Statistica Sinica Preprint No: SS-2019-0425	
Title	A Permutation Test for Two-Sample Means and Signal Identification of High-dimensional Data
Manuscript ID	SS-2019-0425
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202019.0425
Complete List of Authors	Efang Kong, Lengyang Wang, Yingcun Xia and Jin Liu
Corresponding Author	Yingcun Xia
E-mail	staxyc@nus.edu.sg
Notice: Accepted version subject to English editing.	

A Permutation Test for Two-Sample Means and Signal Identification of High-dimensional Data

Efang Kong[a], Lengyang Wang[b], Yingcun Xia[b,a] and Jin Liu[b]

[a] *University of Electronic Science and Technology of China*

[b] *National University of Singapore, Singapore*

Abstract: Permutation tests are widely used in practice. However, these tests either need restrictive assumptions for the validity, or are not applicable to high-dimensional data. This paper considers permutation tests for high-dimensional mean comparison, where in order to get round those restrictions, the test statistics are calculated based on pseudo samples that are generated through a "binning" procedure. The corresponding permutation tests are proved to be asymptotically consistent. We also consider a related problem for signal identification and establish the asymptotic properties. Simulation studies demonstrate favorable performance of our methods in comparison to existing tests. Finally, the proposed method is applied to a genome-wide association study (GWAS) for seven complex human diseases to identify possible single nucleotide polymorphisms (SNPs) associated with the diseases.

Key words and phrases: consistency of test; high-dimensional data; permutation tests; signal identification; test of mean-difference

1. Introduction

Testing the equality of means of two random vectors based on random samples has long been one of the statutory issues in multivariate analysis. The past two decades have witnessed increasing interest in this problem in high-dimensional settings. Existing methods could largely be divided into two categories. Some are based on the sum-of-squares of the sample mean differences, e.g., Bai and Saranadasa (1996) and Chen and Qin (2010), and are generally more powerful against dense alternatives, in the sense that there is a large proportion of small to moderate component-wise differences. The others are based on the infinity norm of the mean differences, e.g., Cai et al. (2014), Xu et al. (2016), Chang et al. (2017) and Xue and Yao (2018), and these are better suited for testing against sparse alternatives, i.e. when there are only a few but significant component-wise differences.

The focus of this paper is permutation methods, which have always served as a very useful alternative to traditional methods for hypothesis testing; see Good (2005) and Ernest (2004) for a comprehensive review. The basic idea is to generate a reference distribution by recalculating a statistic for many permutations of the data. To illustrate, suppose p-dimensional random vectors $X_1, \cdots, X_m$ are $\overset{i.i.d.}{\sim} P_1(.)$ with mean μ^X and variance

Σ^X, while $Y_1, \cdots, Y_n$ are $\overset{i.i.d.}{\sim} P_2(.)$ with mean μ^Y and variance Σ^Y. Write $N = m + n$ and suppose that $m/N \to c$ for some constant $c \in (0, 1)$. Our interest is to test the null hypothesis

$$H_0 : \mu^X = \mu^Y.$$

Chung and Romano (2013) considered the testing of H_0 using permutation methods for $p = 1$. The procedure is as follows. Write $Z^N = \{Z_1, \cdots, Z_N\}$, with $Z_i = X_i$, $1 \le i \le m$; $Z_{m+j} = Y_j$, $1 \le j \le n$. Consider the standardized statistic

$$S_N(Z^N) = \frac{N^{1/2}(\bar{X}_m - \bar{Y}_n)}{\sqrt{\frac{N}{m}\hat{\sigma}_m^2(Z^N) + \frac{N}{n}\hat{s}_n^2(Z^N)}}, \tag{1.1}$$

where, $\bar{X}_m$, $\bar{Y}_n$ are the sample means of $\{Z_1, \cdots, Z_m\}$ and $\{Z_{m+1}, \cdots, Z_N\}$, respectively, while $\hat{\sigma}_m^2(Z^N)$ and $\hat{s}_n^2(Z^N)$ are the corresponding sample variances. Let G_N be the set of all permutations of $\{1, \cdots, N\}$. For any $\pi \in G_N$, let Z_π^N denote the rearranged Z^N through permutation π, and $Z_{\pi(i)}^N$, $i = 1, \cdots, N$, be the ith entry of Z_π^N. Recompute $S_N(Z_\pi^N) \equiv S_N(Z_{\pi(1)}^N, \cdots, Z_{\pi(N)}^N)$, and let $\hat{R}_N^S(\cdot)$ denote the empirical distribution of $S_N(Z_\pi^N)$ evaluated at all $N!$ permutations of Z^N, i.e.,

$$\hat{R}_N^S(t) = \frac{1}{N!}\sum_{\pi \in G_N} I\{S_N(Z_\pi^N) \le t\}.$$

This empirical distribution $\hat{R}_N^S(\cdot)$, also referred to as the permutation distribution, is used as an approximation of the null (limiting) distribution of

statistic (1.1), which in this case is given by $\Phi(\cdot)$, the distribution function of the standard normal $N(0,1)$. Reject H_0, if $\hat{R}_N^S(S_N(Z^N)) \geq 1-\alpha$. Chung and Romano (2013) proved that

$$\sup_{t\in R} |\hat{R}_N^S(t) - \Phi(t)| \to 0 \quad \text{in probability,}$$

and in this sense the permutation procedure based on statistic (1.1) is considered to be consistent (valid). In general, however, the consistency of permutation tests should not be taken for granted. Indeed, Chung and Romano (2013) showed that the permutation test based on $S_N(Z^N) = \bar{X}_m - \bar{Y}_n$, i.e. (1.1) without standardization, is inconsistent, unless $c = 1/2$ or $\Sigma^X = \Sigma^Y$.

Clearly, in the high-dimensional cases where the dimension p could far exceed the sample sizes, permutation tests based on standardized statistic (1.1) are no longer applicable. Nor is the prepivoting method of Chung and Romano (2016) proposed for a multivariate setting computationally feasible. This paper intends to fill this gap and we shall propose a permutation procedure which is both asymptotically consistent and easy to implement even for ultra-high dimensional data.

The rest of this paper is organized as follows. Section 2 begins with the basic formulation of the problem, followed by the results concerning the consistency of permutation tests based on the Hotelling's T^2-type of statistics. As these statistics require the estimation of the inverse of covariance matrix, which renders their use impractical in high-dimensional setting, an

alternative is described in details in Section 2.2, where we propose a "binning" procedure to produce pseudo samples from which the test statistics are then derived. Section 3 applies the proposed tests to identify variables which are the sources of the difference of two high-dimensional means, referred to as signal identification hereafter. Some related theoretical results are also given. Numerical performance of the proposed methods and other existing methods are examined in Section 5 through simulation studies. Section 6 contains an empirical study of genome-wide association for seven complex diseases using the data from Wellcome Trust Case Control Consortium (WTCCC). Assumptions needed for asymptotic studies are given in the Appendix, while technical proofs are delegated to a separate online supplementary file.

2. Permutation tests for high-dimensional mean comparison

We first introduce some notations. For any $v = (v_1, \cdots, v_p)^\top \in R^p$, let $|v|_\gamma = \{(|v_1|^\gamma + \cdots + |v_p|^\gamma)/p\}^{1/\gamma}$ for any $\gamma > 0$. In particular, $|v|_1 = (|v_1| + \cdots + |v_p|)/p$, stands for the L_1-norm, and $|v|_\infty = \max_{k=1,\cdots,p} |v_k|$, the L_∞-norm. Write $\bar{X}_m = m^{-1}\sum_i X_i$, $\bar{Y}_n = n^{-1}\sum_j Y_j$, $\delta_N = (\delta_{N,1}, \cdots, \delta_{N,p})^\top = N^{1/2}(\bar{X}_m - \bar{Y}_n)$, and

$$\hat{\Sigma}_m^X = \frac{1}{m}\sum_i (X_i - \bar{X}_m)(X_i - \bar{X}_m)^\top, \quad \hat{\Sigma}_n^Y = \frac{1}{n}\sum_j (Y_j - \bar{Y}_n)(Y_j - \bar{Y}_n)^\top. \tag{2.2}$$

Denote by $\hat{\sigma}^2_{m,k}(X_1, \cdots, X_m)$, $\hat{s}^2_{n,k}(Y_1, \cdots, Y_n)$, $k = 1, \cdots, p$, the diagonal elements of $\hat{\Sigma}^X_m$ and $\hat{\Sigma}^Y_n$, respectively. Write $\Sigma(\bar{P}) = c\Sigma^X + (1-c)\Sigma^Y$.

2.1 Permutation tests based on Hotelling's T^2-type statistics

Write $\tilde{\Sigma} = c^{-1}\Sigma^X + (1-c)^{-1}\Sigma^Y$, the variance of δ_N, and suppose $\tilde{\Omega}_N = \tilde{\Omega}_N(Z^N)$ is an estimator of $\tilde{\Omega} = \tilde{\Sigma}^{-1}$. Then in a manner similar to (1.1), define $e_N(Z^N) = \{\tilde{\Omega}_N\}^{1/2}\delta_N$ and

$$H^\gamma(Z^N) \equiv |e_N(Z^N)|_\gamma, \quad \gamma = 1 \text{ or } \infty. \tag{2.3}$$

Xu et al. (2016) considered the use of other values for γ, but in the present study of permutation tests for high dimensions, we only focus on the cases where $\gamma = 1$ or ∞. In practice these two choices should serve the purposes well enough, as the use of $H^1(\cdot)$ is expected to be more powerful against dense alternatives, while $H^\infty(\cdot)$ works better against sparse alternatives. The latter also finds important applications in signal identification; see, e.g, Benjamini and Hochberg (1995) and Jin and Cai (2007). Regarding the permutation tests based on test statistics (2.3), we have

Theorem 1. *Suppose conditions (C1)-(C5) of Section 4 hold. Then*

$$\text{under } H_0, \ \sup_{t\in R}\Big|\frac{1}{N!}\sum_{\pi\in G_N} I\{H^\infty(Z^N_\pi) < t\} - Pr\Big(H^\infty(Z^N) \le t\Big)\Big| \xrightarrow{p} 0, \tag{2.4}$$

where $\xrightarrow{p}$ stands for convergence in probability. Parallel results hold for $H^1(.)$, if conditions (C1)-(C2), (C3'), (C4)-(C5) of Section 4 hold.

In other words, the permutation tests based on (2.3) with $\gamma = 1$ or $\gamma = \infty$ are both consistent. However, in high-dimensional settings, these tests are difficult to implement due to the challenges with estimating the high-dimensional precision matrix $\tilde{\Omega}$, if at all possible. A naive solution is to standardize (divide) the entries of δ_N by their marginal standard error. Namely, with

$$v_{N,k}^2 = \frac{N}{m}\hat{\sigma}_{m,k}^2(X_1, \cdots, X_m) + \frac{N}{n}\hat{s}_{n,k}^2(Y_1, \cdots, Y_n), \tag{2.5}$$

consider the following test statistics

$$S^1(Z^n) = p^{-1}\sum_{k=1}^{p}|\delta_{N,k}/v_{N,k}|, \quad S^\infty(Z^n) = \max_{1\le k\le p}|\delta_{N,k}/v_{N,k}|. \tag{2.6}$$

Theorem 2. *If conditions (C1)-(C3) and (C6) of Section 4 hold, then*

$$\sup_{t\in R}\left|\frac{1}{N!}\sum_{\pi\in G_N} I\{S^\infty(Z_\pi^N) < t\} - Pr(|\Xi|_\infty < t)\right| \xrightarrow{p} 0, \tag{2.7}$$

where Ξ is a p-dimensional Gaussian, with covariance matrix given by $[diag(\Sigma(\bar{P}))]^{-1/2}\Sigma(\bar{P})[diag(\Sigma(\bar{P}))]^{-1/2}$, the correlation matrix associated with $\Sigma(\bar{P})$; on the other hand,

$$\text{under } H_0, \quad \sup_{t\in R}\left|P\Big(S^\infty(Z^N) \le t\Big) - Pr(|\tilde{\Xi}|_\infty < t)\right| \to 0, \tag{2.8}$$

where $\tilde{\Xi}$ is also a p-dimensional Gaussian, with covariance matrix given by $[diag(\tilde{\Sigma})]^{-1/2}\tilde{\Sigma}[diag(\tilde{\Sigma})]^{-1/2}$, the correlation matrix given by that of $\tilde{\Sigma}$. Parallel results hold for $S^1(.)$ under conditions (C1)-(C2), (C3'), and (C6).

As $\Sigma(\bar{P}) = c\Sigma^X + (1-c)\Sigma^Y$, permutation tests based on $S^\gamma(\cdot)$ are thus in general inconsistent, except when $\Sigma^X = \Sigma^Y$ or $c = 1/2$; this is also noted in Chung and Romano (2016) for the finite-dimension case. To correct the inconsistency associated with statistic (2.6) $S^\infty(.)$, the permutation tests in Chung and Romano (2016) are coupled with a pre-pivoting procedure: for each permutation, bootstrapping is implemented to get an estimate of a 'prepivoted' statistic. Due to the huge amount of computation required, however, this approach is thus not practically feasible. Moreover, the theoretical results therein were only established for the fixed dimensional setting. Our solution is described in the next section.

2.2 A 'binning' procedure and pseudo samples

The purpose of this procedure is to produce two pseudo samples of equal sizes. Without loss of generality, suppose $m > n$ so that $m = K \times n + k$, for some nonnegative integers K and k, with $0 \le k < n$. Thus, $K = [c/(1-c)]$, the integer part of $c/(1-c)$, and $k/n \to c/(1-c) - K$. Define

$$X_i' = X_i - \mu^X, \quad Y_j' = Y_j - \mu^X, \ i = 1, \cdots, m, \ j = 1, \cdots, n; \tag{2.9}$$

in practice, $\bar{X}_m$ could be used as a substitute for μ^X. The pseudo observations are then constructed as follows. If $k = 0$, define

$$X_i^* = \frac{n}{m} \sum_{j=(i-1)K+1}^{i\times K} X_j', \quad i = 1, \cdots, n.$$

If $k > 0$, randomly select k from the above defined $\{X_i^*,\, i = 1, \cdots, n\}$ first, and assign each to one of the left-over $X'_{K\times n+i}$, $i = 1, \cdots, k$. Specifically and without loss of generality, define

$$X_i^* := X_i^* + \frac{n}{m} X'_{K\times n+i}, \quad i = 1, \cdots, k. \tag{2.10}$$

We call $\{X_1^*, ..., X_n^*\}$ and $\{Y_1', ..., Y_n'\}$, the pseudo samples. Note that although some of the pseudo observations X_i^* are derived from K original X_is, while others are derived from $K+1$ original X_is, these X_i^*s are nevertheless identically distributed (more explanation given in the proof of Theorem 3). More importantly, if the null hypothesis H_0 holds for the original observations X_i and Y_j, then it also holds for the pseudo samples, and vice versa. From now on, all steps involved in the permutation test are applied to these pseudo samples instead of the original X_is and Y_js.

Write $Z^n = \{Z_1, \cdots, Z_{2n}\}$, such that $Z_i = X_i^*$, $Z_{n+j} = Y_j'$, $i, j = 1, \cdots, n$. Recall that $X_1^*, \cdots, X_n^*$ stand for the first n elements of Z^n, while $Y_1', \cdots, Y_n'$ are the remaining ones. Let $\bar{X}^* = n^{-1}\sum_{i=1}^n X_i^*$ and $\bar{Y}^* = n^{-1}\sum_{j=1}^n Y_j'$ be the two sample means. Write $\delta_n^* = (\delta_{n,1}^*, \cdots, \delta_{n,p}^*)^\top = n^{1/2}(\bar{X}^* - \bar{Y}^*)$, and consider the following simple test statistics:

$$S_0^1(Z^n) = |\delta_n^*|_1, \quad S_0^\infty(Z^n) = |\delta_n^*|_\infty. \tag{2.11}$$

Apparently these statistics do not take into account the differences in the

variations of variables. Thus an arguably improved alternative is such that

$$S_1^1(Z^n) = p^{-1}\sum_{k=1}^{p}|\delta_{n,k}^*/v_{n,k}^*|, \quad S_1^\infty(Z^n) = \max_{k=1,\cdots,p}|\delta_{n,k}^*/v_{n,k}^*|, \tag{2.12}$$

where $v_{n,k}^* = \{\hat{\sigma}_{n,k}^2(X_1^*,\cdots,X_n^*) + \hat{s}_{n,k}^2(Y_1',\cdots,Y_n')\}^{1/2}$ is the estimator of the variance of $\delta_{n,k}^*$. Denote by Z_π^n the rearranged Z^n through any given permutation $\pi \in G_{2n}$, and $S_1^\gamma(Z_\pi^n) \equiv S(Z_{\pi(1)}^n,\cdots,Z_{\pi(2n)}^N)$. The distribution of $S_1^\gamma(Z^n)$ is then given by the empirical distribution of $S_1^\gamma(Z_\pi^n)$ evaluated at all $(2n)!$ permutations of Z^n.

Theorem 3. *The permutation tests based on $S_0^\infty(\cdot)$ of (2.11) are consistent, under conditions (C1)-(C3) of Section 4. Similarly, the permutation test based on $S_0^1(\cdot)$ is consistent under conditions (C1)-(C2), and (C3'). The same conclusions hold for permutation tests based on $S_1^\infty(\cdot)$ or $S_1^1(\cdot)$, if condition (C6) of Section 4 is also true.*

Numerical evidence suggests that in terms of Type-I error control, the tests based on $S_0^\gamma(\cdot)$ are more stable than those based on $S_1^\gamma(\cdot)$, especially when p is large. However, it should also be noted that the latter in general posses better power as they take into account the possibility of different marginal standard errors.

3. Signal identification

Write $\delta_0 = (\delta_{01}, \cdots, \delta_{0p})^\top = \mu^X - \mu^Y$. Denote by $I_0 \subseteq \{1, ..., p\}$, such that

$$|\delta_{0k}| > 0, \ \forall k \in I_0; \ quad|\delta_{0k}| = 0, \ \forall k \notin I_0.$$

This is referred to as the set of signals. The number of signals, i.e., the cardinality of I_0, could increase with p.

Let $\tilde{t}_{n,p}(\cdot)$ stand for the permutation distribution function of $S_1^\infty = \max_{1\le k\le p} |\delta^*_{n,k}/v^*_{n,k}|$, and $\tilde{t}^{-1}_{n,p}(.)$, its inverse. The significance level α_n is chosen so that $q_{\alpha_n}/(2\ln p)^{1/2} \to 1$, where $q_\alpha = -\ln(\pi) - 2\ln(-\ln(1-\alpha))$, is the $(1-\alpha)$ quantile of the type-I extreme value distribution $F(x) = \exp(-\exp\{-(\ln\pi + x)/2\})$. In other words, α_n is such that

$$\ln\{-\ln(1-\alpha_n)\}/(\ln p)^{1/2} \to -\sqrt{2}/2. \tag{3.13}$$

Consequently, the estimated set of signals is defined as

$$\hat{I}_n = \{k : |\delta^*_{n,k}/v^*_{n,k}| > \tilde{t}^{-1}_{n,p}(1-\alpha_n), \ k = 1, \cdots, p\}.$$

Theorem 4. *Suppose conditions (C1)-(C3) and (C6) in Section 4 hold. If*

$$\liminf_{n,p\to\infty} (c/s_1)^{1/2} n^{1/2} (\ln p)^{-1/2} \min_{k\in I_0} |\delta_{0k}| \ge 2\sqrt{2}, \tag{3.14}$$

where s_1 is as given in (C2), and α_n satisfies (3.13), then as $n, p \to \infty$,

$$Pr(\hat{I}_n = I_0) \to 1.$$

In other words, if the strength of the signals, measured through $\min_{k\in I_0}|\delta_{0k}|$, is strong enough, the set of signals could be correctly identified in probability.

4. Notations and assumptions

For any square matrix $M=[m_{ij}]$, $\|M\|_{(1,1)}=\max_j\sum_i|m_{ij}|$, while $\lambda_{max}(M)$ and $\lambda_{min}(M)$ denote the largest and smallest absolute eigenvalue of M, respectively. We assume the following conditions

(C1) $\lim_{m\to\infty} m/N=c\in(0,1)$ and $c-m/N=O(N^{-1/2})$.

(C2) There exists some constant $s_1>s_0>0$, such that $s_0\le\sigma_{kk}^2,\ s_{kk}^2\le s_1$.

(C3) $\ln(p)=O(n^\alpha)$, $\alpha<1/7$; there exist finite constants $c_1,\ c_2>0$ such that

$$E[|X_{i,k}|^{2+l})]\le c_1^l,\ E[|Y_{j,k}|^{2+l})]\le c_2^l,\ k=1,\cdots,p,\ l=1,2;$$
$$E\{\exp(X_{i,k}/c_1)\}\le 2,\quad E\{\exp(Y_{j,k}/c_2)\}\le 2,\ k=1,\cdots,p.$$

(C3') $p=O(n^\alpha)$, $\alpha<1/7$; for $\nu=\{p^{-1/2}(v_1,v_2,\cdots,v_p)^\top: v_j=1 \text{ or } -1\}$, and $\tilde X_i=(v^\top X_i)_{v\in\nu}$, $\tilde Y_j=(v^\top Y_j)_{v\in\nu}$, $i=1,\cdots,m$, $j=1,\cdots,n$, there exist finite constants $\tilde c_1>0$, $\tilde c_2>0$ such that

$$E[|\tilde X_{i,k}|^{2+l})]\le\tilde c_1^l,\ E[|\tilde Y_{j,k}|^{2+l})]\le\tilde c_2^l,\ k=1,\cdots,2^{p-1},\ l=1,2;$$
$$E\{\exp(\tilde X_{i,k}/\tilde c_1)\}\le 2,\quad E\{\exp(\tilde Y_{j,k}/\tilde c_2)\}\le 2,\ k=1,\cdots,2^{p-1}.$$

(C4) The eigen-values of Σ^X and Σ^Y are bounded from both below and above by some constants $0 < c_3 < c_4$.

(C5) $\tilde{\Omega}_N$ is an estimate of $\tilde{\Omega} = \tilde{\Sigma}^{-1}$ which satisfies the following condition

$$\|\{\tilde{\Omega}_N\}^{1/2} - \{\tilde{\Omega}\}^{1/2}\|_{(1,1)} = o_p(\{\ln p\}^{-1}); \tag{4.15}$$

similarly for $\tilde{\Omega}_N = \tilde{\Omega}_N(Z_1, \cdots, Z_N)$, with $Z_1, \cdots, Z_N \overset{i.i.d.}{\sim} \bar{P} = cP_1(.) + (1-c)P_2(.)$ (the mixture distribution), we have

$$\|\{\tilde{\Omega}_N\}^{1/2} - \{\Sigma(\bar{P})/c(1-c)\}^{-1/2}\|_{(1,1)} = o_p(\{\ln p\}^{-1}) \tag{4.16}$$

(C6) $\hat{\sigma}^2_{m,k}$ and $\hat{s}^2_{n,k}$, $k = 1, \cdots, p$, defined in (2.2) are consistent and

$$\max_{1\le k\le p} |\frac{\hat{\sigma}^2_{m,k}}{\sigma^2_{kk}} - 1| = o_p(\frac{1}{\ln p}), \ \max_{1\le k\le p} |\frac{\hat{s}^2_{n,k}}{s^2_{kk}} - 1| = o_p(\frac{1}{\ln p}); \tag{4.17}$$

in a sense similar to (4.16), (4.17) also holds for the same statistic based on i.i.d. observation from the mixture distribution $\bar{P} = cP_1(.) + (1-c)P_2(.)$.

Remarks. (C1) is taken from Chung and Romano (2013). (C2) and (C3) are found in Chernozhukov et al. (2017) to obtain a uniform bound over probabilities concerning hyperrectangles (see Proposition 2.1 therein); while assumption (C3') corresponds to those conditions in their Proposition 3.1 which concerns a uniform bound for probabilities over simple convex sets. Note that for the latter case, it requires a stricter rate on how large

p could be relative to n. For simplicity, c_1 and c_2 are taken to be finite here, but it is possible to allow for infinite c_1 and c_2, but then compromise must be made on how large $\ln p$ could be relative to n; refer to equation (9) of Chernozhukov et al. (2017) for an explicit expression which relates these two cases. (C4) is necessary for the anti-concentration inequality, e.g., Proposition 4. Conditions (4.15) and (4.17) have been adopted in Cai et al. (2014) to derive the asymptotic power of the data-driven statistics including $H^{\infty}(\cdot)$ of (2.3) and $S^{\infty}(\cdot)$ of (2.11) for two Gaussian populations. Kosorok and Ma (2007) gave sufficient conditions for (4.17) to hold, among them $\ln(p) = o(n^{\alpha})$ with $\alpha \in (0, 1/3]$.

5. Simulation study

As far as permutation tests are concerned, we choose to exclude those based on the Hotelling's T^2-type statistics of (2.3) from our numerical studies due to the heavy computational burden. The method of Chung and Romano (2016) is also excluded for the same reason. Instead, we focus on the permutation tests based on statistics calculated for pseudo samples generated through the binning procedures: $S_1^{\gamma}(\cdot)$ and $S_0^{\gamma}(\cdot)$ in (2.12) and (2.11), respectively. Other existing methods included in our comparison studies are : Chen and Qin (2010) (CQ), Cai et al. (2014)(CAI), Xu et al. (2016) (XLWP) and Xue and Yao (2018) (XY). R package *'highmean'* is used for

computations related to CQ, CAI, XLWP and XY. Note that CQ only uses L_2-norm and CAI only L_∞-norm. For signal identification, our method based on S_1^∞ is also compared with Benjamini and Yekutieli (2001).

Sample sizes range from relatively small ($m = 75$, $n = 50$), to medium ($m = 300$, $n = 200$), to large ($m = 600$, $n = 400$); while for dimensionality, $p = 10$, 100, or 1,000. As it is computationally infeasible to evaluate for all possible permutations, random permutations are usually used in practice, which is first proposed by Dwass (1957). In our case, the permutation distribution is evaluated based on 2,500 (random) permutations. Also, the empirical sizes of tests are calculated based on 10,000 replications, while the empirical powers are based on 2,000 replications.

The simulated data are generated according to the following model,

$$X_i = (x_{i,1}, ..., x_{ip})^\top + \mu^X, \text{ and } Y_j = (y_{i,1}, ..., y_{ip})^\top + \mu^Y; \tag{5.18}$$

here μ^X and μ^Y are two constant vectors; and for any given $i = 1, \cdots, m$, $j = 1, \cdots, n$, $\{x_{i,k}, k = 1, 2, \cdots\}$ and $\{y_{j,k}, k = 1, 2, \cdots\}$ are stationary times series such that

$$x_{i,k+1} = a_i x_{i,k} + \xi_k, \quad y_{j,k+1} = b_j y_{j,k} + \eta_k, \quad k = 1, 2, \cdots, \tag{5.19}$$

where ξ_k, η_k are independent random errors, $\{a_i\}_{i=1}^m$, $\{b_j\}_{j=}^n$ are hyper parameters, either fixed or random. This is implemented independently for all $i = 1, \cdots, m$, $j = 1, \cdots, n$. With different specifications for a_i, b_j and

ξ_k, η_k , we come up with the following three models.

Model 1. $a_i, b_k, i = 1, \cdots, m,\ k = 1, \cdots, n$ are i.i.d., following a uniform distribution on $[0, 0.95]$; $\xi_k \overset{i.i.d.}{\sim} N(0,1), \eta_k \overset{i.i.d.}{\sim} N(0,4)$. In this model, X_is are distinctly distributed and so do the Y_is. However, The elements in X_i still have the same variance, and so do Y_j.

Model 2. The same as Model 1, but the even-indexed elements of X_i and Y_i are multiplied by 2. Thus elements in X_i have different variances, and so do Y_i.

Model 3. $a_i \equiv -0.2, b_j \equiv 0.7$, and $\xi_k \sim t(3)$, $\eta_k \sim 2t(3)$, where $t(3)$ is the t-distribution. Thus the generated data are heavy-tailed.

In the study of empirical sizes, $\mu^X = \mu^Y = 0$; when comparing the empirical power of various tests, we keep $\mu^Y = 0$ and consider two different designs for $\mu^X = (\mu_1^X, ..., \mu_p^X)^\top$.

(i) Dense alternatives: $\mu_1^X, ..., \mu_p^X \overset{i.i.d.}{\sim}$ uniform $[0, c_{n,p}]$, with $c_{n,p} = s/(p^{0.25} \times \min(m,n)^{0.5})$, and $s = 6, 9, 11, 14$, which specifies the overall signal-to-noise ratio.

(ii) Sparse alternatives: with $s = 7, 8, 9, 10$, randomly select $0.2 \times p^{0.5}$ elements from $\{\mu_1^X, ..., \mu_p^X\}$, and assign to them the value $c_{n,p} = s/\min(m,n)^{0.5}$, while the entries unselected remain zero.

Note that the strength of the signals varies with sample sizes as well as the dimension; we adopt such design so that we could evaluate how the empirical power of various tests are affected by different sample sizes and dimension.

The empirical sizes of various tests, are summarized in Tables 1 (significance level 1%). For the two columns under the label $S_0^\gamma(\cdot)$, L_1 corresponds to $\gamma = 1$, while L_∞ corresponds to $\gamma = \infty$. The same format also applies to the columns under XLWP and $S_1^\gamma(\cdot)$. In both tables, numbers in small bold font stand for empirical sizes which deviate from the nominal level by more than 20%. The first thing to note is that permutation tests, based on either $S_1^\gamma(\cdot)$ or $S_0^\gamma(\cdot)$, are able to control the type-I error better than all other methods for nearly all models, and especially so when the sample size is small ($n = 50$ or $n = 75$). Furthermore, we also observe that the performance of $S_1^\gamma(\cdot)$ is slightly hampered by low efficiency in variance estimation when p is large while n is small; this is consistent with the remarks we made after Theorem 3.

On the other hand, CQ is also able to control type-I error quite well at 5% significance level (not reported here), but much less so with nominal level at 1% unless the sample size is big enough; see Table 1. As p increases to 1,000, the performance of CQ improves, which is consistent with the fact that its asymptotic (null) distribution is derived for when $p \to \infty$. The performance of XLWP with the L_2-norm is similar to that of CQ while

Table 1: Empirical sizes (%) of different methods (nominal sizes = 1%)

			CQ	XLWP		XY	S_0^γ		S_1^γ	
model	n	p	L_2	L_2	L_∞	L_∞	L_1	L_∞	L_1	L_∞
		10	**2.09**	**3.10**	**0.60**	0.91	0.94	0.95	0.97	0.98
	50	100	1.20	**1.64**	**0.67**	**0.63**	1.01	0.82	1.01	0.98
		1,000	1.04	**1.68**	0.81	**0.52**	**1.22**	0.96	**1.21**	0.95
		10	**2.02**	**2.91**	**0.60**	1.07	1.05	0.99	1.07	0.99
1	200	100	**1.23**	**1.55**	**0.60**	0.86	0.94	0.96	0.98	0.84
		1,000	0.98	1.12	**0.52**	0.82	1.11	1.04	1.08	0.97
		10	**1.79**	**3.59**	**0.55**	**0.72**	0.95	**0.76**	0.92	**0.76**
	400	100	1.19	**2.19**	**0.62**	0.97	0.94	0.96	0.91	1.03
		1,000	1.10	**1.28**	0.84	0.91	1.11	0.92	1.08	0.98
		10	**2.42**	**4.63**	0.80	0.90	1.03	0.86	1.04	0.93
	50	100	**1.47**	**3.30**	0.98	0.83	1.13	0.99	**1.32**	**1.30**
		1,000	1.17	**5.34**	**1.43**	**0.62**	**1.27**	0.84	**1.68**	**1.39**
		10	**2.46**	**3.05**	**0.54**	1.06	0.87	0.95	1.06	1.08
2	200	100	**1.46**	**2.31**	**0.64**	0.95	1.04	0.96	0.99	1.02
		1,000	1.10	**1.21**	1.09	1.03	1.09	1.17	1.05	1.00
		10	**2.10**	**3.63**	**0.55**	**0.73**	0.86	**0.72**	0.87	**0.71**
	400	100	**1.43**	**2.22**	**0.62**	0.88	0.98	0.86	1.01	0.96
		1,000	1.12	**1.26**	0.84	1.05	1.18	1.01	1.12	1.07
		10	**2.55**	**3.18**	**0.65**	**0.76**	1.08	1.02	1.18	1.18
	50	100	**1.54**	**1.84**	1.12	**0.19**	1.20	1.08	**1.51**	**1.35**
		1,000	0.97	**3.56**	**2.14**	**0.01**	1.02	1.03	**1.55**	**1.67**
		10	**2.22**	**3.26**	**0.64**	0.80	0.95	1.07	0.88	0.94
3	200	100	**1.46**	**2.20**	**0.74**	**0.31**	1.04	0.98	1.07	1.01
		1,000	1.04	1.16	1.09	**0.01**	1.00	0.99	**1.25**	**1.23**
		10	**2.38**	**3.12**	**0.49**	0.94	1.01	1.12	1.01	1.06
	400	100	**1.42**	**2.36**	**0.52**	**0.56**	1.04	1.12	0.95	1.02
		1,000	0.87	1.03	**0.63**	**0.00**	0.91	0.87	1.01	1.18

* CAI and L_∞ of XLWP are almost identical and thus are not reported. Values that deviate more than 20% from the nominal level are highlighted in small bold font.

the performance of XLWP with L_∞-norm is mostly too conservative. Even though XY can produce reasonable type-I error, it does not fare well with heavy-tailed distributions (Model 3) and moderate dimensions. In addition, one can observe from Table 1 that the type-I error of XLWP with the L_2-norm tends to be inflated when p is small or moderate, e.g., 10 or 100. Thus, it is no surprising XLWP possesses a higher empirical power than other methods do in these settings as seen in Figures 1 and 2 for Model 2.

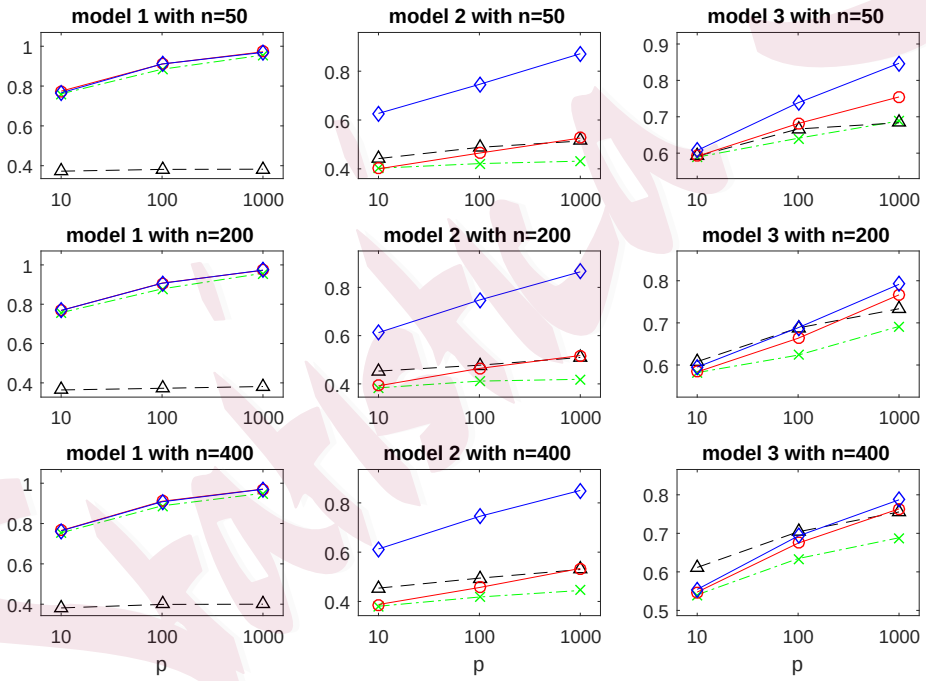

Figure 1: Simulation results with dense signals. In each panel, grey dash-dot line with cross represents CQ, black dash line with triangle represents XLWP with L_2-norm, red solid-line with circle S_0^1, and blue solid-line with diamond S_1^1.

Figure 1 shows the statistical power of various tests against dense alternatives with p varying from 10 to 1000 at significance level of 5%. We could

draw the following conclusions. The performance of XLWP differs dramatically across different models. Specifically, it has decent powers with Models 2 and 3, but has very low power for Model 1. Recall that XLWP incurs excessive type-I errors when dimension p is small (10 or 100). Therefore, we should be cautious with the high power of XLWP with Models 2 and 3, as very likely this comes at the price of an inflated type-I error. In contrast, the permutation test based on $S_1^1(\cdot)$ is always among the best performers across all the models.

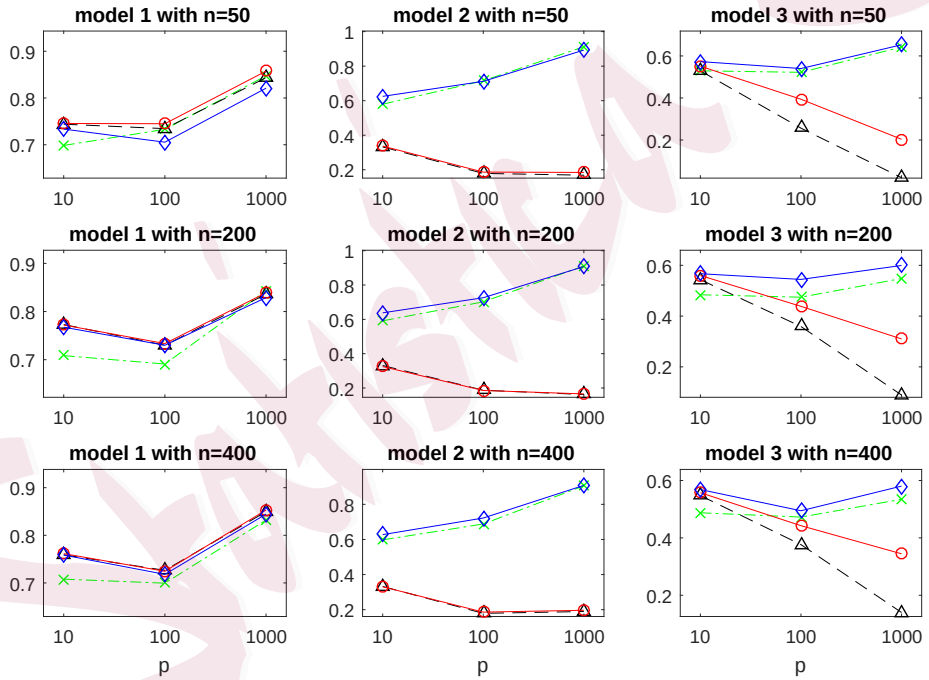

Figure 2: Simulation results with sparse signals. In each panel, grey dash-dot line with cross represents XLWP with L_∞-norm (or CLX), black dash line with triangle represents XY, red solid line with circle S_0^∞, and blue solid line with diamond S_1^∞.

Figure 2 depicts the changes in powers for all four tests against sparse alternatives, for which the L_∞-norm is expected to fare better. Since the powers of CAI and XLWP with the L_∞-norm are very similar in all the settings considered, we only report those of XLWP. For Models 1 and 2, the performances of the various methods are nearly indistinguishable, except for XY which is significantly worse than the other three when p is less than 100. Note that the high statistical power of XLWP is the consequence of the aforementioned unduly high type-I error. Similar to $S_1^\gamma(\cdot)$, CAI and XLWP also take into account of the possible difference across the variances; yet to our surprise, their performances for Model 3 seems to contradict conclusions drawn about their theoretical properties, especially when they are compared with $S_1^\gamma(\cdot)$. CAI and XLWP also suffer from the low power for Model 3, possibly due to the difficulty in estimating covariance matrices for heavy-tailed data. As for Model 3, the powers of XY and S_1^∞ substantially outperform that of XLWP and S_0^∞. However, XY achieves the same statistical power compared to S_1^∞ at the expense of inflated type-I error.

Overall, the permutation tests based on $S_0^\gamma(\cdot)$ and $S_1^\gamma(\cdot)$, with $\gamma = 1$ against dense alternatives and $\gamma = \infty$ against sparse alternatives, deliver better results than the other methods in terms of both empirical sizes and powers. Between $S_0^\gamma(\cdot)$ and $S_1^\gamma(\cdot)$, the former has a better control over the type-I error especially when the sample size is small, but the later usually achieves higher power.

Table 2: The average of true discovers (and FDR in the parenthesis) based on 10,000 replications at significance level 5%.

			$\#I_0 = 0$		$\#I_0 = 8$	
model	p	n	BY	S_1^∞	BY	S_1^∞
		200	–(0.0097)	–(0.0502)	0.8760(0.0098)	1.2690(0.0322)
	100	500	–(0.0105)	–(0.0486)	0.8710(0.0074)	1.2610(0.0330)
1		1,000	–(0.0104)	–(0.0486)	0.8700(0.0086)	1.2220(0.0232)
($\delta = 5$)		200	–(0.0056)	–(0.0544)	0.8070(0.0082)	1.2900(0.0320)
	10,000	500	–(0.0054)	–(0.0530)	0.8020(0.0080)	1.3170(0.0252)
		1,000	–(0.0050)	–(0.0494)	0.7810(0.0024)	1.2570(0.0239)
		200	–(0.0079)	–(0.0495)	1.6180(0.0081)	2.0770(0.0323)
	100	500	–(0.0076)	–(0.0503)	1.3880(0.0097)	1.8670(0.0311)
3		10,000	–(0.0079)	–(0.0511)	1.3880(0.0088)	1.8540(0.0228)
($\delta = 10$)		200	–(0.0039)	–(0.0596)	1.7910(0.0049)	2.5190(0.0205)
	10,000	500	–(0.0044)	–(0.0520)	1.5750(0.0028)	2.2860(0.0214)
		1,000	–(0.0042)	–(0.0505)	1.5380(0.0033)	2.2620(0.0233)

– no true signals

Next, we evaluate the performance of the permutation tests based on S_1^∞ in signal identification as described in Section 3. If the set of signals is empty, the empirical false discovery rate should be no more than a pre-set value, α (set to be 5% in this case). Otherwise, fix $\#I_0 = 8$ with the exact locations of the eight signals randomly distributed amongst $\{1, ..., p\}$ and the strength of signals given by

$$\delta \times \ln(\log(p))/\sqrt{n},$$

which changes with n and p. We compare our method S_1^∞ with Benjamini and Yekutieli (2001), denoted as BY. When $\#I_0 = 0$ (the first two columns of Table 2), it is obvious that our method S_1^∞ has a very good control over the false discovery rate, while BY tends to be over conservative. On the other hand, when $\#I_0 = 8$ (the last two columns of Table 2), our method S_1^∞ is able to identify more true signals than BY.

6. Analysis of WTCCC dataset

Genome-wide association studies (GWAS) are widely used to identify the risk genetic variants by genotyping millions of single nucleotide polymorphisms (SNPs) in large cohorts. The traditional GWAS analysis proceeds by a single-variant analysis that does account LD structure among SNPs and suffers from the heavy burden of multiple testing. Thus, the results from such analyses are usually conservative. As the proposed method is based on S_1^∞, it can explicitly control false discovery rate (FDR). We applied S_1^∞ to seven traits from Wellcome Trust Case Control Consortium (WTCCC) including bipolar disorder (BPD), coronary artery disease (CAD), Crohn's disease (CD), hypertension (HT), rheumatoid arthritis (RA), type 1 diabetes (T1D), and type 2 diabetes (T2D) (WTCCC, 2007). We performed strictly quality control on the samples from WTCCC using PLINK (Purcell et al., 2007) and GCTA (Yang et al., 2011). First, we removed individuals

with missing genotypes higher than 0.02. For each trait case and two shared control datasets, we removed SNPs with minor alleles frequencies less than 0.05 and SNPs with missing rate larger than 0.01. Then, we combined cases with controls for each trait and removed SNPs with p-values less than 0.001 for Hardy-Weinberg equilibrium test. Pairs of subjects with estimated relatedness greater than 0.025 were identified and one subject from these pairs is removed. After quality control, we have 1,959 cases and 2,992 controls over 308,093 SNPs for CAD, 1,970 cases and 2,992 controls over 307,741 SNPs for CD, 1,994 cases and 2,992 controls over 307,357 SNPs for T1D, and 1,969 cases and 2,992 controls over 305,394 SNPs for T2D. We applied the permutation test with S_1^∞ to the data, and the resulting Manhattan plots are shown in Figure 3. The analysis for each disease can be done around 16 minutes on a Windows console with 2.30GHz intel Xeron CPU E5-2697.

With significance level 1%, we summarize our findings as follows.For CAD, S_1^∞ identified 15 SNPs and all of them are from genes *AL359922.1* and *CDKN2B-AS1* within band 9p21.3. These two genes have previously been reported to be associated with CAD (van der Harst and Verweij , 2018; Lee et al., 2013). For CD, S_1^∞ identified 39 SNPs. Among these identified SNPs, 21 SNPs are within six gene regions, where all six genes were reported to be associated with CD in previous studies (Julià et al., 2013; de Lange et al., 2017; Liu et al., 2015). For T1D, S_1^∞ identified 369 SNPs and 173 SNPs

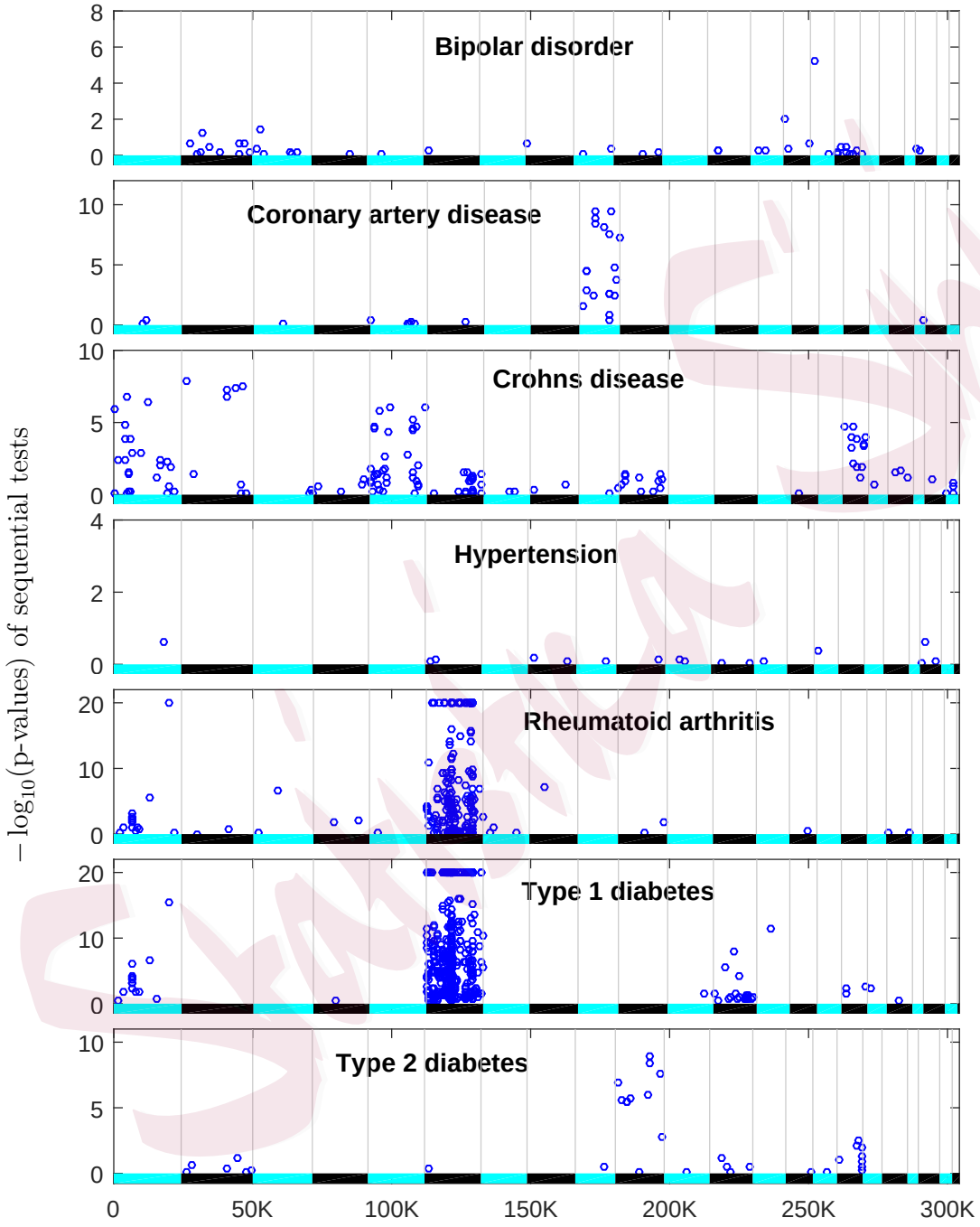

Figure 3: For each of seven diseases -log10 of the test p-value for quality-control-positive SNPs, (values bigger than 20 are censored at 20) are plotted against position of SNPs that arranged in according to the chromosomes in black and grey.

were within 83 genes, among which 23 genes were previously reported to be associated with T1D, including *ERBB3*, *CLEC16A* and *DDR1* (Plagnol et al., 2011; Hakonarson et al., 2007; Tomer et al., 2015). For T2D, S_1^{∞} identified 13 SNPs within two gene regions and both genes have previously been reported, i.e., *TCF7L2* and *FTO* (Hackinger et al., 2018; Tabassum et al., 2013).

Interestingly, in the analysis of the seven diseases using WTCCC data, we identified many "new" SNPs which were not reported in the original study of WTCCC (2007), but were detected in later studies. These SNPs and their corresponding studies are listed in Table S1 in the supplementary file. Statistically, it is more interesting to notice that those studies are based on either much larger cohorts or other populations. This indicates clearly the efficiency of our method in identifying the weak signals (SNPs) associated with the diseases.

ACKNOWLEDGMENTS. We are most grateful to the AE and three referees for their meticulous review, valuable comments and constructive suggestions. The rst two authors contributed equally to this work. EK was supported by a grant of National Natural Science Foundation of China (NNSFC) 11771066. YX was supported partially by National Natural Science Foundation of China 11931014 and AcRF grant R-155-000-220-114 of National University of Singapore; and JL was supported partially by AcRF

Tier 2 [MOE2016-T2-2-029, MOE2018-T2-1-046, MOE2018-T2-2-006] from Ministry of Education, Singapore and block fund R-913-200-098-263 from Duke-NUS Medical School.

References

Bai, Z. and Saranadasa, H. (1996) Effect of high dimension: By an example of a two sample problem. *Statistica Sinica*, **6**, 311-329.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B.* 57, 289-300.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics.* 29 1165-1188.

Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. *The Annals of Statistics*, **36**, 2577-2604.

Cai, T. and Liu, W. D. (2011) Adaptive thresholding for sparse covariance matrix estimation. *J. Amer. Statist. Assoc.*, **106**, 672-684.

Cai, T., Liu, W., and Xia, Y. (2014) Two-sample test of high dimensional means under dependence. *Journal of the Royal Statistical Society, Series B*, **76**, 349-372.

Chang, J., Zheng, C., Zhou, W.-X. and Zhou, W. (2017) Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics*, **73**, 1300-1310.

Chen, S., and Qin, Y. (2010) A two sample test for high dimensional data with applications to gene-set testing. *The Annals of Statistics*, **38**, 808-835.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017) Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, **45**, 2309-2352.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2017) Comparison and anti-concentration bounds for maxima of Gaussian random vectors. *Probability Theory and Related Fields*, **162**, 47-70.

Chung, E. and Romano, J. P. (2013) Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41**, 484-507.

Chung, E. and Romano, J. P. (2016) Multivariate and multiple permutation tests. *Journal of Econometrics*, **193**,76-91.

de Lange, K. M. and Moutsianas, L. and others (2017) Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nature Genetics*, **49**, 256.

Dizier, M-H and Demenais, F. and Mathieu, F. (2017) Gain of power of the general regression model compared to Cochran-Armitage Trend tests: simulation study and application to bipolar disorder *BMC Genetics*, **18**, 24.

Dwass, M. (1957) Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, **28**, 181-187.

Ernest, N. (2004) Permutation methods: a basis for exact inference. *Statistical Science*, **19**, 676-685.

Good, P. (2005) *Permutation, Parametric and Bootstrap Tests of Hypotheses*. Springer.

Hackinger, S., Prins, B. and others (2018) Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia. *Translational Psychiatry*, **8**, 252.

Hakonarson, H., Grant, S. and others (2007) A genome-wide association study identifies KIAA0350 as a type 1 diabetes gene. *Nature*, **448**, 591.

Hoeffding, W. (1952) The Large-Sample Power of Tests Based on Permutations of Observations. *Annals of Mathematical Statistics*, **23**, 169-192.

Jin, J. and Cai, T. (2007) Estimating the null and the proportion of non-null effects in large-scale multiple comparisons. *J. American Statistical Association*, 102, 495-506.

Julià, A. and Domènech, E. and others (2013) A genome-wide association study on a southern European population identifies a new Crohn's disease susceptibility locus at RBX1-EP300. *Gut*, **62**, 1440–1445.

Kosorok, M. and Ma, S. (2007) Marginal asymptotics for the large p, small n paradigm: with applications to microarray data. *The Annals of Statistics*, **35**, 1456-1486.

Lee, J.-Y. and Lee, B. and others (2013) A genome-wide association study of a coronary artery disease risk variant. *Journal of Human Genetics*, **58**, 120.

Liu, J. and van Sommeren, S. and others (2018) Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nature Genetics*, **47**, 979.

Plagnol, V., Howson, J. and others (2011) Genome-wide association analysis of autoantibody positivity in type 1 diabetes cases. *PLoS Genetics*, **7**, e1002216.

Purcell, S., Neale, B. and Todd-Brown, K. and Thomas, L. and others (2007) PLINK: a tool set for whole-genome association and population-

based linkage analyses. *The American Journal of Human Genetics*, **81**, 559–575.

Romano, J. P. (1990) On the behavior of randomization tests without a group invariance assumption. *J. Amer. Statist. Assoc.*, **85**, 686-692.

Tabassum, R. and Chauhan, G. and Dwivedi, O. P. and Mahajan, A. and others (2013) Genome-wide association study for type 2 diabetes in Indians identifies a new susceptibility locus at 2q21. *Diabetes*, **62**, 977–986.

Talagrand, M. (2003) *Spin Glasses: A Challenge for Mathematicians.* Springer.

Tomer, Y. and Dolan, L. M. and Kahaly, G. and others (2015) Genome wide identification of new genes and pathways in patients with both autoimmune thyroiditis and type 1 diabetes. *Journal of Autoimmunity*, **60**, 32–39.

van der Harst, P. and Verweij, N. (2018) Identification of 64 novel genetic loci provides an expanded view on the genetic architecture of coronary artery disease. *Circulation research*, **122**, 433–443.

van Hulzen, K. and Scholz, C. J. Franke, B., Ripke, S. and others (2017) Genetic overlap between attention-deficit/hyperactivity disorder and

bipolar disorder: evidence from genome-wide association study meta-analysis. *Biological Psychiatry*, **82**, 634–641.

Wellcome Trust Case Control Consortium and others (2011) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661.

Xu, G., Lin, L., Wei, P. and Pan, W. (2016) An adaptive two-sample test for high dimensional means. *Biometrika*, **103**, 609-624.

Xue, K. and Yao, F. (2018) Distribution and correlation free two-sample test high dimensional means. *Annals of Statistics* (to appear).

Yang, J. and Lee, S. H. and Goddard, M. E. and Visscher, P. M. (2011) GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics*, **88**, 76–82.

Supplementary Material. All technical proofs are given in a separate supplemental file.