these simulation settings, the higher criticism statistic is asymptotically adaptively optimal among all single-sequence detection methods. Its null distribution was approximated using 200 simulated realizations of  $p$  standard normals, and this distribution was used to provide  $p$ -values.

Table 4 reports the empirical type I errors and powers for different values of  $\beta_1$ . The type I error refers to the null hypothesis of independence between  $T_{1j}$  and  $T_{2j}$ , so it does not apply to higher criticism because it does not test independence. The proposed  $\hat{D}$  and the max test of Zhao et al. [56] both had substantial power to detect dependence, and thus to detect signal in  $T_{1j}$ , even when higher criticism did not.

#### 4.6. Application to pediatric autoimmune disease

Different autoimmune diseases can be genetically related, meaning that there are genetic variants which are associated with more than one disease. The proposed  $\hat{D}$  can be used to rigorously test the degree to which a pair of conditions are genetically related. Let  $P_{kj}$  be the  $p$ -value for association between the  $j$ th variant and the  $k$ th disease,  $k = 1, 2$ . Testing for weak positive latent dependence (2) between the  $P_{kj}$  is equivalent to testing whether there are more markers that affect both diseases than expected by chance.

**Table 4**

Empirical type I errors and powers for single-sequence signal detection for  $p = 10^5$  tests at nominal significance level  $\alpha = 0.05$  over 400 replications. HC = higher criticism method of Donoho and Jin [15]; GPA = method of Chung et al. [11]; Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

$\beta_1$ :	Type I errors			Powers		
	0.51	0.6	0.7	0.51	0.6	0.7
HC	—	—	—	0.04	0.07	0.14
Spearman	0.04	0.04	0.06	0.04	0.07	0.07
GPA	0.00	0.00	0.00	0.00	0.46	0.20
Max	0.06	0.05	0.05	0.13	0.63	0.80
$\hat{D}$	0.06	0.06	0.05	0.14	0.61	0.72

**Table 5**

Pairs of pediatric autoimmune diseases for which at least one testing method was significant at the 0.05 level after Bonferroni or Benjamini–Hochberg (BH) correction. Bold  $p$ -values are less than 0.05/45, and bold BH-corrected  $p$ -values are less than 0.05. Disorders: AS = ankylosing spondylitis; CEL = Celiac's disease; CD = Crohn's disease; CVID = common variable immunodeficiency; JIA = juvenile idiopathic arthritis; SLE = systemic lupus erythematosus; T1D = type I diabetes; THY = thyroiditis; UC = ulcerative colitis. Methods: GPA = method of Chung et al. [11]; Max = method of Zhao et al. [56];  $\hat{D}$  = proposed method.

Disorders	p-values				BH-corrected p-values			
	Spearman	GPA	Max	$\hat{D}$	Spearman	GPA	Max	$\hat{D}$
UC–CD	<b>0.0000</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0001</b>	<b>0.0000</b>	<b>0.0000</b>	<b>0.0004</b>	<b>0.0015</b>
CVID–JIA	0.0137	<b>0.0000</b>	<b>0.0000</b>	<b>0.0001</b>	0.0883	<b>0.0000</b>	<b>0.0002</b>	<b>0.0015</b>
UC–JIA	0.8101	<b>0.0004</b>	<b>0.0000</b>	<b>0.0001</b>	0.8867	<b>0.0033</b>	<b>0.0002</b>	<b>0.0015</b>
UC–T1D	<b>0.0007</b>	<b>0.0000</b>	<b>0.0002</b>	<b>0.0002</b>	<b>0.0080</b>	<b>0.0000</b>	<b>0.0022</b>	<b>0.0022</b>
T1D–JIA	0.0519	0.0060	<b>0.0006</b>	<b>0.0003</b>	0.1826	<b>0.0386</b>	<b>0.0042</b>	<b>0.0027</b>
JIA–CD	0.0568	0.0014	<b>0.0001</b>	<b>0.0004</b>	0.1826	<b>0.0102</b>	<b>0.0008</b>	<b>0.0030</b>
T1D–CD	<b>0.0000</b>	<b>0.0000</b>	0.0033	0.0013	<b>0.0000</b>	<b>0.0000</b>	<b>0.0211</b>	<b>0.0079</b>
THY–T1D	0.6735	0.9232	0.0618	0.0014	0.8022	1.0000	0.1987	<b>0.0079</b>
AS–CVID	0.1494	1.0000	0.2023	0.0046	0.3169	1.0000	0.4791	<b>0.0225</b>
AS–JIA	0.0477	0.9631	0.2369	0.0050	0.1826	1.0000	0.5330	<b>0.0225</b>
THY–SLE	0.2059	0.9864	0.0155	0.0087	0.4028	1.0000	0.0871	<b>0.0356</b>
THY–JIA	0.0036	1.0000	0.6666	0.5730	<b>0.0273</b>	1.0000	0.8228	0.8318
CVID–CD	<b>0.0000</b>	1.0000	0.9905	0.7264	<b>0.0006</b>	1.0000	0.9905	0.8795
CEL–CD	<b>0.0026</b>	0.9005	0.9330	0.7965	<b>0.0230</b>	1.0000	0.9905	0.8795

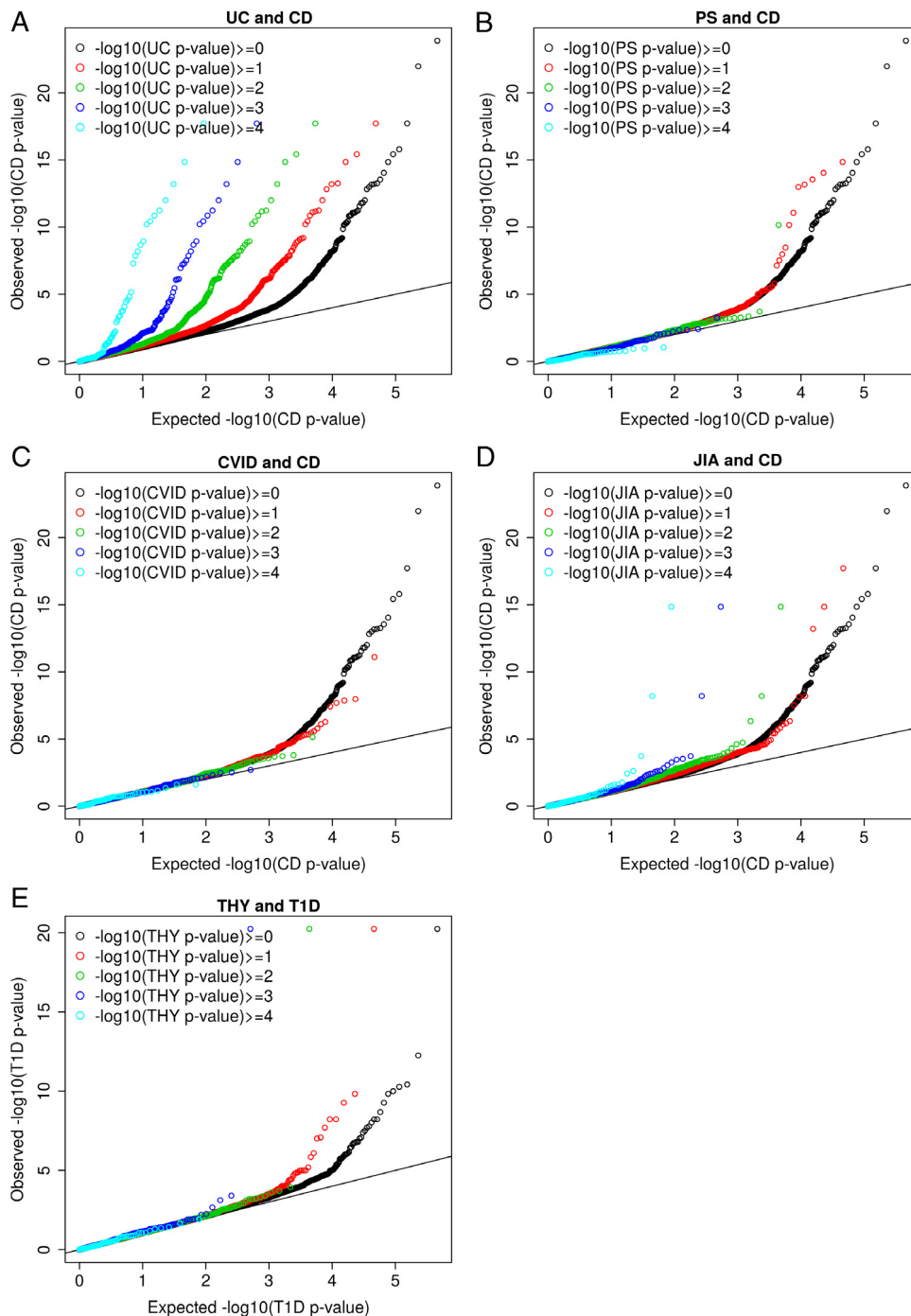
Hakonsarson and colleagues at the Children's Hospital of Pennsylvania conducted separate genome-wide association studies in 10,718 shared controls and over 5,000 cases across ten different diseases: ankylosing spondylitis, Celiac's disease, common variable immunodeficiency, Crohn's disease, juvenile idiopathic arthritis, psoriasis, systemic lupus erythematosus, thyroiditis, type I diabetes, and ulcerative colitis [41,42]. Subjects were genotyped on Illumina Infinium HumanHap550 and Human610 BeadChip array platforms, and only variants common to both arrays and surviving quality control were used for analysis.

Only autosomal chromosomes were considered, and variants in the major histocompatibility complex region, defined as the 25,500,000 to 34,000,000 base pair region of chromosome 6, were not considered because they are known to be highly associated with all autoimmune diseases. This resulted in roughly 450,000 typed variants for each disorder. Genome-wide association  $p$ -values  $P_{kj}$  were calculated for each variant. The correlation between test statistics from different studies due to the shared controls was found, using the method of Zaykin and Kozbur [54], to be at most only 0.019.

The proposed  $\hat{D}$  (3) was implemented with  $m_1 = m_2 = 1000$ , with the truncation parameters  $m_k$  defined in Section 2.3. The results were compared to those of Spearman's correlation, the GPA method of Chung et al. [11], and the max test of Zhao et al. [56]. For computational reasons, the distance covariance method of Székely et al. [51] and the  $M_{m \times m}^{DDP}$  test of Heller et al. [28] were omitted. For  $\hat{D}$  and the max test, the  $P_{kj}$  were converted to  $-\log_{10} P_{kj}$  in order to satisfy the stochastic ordering condition of Assumption 1.

These methods were applied to test for weak positive latent dependence (2) between all 45 unique pairs of the 10 disorders. Permutation  $p$ -values for  $\hat{D}$  were calculated using 10,000 random permutations; this procedure still maintain type I error in the presence of linkage disequilibrium, as shown in simulations in Section 4.3. A Bonferroni correction can be applied to adjust for multiple comparisons, but this may be overly conservative because the pairwise nature of the 45 tests makes them highly dependent. As an alternative, it has been found that in this pairwise testing setting, the Benjamini–Hochberg procedure [5] can still maintain false discovery rate control in practice [53].

Table 5 presents disease pairs for which at least one dependency detection method was significant at an error rate of 0.05 after either Bonferroni or Benjamini–Hochberg correction. The proposed  $\hat{D}$  and the max test of Zhao et al. [56] identified the most disease pairs after Bonferroni correction, while  $\hat{D}$  alone gave the most findings after Benjamini–Hochberg correction. These results suggest that for detecting weak dependence, the test proposed in this paper is a valuable alternative to existing methods.



**Fig. 7.** Stratified QQ-plots of selected pairs of diseases. Diseases: CD = Crohn's disease; CVID = common variable immunodeficiency; PS = psoriasis; T1D = type I diabetes; THY = thyroiditis; UC = ulcerative colitis.

Fig. 7 illustrates some selected results. For each disorder pair, it depicts different QQ-plots of the  $-\log_{10}p$ -values of one of the disorders, for those variants that have  $-\log_{10}p$ -values of at least certain sizes in the other disorder. Panel A illustrates the ulcerative colitis–Crohn's disease pair. As both are inflammatory bowel diseases, it is no surprise that these were found by all methods to exhibit genetic sharing. Panel A indeed shows that genetic variants that are more significant in the ulcerative colitis genome-wide association study also tend to be more significant in the Crohn's disease study.

In contrast, panel B illustrates the psoriasis–Crohn’s disease pair, which was one of the pairs not found to be significant by any method. The QQ-plots reflect the fact that variants more significant in one study are not always more significant in the other. Panel C illustrates the common variable immunodeficiency–Crohn’s disease pair, which was found to be significant only by Spearman’s test. The QQ-plots show a negative dependence, which is not of interest here.

Finally, panel D illustrates the juvenile idiopathic arthritis–Crohn’s disease pair, which was detected only by the max test of Zhao et al. [56] and the proposed  $\widehat{D}$  after Bonferroni correction. Panel E illustrates the thyroiditis-type I diabetes pair, which was detected only by  $\widehat{D}$  after Benjamini–Hochberg correction. Both pairs of QQ-plots show the presence of positive dependence, but in contrast to the type of strong positive dependence present in panel A, the dependence here appears to be heavily driven by small number of variants that are significant in both disorders. This is exactly the type of weak dependence that is difficult for existing methods to detect, and exactly the motivation behind the method proposed in this paper.

## 5. Discussion

The simulations in Section 4 considered only the particular type of dependence described in Eqs. (1) and (2). The proposed method and the competing procedures have different properties otherwise. For example, when the  $(T_{1j}, T_{2j})$  are dependent in such a way so as to form a circle when plotted in  $\mathbb{R}^2$ , a small scale simulation study with 100 replications showed that both  $M_{m \times m}^{DDP}$  and  $S_{m \times m}^{DDP}$  had 100% power,  $\widehat{D}$  (10) had 76% power, and the proposed  $\widehat{D}$  had only 6% power. Thus while  $\widehat{D}$  was the best performer under the dependence alternative considered in this paper, it is not an omnibus test for independence. Interestingly, its generalization  $\widetilde{D}$  had good all-around performance, making it a potentially good candidate for detecting general dependence alternatives.

The asymptotic properties of the proposed method were derived in Section 3 under the assumption that the test statistics  $T_{kj}$  are independent across  $j$ . When the  $T_{kj}$  are correlated,  $\widehat{D}$  (3) will likely no longer be asymptotically optimal, though the simulations in Section 4.3 indicate that it can still have good power. Hall and Jin [23] studied the asymptotic properties of the higher criticism procedure for single-sequence signal detection with correlated tests, and Hall and Jin [24] proposed the innovated higher criticism method that can achieve optimality for certain correlation structures. However, their results do not immediately extend to testing for dependence (2), and further work is necessary to determine the fundamental limits of detection as well as to develop optimal methods.

Some alternatives to the proposed  $\widehat{D}$  may have better finite-sample performance. Recently, Li and Siegmund [40] showed that for the single-sequence detection problem, a test based on the Berk–Jones goodness-of-fit statistic can be dramatically more powerful than the higher criticism statistic, on which  $\widehat{D}$  is based. Previously, Jager and Wellner [36] showed that the Berk–Jones based test has the same asymptotic optimality properties as higher criticism. A similar statistic for testing (2) would be a useful alternative to  $\widehat{D}$ .

This paper assumes that the  $T_{kj}$  are two-tailed test statistics, and as such ignores the directions of effect of the non-null signals. However, it may be desirable to require variants to exhibit the same directions of effect in order to be considered as evidence for genetic sharing. There exist methods that can test for this more stringent condition [30], and it would be interesting to study their asymptotic properties.

## Acknowledgments

The authors thank Drs. Hakon Hakonarson, Brendan J. Keating, Yun Li, and Julie Kobie for providing the pediatric autoimmune disease data and helping with its analysis, Dr. Yihong Wu for helpful discussions, and the anonymous referees for excellent suggestions. The research of Tony Cai was supported in part by National Science Foundation grants DMS-1208982 and DMS-1403708, and the National Institutes of Health grant R01 CA127334. The research of Hongzhe Li was supported in part by the National Institutes of Health grants R01 GM097505 and R01 CA127334. The research of Dave Zhao was supported in part by National Science Foundation grant DMS-1613005 and the Simons Foundation grant SFLife 291812.

## Appendix A. Supplementary data

Supplementary material related to this article can be found online at <http://dx.doi.org/10.1016/j.jmva.2017.06.009>.

## References

- [1] E. Arias-Castro, E.J. Candès, Y. Plan, Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism, *Ann. Statist.* 39 (5) (2011) 2533–2556.
- [2] E. Arias-Castro, M. Wang, Distribution-free tests for sparse heterogeneous mixtures, *TEST* 26 (1) (2017) 71–94.
- [3] I. Barnett, R. Mukherjee, X. Lin, The generalized higher criticism for testing SNP-set effects in genetic association studies, *J. Amer. Statist. Assoc.* 112 (517) (2017) 64–76.
- [4] I.J. Barnett, X. Lin, Analytical  $p$ -value calculation for the higher criticism test in finite- $d$  problems, *Biometrika* 101 (4) (2014) 964–970.
- [5] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 57 (1995) 289–300.
- [6] P. Bickel, H. Chernoff, Asymptotic distribution of the likelihood ratio statistic in a prototypical nonregular problem, in: J. Ghosh, S. Mitra, K. Parthasarathy, B. Prakasa Rao (Eds.), *Statistics and Probability: A Raghu Raj Bahadur Festschrift*, Wiley Eastern, New Delhi, 1983, pp. 83–96.
- [7] J. Blum, J. Kiefer, M. Rosenblatt, Distribution free tests of independence based on the sample distribution function, *Ann. Math. Stat.* 32 (1961) 485–498.



- [8] B.C. Brown, C.J. Ye, A.L. Price, N. Zaitlen, Asian Genetic Epidemiology Network Type 2 Diabetes Consortium, et al., Transethnic genetic-correlation estimates from summary statistics, *Am. J. Hum. Genet.* 99 (1) (2016) 76–88.
- [9] T.T. Cai, X.J. Jeng, J. Jin, Optimal detection of heterogeneous and heteroscedastic mixtures, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73 (5) (2011) 629–662.
- [10] T.T. Cai, Y. Wu, Optimal detection of sparse mixtures against a given null distribution, *IEEE Trans. Inform. Theory* 60 (4) (2014) 2217–2232.
- [11] D. Chung, C. Yang, C. Li, J. Gelernter, H. Zhao, GPA: A statistical approach to prioritizing GWAS Results by integrating pleiotropy and annotation, *PLoS Genet.* 10 (11) (2014) e1004787.
- [12] Cross-Disorder Group of the Psychiatric Genomics Consortium, et al., Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs, *Nature Genet.* 45 (9) (2013a) 984–994.
- [13] Cross-Disorder Group of the Psychiatric Genomics Consortium, et al., Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis, *Lancet* 381 (9875) (2013b) 1371.
- [14] A. Delaigle, P. Hall, J. Jin, Robustness and accuracy of methods for high dimensional data analysis based on student's *t*-statistic, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73 (3) (2011) 283–301.
- [15] D. Donoho, J. Jin, Higher criticism for detecting sparse heterogeneous mixtures, *Ann. Statist.* 32 (3) (2004) 962–994.
- [16] B. Efron, Large-scale inference: Empirical bayes methods for estimation, testing, and prediction, Cambridge University Press, Cambridge, 2010.
- [17] J.H. Einmahl, Extension to higher dimensions of the Jaeschke–Eicker result on the standardized empirical process, *Commun. Stat. Theory Methods* 25 (4) (1996) 813–822.
- [18] J.H. Einmahl, D.M. Mason, Bounds for weighted multivariate empirical distribution functions, *Probab. Theory Relat. Fields* 70 (4) (1985) 563–571.
- [19] Y. Fan, P.L. de Micheaux, S. Penev, D. Salopek, Multivariate nonparametric test of independence, *J. Multivariate Anal.* 153 (2017) 189–210.
- [20] C. Genest, B. Rémillard, Tests of independence and randomness based on the empirical copula process, *Test* 13 (2004) 335–370.
- [21] C. Genovese, L. Wasserman, Operating characteristics and extensions of the false discovery rate procedure, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 64 (3) (2002) 499–517.
- [22] Z. Guo, W. Wang, T.T. Cai, H. Li, Optimal estimation of co-heritability in high-dimensional linear models, 2016. ArXiv preprint arXiv:1605.07244.
- [23] P. Hall, J. Jin, Properties of higher criticism under strong dependence, *Ann. Statist.* 36 (2008) 381–402.
- [24] P. Hall, J. Jin, Innovated higher criticism for detecting sparse signals in correlated noise, *Ann. Statist.* 38 (3) (2010) 1686–1732.
- [25] J. Hartigan, A failure of likelihood asymptotics for normal mixtures, *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, Wadsworth, Belmont, CA, 1985, pp. 807–810.
- [26] X. He, C.K. Fuller, Y. Song, Q. Meng, B. Zhang, X. Yang, H. Li, Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS, *Am. J. Hum. Genet.* 92 (5) (2013) 667–680.
- [27] R. Heller, M. Bogomolov, Y. Benjamini, Deciding whether follow-up studies have replicated findings in a preliminary large-scale omics study, *Proc. Natl. Acad. Sci.* 111 (46) (2014a) 16262–16267.
- [28] R. Heller, Y. Heller, S. Kaufman, B. Brill, M. Gorfine, Consistent distribution-free *k*-sample and independence tests for univariate random variables, *J. Mach. Learn. Res.* 17 (29) (2016) 1–54.
- [29] R. Heller, S. Yaacoby, D. Yekutieli, repfdr: A tool for replicability analysis for genome-wide association studies, *Bioinformatics* 30 (2014b) 2971–2972.
- [30] R. Heller, D. Yekutieli, et al., Replicability analysis for genome-wide association studies, *Ann. Appl. Stat.* 8 (1) (2014c) 481–498.
- [31] W. Hoeffding, A non-parametric test of independence, *Ann. Math. Stat.* 19 (1948) 546–557.
- [32] D.W. Huang, B.T. Sherman, R.A. Lempicki, Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists, *Nucleic Acids Res.* 37 (1) (2009) 1–13.
- [33] Y.I. Ingster, Some problems of hypothesis testing leading to infinitely divisible distributions, *Math. Methods Statist.* 6 (1) (1997) 47–69.
- [34] Y.I. Ingster, Adaptive detection of a signal of growing dimension, I, *Math. Methods Statist.* 10 (2002a) 395–421.
- [35] Y.I. Ingster, Adaptive detection of a signal of growing dimension, II, *Math. Methods Statist.* 11 (1) (2002b) 37–68.
- [36] L. Jager, J.A. Wellner, Goodness-of-fit tests via phi-divergences, *Ann. Statist.* 35 (5) (2007) 2018–2053.
- [37] T. Ledwina, G. Wylupek, Validation of positive quadrant dependence, *Insurance Math. Econom.* 56 (2014) 38–47.
- [38] S.H. Lee, T.R. DeCandia, S. Ripke, J. Yang, P.F. Sullivan, M.E. Goddard, M.C. Keller, P.M. Visscher, N.R. Wray, S.P. G.-W. A.S. Consortium, et al., Estimating the proportion of variation in susceptibility to schizophrenia captured by common SNPs, *Nature Genet.* 44 (3) (2012) 247–250.
- [39] E.E.L. Lehmann, J.P. Romano, Testing statistical hypotheses, Springer Science+Business Media, New York, 2005.
- [40] J. Li, D. Siegmund, Higher criticism: *p*-values and criticism, *Ann. Statist.* 43 (3) (2015) 1323–1350.
- [41] Y.R. Li, J. Li, S.D. Zhao, J.P. Bradfield, F.D. Mentch, S.M. Maggadottir, C. Hou, D.J. Abrams, D. Chang, F. Gao, D. Guo, Z. Wei, J.J. Connolly, C. C., M. Bakay, J. and Kao, C. Glessner, K.A. Thomas, H. Qiu, R. Chiavacci, C. Kim, F. Wang, J. and Richie, M.D. Snyder, B. Flatø, Ø. Førre, L. Denson, S.D. Thompson, M.L. Becker, S.L. Guthery, A. Latiano, E. Perez, E. Resnick, R.D. Russell, D. Wilson, M.S. Silverberg, V. Annese, B.A. Lie, M. Punaro, M.C. Dubinsky, D.S. Monos, C. Strisciuglio, A. Staiano, E. Miele, S. Kugathasan, J.A. Ellis, J. Munro, K. Sullivan, C. Wise, H. Chapel, C. Cunningham-Rundles, S.F.A. Grant, J. Orange, P.M.A. Sleiman, E. Behrens, A. Griffiths, J. Satsangi, T. Finkel, A. Keinan, E.T. Luning Prak, C. Polychronakos, B. Baldassano, H. Li, B.J. Keating, H. Hakonarson, Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases, *Nature Med.* 21 (2015a) 1018–1027.
- [42] Y.R. Li, S.D. Zhao, M. Mohebasab, J. Li, J. Bradfield, L. Steel, D. Abrams, J. Kobbie, F. Mentch, J. Glessner, Y. Guo, Z. Wei, C. Cardinale, M. Bakay, J. Connolly, D. Li, S.M. Maggadottir, K.A. Thomas, H. Qiu, R. Chiavacci, C. Kim, F. Wang, J. Snyder, B. Flatø, Ø. Førre, L. Denson, S.D. Thompson, M. Becker, S.L. Guthery, A. Latiano, E. Perez, E. Resnick, C. Strisciuglio, A. Staiano, E. Miele, M. Silverberg, B.A. Lie, M. Punaro, R. Russell, D. Wilson, M.C. Dubinsky, D.S. Monos, V. Annese, J. Munro, C. Wise, H. Chapel, C. Cunningham-Rundles, J. Orange, E.M. Behrens, K. Sullivan, S. Kugathasan, A. Griffiths, J. Satsangi, S. Grant, P. Sleiman, T. Finkel, C. Polychronakos, R.N. Baldassano, E. Luning Prak, J. Ellis, H. Li, B.J. Keating, H. Hakonarson, Genetic sharing and heritability of paediatric age of onset autoimmune diseases, *Nature Commun.* 6 (2015b).
- [43] R. Mukherjee, N.S. Pillai, X. Lin, Hypothesis testing for high-dimensional sparse binary regression, *Ann. Statist.* 43 (1) (2015) 352.
- [44] D.L. Nicolae, E. Gamazon, W. Zhang, S. Duan, M.E. Dolan, N.J. Cox, Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS, *PLoS Genet.* 6 (4) (2010) e1000888.
- [45] D. Phillips, D. Ghosh, Testing the disjunction hypothesis using voronoi diagrams with applications to genetics, *Ann. Appl. Stat.* 8 (2) (2014) 801–823.
- [46] D.N. Reshef, Y.A. Reshef, H.K. Finucane, S.R. Grossman, G. McVean, P.J. Turnbaugh, E.S. Lander, M. Mitzenmacher, P.C. Sabeti, Detecting novel associations in large data sets, *Sci.* 334 (6062) (2011) 1518–1524.
- [47] I. Rivals, L. Personnaz, L. Taing, M.-C. Potier, Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* 23 (4) (2007) 401–407.
- [48] O. Scaillet, A Kolmogorov–Smirnov type test for positive quadrant dependence, *Canad. J. Statist.* 33 (2005) 415–427.
- [49] J.D. Storey, R. Tibshirani, Statistical significance for genomewide studies, *Proc. Natl. Acad. Sci.* 100 (16) (2003) 9440–9445.
- [50] W. Sun, T.T. Cai, Oracle and adaptive compound decision rules for false discovery rate control, *J. Amer. Statist. Assoc.* 102 (479) (2007) 901–912.
- [51] G.J. Székely, M.L. Rizzo, et al., Brownian distance covariance, *Ann. Appl. Stat.* 3 (4) (2009) 1236–1265.

- [52] O. Thas, J.-P. Ottoy, A nonparametric test for independence based on sample space partitions, *Commun. Stat. Simul. Comput.* 33 (3) (2004) 711–728.
- [53] D. Yekutieli, False discovery rate control for non-positively regression dependent test statistics, *J. Statist. Plann. Inference* 138 (2) (2008) 405–415.
- [54] D.V. Zaykin, D.O. Kozbur, P-value based analysis for shared controls design in genome-wide association studies, *Genet. Epidemiol.* 34 (7) (2010) 725–738.
- [55] S.D. Zhao, False discovery rate control for identifying simultaneous signals, 2017. ArXiv preprint [ArXiv:1512.04499](https://arxiv.org/abs/1512.04499).
- [56] S.D. Zhao, T.T. Cai, T.P. Cappola, K.B. Margulies, H. Li, Sparse simultaneous signal detection for identifying genetically controlled disease genes, *J. Amer. Statist. Assoc.* in press (2017).