Biometrika (2016), **103**, 3, *pp.* 609–624
Printed in Great Britain
doi: 10.1093/biomet/asw029

An adaptive two-sample test for high-dimensional means

By GONGJUN XU

School of Statistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
xuxxx360@umn.edu

LIFENG LIN

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
linl@umn.edu

PENG WEI

Division of Biostatistics, University of Texas School of Public Health, Houston, Texas 77030, U.S.A.
peng.wei@uth.tmc.edu

and WEI PAN

Division of Biostatistics, University of Minnesota, Minneapolis, Minnesota 55455, U.S.A.
weip@biostat.umn.edu

Summary

Several two-sample tests for high-dimensional data have been proposed recently, but they are powerful only against certain alternative hypotheses. In practice, since the true alternative hypothesis is unknown, it is unclear how to choose a powerful test. We propose an adaptive test that maintains high power across a wide range of situations and study its asymptotic properties. Its finite-sample performance is compared with that of existing tests. We apply it and other tests to detect possible associations between bipolar disease and a large number of single nucleotide polymorphisms on each chromosome based on data from a genome-wide association study. Numerical studies demonstrate the superior performance and high power of the proposed test across a wide spectrum of applications.

Some key words: Genome-wide association study; Single nucleotide polymorphism; Sum-of-powers test.

1. Introduction

Two-sample testing on the equality of two high-dimensional means is common in genomics and genetics. For instance, Chen & Qin (2010) considered analysis of differential expressions for gene sets based on microarray data. In our motivating example and other genome-wide association studies (The International Schizophrenia Consortium, 2009), polygenic testing is of interest: one would like to test whether there is any association between a disease and a large number of genetic variants, mostly single nucleotide polymorphisms. In these applications, the dimension of the data, p, is often much larger than the sample size n. Traditional multivariate two-sample tests, such as the T^2-test of Hotelling (1931), either cannot be directly applied or have too low power. As shown theoretically in Fan (1996), as the dimension p increases, even for simple one-sample

testing on the mean of a normal distribution with a known covariance matrix $\sigma^2 I$, the standard Wald, score or likelihood ratio tests may have power that decreases to the Type I error rate as the departure from the null hypothesis increases. Several two-sample tests for high-dimensional data have been proposed (Bai & Saranadasa, 1996; Srivastava & Du, 2008; Chen & Qin, 2010; Cai et al., 2014; Gregory et al., 2015; Srivastava et al., 2015). There are two common types of testing approach when $p > n$: one based on the sum-of-squares of the sample mean differences and the other based on the maximum componentwise sample mean difference. The two types of tests are powerful against different alternatives: if the true mean differences are dense in the sense that there is a large proportion of small to moderate componentwise differences, then the former type is more powerful; in contrast, if the true mean differences are sparse in the sense that there are only few but large nonzero componentwise differences, the latter type of test is more powerful. In practice, however, it is unclear which should be applied. Furthermore, as will be shown in the simulation study, there are denser and intermediate situations in which neither type of test is powerful.

In this paper, we develop an adaptive testing procedure which yields high testing power against various alternative hypotheses in the high-dimensional setting. This is achieved through combining information across a class of sum-of-powers tests, including tests based on the sum-of-squares of the mean differences and the supremum mean difference. The main idea is to incorporate multiple tests in the procedure so that at least one of them would yield a high power for a particular application with unknown truth. The proposed adaptive sum-of-powers test then selects the most powerful of the candidate tests. To perform the proposed test, we establish the asymptotic null distribution of the adaptive test statistic. In particular, we derive the joint asymptotic distribution for a set of the sum-of-powers test statistics. The marginal distributions of the test statistics converge to the normal distribution or the extreme value distribution, depending on the power parameters. Based on the theoretical results, we develop a new way to calculate asymptotic p-values for the adaptive test.

We further demonstrate the superior performance of the proposed adaptive test in the context of large p and small n. We compare its performance with several existing tests which have not yet been applied to single nucleotide polymorphism data. Due to the discrete nature of single nucleotide polymorphism data, normal-based parametric tests are not suitable. In addition, although the sparsity assumption has been so widely adopted, the nonzero differences in single nucleotide polymorphism data may not be sparse, as predicted by the polygenic theory of Fisher (1918). The problem of nonsparse signals has begun to attract the attention of statisticians (e.g., Hall et al., 2014). It is highly relevant here because the performance of a test, especially a non-adaptive one, may depend on how sparse the signals are, as illustrated in the real-data analysis. An R (R Development Core Team, 2016) package highmean that implements the tests studied here is available from the Comprehensive R Archive Network, CRAN.

2. Some existing tests

Suppose that we observe two groups of p-dimensional independent and identically distributed samples $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$; we consider high-dimensional data with $p \gg n = n_1 + n_2 - 2$. Let μ_1 and μ_2 denote the true mean vectors of the groups, and assume throughout that the two groups share a common covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$. Our primary object is to test $H_0 : \mu_1 = \mu_2$ versus $H_A : \mu_1 \neq \mu_2$. In this section, we review existing two-sample tests for high-dimensional data. For $k = 1, 2$ let $\bar{X}_k$ be the sample mean for group k, and let $S = n^{-1}\sum_{k=1}^{2}\sum_{i=1}^{n_k}(X_{ki} - \bar{X}_k)(X_{ki} - \bar{X}_k)^{\mathrm{T}}$ be the pooled sample covariance matrix. The precision matrix, i.e., the inverse of the covariance matrix, is written as $\Omega = \Sigma^{-1}$. Moreover, for a vector v, we denote by $v^{(i)}$ its ith element.

The best-known two-sample test for low-dimensional data is the T^2-test of Hotelling (1931), which is a generalization of the two-sample t-test for $p = 1$ to multivariate data with $p > 1$ but $p \ll n$: $T_{\mathrm{H}} = (\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} S^{-1} (\bar{X}_1 - \bar{X}_2)$. The T^2-test, however, is not applicable to high-dimensional data because S is singular. Accordingly, some modifications have been proposed in which S is replaced by a known quantity or another estimate. A straightforward procedure is to substitute an identity matrix I for S, forming a sum-of-squares-type test, which is directly based on the L_2-norm of the sample mean differences, $\|\bar{X}_1 - \bar{X}_2\|_2^2 = (\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} (\bar{X}_1 - \bar{X}_2)$, or its weighted version (Bai & Saranadasa, 1996; Srivastava & Du, 2008; Chen & Qin, 2010). Bai & Saranadasa (1996) proposed a test statistic

$$T_{\mathrm{BS}} = \frac{(n_1^{-1} + n_2^{-1})^{-1} (\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} (\bar{X}_1 - \bar{X}_2) - \operatorname{tr} S}{[2n(n+1)(n-1)^{-1}(n+2)^{-1}\{\operatorname{tr} S^2 - n^{-1}(\operatorname{tr} S)^2\}]^{1/2}}$$

and established its asymptotic normal null distribution. Chen & Qin (2010) noticed some theoretical difficulties due to the presence of the cross-product terms $\sum_{i=1}^{n_k} X_{ki}^{\mathrm{T}} X_{ki}$ in T_{BS}, and proposed removing them to obtain a new test statistic

$$T_{\mathrm{CQ}} = \frac{\sum_{i \neq j}^{n_1} X_{1i}^{\mathrm{T}} X_{1j}}{n_1(n_1-1)} + \frac{\sum_{i \neq j}^{n_2} X_{2i}^{\mathrm{T}} X_{2j}}{n_2(n_2-1)} - 2\frac{\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} X_{1i}^{\mathrm{T}} X_{2j}}{n_1 n_2},$$

whose asymptotic properties were established under much weaker conditions.

To account for possibly varying variances of the components of the data, one may replace S by a diagonal version $D_S = \operatorname{diag}(s_{11}, \ldots, s_{pp})$, where s_{ii} are the diagonal elements of S; the matrix D_S is in general nonsingular. Srivastava & Du (2008) introduced such a weighted version of the sum-of-squares-type test of Bai & Saranadasa (1996):

$$T_{\mathrm{SD}} = \frac{(n_1^{-1} + n_2^{-1})^{-1} (\bar{X}_1 - \bar{X}_2)^{\mathrm{T}} D_S^{-1} (\bar{X}_1 - \bar{X}_2) - (n-2)^{-1} np}{\{2(\operatorname{tr} R^2 - p^2 n^{-1}) c_{p,n}\}^{1/2}},$$

where $R = D_S^{-1/2} S D_S^{-1/2}$ is the sample correlation matrix and $c_{p,n} = 1 + \operatorname{tr} R^2 p^{-3/2}$.

All of the above sum-of-squares-type test statistics are asymptotically distributed as normal under H_0. These tests are usually powerful against moderately dense alternative hypotheses, where there is a large proportion of nonzero components in the true mean differences $\mu_1 - \mu_2$. However, if the nonzero signals are sparse, these tests lose substantial power (Cai et al., 2014). Accordingly, Cai et al. (2014) proposed a supremum-type statistic using the L_∞-norm of the sample mean differences, i.e.,

$$T_{\mathrm{CLX}} = (n_1^{-1} + n_2^{-1})^{-1} \max_{1 \leqslant i \leqslant p} \big(\bar{X}_1^{(i)} - \bar{X}_2^{(i)}\big)^2 / \sigma_{ii},$$

where σ_{ii} are the diagonal elements of the covariance matrix Σ. In practice, we use the sample variances s_{ii} to estimate the σ_{ii}.

A supremum-type statistic and a sum-of-squares-type statistic represent two extremes: the former uses only a single component as evidence against the null hypothesis, while the latter uses all of the components. Neither of the statistics will be uniformly better; they are more powerful for sparse and dense nonzero signals, respectively (Gregory et al., 2015). However, for more dense or only weakly dense nonzero signals, neither may be powerful: there may not be a single component to represent a strong departure from H_0, whereas a sum-of-squares statistic may accumulate too much noise through summing over the zero components. To boost the power when nonzero signals are neither too dense nor too sparse, Chen et al. (2014) proposed removing estimated zero

components through thresholding; since zero components are expected to give small squared sample mean differences, those smaller than a given threshold would be ignored, leading to a test statistic

$$T_{\mathrm{CLZ}}(s) = \sum_{i=1}^{p} \left\{ \frac{n_1 n_2 (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2}{(n_1+n_2)\sigma_{ii}} - 1 \right\} I\left\{ \frac{n_1 n_2 (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2}{(n_1+n_2)\sigma_{ii}} > \lambda_p(s) \right\},$$

where the threshold level is $\lambda_p(s) = 2s \log p$ and $I(\cdot)$ is the indicator function. Since an optimal choice of the threshold is unknown, Chen et al. (2014) proposed trying all possible threshold values and then choosing the most significant one as the final test statistic:

$$T_{\mathrm{CLZ}} = \max_{s \in (0, 1-\eta)} \left\{ T_{\mathrm{CLZ}}(s) - \hat{\mu}_{T_{\mathrm{CLZ}}(s),0} \right\} / \hat{\sigma}_{T_{\mathrm{CLZ}}(s),0},$$

where $\hat{\mu}_{T_{\mathrm{CLZ}}(s),0}$ and $\hat{\sigma}_{T_{\mathrm{CLZ}}(s),0}$ are estimates of the mean and standard deviation of $T_{\mathrm{CLZ}}(s)$ under the null hypothesis. The asymptotic null distribution of T_{CLZ} is an extreme value distribution. Because of the slow convergence to the asymptotic null distribution, Chen et al. (2014) proposed using the parametric bootstrap to calculate its p-values. The test T_{CLZ} can be regarded as an adaptive test: it uses thresholding to adapt to unknown signal sparsity. It is closely related to another adaptive test for association analysis of rare variants in genetics (Pan & Shen, 2011).

Remark 1. Sum-of-squares-type tests and supremum-type tests have also been used in analyses of genome-wide association studies with large n and small p. For example, in the framework of generalized linear models, the sum-of-squared-score test in Pan (2009) for association analysis of multiple single nucleotide polymorphisms can be regarded as a sum-of-squares-type test, while another widely used test in single nucleotide polymorphism analysis is similar to the supremum-type test of Cai et al. (2014). As shown in Pan (2011), the sum-of-squared-score test is equivalent to the variance-component-score test with a linear kernel (Wu et al., 2010) and a nonparametric multivariate analysis of variance (Wessel & Schork, 2006), both used in genetics, as well as to an empirical Bayes test for high-dimensional data (Goeman et al., 2006).

Remark 2. Cai et al. (2014) also introduced test statistics based on linearly transformed sample mean differences. Although they discussed the transformation only for their supremum-type statistic, the same transformation can be applied to other test statistics (Chen et al., 2014). However, the transformation may not work for very dense signals; for example, in some cases with more than p^r nonzero signals for $r > 1/2$, a test using the precision matrix transformation could be outperformed by that without transformation (Cai et al., 2014). Furthermore, conducting the Ω-transformation requires an estimate of the $p \times p$ precision matrix, which is time-consuming for large p (Gregory et al., 2015). More importantly, any test can be conducted on either the original or the transformed data, which is not a focus here. Therefore, this article does not consider data transformations.

3. Main results

3·1. *Test statistics*

We first propose a family of sum-of-powers tests, indexed by a positive integer γ. For any $1 \leqslant \gamma < \infty$, we define a sum-of-powers test statistic with power index γ as

$$L(\gamma) = \sum_{i=1}^{p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^{\gamma}.$$

When $\gamma = 2$, this yields a sum-of-squares-type test statistic equivalent to that of Bai & Saranadasa (1996). Since, as an even $\gamma \to \infty$,

$$L(\gamma) \propto \|\bar{X}_1 - \bar{X}_2\|_\gamma \to \|\bar{X}_1 - \bar{X}_2\|_\infty = \max_{1 \leqslant i \leqslant p} |\bar{X}_1^{(i)} - \bar{X}_2^{(i)}|,$$

following the supremum-type test statistic in Cai et al. (2014) we define

$$L(\infty) = \max_{1 \leqslant i \leqslant p} (\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^2 / \sigma_{ii}.$$

Thus, the class of the sum-of-powers tests includes both a sum-of-squares test and a supremum-type test as special cases. Furthermore, $L(1)$ is like a burden test widely studied in genetic association analysis of rare variants for large n and small p (Pan & Shen, 2011; Lee et al., 2012). If nonzero signals are extremely dense with almost the same sign, then a burden test like $L(1)$ can be more powerful than both the sum-of-squares and the supremum-type tests; see our numerical examples and § 3·3. Similarly, there are situations with only weakly dense signals, in which an $L(\gamma)$ test with $2 < \gamma < \infty$ may be more powerful than both the sum-of-squares and the supremum-type tests.

Which $L(\gamma)$ is most powerful depends on the unknown pattern of nonzero signals, such as sparsity and signal strength. Hence, we propose the following adaptive test to combine the sum-of-powers tests and improve the test power:

$$T_{\text{aSPU}} = \min_{\gamma \in \Gamma} P_{\text{SPU}(\gamma)}, \tag{1}$$

where $P_{\text{SPU}(\gamma)}$ is the p-value of $L(\gamma)$ test. The idea of taking the minimum p-value to approximate the maximum power has been widely used (e.g., Yu et al., 2009), but T_{aSPU} is no longer a genuine p-value. In order to perform the proposed adaptive test, in the next section we derive the asymptotic distribution. In practice, one has to decide what candidate values of γ are to be used. From the theoretical power study in § 3·3 and the simulation study, we suggest using $\gamma \in \Gamma = \{1, 2, \ldots, \gamma_u, \infty\}$ with $\gamma_u = 6$ or a little bigger for a larger p/n ratio.

Remark 3. Our tests for small n and large p are in the same spirit as those proposed for analysis of rare variants with large n and small p (Pan et al., 2014). Specifically, Pan et al. (2014) defined $\text{SPU}(\gamma) = \sum_{i=1}^{p} U_i^\gamma$, where $U = (U_1, \ldots, U_p)^{\mathrm{T}}$ is the score vector for a parameter, say β, in a generalized linear model under a null hypothesis $H_0' : \beta = 0$. The $\text{SPU}(\gamma)$ test can be regarded as a weighted score test (Lin & Tang, 2011) with weights $w_i = U_i^{\gamma-1}$. In the current context, the score U becomes the sample mean difference, so we use the same name and denote the adaptive test statistic by T_{aSPU} in (1). Apart from, the difference between the small n large p and large n small p scenarios, asymptotic results for the adaptive test have not yet been described in the literature. In this paper we derive asymptotics of the test statistics $L(\gamma)$ in the high-dimensional setting, based on which we can calculate the asymptotic p-values of $L(\gamma)$ and T_{aSPU}.

3·2. *Asymptotic theory*

For simplicity, we present our results under the assumption of a common covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma$, although our derivations and proofs in the Supplementary Material are established without this assumption. In the following we write $\Sigma = (\sigma_{ij})_{p \times p}$. Under $H_0 : \mu_1 = \mu_2$, we first derive asymptotic approximations to the mean and variance of $L(\gamma)$ for $\gamma < \infty$, denoted by $\mu(\gamma)$ and $\sigma^2(\gamma)$, respectively. We assume that $n_1/(n_1 + n_2) \to \rho \in (0, 1)$ as $n \to \infty$. We write $a_n \sim b_n$ if $a_n/b_n = 1 + o(1)$ and let $\lfloor x \rfloor$ denote the largest integer not greater than x.

We need the following assumptions.

Condition 1 (Covariance assumption). There exists some constant B such that $B^{-1} \leqslant \lambda_{\min}(\Sigma) \leqslant \lambda_{\max}(\Sigma) \leqslant B$, where $\lambda_{\min}(\Sigma)$ and $\lambda_{\max}(\Sigma)$ denote the minimum and maximum eigenvalues of the covariance matrix Σ. In addition, all correlations are bounded away from -1 and 1, i.e., $\max_{1\leqslant i\neq j\leqslant p} |\sigma_{ij}|/(\sigma_{ii}\sigma_{jj})^{1/2} < 1-\eta$ for some $\eta > 0$.

Condition 2 (Mixing assumption). For a set of multivariate random vectors $Z = \{Z^{(j)} : j \geqslant 1\}$ and integers $a < b$, let $\mathcal{Z}_a^b$ be the σ-algebra generated by $\{Z^{(j)} : j \in [a, b]\}$. For each $s \geqslant 1$, define the α-mixing coefficient $\alpha_Z(s) = \sup_{t\geqslant 1}\{|\mathrm{pr}(A \cap B) - \mathrm{pr}(A)\mathrm{pr}(B)| : A \in \mathcal{Z}_1^t,\ B \in \mathcal{Z}_{t+s}^{\infty}\}$. We assume that $\{(X_{ki}^{(j)},\ i = 1, \ldots, n_k) : j \geqslant 1\}$ is α-mixing for $k = 1, 2$ and that $\alpha_X(s) \leqslant M\delta^s$, where $\delta \in (0, 1)$ and M is some constant.

Condition 3 (Moment assumption). We assume that $\log p/n^{1/4} = o(1)$ and

$$\max_{1\leqslant i\leqslant p} E[\exp\{h(X_{k1}^{(i)} - \mu_k^{(i)})^2\}] < \infty$$

for $h \in [-M, M]$ and $k = 1, 2$.

Remark 4. Conditions 1 and 3 were also assumed in Cai et al. (2014), and they are needed to establish the weak convergence of $L(\infty)$. When $\gamma < \infty$, asymptotic normality can be established under weaker assumptions on the eigenvalues and correlations. However, in order to establish weak convergence of $L(\gamma)$ for $\gamma > 2$, stronger moment assumptions may be needed than those in Chen & Qin (2010), whose test statistic is similar to $L(2)$. Condition 2 imposes weak dependence on the data. A similar mixing condition is considered in Chen et al. (2014), and such weak dependence is also commonly assumed in time series and spatial statistics. Alternatively, we may consider the weak dependence structure introduced in Bai & Saranadasa (1996) and Chen & Qin (2010), where a factor-type model for X is assumed. Since the variables in the motivating genome-wide association studies have a local dependence structure, with their correlations often decaying to zero as their physical distances on a chromosome increase, we focus on mixing-type weak dependence in this paper.

We write $\mu(\gamma) = \sum_{i=1}^p \mu^{(i)}(\gamma)$, where $\mu^{(i)}(\gamma) = E\{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^\gamma\}$. Then the following approximation holds for $\mu(\gamma)$ and $\sigma^2(\gamma)$ with $\gamma < \infty$.

PROPOSITION 1. *Under $H_0 : \mu_1 = \mu_2$, we have $\mu^{(i)}(1) = 0$ and*

$$\mu^{(i)}(\gamma) = \begin{cases} \dfrac{\gamma!}{2^{\gamma/2}} \displaystyle\sum_{d=0}^{\gamma/2} \frac{1}{d!\,(\gamma/2-d)!\,n_1^d n_2^{\gamma/2-d}}\,\sigma_{ii}^{\gamma/2} + o(n^{-\gamma/2}), & \gamma \text{ even}, \\ \displaystyle\sum_{d=1}^{\lfloor\gamma/2\rfloor} \frac{\gamma!}{(d-1)!\,(\lfloor\gamma/2\rfloor-d)!\,3!\,2^{\lfloor\gamma/2\rfloor-1}} \\ \quad \times \left(\dfrac{m_{1i}}{n_1^{d+1} n_2^{\lfloor\gamma/2\rfloor-d}} - \dfrac{m_{2i}}{n_1^{\lfloor\gamma/2\rfloor-d} n_2^{d+1}}\right) \sigma_{ii}^{\lfloor\gamma/2\rfloor-1} + o(n^{-\lfloor\gamma/2\rfloor-1}), & \gamma \geqslant 3 \text{ odd}, \end{cases}$$

where m_{ki} is the third central moment of the random variable in component i from group k, i.e., $m_{ki} = E\{(X_k^{(i)} - \mu_k^{(i)})^3\}$.

For any positive integers s and t with $s+t$ even, define a set $\mathcal{A}(c_1, c_2, c_3, d_1, d_2, d_3; s, t)$ of integers $(c_1, c_2, c_3, d_1, d_2, d_3)$ such that $c_1 \geqslant 0$, $c_2 \geqslant 0$, $d_1 \geqslant 0$, $d_2 \geqslant 0$, $c_3 + d_3 > 0$, $2c_1 + c_3 + 2d_1 + d_3 = s$ and $2c_2 + c_3 + 2d_2 + d_3 = t$. For simplicity, we write the set as $\mathcal{A}(s, t)$.

PROPOSITION 2. *Under Conditions 1–3 and H_0, $\sigma^2(1) = (n_1{}^{-1} + n_2{}^{-1}) 1_p^{\mathrm{T}} \Sigma 1_p$ where 1_p is a $p \times 1$ vector whose elements are all 1, and, for $\gamma \geqslant 2$,*

$$\sigma^2(\gamma) = \mu(2\gamma) - \sum_{i=1}^{p} \{\mu^{(i)}(\gamma)\}^2 + \sum_{\mathcal{A}(\gamma,\gamma)} \frac{(\gamma!)^2 \sum_{i \neq j} \sigma_{ii}^{c_1+d_1} \sigma_{jj}^{c_2+d_2} \sigma_{ij}^{c_3+d_3}}{n_1^{c_1+c_2+c_3} n_2^{d_1+d_2+d_3} c_1!\, c_2!\, c_3!\, d_1!\, d_2!\, d_3!\, 2^{c_1+c_2+d_1+d_2}} + o(pn^{-\gamma}). \tag{2}$$

Because $\mu(2\gamma) - \sum_{i=1}^{p} \{\mu^{(i)}(\gamma)\}^2 = \sum_{i=1}^{p} \mathrm{var}\{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^\gamma\}$, we have that $\sigma^2(\gamma) \sim \sum_{i=1}^{p} \mathrm{var}\{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})^\gamma\}$ if $\sigma_{ij} = 0$ for any $i \neq j$. Since the boundedness condition on the eigenvalues, Condition 1, implies the boundedness of the variances σ_{ii}, $\sigma^2(\gamma)$ is of order $pn^{-\gamma}$.

To derive the asymptotic joint distribution of the test statistics $L(\gamma)$, we also need the following result to approximate their correlations: $\mathrm{corr}\{L(s), L(t)\} = \mathrm{cov}\{L(s), L(t)\}/\{\sigma(s)\sigma(t)\}$.

PROPOSITION 3. *Under Conditions 1–3 and $H_0 : \mu_1 = \mu_2$, for finite $s, t \in \Gamma$, if $s+t$ is even then*

$$\mathrm{cov}\{L(s), L(t)\} = \mu(s+t) - \sum_{i=1}^{p} \mu^{(i)}(t)\mu^{(i)}(s) + \sum_{\mathcal{A}(s,t)} \frac{t!\, s! \sum_{i \neq j} \sigma_{ii}^{c_1+d_1} \sigma_{jj}^{c_2+d_2} \sigma_{ij}^{c_3+d_3}}{n_1^{c_1+c_2+c_3} n_2^{d_1+d_2+d_3} c_1!\, c_2!\, c_3!\, d_1!\, d_2!\, d_3!\, 2^{c_1+c_2+d_1+d_2}} + o(pn^{-(s+t)/2});$$

and if $s+t$ is odd, $\mathrm{cov}\{L(s), L(t)\} = o(pn^{-(s+t)/2})$.

Now we are ready to introduce the asymptotic joint distributions for the test statistics $L(\gamma)$.

THEOREM 1. *Let Γ be a candidate set of γ values containing ∞. Assume that $\liminf_{n\to\infty} \sigma^2(\gamma)/(pn^{-\gamma}) > 0$ for $\gamma < \infty$. Under Conditions 1–3 and the null hypothesis $H_0 : \mu_1 = \mu_2$, the following properties hold*:

(i) *for the set $\Gamma' = \Gamma \backslash \{\infty\}$, $[\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma)]_{\gamma\in\Gamma'}^{\mathrm{T}}$ converges weakly to a normal distribution $N(0, \mathcal{R})$, where $\mathcal{R} = (\rho_{st})$ satisfies $\rho_{ss} = 1$ for $s \in \Gamma'$ and $\rho_{st} = \mathrm{corr}\{L(s), L(t)\}$ for $s \neq t \in \Gamma'$. In particular, $\rho_{st} = o(1)$ when $s+t$ is odd*;

(ii) *when $\gamma = \infty$, $\mathrm{pr}\{n_1 n_2/(n_1+n_2) L(\infty) - a_p \leqslant x\} \to \exp\{-\pi^{-1/2} \exp(-x/2)\}$ for any $x \in \mathbb{R}$, where $a_p = 2\log p - \log\log p$*;

(iii) *$[\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma)]_{\gamma\in\Gamma'}^{\mathrm{T}}$ and $\{n_1 n_2/(n_1+n_2) L(\infty) - a_p\}$ are asymptotically independent.*

We can use Propositions 1–3 to approximate $\mu(\gamma)$, $\sigma(\gamma)$ and ρ_{st}, respectively, and then calculate the p-value for the proposed adaptive test. Define L_{O} and L_{E} as the sets consisting of standardized $L(\gamma)$ with γ odd and even, respectively, i.e., $L_{\mathrm{O}} = \{\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma) : \text{odd } \gamma \in \Gamma\}$ and $L_{\mathrm{E}} = \{\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma) : \text{even } \gamma \in \Gamma\}$. By Theorem 1, L_{O} and L_{E} are asymptotically independent, and each is asymptotically independent of $L(\infty)$. Thus we can obtain the p-value of the adaptive test from these three sets of statistics. Consider the realizations of the test statistics, $T_{\mathrm{O}} = \max_{\text{odd } \gamma \in \Gamma} |\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma)|$ and $T_{\mathrm{E}} = \max_{\text{even } \gamma \in \Gamma} \{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma)$. We calculate the p-values for T_{O} and T_{E} as $p_{\mathrm{O}} = \mathrm{pr}[\max_{\text{odd } \gamma \in \Gamma} |\{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma)| > T_{\mathrm{O}}]$ and $p_{\mathrm{E}} = \mathrm{pr}[\max_{\text{even } \gamma \in \Gamma} \{L(\gamma) - \mu(\gamma)\}/\sigma(\gamma) > T_{\mathrm{E}}]$. We use the function pmvnorm in the R package mvtnorm to calculate the multivariate normal tail probabilities p_{O} and p_{E} (R Development Core Team, 2016). Finally, we take the minimum p-value from the odd, even and infinity tests, i.e., $p_{\min} = \min(p_{\mathrm{O}}, p_{\mathrm{E}}, p_{\infty})$; then, by the asymptotic independence of L_{O}, L_{E} and $L(\infty)$, the asymptotic p-value for the adaptive test is $p_{\mathrm{aSPU}} = 1 - (1 - p_{\min})^3$.

The above discussion focuses on the case where the covariance matrix Σ is known. In practice, Σ must be estimated. We can apply existing methods, such as banding and thresholding techniques, to estimate a high-dimensional sparse covariance matrix (Bickel & Levina, 2008; Rothman et al., 2010; Cai & Liu, 2011; Xue et al., 2012). In the simulation study and real-data analysis, we used the banding approach of Bickel & Levina (2008): for a sample covariance matrix $S = (s_{ij})$, the banded matrix with bandwidth k_n is defined as $\hat{\Sigma}_{k_n} = \{s_{ij} I(|i - j| \leqslant k_n)\}$. Theoretical properties of $\hat{\Sigma}_{k_n}$ have been studied in Bickel & Levina (2008). We used five-fold crossvalidation to select an optimal bandwidth in our simulations and real-data analysis (Bickel & Levina, 2008; Cai & Liu, 2011). Under the conditions in Theorem 1, we can show that $\hat{\sigma}^2(\gamma)$ estimated based on the banded matrix $\hat{\Sigma}_{k_n}$ satisfies $\hat{\sigma}^2(\gamma) = \{1 + o(1)\}\sigma^2(\gamma)$ for properly chosen k_n. Consider the approximation of $\hat{\sigma}(\gamma)$ in (2). Under the weak dependence condition, Condition 2, for any i, j and $\epsilon > 0$, there is a constant C such that $\sigma_{ij} \leqslant C\delta^{|i-j|\epsilon/(2+\epsilon)}$ (see, e.g., Guyon, 1995). Therefore, for $k_n \to \infty$ as $n \to \infty$, the sum of terms with $|i - j| > k_n$, i.e., $\sum_{i \neq j; |i-j| > k_n} \sigma_{ii}^{c_1+d_1} \sigma_{jj}^{c_2+d_2} \sigma_{ij}^{c_3+d_3}$, is ignorable. On the other hand, in $\sum_{i \neq j; |i-j| \leqslant k_n} \sigma_{ii}^{c_1+d_1} \sigma_{jj}^{c_2+d_2} \sigma_{ij}^{c_3+d_3}$ there are $O(k_n p)$ summands in total. Since $s_{ij} = \sigma_{ij} + O_{\mathrm{p}}(n^{-1/2})$, we can obtain $\hat{\sigma}^2(\gamma) - \sigma^2(\gamma) = o_{\mathrm{p}}(pn^{-\gamma})$ if $k_n = o(n^{1/2})$. By the result that $\sigma^2(\gamma)$ is of order $pn^{-\gamma}$, we obtain $\hat{\sigma}^2(\gamma) = \{1 + o(1)\}\sigma^2(\gamma)$. Similarly, we can show that $\hat{\mu}(\gamma) = \{1 + o(1)\}\mu(\gamma)$ and the estimators of the correlations are consistent.

In applications, the components of the observations may be measured on different scales. Therefore, we could consider an inverse variance weighted test statistic $W(\gamma) = \sum_{i=1}^{p} \{(\bar{X}_1^{(i)} - \bar{X}_2^{(i)})/\sqrt{\sigma_{ii}}\}^{\gamma}$ $(1 \leqslant \gamma < \infty)$. For $\gamma = \infty$, $L(\infty)$ is already weighted by the inverse variances and we let $W(\infty) = L(\infty)$. To calculate the p-values for $W(\gamma)$ and the corresponding adaptive test, it is straightforward to use the asymptotic properties on the weighted samples $\{Y_{ki}\}_{i=1}^{n_k}$, where $Y_{ki} = D^{-1/2} X_{ki}$ and $D = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. In practice, we replace the unknown σ_{ii} with the sample variances s_{ii}. Results similar to Theorem 1 can be established.

Remark 5. For simplicity, this paper focuses on the case where two groups of samples share a common covariance matrix. More generally, the two groups may have different covariance matrices, $\Sigma_1 \neq \Sigma_2$. In this situation, one may apply a two-sample test without assuming a common covariance matrix: the definitions of the tests remain the same, while for the weighted tests, the weights for the sample mean differences become the reciprocals of the diagonal elements in $\Sigma_1/n_1 + \Sigma_2/n_2$. The asymptotic properties of the proposed tests are still valid in this situation; see the Supplementary Material.

Remark 6. The asymptotic independence of the sum-of-squares- and supremum-type statistics has been studied in Hsing (1995) for weakly dependent observations. Under the sparse signal alternative with $1/2 < \beta < 1$, similar tests to the proposed $L(1)$ and $L(2)$ have also been studied in Zhong et al. (2013) with an additional higher criticism thresholding of the means; the asymptotic independence between the sum-of-squares-type statistics and a screening statistic by higher criticism thresholding has been studied in Fan et al. (2015). However, our study differs from theirs in several respects. First, our proposed method is adaptive and powerful for both sparse and dense signal alternatives, as shown by the theoretical and numerical results, whereas Zhong et al. (2013) and Fan et al. (2015) focus on sparse alternatives. As illustrated in the simulation, when the signals are dense, the proposed test performs better than the thresholding-type test in Chen et al. (2014). Second, we theoretically study a family of power statistics $L(\gamma)$ with different finite and infinite values of γ and establish their joint distribution; Zhong et al. (2013), on the other hand, focused on $L(1)$- and $L(2)$-type statistics and studied their performance separately, while Fan et al. (2015) considered the limiting behaviour of the summation of a sum-of-squares-type statistic and a screening statistic by higher criticism thresholding.

3·3. *Asymptotic power analysis*

In this section, we analyse the asymptotic power of the proposed adaptive test. Under the alternative $H_{\mathrm{A}} : \mu_1 \neq \mu_2$, we first derive approximations for the mean, variance and covariance functions for $L(\gamma)$ with $\gamma < \infty$, denoted respectively by $\mu_{\mathrm{A}}(\gamma)$, $\sigma_{\mathrm{A}}(\gamma)$ and $\mathrm{cov}_{\mathrm{A}}\{L(s), L(t)\}$ for $s, t < \infty$. We write $\delta_i = \mu_1^{(i)} - \mu_2^{(i)}$ $(i = 1, \ldots, p)$.

PROPOSITION 4. *Under the regularity conditions in Theorem 1 and* $H_{\mathrm{A}} : \mu_1 \neq \mu_2$,

$$\mu_{\mathrm{A}}(\gamma) = \mu(\gamma) + \sum_{i=1}^{p} \sum_{c=1}^{\gamma} \binom{\gamma}{c} \delta_i^c \mu^{(i)}(\gamma - c) \quad (\gamma < \infty),$$

where approximations for $\mu(\cdot)$ *and* $\mu^{(i)}(\cdot)$ *are given in Proposition 1. In particular,* $\mu_{\mathrm{A}}(1) = \sum_{i=1}^{p} \delta_i$, $\mu_{\mathrm{A}}(2) = \mu(2) + \sum_{i=1}^{p} \delta_i^2$, $\mu_{\mathrm{A}}(3) = \mu(3) + \sum_{i=1}^{p} \delta_i^3 + 3(n_1{}^{-1} + n_2{}^{-1}) \sum_{i=1}^{p} \delta_i \sigma_{ii}$, *and*

$$\mu_{\mathrm{A}}(4) = \mu(4) + \sum_{i=1}^{p} \delta_i^4 + 6\big(n_1{}^{-1} + n_2{}^{-1}\big) \sum_{i=1}^{p} \delta_i^2 \sigma_{ii} + 4 \sum_{i=1}^{p} \delta_i \big(m_{1i} n_1^{-2} - m_{2i} n_2^{-2}\big).$$

PROPOSITION 5. *Under the conditions in Theorem 1 and* H_{A},

$$\mathrm{cov}_{\mathrm{A}}\{L(s), L(t)\} \sim \mu_{\mathrm{A}}(t+s) - \sum_{i=1}^{p} \mu_{\mathrm{A}}^{(i)}(t) \mu_{\mathrm{A}}^{(i)}(s) + \sum_{i \neq j} \sum_{\substack{0 \leqslant h \leqslant l \\ 0 \leqslant l \leqslant s}} \binom{t}{h} \binom{s}{l} \delta_i^h \delta_j^l r_{ij}(t-h, s-l),$$

where for $s=0$ or $t=0$, $r_{ij}(s,t)=0$, and for $s>0$ and $t>0$,

$$r_{ij}(s,t)=\begin{cases}\displaystyle\sum_{\mathcal{A}(s,t)}\frac{t!\,s!\,\sigma_{ii}^{c_1+d_1}\sigma_{jj}^{c_2+d_2}\sigma_{ij}^{c_3+d_3}}{n_1^{c_1+c_2+c_3}n_2^{d_1+d_2+d_3}c_1!\,c_2!\,c_3!\,d_1!\,d_2!\,d_3!\,2^{c_1+c_2+d_1+d_2}}, & s+t \text{ even},\\ \displaystyle\sum_{\mathcal{B}(s,t)}\frac{t!\,s!\,\sigma_{ii}^{c_1+d_1}\sigma_{jj}^{c_2+d_2}\sigma_{ij}^{c_3+d_3}\left(\frac{m_{1,i^aj^b}}{a!\,b!\,n_1^2}-\frac{m_{2,i^aj^b}}{a!\,b!\,n_2^2}\right)}{n_1^{c_1+c_2+c_3}n_2^{d_1+d_2+d_3}c_1!\,c_2!\,c_3!\,d_1!\,d_2!\,d_3!\,2^{c_1+c_2+d_1+d_2}}, & s+t \text{ odd},\end{cases}$$

where $m_{k,i^aj^b}=E\{(X_k^{(i)}-\mu_k^{(i)})^a(X_k^{(j)}-\mu_k^{(j)})^b\}$ for $a+b=3$ and $\mathcal{B}(s,t)$ is the set of non-negative integers $(a,b,c_1,c_2,c_3,d_1,d_2,d_3)$ such that $a+b=3$, $2c_1+c_3+2d_1+d_3=t-a$, $2c_2+c_3+2d_2+d_3=s-b$ and $ab>0$ or $c_3+d_3>0$.

The variance function is $\sigma_{\mathrm{A}}^2(\gamma)=\mathrm{cov}_{\mathrm{A}}\{L(\gamma),L(\gamma)\}$. In particular, $\sigma_{\mathrm{A}}^2(1)=(n_1^{-1}+n_2{}^{-1})1_p^{\mathrm{T}}\,\Sigma\,1_p$ and $\sigma_{\mathrm{A}}^2(2)\sim\sigma^2(2)+4(n_1^{-1}+n_2{}^{-1})\sum_{i,j}\sigma_{ij}\delta_i\delta_j$.

We now analyse the power of the test. For the testing statistic in (1), let p_α^* be the critical threshold under H_0 with significance level α. The test power under H_{A} then satisfies $\mathrm{pr}(T_{\mathrm{aSPU}}=\min_{\gamma\in\Gamma}P_{\mathrm{SPU}(\gamma)}<p_\alpha^*)\geqslant\mathrm{pr}(P_{\mathrm{SPU}(\gamma)}<p_\alpha^*)$ for any $\gamma\in\Gamma$. Therefore, the asymptotic power of the proposed adaptive test is 1 if there exists $\gamma\in\Gamma$ such that $\mathrm{pr}(P_{\mathrm{SPU}(\gamma)}<p_\alpha^*)\to1$, that is, if $L(\gamma)$ has asymptotic power equal to 1. Hence, to study the asymptotic power of the adaptive test, we only need to focus on the power of $L(\gamma)$ for $\gamma\in\Gamma$.

Under the alternative, we denote the set of locations of the signals by $\mathcal{S}_\beta=\{i:\delta_i\neq0\}$ and the cardinality of $\mathcal{S}_\beta$ by $p^{1-\beta}$, where $\beta\in(0,1]$ is the sparsity parameter. In the following, we consider two cases: the dense signal case with $\beta<1/2$ and the sparse signal case with $\beta\geqslant1/2$.

Case 1: $0<\beta<1/2$. To study the asymptotic power, we consider the local alternative with small δ_i. Consider the set $\Gamma'=\Gamma\backslash\{\infty\}$, and for any finite γ define the corresponding average standardized signal as $\bar\delta(\gamma)=\sum_{i\in\mathcal{S}_\beta}n^{\gamma/2}\delta_i^\gamma/p^{1-\beta}$. If $\delta_i=O\{n^{-1/2}(\log p)^\epsilon\}$ with $\epsilon>0$, then $\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)=o(pn^{-\gamma/2})$ and $\sigma_{\mathrm{A}}^2(\gamma)-\sigma^2(\gamma)=o(pn^{-\gamma})$. A proof similar to that of Theorem 1 gives the following result.

THEOREM 2. *Under the conditions in Theorem 1 and the alternative H_{A} with $0<\beta<1/2$ and $\delta_i=O\{n^{-1/2}(\log p)^\epsilon\}$ for $\epsilon>0$, $[\{L(\gamma)-\mu_{\mathrm{A}}(\gamma)\}/\sigma_{\mathrm{A}}(\gamma)]_{\gamma\in\Gamma'}^{\mathrm{T}}$ converges weakly to a multivariate normal distribution with mean zero and covariance matrix $\mathcal{R}_{\mathrm{A}}$ given in Proposition 5.*

Theorem 2 gives the asymptotic test power of $L(\gamma)$ at significance level p_α^* as

$$\mathrm{pr}(P_{\mathrm{SPU}(\gamma)}<p_\alpha^*)=\begin{cases}\Phi\left\{\dfrac{\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)-z_{p_\alpha^*}\sigma(\gamma)}{\sigma_{\mathrm{A}}(\gamma)}\right\}, & \gamma \text{ even},\\ \Phi\left\{\dfrac{\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)-z_{p_{\alpha/2}^*}\sigma(\gamma)}{\sigma_{\mathrm{A}}(\gamma)}\right\}+\Phi\left\{-\dfrac{\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)+z_{p_{\alpha/2}^*}\sigma(\gamma)}{\sigma_{\mathrm{A}}(\gamma)}\right\}, & \gamma \text{ odd},\end{cases}$$

where Φ is the standard normal cumulative distribution function and $z_{p_\alpha^*}$ is its $(1-p_\alpha^*)$th quantile. Since $\sigma(\gamma)/\sigma_{\mathrm{A}}(\gamma)$ is bounded under the alternative considered, the asymptotic power is mainly dominated by $\{\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)\}/\sigma_{\mathrm{A}}(\gamma)$. In addition, $\sigma_{\mathrm{A}}(\gamma)$ is of order $p^{1/2}n^{-\gamma/2}$ and therefore the power goes to 1 if $n^{\gamma/2}\{\mu_{\mathrm{A}}(\gamma)-\mu(\gamma)\}/p^{1/2}\to\infty$. Intuitively speaking, the power of the adaptive test converges to 1 if some of the average standardized signals are of order higher

than $p^{\beta-1/2}$, which is $o(1)$. For example, when $\gamma = 1$ or 2, from the derivations in Proposition 4 we have that the asymptotic power of $L(1)$ or $L(2)$ goes to 1 if $n^{1/2}\sum_i \delta_i/p^{1/2} \to \infty$ or $n\sum_i \delta_i^2/p^{1/2} \to \infty$, that is, if $\bar{\delta}(1)$ or $\bar{\delta}(2)$ is of order higher than $p^{\beta-1/2}$.

For different values of γ, the test statistic $L(\gamma)$ that achieves the highest power depends on the specific dense alternative. To further study the power of different test statistics $L(\gamma)$ and how to choose the set Γ, we consider a special case where the signal strength is fixed at the same level, $n_1 = n_2$, $\sigma_{ii} = 1$ and $\sigma_{ij} \geqslant 0$. In this case, we show in the Supplementary Material that under the alternative hypothesis with small δ, the $L(1)$ test is asymptotically more powerful than the other $L(\gamma)$ tests. On the other hand, because of the slow convergence rate to the asymptotic distribution, which depends on the value of $p^{1/2-\beta}$, the performance of $L(1)$ for a finite sample may not be as good as that of $L(\gamma)$ tests with $\gamma > 1$, especially when the sparsity parameter β is close to $1/2$ and p is not large enough; see the Supplementary Material. Similarly, we can show that $L(2)$ is asymptotically more powerful if the absolute values of the δ_i have the same level but the signs are random with about half being positive.

Case 2: $\beta \geqslant 1/2$. The result in Case 1 implies that when $\beta < 1/2$ and $\gamma < \infty$, the test power of $L(\gamma)$ goes to 1 if $n^{\gamma/2}\{\mu_{\mathrm{A}}(\gamma) - \mu(\gamma)\}/p^{1/2} \to \infty$, which is satisfied in most cases if some average standardized signal is of order higher than $p^{\beta-1/2} = o(1)$. However, in the sparse setting with $\beta \geqslant 1/2$ and $\gamma < \infty$, $L(\gamma)$ loses power. To illustrate this, take $\gamma = 1$ and 2. For any $\beta < 1/2$, the powers of $L(1)$ and $L(2)$ converge to 1 if $\bar{\delta}(1)$ and $\bar{\delta}(2)$ are of order higher than $o(1)$. However, when $\beta > 1/2$, $p^{\beta-1/2} \to \infty$ and the asymptotic powers of $L(1)$ and $L(2)$ are strictly less than 1 even if $\bar{\delta}(1) = O(p^{\beta-1/2})$ and $\bar{\delta}(2) = O(p^{\beta-1/2})$.

On the other hand, $L(\infty)$ is known to be powerful against sparse alternatives; therefore, the proposed adaptive sum-of-powers test still has asymptotic power equal to 1 if that of $L(\infty)$ converges to 1. The asymptotic power of $L(\infty)$ has been studied in Cai et al. (2014); from their Theorem 2, the power of $L(\infty)$ converges to 1 if $\max_i |\delta_i| \geqslant c(\log p)^{1/2}n^{-1/2}$ for a certain constant c and if the nonzero δ_i are randomly uniformly sampled with sparsity level $\beta > 3/4$. The condition that $\beta > 3/4$ was assumed by the authors because of the technical difficulty in proving the asymptotic results. It is expected that the asymptotic power is still 1 for $1/2 < \beta \leqslant 3/4$ but the proof would be more challenging (Cai et al., 2014).

Combining the above theoretical arguments and simulation results, we recommend including small γ values such as 1, 2 and medium γ values such as $3, \ldots, 6$ in Γ to achieve balance between the asymptotic and finite-sample performances when the signals are dense; in addition, we also recommend including ∞ in Γ, as $L(\infty)$ is more powerful when the signals are sparse. See the Supplementary Material for more details and simulation studies.

Remark 7. When the signal is dense, $\beta < 1/2$, the $L(2)$ test performs similarly to the tests in Bai & Saranadasa (1996) and Chen & Qin (2010). As discussed above, there are alternatives under which $L(2)$ is not as powerful as other $L(\gamma)$ tests, and therefore in these dense signal cases, the proposed test is more powerful than those of Bai & Saranadasa (1996) and Chen & Qin (2010), as illustrated by the simulation study. When the signal is sparse, $\beta > 1/2$, the $L(\infty)$ test is equivalent to the supremum test in Cai et al. (2013), so the proposed adaptive test would perform similarly to that of Cai et al. (2013). On the other hand, under certain sparse alternatives, the $L(\infty)$ test may not be as powerful as the thresholding tests in the literature, such as the test proposed in Chen et al. (2014). To illustrate this, consider the oracle case, where the signal set $\mathcal{S}_\beta$ is known and has order $p^{1-\beta}$ with $1/2 < \beta < 1$. Suppose that the X_k are independent standard normal and signals are at the same level $\delta = (r \log p/n)^{1/2}$ for some large constant r. Then, the oracle test statistic with power index $\gamma = 1$, namely $O_n = \sum_{i\in\mathcal{S}_\beta}\{\bar{X}_1^{(i)} - \bar{X}_2^{(i)}\}$, has test power

going to 1 if $p^{1-\beta} \to \infty$. In particular, the log of the Type II error is of the order of $p^{1-\beta} \log p$. For the $L(\infty)$ test, the log of the Type II error is of the order of $\log p$. Therefore, in this ideal case, the O_n test, which excludes nonsignal locations, is more powerful than the supremum-type test $L(\infty)$.

4. Simulations

In this section we compare, through simulations, the performance of the proposed adaptive method and the existing tests described in § 2. The candidate set of γ for the sum-of-powers tests $L(\gamma)$ was taken to be $\Gamma = \{1, \ldots, 6, \infty\}$. We generated two groups of random samples, $\{X_{1i}\}_{i=1}^{n_1}$ and $\{X_{2j}\}_{j=1}^{n_2}$, with sample sizes $n_1 = n_2 = 50$, from two multivariate normal distributions of dimension $p = 200$, so $X_{ki} \sim N(\mu_k, \Sigma)$ for $k = 1, 2$. Without loss of generality, we let $\mu_1 = 0$. Under the null hypothesis, $\mu_2 = 0$; under the alternative hypothesis, $\lfloor p^{1-\beta} \rfloor$ elements in μ_2 were set to nonzero values, where $\beta \in [0, 1]$ controls the signal sparsity. In our simulations we used $\beta = 0{\cdot}1, \ldots, 0{\cdot}9$, covering very dense signals for an alternative hypothesis at $\beta = 0{\cdot}1$, to dense and then only moderately dense signals at $\beta = 0{\cdot}2$ and $0{\cdot}5$, and finally to moderately sparse and very sparse signals at $\beta = 0{\cdot}7$ and $0{\cdot}9$, respectively. The nonzero elements of μ_2 were assumed to be uniformly distributed in $\{1, \ldots, p\}$, and their values were constant at $\{2r(1/n_1 + 1/n_2) \log p\}^{1/2}$, where r controls the signal strength. The common covariance matrix is $\Sigma = D^{1/2} R D^{1/2}$, where R is the correlation matrix and the diagonal matrix D contains the variances. We considered various structures of $R = (r_{ij})$ and D, as detailed in the Supplementary Material. To save space, here we only show results for a first-order autoregressive correlation matrix $R = (0{\cdot}6^{|i-j|})$ and an equal-variance case with $D = I_p$. Although this covariance matrix is only approximately bandable, we applied the banding estimator of Bickel & Levina (2008) to show the robustness of the proposed tests. For each setting, 1000 replicates were simulated to calculate the empirical Type I error and power of each test. The p-values were calculated based on both the asymptotic distributions of the tests and the permutation method with $B = 1000$ iterations. The nominal significance level was set to $\alpha = 0{\cdot}05$.

Table 1 presents empirical Type I error rates and powers for $\beta = 0{\cdot}1$. The results of most tests based on the asymptotics are very close to those based on permutations. This validates the results in Theorem 1. The Type I error rate and power of the thresholding test T_{CLZ} were overestimated by the corresponding asymptotic approximation, probably due to the slow convergence to its asymptotic distribution.

Since the Type I error rates of all the tests were well controlled by their permutation-based p-values, we present the permutation-based powers in Fig. 1 to offer a fair comparison between the tests. The proposed adaptive sum-of-powers test T_{aSPU} was much more powerful than the other tests when the signals were highly dense, with $\beta = 0{\cdot}1$. When the signal sparsity increased from $0{\cdot}2$ to $0{\cdot}4$, the adaptive sum-of-powers test performed similarly to the sum-of-squares-type tests in Bai & Saranadasa (1996), Srivastava & Du (2008) and Chen & Qin (2010), and it was slightly more powerful than the thresholding test in Chen et al. (2014) and much more powerful than the supremum-type test in Cai et al. (2014). As the signals became less dense at $\beta = 0{\cdot}5$, the adaptive sum-of-powers and thresholding tests were the most powerful, closely followed by the sum-of-squares-type tests and then the supremum-type test. At $\beta = 0{\cdot}6$, although the adaptive sum-of-powers and thresholding tests remained the winners, the supremum-type test was more powerful than the sum-of-squares-type tests. When the signals were moderately sparse at $\beta = 0{\cdot}7$, the adaptive sum-of-powers and supremum-type tests were the most powerful, closely followed by the thresholding test; they were much more powerful than the sum-of-squares-type tests. When the signals were highly sparse at $\beta = 0{\cdot}9$, as expected, the supremum-type test became the sole

Table 1. *Empirical Type I errors and powers* (%) *of various tests for normal samples with* $n_1 = n_2 = 50$, $p = 200$ *and covariance matrix* $\Sigma = (0{\cdot}6^{|i-j|})$. *Zero signal strength* $r = 0$ *represents Type I errors, while* $r \neq 0$ *represents powers; the results outside and inside parentheses were calculated from asymptotics- and permutation-based p-values, respectively. The sparsity parameter was* $\beta = 0{\cdot}1$, *leading to* 117 *nonzero elements in* μ_2 *with a constant value of* $\{2r(1/n_1 + 1/n_2)\log p\}^{1/2}$

Test	$r = 0$	$r = 0{\cdot}02$	$r = 0{\cdot}04$	$r = 0{\cdot}06$	$r = 0{\cdot}08$
SPU(1)	5 (5)	50 (46)	78 (76)	92 (91)	98 (97)
SPU(2)	5 (5)	22 (20)	47 (46)	69 (67)	87 (85)
SPU(3)	4 (4)	40 (40)	71 (70)	88 (89)	97 (97)
SPU(4)	5 (5)	19 (18)	38 (37)	61 (60)	79 (78)
SPU(5)	4 (5)	24 (25)	47 (49)	70 (72)	84 (86)
SPU(6)	4 (4)	13 (14)	26 (29)	42 (45)	60 (64)
SPU(∞)	6 (5)	12 (9)	18 (15)	25 (21)	35 (28)
aSPU	6 (5)	33 (34)	66 (66)	85 (85)	94 (94)
CLZ	12 (5)	33 (15)	56 (34)	77 (57)	91 (76)
CLX	6 (5)	12 (9)	18 (15)	25 (21)	35 (28)
BS	6 (5)	23 (20)	48 (46)	70 (67)	88 (85)
CQ	6 (5)	23 (20)	48 (46)	70 (67)	88 (85)
SD	4 (5)	19 (19)	43 (45)	67 (68)	85 (86)

SPU, the proposed sum-of-powers tests with different values of γ; aSPU, the adaptive sum-of-powers test; CLZ, test of Chen et al. (2014); CLX, test of Cai et al. (2014); BS, test of Bai & Saranadasa (1996); CQ, test of Chen & Qin (2010); SD, test of Srivastava & Du (2008).

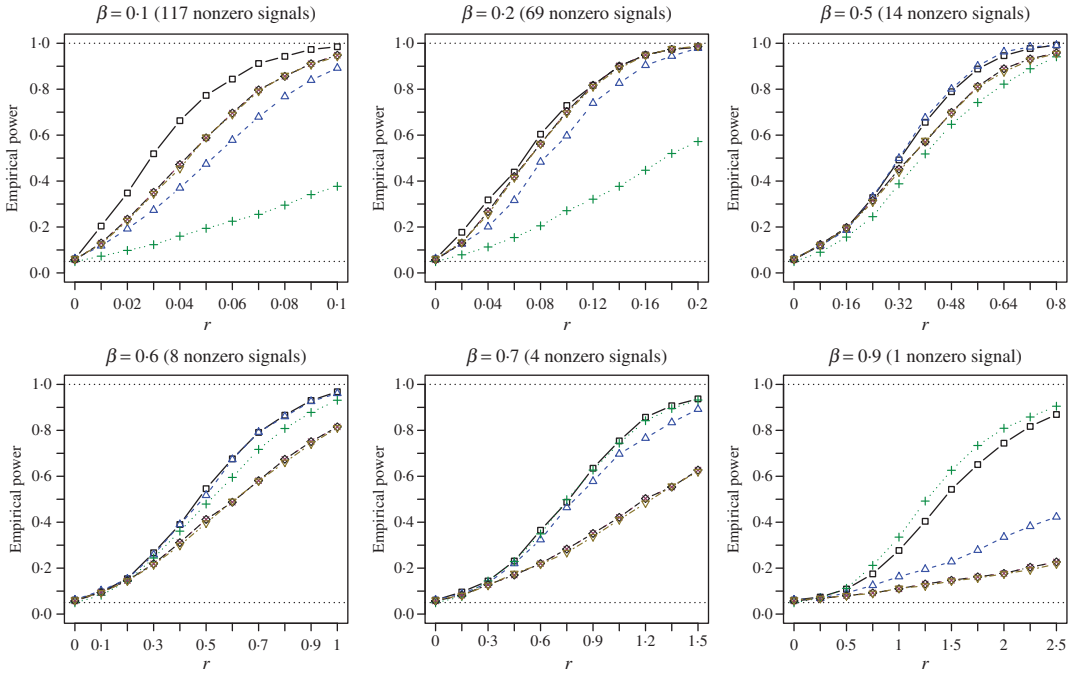

Fig. 1. Empirical powers of the adaptive sum-of-powers test (squares) and the tests of Chen et al. (2014) (triangles point up), Cai et al. (2014) (plus signs), Bai & Saranadasa (1996) (crosses), Chen & Qin (2010) (diamonds), and Srivastava & Du (2008) (triangles point down). The signal sparsity parameter β varies from 0·1 to 0·9.

winner, and the powers of the sum-of-squares-type and thresholding tests dropped substantially; however, the power of the adaptive sum-of-powers test remained high, close to that of the winner, the supremum-type test.

We obtained similar results for other simulation settings, including a more extreme case with a compound symmetric R and unequal variances D for multivariate normal data, and for simulated single nucleotide polymorphism data; see the Supplementary Material. In summary, owing to its adaptivity, the adaptive sum-of-powers test either achieved the highest power or had power close to that of the winner in any setting; it performed consistently well across all the situations. The banding estimator performed well, although occasionally the asymptotic adaptive sum-of-powers test would have slightly inflated Type I error rates when the assumptions in § 3·2 were severely violated.

5. Real-data analysis

We applied the various tests to the bipolar disorder dataset from a genome-wide association study collected by The Wellcome Trust Case Control Consortium (2007). We used their quality control procedure to screen the subjects and obtained $n_1 = 2938$ controls and $n_2 = 1868$ cases. We filtered out all the single nucleotide polymorphisms with minor allele frequency lower than 0·05 and those with Hardy–Weinberg equilibrium test p-value less than 10^{-5} in either cases or controls, giving 354 796 variables in total. To obtain a set of single nucleotide polymorphisms in approximate linkage equilibrium, as in the work of The International Schizophrenia Consortium (2009), we used the software plink (Purcell et al., 2007) to prune them with a criterion of linkage disequilibrium $r^2 \leqslant 0{\cdot}1$, a sliding window covering 200 single nucleotide polymorphisms, and a moving step of 20; this yielded 42 092 remaining single nucleotide polymorphisms. As The International Schizophrenia Consortium (2009) has shown that for bipolar disorder there is strong evidence of polygenic effects, we applied the various tests to the single nucleotide polymorphisms in each of the 22 autosomes separately to better demonstrate the possible power differences between the tests. The familywise nominal significance level was set at 0·05, and it would be $0{\cdot}05/22 = 0{\cdot}00227$ for each chromosome after Bonferroni adjustment. This indicates that 10 000 permutations should be sufficient to yield a possibly significant p-value to reject the null hypothesis.

We calculated both asymptotics- and permutation-based p-values for each test. To save space, Table 2 shows only some representative results. Most of the asymptotics-based p-values of the proposed sum-of-powers and adaptive tests were similar to their permutation-based ones, indicating good approximations. Again, the thresholding test T_{CLZ} produced asymptotics-based p-values that were far more significant than the permutation-based ones for most chromosomes, indicating its poor approximation. The test of Srivastava & Du (2008) also performed poorly; it always gave asymptotic p-values less than 0·0001. To avoid potentially poor asymptotic approximations, we use the permutation-based p-values to compare the various tests. In chromosomes 1, 2, 3, 6, 7, 9, 14, 15 and 16, both the sum-of-squares-type tests and the adaptive sum-of-powers test gave p-values less than $0{\cdot}05/22 = 0{\cdot}00227$. In contrast, the thresholding test yielded significant p-values for only five of those chromosomes, while the supremum-type test was not significant for any chromosome. These results were presumably due to dense signals in these chromosomes, thus favouring the sum-of-squares-type tests. However, in other situations the sum-of-squares-type tests might not perform well. For example, for chromosome 13, only the sum-of-powers test $L(\gamma)$ with $\gamma = 4$ gave a significant p-value. Another example is chromosome 18: perhaps due to sparse signals, the supremum-type test gave the most significant p-value, but none of the sum-of-squares-type tests yielded even marginal significance; borrowing strength from the

Table 2. *The p-values* (%) *of various tests applied to the Wellcome Trust Case Control Consortium bipolar disease data; the p-values outside parentheses were calculated from asymptotic distributions, and those inside parentheses were based on permutations*

	Chromosome (number of single nucleotide polymorphisms)				
Test	1 (3340)	2 (3194)	4 (2617)	13 (1592)	18 (1421)
SPU(1)	63·6 (64·3)	17·0 (17·8)	0·2 (0·2)	3·7 (3·7)	33·0 (32·3)
SPU(2)	<0·1 (<0·1)	<0·1 (<0·1)	1·5 (1·7)	2·7 (2·9)	28·9 (28·7)
SPU(3)	73·8 (74·5)	0·6 (0·7)	3·1 (3·1)	12·9 (12·6)	18·7 (17·4)
SPU(4)	<0·1 (<0·1)	<0·1 (<0·1)	2·0 (2·7)	<0·1 (0·2)	35·3 (33·1)
SPU(5)	74·2 (73·2)	0·2 (0·3)	37·5 (36·1)	39·4 (37·1)	25·9 (23·4)
SPU(6)	<0·1 (<0·1)	<0·1 (0·1)	2·7 (4·1)	<0·1 (0·4)	44·8 (38·6)
SPU(∞)	13·1 (11·8)	4·5 (4·3)	12·1 (11·9)	8·8 (8·0)	0·5 (0·4)
aSPU	<0·1 (<0·1)	<0·1 (<0·1)	1·0 (1·2)	<0·1 (1·3)	1·4 (1·9)
CLZ	<0·1 (<0·1)	<0·1 (0·3)	9·6 (10·2)	0·2 (0·5)	5·6 (6·6)
CLX	13·1 (11·8)	4·5 (4·3)	12·1 (11·9)	8·8 (8·0)	0·5 (0·4)
BS	<0·1 (<0·1)	<0·1 (<0·1)	1·5 (1·7)	2·6 (2·9)	28·8 (28·7)
CQ	<0·1 (<0·1)	<0·1 (<0·1)	1·5 (1·7)	2·7 (2·9)	29·0 (28·7)
SD	<0·1 (<0·1)	<0·1 (<0·1)	<0·1 (1·0)	<0·1 (11·4)	<0·1 (9·7)

SPU, the proposed sum-of-powers tests with different values of γ; aSPU, the adaptive sum-of-powers test; CLZ, test of Chen et al. (2014); CLX, test of Cai et al. (2014); BS, test of Bai & Saranadasa (1996); CQ, test of Chen & Qin (2010); SD, test of Srivastava & Du (2008).

supremum-type test, i.e., $L(\infty)$, the p-value of the adaptive sum-of-powers test was marginally significant. In summary, owing to its adaptivity, the proposed adaptive test retained high power across various chromosomes with varying association patterns.

Acknowledgement

We thank the editor, an associate editor and two reviewers for many helpful and constructive comments. This research was supported by the U.S. National Institutes of Health. This study makes use of data generated by The Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available at www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust. Peng Wei is also affiliated with the Human Genetics Center at the University of Texas.

Supplementary material

Supplementary material available at *Biometrika* online includes additional numerical results and proofs of the main theoretical results.

References

Bai, Z. D. & Saranadasa, H. (1996). Effect of high dimension: By an example of a two sample problem. *Statist. Sinica* **6**, 311–29.

Bickel, P. J. & Levina, E. (2008). Regularized estimation of large covariance matrices. *Ann. Statist.* **36**, 199–227.

Cai, T. T. & Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *J. Am. Statist. Assoc.* **106**, 672–84.

Cai, T. T., Liu, W. & Xia, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Statist. Assoc.* **108**, 265–77.

Cai, T. T., Liu, W. & Xia, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Statist. Soc.* B **76**, 349–72.

CHEN, S. X., LI, J. & ZHONG, P.-S. (2014). Two-sample tests for high dimensional means with thresholding and data transformation. arXiv:1410.2848.
CHEN, S. X. & QIN, Y.-L. (2010). A two-sample test for high-dimensional data with applications to gene-set testing. *Ann. Statist.* **38**, 808–35.
FAN, J. (1996). Test of significance based on wavelet thresholding and Neyman's truncation. *J. Am. Statist. Assoc.* **91**, 674–88.
FAN, J., LIAO, Y. & YAO, J. (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica* **83**, 1497–541.
FISHER, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin.* **52**, 399–433.
GOEMAN, J. J., VAN DE GEER, S. A. & VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *J. R. Statist. Soc.* B **68**, 477–93.
GREGORY, K. B., CARROLL, R. J., BALADANDAYUTHAPANI, V. & LAHIRI, S. N. (2015). A two-sample test for equality of means in high dimension. *J. Am. Statist. Assoc.* **110**, 837–49.
GUYON, X. (1995). *Random Fields on a Network: Modeling, Statistics, and Applications*. New York: Springer.
HALL, P., JIN, J. & MILLER, H. (2014). Feature selection when there are many influential features. *Bernoulli* **20**, 1647–71.
HOTELLING, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2**, 360–78.
HSING, T. (1995). A note on the asymptotic independence of the sum and maximum of strongly mixing stationary random variables. *Ann. Prob.* **23**, 938–47.
LEE, S., EMOND, M. J., BAMSHAD, M. J., BARNES, K. C., RIEDER, M. J., NICKERSON, D. A., NHLBI GO EXOME SEQUENCING PROJECT ESP LUNG PROJECT TEAM, CHRISTIANI, D. C., WURFEL, M. M. & LIN, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–37.
LIN, D.-Y. & TANG, Z.-Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. *Am. J. Hum. Genet.* **89**, 354–67.
PAN, W. (2009). Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet. Epidemiol.* **33**, 497–507.
PAN, W. (2011). Relationship between genomic distance-based regression and kernel machine regression for multi-marker association testing. *Genet. Epidemiol.* **35**, 211–6.
PAN, W., KIM, J., ZHANG, Y., SHEN, X. & WEI, P. (2014). A powerful and adaptive association test for rare variants. *Genetics* **197**, 1081–95.
PAN, W. & SHEN, X. (2011). Adaptive tests for association analysis of rare variants. *Genet. Epidemiol.* **35**, 381–8.
PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DEBAKKER, P. I. W. & DALY, M. J. ET AL. (2007). PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–75.
R DEVELOPMENT CORE TEAM (2016). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. http://www.R-project.org.
ROTHMAN, A. J., LEVINA, E. & ZHU, J. (2010). A new approach to Cholesky-based covariance regularization in high dimensions. *Biometrika* **97**, 539–50.
SRIVASTAVA, M. S. & DU, M. (2008). A test for the mean vector with fewer observations than the dimension. *J. Mult. Anal.* **99**, 386–402.
SRIVASTAVA, R., LI, P. & RUPPERT, D. (2015). RAPTT: An exact two-sample test in high dimensions using random projections. *J. Comp. Graph. Statist.* **25**, 954–70.
THE INTERNATIONAL SCHIZOPHRENIA CONSORTIUM (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–52.
THE WELLCOME TRUST CASE CONTROL CONSORTIUM (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–78.
WESSEL, J. & SCHORK, N. J. (2006). Generalized genomic distance-based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* **79**, 792–806.
WU, M. C., KRAFT, P., EPSTEIN, M. P., TAYLOR, D. M., CHANOCK, S. J., HUNTER, D. J. & LIN, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.* **86**, 929–42.
XUE, L., MA, S. & ZOU, H. (2012). Positive-definite ℓ_1-penalized estimation of large covariance matrices. *J. Am. Statist. Assoc.* **107**, 1480–91.
YU, K., LI, Q., BERGEN, A. W., PFEIFFER, R. M., ROSENBERG, P. S., CAPORASO, N., KRAFT, P. & CHATTERJEE, N. (2009). Pathway analysis by adaptive combination of p-values. *Genet. Epidemiol.* **33**, 700–9.
ZHONG, P.-S., CHEN, S. X. & XU, M. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *Ann. Statist.* **41**, 2820–51.

[*Received June* 2015. *Revised June* 2016]