Statistica Sinica **30** (2020), 1633-1655
doi:https://doi.org/10.5705/ss.202016.0492

TESTING CONSTANCY OF CONDITIONAL VARIANCE IN HIGH DIMENSION

Lu Deng[1], Changliang Zou[1], Zhaojun Wang[1] and Xin Chen[2]

[1]*Nankai University and* [2]*South University of Science and Technology of China*

Abstract: Testing the constancy of a conditional covariance matrix is a fundamental problem, because deviating from this assumption can result in a severely inefficient estimate. We propose a slice-based procedure to test for constant conditional variance in cases where the data dimension is larger than the sample size. We develop a high-order correction that makes the test statistic robust with respect to high dimensionality, and show that the proposed test statistic is asymptotically normal under some mild conditions. The proposed method allows the dimensionality to increase as the square of the sample size. Furthermore, simulations demonstrate that it exhibits good size and power in a wide range of settings.

Key words and phrases: Asymptotic normality, constant variance condition, high dimensional data, inverse regression, sufficient dimension reduction.

1. Introduction

Testing the constancy of a conditional covariance matrix in high-dimensional data has many applications, especially in regression studies. The null hypothesis of such a test can be described as follows:

$$H_0 : \mathrm{cov}(\mathbf{X} \mid \mathbf{Y}) = \mathbf{\Sigma}_0, \tag{1.1}$$

where $\mathbf{X}$ is a p-dimensional variable, $\mathbf{Y}$ is a d-dimensional variable, and $\mathbf{\Sigma}_0$ is an unknown $p \times p$ nonnegative definite matrix that does not vary with $\mathbf{Y}$. In this study, we allow p and the sample size n to go to infinity while the dimension $d \geq 1$ is fixed. We discuss the importance of testing H_0 below.

First, we consider estimating a high-dimensional covariance/precision matrix (Bickel and Levina (2008)). Here, a conventional assumption is that the targeted matrix is static, that is, its entries are all constant. However, in practice, this assumption may not be true, and hence a dynamic estimation procedure is required. Chen and Leng (2016) studied fMRI data collected by New York University Child Study Center. They found that the covariance matrices for the scans varied with time, and therefore suggested using a dynamic covariance matrix for

complex data. A testing procedure for the hypothesis given in (1.1) could serve as a preliminary step before employing a dynamic estimation strategy in such situations.

As another example, in genome-wide association studies (GWAS), X_k could represent gene expression levels, where $\mathbf{Y}$ is a vector of d biomarkers characterizing various properties of the disease condition. Here the goal is to identify disease-associated genetics. If the disease is complex with a heterogeneous etiology, that is, $\operatorname{cov}(\mathbf{X} \mid \mathbf{Y})$ varies, then the power of the tests used to detect influential genes would be compromised if the underlying heterogeneous effect was not taken into account (Yu et al. (2015)).

In this study, we test (1.1) for high-dimensional data, where the dimension p increases to infinity as the number of observations $n \to \infty$. We allow $\mathbf{Y}$ to be a continuous variable, and divide the data into several slices by clustering the values of $\mathbf{Y}$. The constancy is gauged by comparing the sample covariance matrices obtained from these slices. The proposed test statistic is similar to the two-sample and multi-sample tests for high-dimensional covariance matrices (e.g., Li and Chen (2012)). However, their results are not directly applicable. On the one hand, the sample sizes of the slices are random variables rather than constants, as they are in the multi-sample problem. On the other hand, to make the conditional expectation of $\mathbf{X} \mid \mathbf{Y}$ in each slice sufficiently smooth, the number of slices needs to be large and, asymptotically speaking, needs to go to infinity as $n \to \infty$. The technical treatments are therefore not trivial. Thus, a further contribution of this study is that we propose higher-order bias corrections and variance estimates for our test that, in general, outperform several "off-the-shell" methods in finite-sample situations. In the literature on time series, Tse (2000) and Bruno and Timo (2007) proposed Lagrange multiplier tests for the constancy of an error covariance matrix. However, their methods apply to low-dimensional situations only.

2. Methodology

2.1. Constancy test statistic

Let $\{(\mathbf{Y}_i, \mathbf{X}_i), i = 1, \ldots, n\}$ be a sample from the joint distribution $(\mathbf{Y}, \mathbf{X})$, where $\mathbf{Y}$ and $\mathbf{X}$ are d-dimensional and p-dimensional random variables, respectively. We assume that $\mathbf{Y}$ follows a continuous distribution. However, the proposed method can be modified accordingly to accommodate discrete $\mathbf{Y}$. First, we divide the range of $\mathbf{Y}$ into H clusters (slices), as in Li (1991), say $\{I_1, \ldots, I_H\}$,

and compare the covariances of $\mathbf{X}$ of the H clusters. Let $\mathbf{\Sigma}_i = \mathrm{cov}(\mathbf{X} \mid I_i)$. Naturally,

$$\sum_{1\le i<j\le H}\sum \mathrm{tr}\{(\mathbf{\Sigma}_i - \mathbf{\Sigma}_j)^2\} \tag{2.1}$$

could be a reasonable measure to gauge the constancy of the conditional covariance. The validity of the null hypothesis (1.1) implies that an appropriate estimate of (2.1) should not be large.

We let the sample sizes of all H clusters be equal to $l = n/H$, for ease of computation and technical analysis. If Y is univariate, that is, $d = 1$, then we can directly assign $Y_{((i-1)l+1)}, \ldots, Y_{(il)}$ as the ith cluster, denoted by S_{ni}, for $i = 1, \ldots, H$, where $Y_{(1)} \le \cdots \le Y_{(n)}$. This is equivalent to the set

$$S_{ni} = \left\{ Y_k : Y_k \in \left(F_n^{-1}\left(\frac{i-1}{H}\right), F_n^{-1}\left(\frac{i}{H}\right)\right)\right\},$$

where F_n denotes the sample empirical distribution of Y.

When $d > 1$, a natural ordering of the data points does not exist. Several efficient clustering methods exist; however, we suggest the following method for simplicity and technical convenience. Without loss of generality, we assume that $n = h_1 h_2 \ldots h_d l$. Then, we sort the data on the first component of $\mathbf{Y}$, and divide the data into h_1 equal clusters, $B_1, \ldots, B_{h_1}$, using the sample quantiles. Next, for cluster B_k, we sort on the second component of $\mathbf{Y}$, and divide B_k into h_2 equal and smaller clusters, $B_{k,1}, \ldots, B_{k,h_2}$. We repeat this step until we obtain H clusters $\{S_{ni}\}_{i=1}^H$. Clearly, each cluster is simply a hypercube containing exactly l observations. The partition of $\mathbf{Y}$ depends on the order of the response when $\mathbf{Y}$ is multivariate. In general, it is difficult to find an "optimal" partition without knowing the alternative structures. The partition proposed above may be a useful initial step before applying a better dividing method, such as the standard K-means procedure.

Given the above clusters, we can arrange the data as $(\mathbf{X}_{is}, \mathbf{Y}_{is})$, for $i = 1, \ldots, H$ and $s = 1, \ldots, l$, such that $\mathbf{Y}_{is} = \mathbf{Y}_{(l(i-1)+s)}$. Accordingly, we may consider $\sum\sum_{1\le i<j\le H} \mathrm{tr}\{(\mathbf{V}_i - \mathbf{V}_j)^2\}$ as a naive estimator of (2.1), where $\mathbf{V}_i$ is the sample covariance matrix of $\{\mathbf{X}_{i1}, \ldots, \mathbf{X}_{il}\}$. However, in high-dimensional cases, $\mathrm{tr}(\mathbf{V}_i^2)$ is not an effective estimator of $\mathrm{tr}(\mathbf{\Sigma}_i^2)$ owing to its inherent bias. Furthermore, the bias is always difficult, if not impossible, to estimate without further information (Chen, Zhang, and Zhong (2010)). Rather than estimating

$\boldsymbol{\Sigma}_i$, we suggest using

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} = \frac{1}{l(l-1)} \sum\sum_{s \neq t} [\{\mathbf{X}_{is} - \boldsymbol{\theta}_i(\{s,t\})\}^{\mathrm{T}}[\mathbf{X}_{it} - \boldsymbol{\theta}_i(\{s,t\})\}]^2, \tag{2.2}$$

$$\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)} = \frac{1}{l^2} \sum_{s,t}\sum [\{\mathbf{X}_{is} - \boldsymbol{\theta}_i(\{s\})\}^{\mathrm{T}}\{\mathbf{X}_{jt} - \boldsymbol{\theta}_j(\{t\})\}]^2 \tag{2.3}$$

as estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)$, respectively, where $\boldsymbol{\theta}_i(\{s,t\}) = (l-2)^{-1}\sum_{k\neq s,t}\mathbf{X}_{ik}$ and $\boldsymbol{\theta}_i(\{s\}) = (l-1)^{-1}\sum_{k\neq s}\mathbf{X}_{ik}$. Here $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)}$ are of the "leave-two-out" and "leave-one-out" forms, respectively, and are used to remove certain terms that impose unnecessary demands on the dimensionality. Our test statistic is constructed as follows:

$$T_n = (H-1)\sum_i \widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} - 2\sum\sum_{i<j}\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)}.$$

Similar estimators have been proposed and shown to be consistent in the literature on high-dimensional location tests; see, for example, Li and Chen (2012). However, it is clear that a more thorough analysis is required here because these estimators are employed to form a test statistic.

Remark 1. Chen, Zhang, and Zhong (2010) and Li and Chen (2012) proposed the following unbiased estimators for $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)$:

$$\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} = \frac{1}{P_l^2}\sum_{s\neq t}(\mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{it})^2 - \frac{2}{P_l^3}\sum_{s,t,r}^{*}\mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{it}\mathbf{X}_{it}^{\mathrm{T}}\mathbf{X}_{ir} + \frac{1}{P_l^4}\sum_{s,t,r,q}^{*}\mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{it}\mathbf{X}_{ir}^{\mathrm{T}}\mathbf{X}_{iq},$$

$$\begin{aligned}\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)} =& \frac{1}{l^2}\sum_{s,t}(\mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{jt})^2 - \frac{1}{l^2(l-1)}\sum_{s\neq r}\sum_t \mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{jt}\mathbf{X}_{jt}^{\mathrm{T}}\mathbf{X}_{ir}\\ &- \frac{1}{l^2(l-1)}\sum_{s\neq r}\sum_t \mathbf{X}_{js}^{\mathrm{T}}\mathbf{X}_{it}\mathbf{X}_{it}^{\mathrm{T}}\mathbf{X}_{jr} + \frac{1}{l^2(l-1)^2}\sum_{s\neq r}\sum_{t\neq q}\mathbf{X}_{is}^{\mathrm{T}}\mathbf{X}_{jt}\mathbf{X}_{ir}^{\mathrm{T}}\mathbf{X}_{jq},\end{aligned}$$

where $P_l^r = l!/(l-r)!$, and $\sum^*$ implies summation over mutually distinct indices. Here, we use (2.2) and (2.3) for the following reasons. First, the total computational complexity of T_n is $O(n^2p)$, whereas using $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ needs $O(n^2p + n^2l^2)$, which seems more computationally complex. Our numerical experience indicates that the computation time of the latter one is, in general, three or four times longer than that of the former, especially when n or l is large. Second, under some mild conditions, we can accurately correct the biases of T_n from using (2.2)

and (2.3). Third, our asymptotic and numerical results reveal that T_n could be a more powerful test than that based on $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j)}$ for commonly encountered cases.

2.2. Null distribution of the test statistic

To study the asymptotic behavior of T_n under the null hypothesis, we need to make several assumptions. First, denote the conditional expectation function as $\boldsymbol{\mu}(\mathbf{Y}) = E(\mathbf{X} \mid \mathbf{Y})$, and its sliced form as $\boldsymbol{\mu}_{is} = E(\mathbf{X}_{is} \mid \mathbf{Y}_{is})$. If we transform $\mathbf{X}_{is}$ by its regulation $\boldsymbol{\varepsilon}_{is} = \mathbf{X}_{is} - \boldsymbol{\mu}_{is}$, the difference between T_n and the test statistic after the transformation may be asymptotically negligible if $\boldsymbol{\mu}(\mathbf{Y})$ is sufficiently smooth in each cluster. Second, different permutation of the components of $\mathbf{Y}$ result in different partition sets, implying that the clustering results are not unique. Define two quantities as

$$r_{1i} = \max_{s,t} \|\mathbf{Y}_{is} - \mathbf{Y}_{it}\|, \; r_{2i} = \min \sum_{k=1}^{l-1} \|\mathbf{Y}_{i,k+1} - \mathbf{Y}_{ik}\|^{2\alpha}, \tag{2.4}$$

where the minimum in r_{2i} is taken with respect to all permutations of $\{\mathbf{Y}_{ik}\}_{k=1}^{l}$, and $\alpha \in (0, 1]$. These quantities can be viewed as generalizations of the range and spacings, respectively, of $\mathbf{Y}$ when $d = 1$. We need to impose conditions on these quantities to ensure the estimation performance of the conditional variances. Third, several assumptions on $\boldsymbol{\Sigma}_0$ and $\mathbf{X} \mid \mathbf{Y}$ are required in order to obtain the asymptotic normality of T_n. Formally, we need the following conditions:

(C1) $\boldsymbol{\mu}(\mathbf{Y})$ satisfies the Lipschitz condition of order α, say,

$$\|\boldsymbol{\mu}(\mathbf{Y}_1) - \boldsymbol{\mu}(\mathbf{Y}_2)\| \le Mp^{1/2}\|\mathbf{Y}_1 - \mathbf{Y}_2\|^{\alpha} \text{ for } \mathbf{Y}_1, \mathbf{Y}_2 \in \Omega,$$

where Ω is the compact support of $\mathbf{Y}$, and $\alpha \in (0, 1]$.

(C2) $n/l^5 \to 0$, $\sum_i r_{1i}^{4\alpha} = o_p(H^{1/2}(pl)^{-1})$, $\sum_i r_{2i} = o_p(H^{1/2}l^2p^{-1})$.

(C3) $\mathrm{tr}(\boldsymbol{\Sigma}_0^4) = o(\mathrm{tr}^2(\boldsymbol{\Sigma}_0^2))$, as $p = p(n) \to \infty$.

(C4) Given $\mathbf{Y}$, $\mathbf{X} \mid \mathbf{Y}$ follows the model $\mathbf{X} \mid \mathbf{Y} = \boldsymbol{\mu}(\mathbf{Y}) + \boldsymbol{\Gamma}(\mathbf{Y})\mathbf{Z}(\mathbf{Y})$, where $\boldsymbol{\Gamma}(\mathbf{Y})$ is a $p \times m$ matrix, such that $\boldsymbol{\Gamma}(\mathbf{Y})\boldsymbol{\Gamma}(\mathbf{Y})^{\mathrm{T}} = \boldsymbol{\Sigma}_0$ under the null hypothesis, and $\mathbf{Z}(\mathbf{Y}) = (Z_1(\mathbf{Y}), \ldots, Z_m(\mathbf{Y}))^{\mathrm{T}}$ satisfies the following: $E(\mathbf{Z}(\mathbf{Y}) \mid \mathbf{Y}) = 0$; $\mathrm{var}(\mathbf{Z}(\mathbf{Y}) \mid \mathbf{Y}) = \mathbf{I}_m$; $E(Z_i^4(\mathbf{Y}) \mid \mathbf{Y}) = 3 + \Delta(\mathbf{Y})$ is uniformly bounded; $E(Z_i^8(\mathbf{Y}) \mid \mathbf{Y})$ is uniformly bounded; and $E(Z_{i_1}^{\alpha_1}(\mathbf{Y}) \ldots Z_{i_q}^{\alpha_q}(\mathbf{Y}) \mid \mathbf{Y}) =$

$E(Z_{i_1}^{\alpha_1}(\mathbf{Y}) \mid \mathbf{Y}) \ldots E(Z_{i_q}^{\alpha_q}(\mathbf{Y}) \mid \mathbf{Y})$, for any positive integer in the set $\{q : \sum_{k=1}^{q} \alpha_k \leq 8\}$ and any $i_1 \neq i_2 \neq \cdots \neq i_q$.

Remark 2. Condition (C1) guarantees that the effect of using $\boldsymbol{\theta}_i(\{s,t\})$ or $\boldsymbol{\theta}_i(\{s\})$ to approximate $\boldsymbol{\mu}_{is}$ on the null distribution of T_n could be negligible, provided that the cluster partition satisfies Condition (C2). The first part of Condition (C2) imposes a restriction on the lower bounds of the order for l. This is needed because the test statistic T_n is not unbiased. The remaining part of Condition (C2) is related to the quantities r_{1i} and r_{2i}. In practice, these quantities depend mainly on the distribution of $\mathbf{Y}$ and the number of clusters H. In general, the faster the convergence rate in the tails of the density function $f(\mathbf{Y})$ or the larger the number of clusters H, the higher the orders are of these quantities that can be attained. For example, if Y follows a uniform (0,1) distribution, we have $Y_{(i+1)} - Y_{(i)} = O_p(n^{-1})$, such that $r_{1i} = O_p(H^{-1})$ and $r_{2i} = O_p(ln^{-2\alpha})$. Then, Condition (C2) requires that $p = o(n^{4\alpha-1/2}l^{-4\alpha-1/2})$ and $p = o(n^{2\alpha-1/2}l^{3/2})$. Suppose $\alpha = 1$. This condition implies that if we set $l = O(n^{1/3})$, then p is allowed to grow at a rate of $o(n^2)$. If Y follows the standard exponential distribution, which does not have bounded support, we have $Y_{(i+1)} - Y_{(i)} = O_p\{(n-i)^{-1}\}$. Consequently, p should be $o(n^{1/4})$, which is a much lower rate. In this situation, larger n is required to ensure the observations in the tail are sufficiently dense and, thus, reasonably good performance. In fact, Condition (C1) can be relaxed to hold only when $\mathbf{Y}_1, \mathbf{Y}_2$ fall into the same cluster. Moreover, when $\mathbf{Y}$ is discrete, under a natural partition, each cluster contains same value $\mathbf{Y}$, which implies $M = r_{1i} = r_{2i} = 0$. In this case, Conditions (C1) and (C2) trivially hold.

Remark 3. Conditions (C3) and (C4) are common in the literature; for example, see Bai and Saranadasa (1996), Chen, Zhang, and Zhong (2010), and Li and Chen (2012). Condition (C3) is used to satisfy the Lindeberg condition, on which the martingale central limit theorem relies. If all eigenvalues of $\boldsymbol{\Sigma}_0$ are bounded, this is trivially true. It also holds for the popular autoregressive covariance matrix $\boldsymbol{\Sigma}_0(i,j) = \rho^{|i-j|}$, with $|\rho| < 1$. However, if the covariance matrix contains many large entries, neither this condition nor the asymptotic normality of T_n hold. Thus, the asymptotic normality relies on the strength of the dependencies between the variables; here, a certain sparseness on $\boldsymbol{\Sigma}_0$ is needed (Zou et al. (2014)). Condition (C4) implies that the distribution of $\mathbf{X} \mid \mathbf{Y}$ satisfies the linear structure used in the literature on high-dimensional tests (Chen, Zhang, and Zhong (2010)). This condition greatly facilitates the calculation of the mo-

ments of the proposed test statistic. Note that in the condition of Chen, Zhang, and Zhong (2010), $m \geq p$ is required, because it implies that the nonconditional covariance $\mathbf{\Sigma}$ is positive definite. However, in our problem, $m < p$ can be allowed, because we require only that the conditional covariance $\mathrm{cov}(\mathbf{X} \mid \mathbf{Y})$ be nonnegative definite. For example, if $\mathbf{X} \sim N_p(0, I_p), Y = x_1$, then the elements in the first row of $\mathbf{\Sigma}_Y$ are all zero. A proper $\mathbf{\Gamma}$ is $(0, I_{p-1})^{\mathrm{T}}$, with $m = p - 1 < p$.

The following proposition provides approximations for the mean and variance of T_n under H_0.

Proposition 1. *Under H_0 and Conditions* (C1)–(C4)*, if $p = o(l^3)$, we have*

$$\begin{aligned} E(T_n) =& \left\{ \frac{2}{l-2} + \frac{2}{(l-2)^2} - \frac{2}{l-1} - \frac{1}{(l-1)^2} \right\} H(H-1)\mathrm{tr}(\mathbf{\Sigma}_0^2) \\ &+ \frac{1}{(l-2)^2} H(H-1)\mathrm{tr}^2(\mathbf{\Sigma}_0) + o\{\sqrt{\mathrm{var}(T_n)}\}, \\ \mathrm{var}(T_n) =& \left\{ \frac{4}{l^2} H^2(H-1)\mathrm{tr}^2(\mathbf{\Sigma}_0^2) \right\} \{1 + o(1)\}. \end{aligned}$$

Note that in Proposition 1, we need the condition $p = o(l^3)$, which is similar to the condition $p = o(n^3)$ in some two-sample testing problems (e.g., Feng et al. (2015)). However, in the present problem, if $\boldsymbol{\mu}(\mathbf{Y})$ is not sufficiently smooth, large values of l will not be allowed. Considering a special case that $l = O(n^{1/3})$, this condition requires $p/n \to 0$, which clearly contradicts high-dimensional settings.

Our analysis of the high-order expansion of T_n shows that this condition can be much relaxed after we correct the bias term of $\mathrm{tr}^2(\mathbf{\Sigma}_0)$. A simple estimate of $\mathrm{tr}(\mathbf{\Sigma}_0)$ in each slice is $l^{-1} \sum_{s=1}^{l} \mathbf{X}_{is}^{\mathrm{T}} \mathbf{X}_{is} - \{l(l-1)\}^{-1} \sum_{s \neq t} \mathbf{X}_{is}^{\mathrm{T}} \mathbf{X}_{it}$; thus, a pooled estimator is given by

$$\widehat{\mathrm{tr}(\mathbf{\Sigma}_0)} = \frac{1}{H} \sum_{i=1}^{H} \left\{ \frac{1}{l} \sum_{s=1}^{l} \mathbf{X}_{is}^{\mathrm{T}} \mathbf{X}_{is} - \frac{1}{l(l-1)} \sum_{s \neq t} \mathbf{X}_{is}^{\mathrm{T}} \mathbf{X}_{it} \right\}.$$

Let $T_n' = T_n - H(H-1)\{\widehat{\mathrm{tr}(\mathbf{\Sigma}_0)}\}^2/(l-2)^2$. We have the following result.

Proposition 2. *Under H_0 and Conditions* (C1)–(C4)*, if $p = o(l^7)$, we have*

$$\begin{aligned} E(T_n') =& \left\{ \frac{2}{l-2} + \frac{2}{(l-2)^2} - \frac{2}{l-1} - \frac{1}{(l-1)^2} \right\} H(H-1)\mathrm{tr}(\mathbf{\Sigma}_0^2) + o\{\sqrt{\mathrm{var}(T_n')}\} \\ \equiv& \mu_{T_n',0} + o\{\sqrt{\mathrm{var}(T_n')}\}, \\ \mathrm{var}(T_n') =& \left\{ \frac{4}{l^2} H^2(H-1)\mathrm{tr}^2(\mathbf{\Sigma}_0^2) \right\} \{1 + o(1)\} \equiv \sigma_{T_n',0}^2 \{1 + o(1)\}. \end{aligned}$$

The following theorem establishes the asymptotic null distribution of T'_n.

Theorem 1. *Under Conditions* (C1)–(C4) *and the null hypothesis* H_0*, if* $p = o(l^7)$*, as* $(n,p) \to \infty$*,*

$$\sigma_{T'_n,0}^{-1} T'_n - \delta_{n,l} \xrightarrow{d} N(0,1),$$

where $\delta_{n,l} = \mu_{T'_n,0}/\sigma_{T'_n,0}$ *depends only on* H *and* l*.*

To formulate a test procedure, we need to obtain a good estimate for $\mathrm{tr}(\mathbf{\Sigma}_0^2)$. Note that in the construction of the test statistic, we already obtain $\widehat{\mathrm{tr}(\mathbf{\Sigma}_i^2)}$. By the proof of Proposition 1, we know that

$$E\{\widehat{\mathrm{tr}(\mathbf{\Sigma}_i^2)}\} = \left\{1 + \frac{2}{l-2} + \frac{2}{(l-2)^2}\right\}\mathrm{tr}(\mathbf{\Sigma}_0^2) + \frac{1}{(l-2)^2}\mathrm{tr}^2(\mathbf{\Sigma}_0) + O\left(\frac{\mathrm{tr}(\mathbf{\Sigma}_0^2)}{(l-2)^3}\right).$$

This motivates us to use

$$\widehat{\mathrm{tr}(\mathbf{\Sigma}_0^2)} = \left\{1 + \frac{2}{l-2} + \frac{2}{(l-2)^2}\right\}^{-1}\left\{\frac{1}{H}\sum_{i=1}^{H}\widehat{\mathrm{tr}(\mathbf{\Sigma}_i^2)} - \frac{(\widehat{\mathrm{tr}(\mathbf{\Sigma}_0)})^2}{(l-2)^2}\right\}.$$

This is a ratio-consistent estimator of $\mathrm{tr}(\mathbf{\Sigma}_0^2)$, as revealed by the following proposition.

Proposition 3. *Under Conditions* (C1)–(C4) *and* H_0*, if* $p = o(l^7)$*, we have as* $(n,p) \to \infty$*,*

$$\frac{\widehat{\mathrm{tr}(\mathbf{\Sigma}_0^2)} - \mathrm{tr}(\mathbf{\Sigma}_0^2)}{\mathrm{tr}(\mathbf{\Sigma}_0^2)} \xrightarrow{p} 0.$$

The advantage of this estimator is that no additional computation effort is required. This result, together with Theorem 1, suggests we should reject H_0 with α level of significance if

$$\widehat{\sigma}_{T'_n,0}^{-1} T_{n'} - \delta_{n,l} > z_\alpha,$$

where z_α is the upper α-quantile of $N(0,1)$, and $\widehat{\sigma}^2_{T'_n,0}$ is the plug-in estimator of $\sigma^2_{T'_n,0}$.

2.3. Asymptotic power analysis and comparison

We investigate the asymptotic behavior of our test under the alternative hypotheses. For simplicity, we consider only $d = 1$, and the following local

alternative:

$$H_1 : \operatorname{cov}(\mathbf{X} \mid Y) = \boldsymbol{\Sigma}_0, \text{ for all } Y < a, \operatorname{cov}(\mathbf{X} \mid Y) = (1+\theta_n)\boldsymbol{\Sigma}_0, \text{ for all } Y \geq a,$$

where a is a constant value, and θ_n is a sequence tending to zero. This alternative is essentially a two-sample problem, though the mean in each sample varies with the response Y. This not only facilitates our technical analysis, but also mimics real-data behavior. For example, in GWAS, we have measured a set of d biomarkers $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ that are used for disease subtype classification. Our proposed constancy test is used to detect whether a heterogeneous effect exists. This alternative is thus the simplest that could be encountered in practice.

For ease of computation, we assume that $\operatorname{pr}(Y < a) = \operatorname{pr}(Y \geq a) = 1/2$ and $\operatorname{pr}(Y = a) = 0$. We have the following result.

Theorem 2. *Under Conditions* (C1)–(C4) *and the alternative hypothesis* H_1*, if* $p = o(l^7)$*, as* $(n,p) \to \infty$*,* $\sigma_{T_n',1}^{-1}(T_n' - \mu_{T_n',1}) \xrightarrow{d} N(0,1)$*, where*

$$\begin{aligned}
\mu_{T_n',1} =& \frac{b_l}{4} H^2 \theta_n^2 \operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \frac{1}{4(l-2)^2} H(H-1)\theta_n^2 \operatorname{tr}^2(\boldsymbol{\Sigma}_0) \\
&+ \frac{(a_l - b_l)}{2} H(H-1)\{1 + (1+\theta_n)^2\}\operatorname{tr}(\boldsymbol{\Sigma}_0^2), \\
\sigma_{T_n',1}^2 =& \frac{\sigma_{T_n',0}^2}{2}\{1 + (1+\theta_n)^4\},
\end{aligned}$$

with $a_l = 1 + 2(l-2)^{-1} + 2(l-2)^{-2}$*, and* $b_l = 1 + 2(l-1)^{-1} + (l-1)^{-2}$.

Accordingly, the asymptotic power of the T_n' test under H_1 is approximately

$$\beta(\theta_n) = \Phi\left\{ \frac{-z_\alpha \widehat{\sigma}_{T_n',0}}{\sigma_{T_n',1}} + \frac{\mu_{T_n',1} - \delta_{n,l}\widehat{\sigma}_{T_n',0}}{\sigma_{T_n',1}} \right\},$$

where Φ is the standard normal distribution function. By similar arguments to those used in the proof of Propositions 1–2, we can show that under alternative H_1,

$$\widehat{\sigma}_{T_n',0} = 2l^{-1}H^{3/2}\left\{ \frac{1 + (1+\theta_n)^2}{2}\operatorname{tr}(\boldsymbol{\Sigma}_0^2) + \frac{\theta_n^2 \operatorname{tr}^2(\boldsymbol{\Sigma}_0)}{4l^2} \right\}(1 + o_p(1)).$$

In addition, the leading order terms of the last two parts in $\Phi(\cdot)$ of (2.5) have an

explicit expression, resulting in the asymptotic power

$$\beta_{T_n'}(\theta_n) = \Phi\bigg(- L_n z_\alpha + \frac{\sqrt{H} l \theta_n^2}{8K_n} + \frac{\theta_n^2 \mathrm{tr}^2(\boldsymbol{\Sigma}_0)}{8l^2 \mathrm{tr}(\boldsymbol{\Sigma}_0^2) K_n} (\sqrt{H} l - z_\alpha)\bigg), \tag{2.5}$$

where $K_n = \{(1 + (1+\theta_n)^4)/2\}^{1/2}$ and $L_n = \{1 + (1+\theta_n)^2\}/(2K_n)$.

For comparison purposes, we consider a test statistic using the unbiased estimators of $\mathrm{tr}(\boldsymbol{\Sigma}_i^2)$ and $\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)$ proposed by Li and Chen (2012); that is,

$$\tilde{T}_n = (H-1) \sum_i \widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)} - 2 \sum_{i<j} \sum \widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}, \tag{2.6}$$

where $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}$ are given in Remark 1.

Compared with $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}$, $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$ and $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_j)}$ have similar forms, but remove more diagonal terms. If we assume that $\boldsymbol{\mu}(\mathbf{Y})$ is a constant function or is constant in every S_{ni}, the latter two estimators are exactly unbiased, such that under Conditions (C1) and (C2), no bias correction is needed for $\tilde{T}_n$. Under H_0, we can show that $\tilde{\sigma}_{\tilde{T}_n,0}^{-1} \tilde{T}_n \xrightarrow{d} N(0,1)$, where $\tilde{\sigma}_{\tilde{T}_n,0}^2 = 4H^2(H-1)\{\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_0^2)}\}^2/l^2$, with $\widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_0^2)} = H^{-1} \sum_i \widetilde{\mathrm{tr}(\boldsymbol{\Sigma}_i^2)}$. Under the local alternative, its asymptotic power can be written as

$$\beta_{\tilde{T}_n}(\theta_n) = \Phi\bigg(- L_n z_\alpha + \frac{\sqrt{H} l \theta_n^2}{8K_n}\bigg). \tag{2.7}$$

Note that if $\mathrm{tr}^2(\boldsymbol{\Sigma}_0)/\{l^2 \mathrm{tr}(\boldsymbol{\Sigma}_0^2)\} \to \infty$, the last term in the asymptotic power of $\beta_{T_n'}$ becomes dominant. Consequently, our proposed test would be more powerful than $\tilde{T}_n$ in this setting. Our simulation results in the next section concur with this observation. In fact, if we assume that the conditional covariance is fixed as $\boldsymbol{\Sigma}_i$ in each cluster, by a similar discussion to that of Proposition 1, the bias term is $(l-2)^{-2}(H-1)\sum_{i=1}^H \mathrm{tr}^2(\boldsymbol{\Sigma}_i)$. Therefore, correcting the bias using the estimator $\widehat{\mathrm{tr}(\boldsymbol{\Sigma}_0)}$, which averages over all clusters, will enhance the power, unless $\mathrm{tr}(\boldsymbol{\Sigma}_i)$ is the same across all clusters.

3. Practical Guidelines

3.1. More accurate variance estimates

Although $\widehat{\sigma}_{T_n',0}^2$ is a ratio-consistent estimator of $\mathrm{var}(T_n')$, its convergence is usually slow because l cannot be large for a small to moderate sample size

n. For example, when n is 400 or 800, l would in general be less than 20, as required by Condition (C2). Therefore, in practical applications, a more precise estimator of $\mathrm{var}(T_n')$ is definitely desirable. The following result provides us with a higher-order approximation of $\mathrm{var}(T_n')$.

Proposition 4. *Suppose Conditions* (C1)–(C4) *hold. Under* H_0, *if* $p = o(l^5)$, $l\mathrm{tr}(\mathbf{\Sigma}_0^4)/\mathrm{tr}^2(\mathbf{\Sigma}_0^2) \to 0$, $p = o(\{\sum_i r_{1i}^{4\alpha}\}^{-1} n^{1/2} l^{-5/2})$, *and* $p = o(\{\sum_i r_{2i}\}^{-1} n^{1/2} l^{1/2})$, *we have*

$$\begin{aligned}\mathrm{var}(T_n') =& 4l^{-2}H^2(H-1)\mathrm{tr}^2(\mathbf{\Sigma}_0^2)\left\{\frac{l^3}{(l-1)(l-2)^2} + \frac{2l}{(l-1)(l-2)}\right\}\{1+o(1)\} \\ \equiv& c_l \sigma^2_{T_n',0}\{1+o(1)\}.\end{aligned}$$

Clearly, $\lim_{l\to\infty} c_l = 1$. However, c_l is considerably larger than one when l is not large. For example, it is about 1.42 when $l = 20$. Thus, we suggest modifying our test statistic as follows: $\{\widehat{\sigma}^{-1}_{T_n',0} T_n' - \delta_{n,l}\}/\sqrt{c_l}$. Our numerical results in the next section show that, in general, this modified test statistic performs better for small l.

3.2. Choice of H

Like many other smoothing-based or sliced-based tests, the performance of the proposed test depends upon the number of slices, H, which is a smoothing parameter that plays a similar role to that of the bandwidth in a nonparametric regression. It is widely acknowledged that the optimal bandwidth for a nonparametric estimation is usually not optimal for testing (Hart (1997)), and identifying the selection that optimizes the power of the test remains an open problem. Asymptotically, a range of H that satisfies the conditions could maintain the consistency of the test, whereas a specific H may maximize the power. The amount of smoothing applied affects the power of the test. However, we have observed in our simulations that the observed significance changes mildly over a reasonably wide range of values for H. In addition, we found that, in general, a large H leads to better power. This can be understood from the asymptotic power expression of T_n' in (2.5). The term, $|\sqrt{H} l \theta_n^2 \mathrm{tr}^2(\mathbf{\Sigma}_0)/\{8l^2\mathrm{tr}(\mathbf{\Sigma}_0^2)K_n\}|$, is often the leading term and becomes larger with H, resulting in an improvement. However, in practice, Condition (C2) will be violated if l is too small. An inappropriately large H will yield a much larger false alarm rate.

Based on Condition (C2) and our numerical experience, we recommend the empirical $\ell \propto n^{1/\{2+\min(d,2)\}}$. This formula works well for a wide range of models

and sample sizes, as shown in Section 4. How to best utilize the data to select an optimal l for our proposed test is difficult, because it depends not only on the values of (d, p, n), but also on the types of alternatives. A potential remedy is to use a hybrid method that combines a sequence of values of l, similar to that proposed by Horowitz and Spokoiny (2001) in the context of a nonparametric model specification. Another is to consider the maximum of our proposed test statistics over a set of values of l. This is similar to the approach proposed by Zhong, Chen and Xu (2013) in the context of higher criticism. Both are certainly challenging, and warrant future research.

4. Simulation Study

We consider the following two models, with $d = 1$ and $d = 2$. In the first model, Model (I), the vector $\mathbf{X}$ of length p is generated through $\mathbf{X}_y = r\boldsymbol{\mu}_y + \boldsymbol{\Gamma}\mathbf{Z}_y$, where $\boldsymbol{\Gamma} = \mathbf{A}\boldsymbol{\Sigma}^{1/2}$, with $\mathbf{A} = \text{diag}(\mathbf{J}_{p/2}, 2\mathbf{J}_{p/2})$, $\boldsymbol{\Sigma}(i, j) = \rho^{|i-j|}$, $\rho = 0, 0.5$, and $\mathbf{J}_k$ a k-dimensional vector with all components being one. Y is generated from two distributions: (i) $U(2, 4)$; (ii) $N(3, \sigma^2)$, with $\sigma^2 = 0.2$. Two cases of $\boldsymbol{\mu}_y$ are considered. In the first case, Case (I), all components of $\boldsymbol{\mu}_y$ are equal to y. In the second case, Case (II), the first $p/2$ components of $\boldsymbol{\mu}_y$ are equal to y, and the remaining components are equal to y^2. Additionally, two distributions of $\mathbf{Z}_y$ are employed: $N(0, 1)$, and Gamma$(y^2, y^{-1}) - y$. Under this model, linear and nonlinear $E(\mathbf{X} \mid Y)$ are included, and the conditional covariance of $\mathbf{X}$, given Y, is weakly dependent, but a constant matrix.

In Model (II), $d = 2$ and the data-generation process is similar to that in Model (I). $\mathbf{Y} = (y_1, y_2)^{\tau}$ is also generated from two scenarios: (i) two components are independent from a $U(2, 4)$ distribution; and (ii) $y_1 = z_1 + z_2$ and $y_2 = z_1 + z_3$, where z_1, z_2, and z_3 are independent $U(1, 2)$ variables. The first $p/2$ components of $\boldsymbol{\mu}_{\mathbf{Y}}$ are set as y_1, and the remaining components are equal to y_2. Again, $\mathbf{Z}_{\mathbf{Y}}$ follows either $N(0, 1)$ or a gamma distribution. In the gamma setting, the first $p/2$ components are distributed as Gamma$(y_1^2, y_1^{-1}) - y_1$, and the other $p/2$ components follow Gamma$(y_2^2, y_2^{-1}) - y_2$. All simulation results are obtained based on 1,000 repetitions, and the nominal level is fixed as 0.05. We adopt the higher-order expansion form of var(T_n') given in Section 3.1. The first simulation results are intended to support our contention that the asymptotic test based on T_n' can be simple and useful in finite-sample situations, in the sense that the type-I error can be reasonably well controlled. Tables 1–2 show the empirical sizes of our proposed test under the model of $d = 1$ when Y follows a uniform

Table 1. Empirical sizes at 5% significance under the model of $d = 1$ and $y \sim U(2, 4)$.

$\mathbf{X} \mid Y$			Normal				Gamma			
	$\boldsymbol{\mu}_Y$		Case (I)		Case (II)		Case (I)		Case (II)	
methods			T_n'	$\tilde{T}_n$	T_n'	$\tilde{T}_n$	T_n'	$\tilde{T}_n$	T_n'	$\tilde{T}_n$
l	n	p	$\rho = 0.5$							
10	200	20	7.6	11.2	7.2	10.4	8.2	10.9	8.4	11.5
		40	6.0	9.7	6.7	10.3	7.0	10.9	6.7	10.8
		100	4.9	6.8	5.0	7.0	4.2	6.2	4.7	6.9
		200	5.8	7.7	5.8	9.4	6.8	9.8	6.8	10.6
		1,000	4.8	6.8	10.2	14.3	4.8	7.2	11.2	15.6
	600	20	7.3	10.5	7.2	10.3	11.8	14.4	10.2	14.9
		40	6.0	8.5	6.0	8.4	5.9	8.5	5.7	8.3
		100	6.8	10.1	6.3	10.4	5.9	8.4	5.9	8.5
		200	5.1	8.3	4.6	8.2	4.7	5.4	5.3	5.6
		1,000	5.5	7.8	4.8	7.7	6.1	9.0	6.5	9.7
15	200	20	8.5	10.5	8.2	11.6	8.0	9.6	9.1	10.8
		40	7.0	8.8	7.0	10.8	7.6	9.1	8.4	10.5
		100	4.8	6.3	6.2	9.2	6.2	8.2	8.4	12.1
		200	5.2	6.9	8.4	10.8	5.2	7.4	7.5	10.4
		1,000	5.0	6.5	10.8	15.4	4.6	6.7	12.1	15.6
	600	20	7.8	10.0	7.0	9.6	10.9	11.9	10.7	12.2
		40	6.1	7.3	5.5	7.4	5.2	7.2	5.3	6.6
		100	6.0	8.3	6.1	7.8	5.4	7.3	5.5	7.2
		200	6.1	7.6	4.8	7.5	6.0	7.7	6.4	8.1
		1,000	4.9	5.9	4.7	7.2	4.8	7.0	6.5	9.4

and a normal distribution with $r = 1$ and 0.2, respectively. Although we focus on high-dimensional settings, we also present results for small p, such as 20 or 40. We only present the results for $\rho = 0.5$ here. The results for $\rho = 0$ are reported in the Supplementary Material. For comparison, the results for the test statistic $\tilde{T}_n$ given by (2.6) are presented. Here, we do not consider other existing multi-sample tests for high-dimensional covariance matrices, because Li and Chen (2012) have shown that their test performs quite well for a considerable range of dimensionality and distributions, in comparison with some benchmarks. The empirical levels are close to the nominal level in most cases as n and p increase together, which shows the effectiveness of the suggested asymptotic procedure. The performances is not affected by ρ and is insensitive to the choice of l. In contrast, we observe that the sizes of the $\tilde{T}_n$ test appear to be more liberal than those of the proposed test. This demonstrates the benefit of using the high-order variance estimators suggested in Section 3.1. Note that under the same setting, the results given in Table 2 are usually worse than those shown in Table 1. This

Table 2. Empirical sizes at 5% significance under the model of $d = 1$ and $y \sim N(3, 0.2)$.

$\mathbf{X} \mid Y$			Normal				Gamma			
	$\boldsymbol{\mu}_Y$		Case (I)		Case (II)		Case (I)		Case (II)	
methods			T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$
l	n	p	$\rho = 0.5$							
10	200	20	7.4	11.5	8.2	12.7	7.7	10.2	7.4	9.7
		40	5.6	8.4	5.3	8.7	4.9	7.2	5.3	7.9
		100	4.3	6.2	4.3	7.0	4.9	8.4	5.3	8.3
		200	6.2	9.1	6.2	8.7	5.4	7.4	5.6	7.9
		1,000	4.1	6.3	5.1	7.4	4.7	8.2	5.8	8.7
	600	20	8.0	11.0	7.8	11.0	10.0	12.5	10.3	13.0
		40	5.9	9.0	5.9	9.3	6.5	10.0	6.2	9.0
		100	6.2	8.3	6.4	9.0	6.2	8.0	6.3	8.2
		200	5.0	8.6	5.3	9.2	5.5	7.6	5.6	7.8
		1,000	5.1	8.1	5.5	9.0	5.7	8.7	6.8	9.0
15	200	20	7.3	8.5	7.3	8.9	10.2	10.7	9.7	11.7
		40	5.8	7.3	5.6	7.7	8.1	11.4	7.9	10.4
		100	6.4	8.9	6.1	9.1	6.0	7.5	5.7	8.1
		200	5.2	6.7	5.6	7.2	6.2	8.3	6.9	8.8
		1,000	5.0	6.7	8.2	9.7	5.9	7.6	8.4	10.8
	600	20	7.6	9.2	7.8	9.0	9.2	10.8	9.0	10.8
		40	6.7	8.8	7.0	8.6	6.9	9.0	7.1	9.5
		100	5.2	6.6	5.5	7.1	6.2	7.8	6.4	8.1
		200	5.9	7.7	6.5	7.9	6.0	7.1	5.7	6.9
		1,000	4.4	6.1	5.7	7.1	5.4	6.7	6.7	8.7

is not surprising because, unlike the uniform distribution, unbounded support of the normal distribution would yield relatively large r_{1i} and r_{2i}; in such a case, a large n is required to attain desirable empirical levels. Table 3 shows the empirical size of our proposed test under the model with $d = 2$ and $r = 1$. In most cases, the proposed test is still able to maintain the empirical sizes, but with the same sample size, the deviations to the nominal level become more pronounced than those in Tables 1–2. This is expected because as the dimension of $\mathbf{Y}$ increases, r_{1i} and r_{2i} get larger and, accordingly, the convergence of the proposed test statistic becomes slower.

To evaluate the power performance, we consider the alternative $\boldsymbol{\Sigma}(y) = \{1 + \theta_n(y - 2)\}\boldsymbol{\Sigma}_0$ when $d = 1$, and $\boldsymbol{\Sigma}(y) = [1 + \theta_n\{(y_1 + y_2)/2 - 2\}]\boldsymbol{\Sigma}_0$ when $d = 2$, where $\theta_n = 0.1, 0.2, 0.3, 0.4$. Figures 1–2 show the empirical power curves against θ_n when $d = 1$ and $d = 2$, respectively. For a relatively fair comparison, we conduct a size-corrected power evaluation, in the sense that the actual critical values are found using simulations such that all tests have accurate sizes of

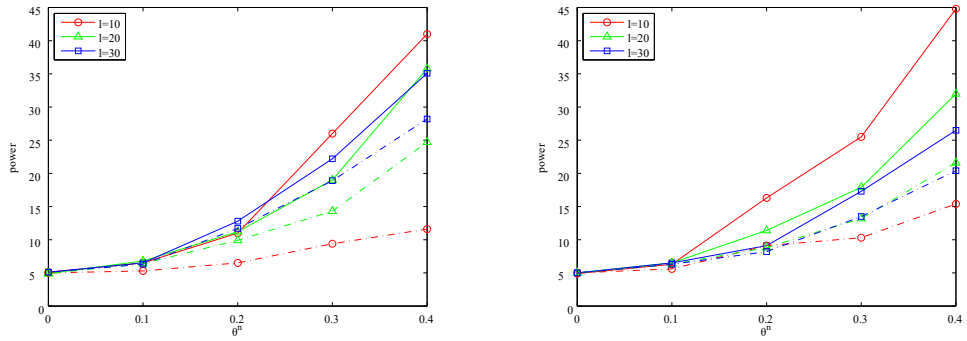
Table 3. Empirical sizes at 5% significance under the model of $d = 2$.

$\mathbf{X} \mid \mathbf{Y}$			Normal				Gamma			
scenarios			(i)		(ii)		(i)		(ii)	
methods			T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$	T'_n	$\tilde{T}_n$
l	n	p	$\rho = 0.5$							
10	200	20	6.0	9.2	7.0	10.1	8.0	10.4	8.2	10.9
		40	7.6	10.4	5.8	8.7	5.9	9.1	6.7	9.9
		100	5.4	8.6	5.3	8.3	5.8	7.8	5.6	8.1
		200	5.5	8.1	4.7	7.7	4.9	8.1	5.9	8.7
		1,000	5.9	8.7	7.2	8.7	6.4	8.6	7.4	8.8
	600	20	9.3	12.4	8.4	11.6	7.8	10.1	8.2	10.8
		40	5.4	8.4	6.9	10.1	6.9	8.9	7.8	10.8
		100	6.4	9.2	4.8	8.4	7.3	10.1	6.5	9.5
		200	5.5	8.8	6.5	9.4	5.5	8.0	5.3	8.2
		1,000	5.5	9.1	5.8	8.1	5.9	7.7	6.7	8.8
15	200	20	5.1	7.5	5.5	8.9	6.3	9.1	7.6	9.9
		40	4.4	7.4	4.8	6.7	4.6	8.2	5.6	8.4
		100	5.4	8.2	5.2	7.9	4.6	7.3	4.9	7.9
		200	4.8	7.2	4.6	7.2	5.9	8.4	6.0	9.3
		1,000	4.0	7.3	6.1	7.4	5.7	8.5	6.8	9.7
	600	20	5.7	8.7	5.5	8.0	8.5	11.5	7.4	9.9
		40	6.4	9.2	4.3	6.9	5.9	8.0	7.7	10.3
		100	4.4	7.6	5.5	9.9	5.5	7.8	4.1	6.7
		200	4.9	7.2	4.6	7.5	5.1	7.2	6.0	8.4
		1,000	5.1	8.3	4.4	8.0	5.7	8.7	6.8	9.2

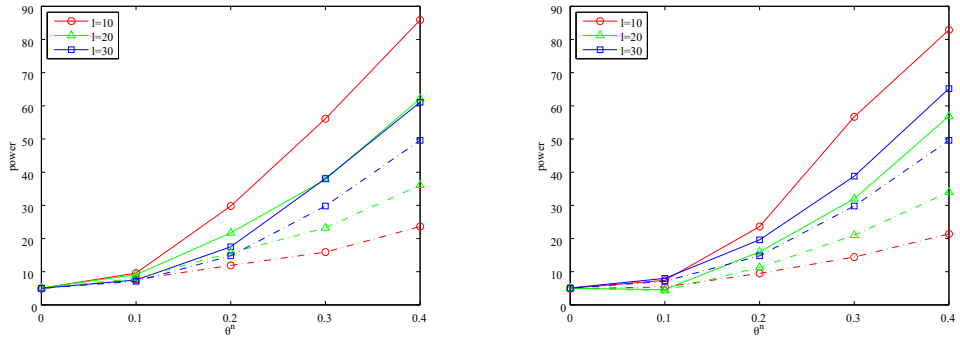
0.05 in each scenario. With the same choice of l, T'_n performs uniformly better than $\tilde{T}_n$ in almost all settings. This finding is consistent with our asymptotic comparison in Section 2.3. The performance of $\tilde{T}_n$ improves as l grows from 10 to 30, whereas T'_n generally performs better with smaller l. These results concur with the asymptotic power expressions (2.5) and (2.7).

For a more comprehensive comparison, we consider a natural benchmark that weights samples using unimodal kernels. Specifically, we extend the estimators of $\mathrm{tr}(\mathbf{\Sigma}_i^2)$ and $\mathrm{tr}(\mathbf{\Sigma}_i\mathbf{\Sigma}_j)$, respectively, as follows:

$$\widehat{\mathrm{tr}(\mathbf{\Sigma}_{\mathbf{Y}}^2)} = \frac{\sum\sum_{s\neq t}\left[\{\mathbf{X}_s - \boldsymbol{\theta}_s^*(\{s,t\})\}^{\mathrm{T}}[\mathbf{X}_t - \boldsymbol{\theta}_t^*(\{s,t\})\}\right]^2 K_h(\mathbf{Y}_s - \mathbf{Y})K_h(\mathbf{Y}_t - \mathbf{Y})}{\sum\sum_{s\neq t} K_h(\mathbf{Y}_s - \mathbf{Y})K_h(\mathbf{Y}_t - \mathbf{Y})},$$



Normal Case with $n = 200$ Gamma Case with $n = 200$



Normal Case with $n = 600$ Gamma Case with $n = 600$

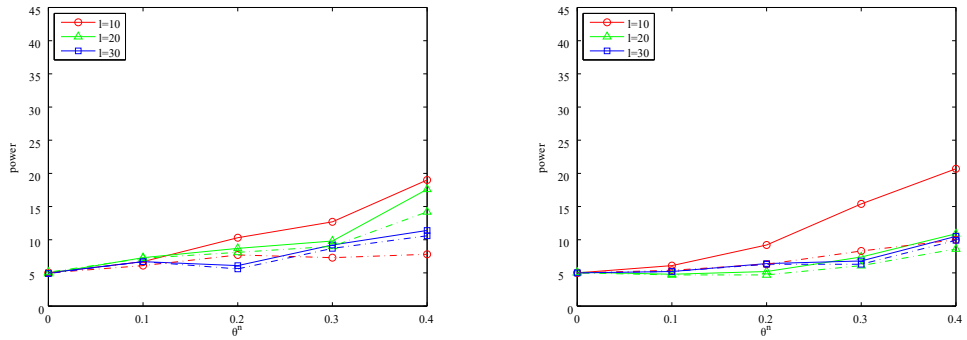
Figure 1. Power performance of T'_n (solid) and $\tilde{T}_n$ (dashed-dot) for the model with $d = 1$ and fixed $p = 400$ in settings where $Y \sim U(2, 4)$, $\boldsymbol{\mu}_Y$ comes from Case (I) and $\rho = 0.5$.

$$\operatorname{tr}(\widehat{\boldsymbol{\Sigma}_{\mathbf{Y}}\boldsymbol{\Sigma}_{\mathbf{Y}'}}) = \frac{\sum\sum_{s\neq t}\left[\{\mathbf{X}_s - \boldsymbol{\theta}^*_s(\{s,t\})\}^{\mathrm{T}}\{\mathbf{X}_t - \boldsymbol{\theta}^*_t(\{s,t\})\}\right]^2 K_h(\mathbf{Y}_s - \mathbf{Y})K_h(\mathbf{Y}_t - \mathbf{Y}')}{\sum\sum_{s\neq t} K_h(\mathbf{Y}_s - \mathbf{Y})K_h(\mathbf{Y}_t - \mathbf{Y}')},$$

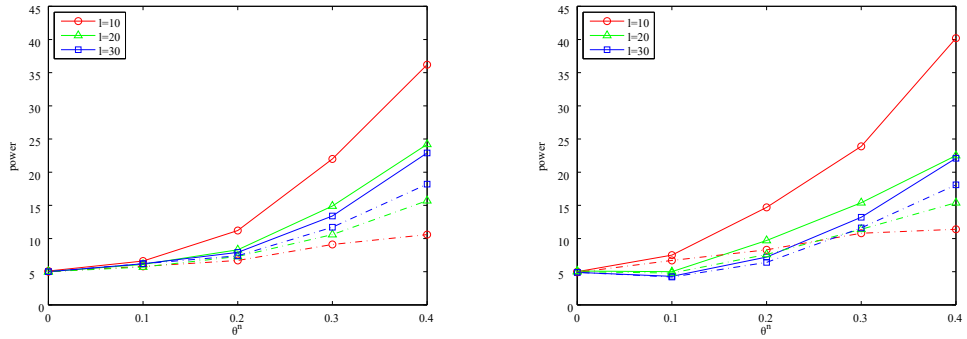
where $\boldsymbol{\theta}^*_s(\{s,t\}) = \sum_{k\neq s,t}\mathbf{X}_k K_h(\mathbf{Y}_k - \mathbf{Y}_s)/\sum_{k\neq s,t} K_h(\mathbf{Y}_k - \mathbf{Y}_s)$ and the function $K_h(y) = h^{-d}K(y/h)$, with a unimodal kernel $K(\cdot)$. Then, the test statistic is constructed by

$$T_{n,h} = (n-1)\sum_{i=1}^{n}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}^2_{\mathbf{Y}_i}}) - 2\sum\sum_{1\leq i<j\leq n}\operatorname{tr}(\widehat{\boldsymbol{\Sigma}_{\mathbf{Y}_i}\boldsymbol{\Sigma}_{\mathbf{Y}_j}}).$$

Note that the computational burden of $T_{n,h}$ is considerably heavier than that of T'_n because it needs to compute all pair differences for n samples. When n is not small, such as 200, $T_{n,h}$ is not easy to obtain within an acceptable time. To make



Normal Case with $n = 200$ Gamma Case with $n = 200$



Normal Case with $n = 600$ Gamma Case with $n = 600$

Figure 2. Power performance of T'_n (solid) and $\tilde{T}_n$ (dashed-dot) for the model with $d = 2$ and fixed $p = 400$ when $\mathbf{Y}$ comes from scenario (i) and $\rho = 0.5$.

the computational cost comparable to that of T'_n, a naive method is to randomly choose k_0 points from n samples, denoted as an index set $\mathcal{M}$. Then, the test statistic is given as

$$T_{k_0,h} = (k_0 - 1) \sum_{i \in \mathcal{M}} \operatorname{tr}(\widehat{\boldsymbol{\Sigma}^2_{\mathbf{Y}_i}}) - 2 \sum_{i<j;i,j \in \mathcal{M}} \sum \operatorname{tr}(\widehat{\boldsymbol{\Sigma}_{\mathbf{Y}_i} \boldsymbol{\Sigma}_{\mathbf{Y}_j}}).$$

To make the tuning parameters comparable, we let $nh^d = l$ and $k_0 = H$ because they basically imply similar quantities of "effective" sample sizes. Here, we use the Epamechnikov kernel as $K(\cdot)$, and derive the critical value of $T_{n,h}$ by numerical simulation.

Table 4 presents the power performance of T'_n, $\tilde{T}_n$, and $T_{n,h}$ for the model with $d = 1$ and $(n, p) = (80, 400)$. Because n is relatively small, we can compute $T_{n,h}$ for comparison purposes. Similarly to the results shown in Figure 1, T'_n

Table 4. Power performances of T'_n, $\tilde{T}_n$ and $T_{n,h}$ for the model of $d = 1$, with $(n, p) = (80, 400)$, where $Y \sim U(2, 4)$, $\boldsymbol{\mu}_Y$ comes from Case (I) and $\rho = 0.5$.

$\rho = 0.5$	$l = 8$			$l = 10$			$l = 20$		
θ_n	T'_n	$\tilde{T}_n$	$T_{n,h}$	T'_n	$\tilde{T}_n$	$T_{n,h}$	T'_n	$\tilde{T}_n$	$T_{n,h}$
Normal case									
0.0	3.9	7.9	6.3	5.2	8.1	6.4	7.2	8.4	6.8
0.1	6.4	9.3	10.8	5.6	8.2	10.1	7.6	9.2	9.1
0.2	11.3	11.6	13.3	9.7	11.0	13.3	12.0	12.7	11.1
0.3	20.1	14.2	17.4	15.8	13.5	17.6	15.6	15.6	14.4
0.4	29.4	14.0	25.1	25.9	15.9	25.8	25.7	23.3	19.7
Gamma case									
0.0	5.7	8.3	6.2	4.5	7.0	6.9	9.0	10.6	5.9
0.1	5.8	8.6	8.6	6.9	9.2	7.9	6.7	8.2	8.3
0.2	10.0	10.1	12.1	9.2	9.8	10.0	11.2	12.4	9.5
0.3	19.1	13.3	16.7	15.0	12.0	14.9	16.0	16.0	12.7
0.4	27.0	12.9	25.5	23.7	14.1	21.0	24.8	22.4	15.7

Table 5. Power performances of T'_n, $\tilde{T}_n$, and $T_{k_0,h}$ for the model of $d = 1$, with $(n, p) = (200, 2, 000)$, where $Y \sim U(2, 4)$, $\boldsymbol{\mu}_Y$ comes from Case (I), and $\rho = 0.5$.

$\rho = 0.5$	$l = 10$			$l = 20$			$l = 30$		
θ_n	T'_n	$\tilde{T}_n$	$T_{k_0,h}$	T'_n	$\tilde{T}_n$	$T_{k_0,h}$	T'_n	$\tilde{T}_n$	$T_{k_0,h}$
Normal case									
0.0	4.2	6.7	5.8	5.1	6.0	6.9	5.7	6.5	5.0
0.1	15.0	9.9	6.3	8.6	8.1	4.7	8.6	8.6	7.2
0.2	63.0	11.6	11.5	23.5	11.9	6.5	21.9	15.5	9.4
0.3	97.9	13.5	10.7	55.7	20.0	7.8	39.6	22.9	10.2
0.4	100	19.9	14.8	82.1	27.8	12.1	66.1	35.6	15.5
Gamma case									
0.0	4.6	7.5	5.5	5.3	7.0	5.5	6.2	7.3	4.5
0.1	13.9	8.6	6.2	8.6	7.9	6.0	9.4	9.2	8.2
0.2	60.5	12.2	7.3	23.8	13.4	7.5	19.1	13.5	11.1
0.3	98.4	16.4	9.2	54.4	19.5	9.7	41.2	24.1	12.3
0.4	100	18.4	13.0	81.8	27.2	14.4	65.2	34.4	19.4

gains power as l decreases. In contrast, $\tilde{T}_n$ gains power as l increases because, in these cases, the last term in the asymptotic power of $\beta_{T'_n}$ becomes dominant. Clearly, the power of $T_{n,h}$ shows no significant improvement. Table 5 reports the simulation results of T'_n, $\tilde{T}_n$, and $T_{k_0,h}$ for the model with $d = 1$ and $(n, p) = (200, 2, 000)$. Because n is relatively large, we only include $T_{k_0,h}$ for comparison. T'_n outperforms $T_{k_0,h}$, as expected.

Previous results show that l can have a significant effect on the power performance of our method. Motivated by the discussion of the power function in

Table 6. Power performances of T_n' for the model of $d = 1$, with $(n, p) = (200, 400)$, in different l under the first alternative.

							l						
θ_n	6	8	10	12	14	16	18	20	22	24	26	28	30
Normal case													
0.0	4.8	5.8	4.4	5.6	5.1	4.8	4.9	5.3	4.7	5.8	7.1	5.0	5.1
0.1	10.2	7.3	6.5	6.5	5.7	5.5	5.4	6.1	7.0	6.9	8.3	7.9	9.0
0.2	20.0	15.8	14.2	14.3	10.7	12.5	10.8	10.0	11.4	12.4	13.5	12.0	14.9
0.3	48.8	34.4	25.8	21.5	21.3	20.3	21.6	21.6	21.5	19.5	22.6	21.7	23.4
0.4	79.4	59.1	47.4	40.3	35.0	34.3	30.6	31.7	33.5	34.2	37.0	36.1	36.8
Gamma case													
0.0	5.1	4.6	6.1	5.6	5.5	5.2	5.9	4.9	5.7	5.8	4.4	6.3	6.2
0.1	10.5	8.2	8.8	6.6	6.8	6.8	7.7	6.2	7.3	8.9	7.4	7.3	5.7
0.2	23.1	19.5	13.3	9.4	11.4	9.9	12.3	12.1	11.2	12.1	11.4	13.5	12.0
0.3	51.2	32.7	25.3	23.9	23.4	21.5	20.3	21.4	19.8	21.2	24.0	22.5	24.2
0.4	79.5	57.3	45.4	38.9	34.8	35.4	30.2	32.6	33.8	32.7	35.4	36.2	38.5

Table 7. Power performances of T_n' for the model of $d = 1$, with $(n, p) = (200, 400)$, in different l under the second alternative.

							l						
θ_n	6	8	10	12	14	16	18	20	22	24	26	28	30
Normal case													
0.0	4.5	4.7	6.3	4.7	3.8	4.4	4.9	6.2	5.5	6.5	5.6	5.9	5.8
0.1	5.3	5.9	5.6	6.8	6.3	6.9	5.6	8.0	9.2	8.9	7.2	9.2	8.9
0.2	7.0	9.6	10.1	12.7	11.9	14.8	16.9	14.1	16.0	18.6	20.5	20.1	21.6
0.3	14.2	16.2	19.4	26.5	29.4	34.3	33.1	39.2	40.8	45.3	48.8	50.1	55.4
0.4	25.9	38.3	46.8	55.6	59.7	67.5	71.7	78.1	81.2	85.4	89.6	88.5	91.6
Gamma case													
0.0	5.2	5.3	4.8	4.4	5.0	3.7	4.6	5.2	5.7	6.3	6.5	5.9	7.5
0.1	6.4	5.9	7.0	6.7	7.7	7.6	6.5	6.4	9.8	9.0	8.8	9.7	7.5
0.2	9.6	9.5	10.1	12.6	11.9	12.1	13.3	16.1	15.5	18.0	19.2	19.0	21.2
0.3	16.5	20.0	24.3	26.1	28.4	32.8	38.2	41.8	45.1	46.7	51.0	50.8	55.0
0.4	30.3	39.1	49.0	52.6	62.1	66.1	72.0	77.2	80.0	83.1	88.0	87.9	90.5

Section 2, two simple experiments were conducted to further examine the role of l. Alternative 1 is generated as in Figure 1, whereas alternative 2 is generated as $\mathbf{\Sigma}(y) = I(y < 3)I_p + I(y \geq 3)\mathbf{\Sigma}^*$, where $\mathbf{\Sigma}^*_{ij} = \theta_n^{|i-j|}$, with no changes on the diagonal. By the asymptotic power expressions (2.5), we expect $\beta_{T_n'}$ (in l) to be a convex function under the first alternative, but a monotone increasing function under the second. This was verified by the simulation results in Tables 6–7.

5. A Real-Data Application: Cardiomyopathy Microarray Data

In this section, we apply our proposed methods to cardiomyopathy microarray data. This data set has been studied by, among others, Segal, Dahlquist and Conklin (2003), Hall and Miller (2009), and Li, Zhong and Zhu (2012). These works typically try to identify the most significant genes for the overexpression of a G protein-coupled receptor (Ro1) in mice. The data set contains 30 samples. Compared with the small sample size, the dimension of the observed real-value vector in each sample is very large (i.e., 6,320). That is, a univariate Y denotes the Ro1 expression level, and its corresponding 6,319 X_k are other gene-expression levels.

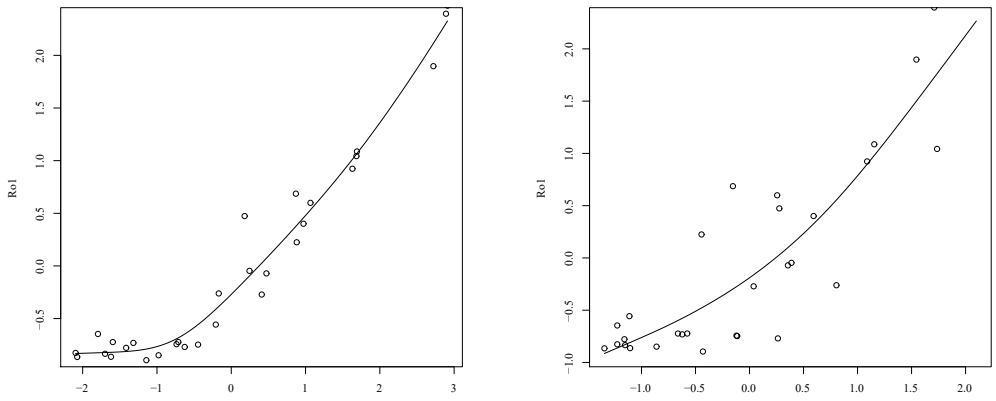
We first test the constancy of $\operatorname{cov}(\mathbf{X} \mid Y)$, which can be regarded as testing whether a heterogeneous effect exists in Ro1. If the underlying heterogeneous effect is ignored, the test for detecting differences in gene expression levels might lose power. After removing obvious outliers in Y, the respective p-values of T_n' and $\tilde{T}_n$ are 0.0645, 0.0713 for $l = 5$; 0.0006, 0.0002 for $l = 6$; and 0.0011, 0.0002 for $l = 10$, which imply that $\operatorname{cov}(\mathbf{X} \mid Y)$ is stochastic and that the heterogeneous effect exists.

Furthermore, by applying the standard SIS procedure (Fan and Lv (2008)), we select the top $\lfloor n/\log n \rfloor$ X and then fit a single-index model,

$$Y = \ell(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) + \epsilon.$$

Here, we consider two estimates of $\boldsymbol{\beta}$: $\hat{\boldsymbol{\beta}}$, from the sliced inverse regression (SIR) procedure (Li (1991)); and $\tilde{\boldsymbol{\beta}}$, from the sliced average variance estimation (SAVE) procedure. In both cases, we use the R package *dr*. Because the result of the hypothesis test for the dimension of the central subspace revealed that there is only one dimension-reduction direction, a single-index model suffices for our study. We next test whether $\operatorname{cov}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is nonrandom. Our motivation is that if the constant conditional variance (CCV) assumption is violated for these data, that is, $\operatorname{cov}(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is random, then the fit between Y and $\tilde{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ would be worse than that between Y and $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$, because SAVE relies on CCV, whereas SIR does not.

Figure 3 shows a scatter plot of Y versus a linear combination of the gene-expression levels $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ identified by SIS-SIR and the top one X ranked by SIS. The figure shows that $\hat{\boldsymbol{\beta}}$ from SIR is closer to the true $\boldsymbol{\beta}$, and thus we can test $\operatorname{cov}(\mathbf{X} \mid \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ instead. This intuitive reasoning is checked further in additional simulation results in the Supplementary Material. The respective P-values of T_n'



$\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ Top one X ranked by SIS

Figure 3. The scatter Y versus the linear combination of gene expression levels $\hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}$ identified by SIS-SIR and the top one X ranked by SIS.

and $\tilde{T}_n$ for testing $\mathrm{cov}(\mathbf{X} \mid \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ are 0.0278, 0.0079 for $l = 5$; 0.0781, 0.0417 for $l = 6$; and 0.0018, 0.0004 for $l = 10$, which implies that $\mathrm{cov}(\mathbf{X} \mid \hat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ varies and the CCV assumption is indeed violated for these data. The R^2 values based on the estimators from SIR and SAVE are 0.9628 and 0.0135, respectively, both computed by the R package *dr*. As expected, SIR performs far better because it does not require the constant variance condition on which the validity of SAVE so heavily relies.

Supplementary Material

The online Supplementary Material contains all technical proofs, as well as several additional simulation results.

Acknowledgments

The authors thank the editor, the associate editor, and three anonymous referees for their many helpful comments and suggestions. This research was supported by NNSF of China Grants 11925106, 11690015, 11622104 and 11431006, and NSF of Tianjin 18JCJQJC46000.

References

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

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Statist.* **36**, 2577–2604.

Bruno, E. and Timo, T. (2007). Testing constancy of the error covariance matrix in vector models. *J. Econometrics* **140**, 753–780

Chen, Z. and Leng, C. (2016). Dynamic covariance models. *J. Amer. Statist. Assoc.* **111**, 1196–1207.

Chen, S. X., Zhang, L,-X. and Zhong, P,-S. (2010). Tests for high-dimensional covariance matrices. *J. Amer. Statist. Assoc.* **105**, 810–819.

Fan J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B (Stat. Methodol)* **70**, 849–911.

Feng, L., Zou, C., Wang, Z. and Zhu, L. (2015). Two-sample behrens-fisher problem for high-dimensional data. *Statist. Sinica* **25**, 1297–1312.

Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comp. Graph. Stat.* **18**, 533–550.

Hart, J. D. (1997). *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer, New York.

Horowitz, J. L. and Spokoiny, V. G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69**, 599–631.

Li, J. and Chen, S. X. (2012). Two sample test for high-dimensional covariance matrices. *Ann. Statist.* **40**, 908–940.

Li, K. C. (1991). Sliced inverse regression for dimension reduction(with discussion). *J. Amer. Statist. Assoc.* **86**, 316–342.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107**, 1129–1139.

Segal, M. R., Dahlquist, K. D. and Conklin, B. R. (2003). Regression approach for microarray data analysis. *J. Comput. Biol.* **10**, 961–980.

Tse, Y. (2000). A test for constant correlations in a multivariate GARCH model. *J. Econometrics* **98**, 107–127.

Yu, K., Zhang, H., Wheeler, W., Horne, H. N., Chen, J. and Figueroa, J. D. (2015). A robust association test for detecting genetic variants with heterogeneous effects. *Biostatistics* **16**, 5–16.

Zhong, P.-S., Chen, S. X. and Xu, M. Y. (2013). Tests alternative to higher criticism for high-dimensional means under sparsity and column-wise dependence. *Ann. Statist.* **41**, 2820–2851.

Zou, C., Peng, L., Feng, L. and Wang, Z. (2014). Multivariate-sign-based high-dimensional tests for sphericity. *Biometrika* **101**, 229–236.

School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China, 30007.

E-mail: denglu014@mail.nankai.edu.cn

School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China, 30007.

E-mail: nk.chlzou@gmail.com

School of Statistics and Data Science, LPMC and KLMDASR, Nankai University, Tianjin, China, 30007.

E-mail: zjwang@nankai.edu.cn

Department of Mathematics, South University of Science and Technology of China, Shenzhen, China, 518055.

E-mail: xchen2006@gmail.com

(Received November 2016; accepted September 2018)