*Biometrika* (2023), **00**, 0, *pp.* 1–21 https://doi.org/10.1093/biomet/asad001
Advance Access Publication 18 January 2023

# High-dimensional analysis of variance in multivariate linear regression

By ZHIPENG LOU

*Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Charlton Street, Princeton, New Jersey 08544, U.S.A.*
zhipengprob@gmail.com

XIANYANG ZHANG

*Department of Statistics, Texas A&M University, Blocker 435, College Station, Texas 77843, U.S.A.*
zhangxiany@stat.tamu.edu

and WEI BIAO WU

*Department of Statistics, University of Chicago, 5734 S. University Avenue, Chicago 60637, U.S.A.*
wbwu@galton.uchicago.edu

## Summary

In this paper, we develop a systematic theory for high-dimensional analysis of variance in multivariate linear regression, where the dimension and the number of coefficients can both grow with the sample size. We propose a new $U$-type statistic to test linear hypotheses and establish a high-dimensional Gaussian approximation result under fairly mild moment assumptions. Our general framework and theory can be used to deal with the classical one-way multivariate analysis of variance, and the nonparametric one-way multivariate analysis of variance in high dimensions. To implement the test procedure, we introduce a sample-splitting-based estimator of the second moment of the error covariance and discuss its properties. A simulation study shows that our proposed test outperforms some existing tests in various settings.

*Some key words*: Data-splitting; Gaussian approximation; Multivariate analysis of variance; One-way layout; $U$ statistic.

## 1. Introduction

In statistical inference of multivariate linear regression, a fundamental problem is to investigate the relationships between the covariates and the responses. In this article, we aim to test whether a given set of covariates are associated with the responses by multivariate analysis of variance, MANOVA. To fix ideas, we consider the following multivariate linear regression model with $p$ predictors:

$$Y_i = B^{\mathrm{T}} X_i + V_i \quad (i = 1, \ldots, n), \tag{1}$$

where $Y_i = (Y_{i1}, \ldots, Y_{id})^{\mathrm{T}}$ and $X_i = (X_{i1}, \ldots, X_{ip})^{\mathrm{T}}$ are respectively the response vector and the predictor vector for the $i$th sample, $B^{\mathrm{T}} = (B_1, \ldots, B_p)$ with $B_k \in \mathbb{R}^d$ is the unknown coefficient matrix consisting of coefficients on the $k$th covariate, and the innovation vectors $V_1, \ldots, V_n \in \mathbb{R}^d$ are independent and identically distributed random vectors with $E(V_1) = 0$ and $\mathrm{cov}(V_1) = \Sigma$. The first element of $X_i$ can be set to 1 to reflect an intercept term. Equivalently, we can write (1) in compact matrix form as

$$Y = XB + V, \tag{2}$$

where $Y = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ and $V = (V_1, \ldots, V_n)^{\mathrm{T}}$. Let $\mathcal{C} \in \mathbb{R}^{m \times p}$ be a matrix of rank $m$, where $m \in \{1, \ldots, p\}$. We are interested in testing a collection of linear constraints on the coefficient matrix,

$$H_0 : \mathcal{C}B = 0 \quad \text{versus} \quad H_1 : \mathcal{C}B \neq 0. \tag{3}$$

This testing problem has been extensively studied in the low-dimensional setting where both the number of predictors and the dimension of the response are small relative to the sample size. A natural and popular choice of method is the classical likelihood ratio test in the case where the errors are normally distributed; see Anderson (2003, Ch. 8) for a review of theoretical investigations. In recent years, high-dimensional data are increasingly encountered in various applications, and tremendous efforts have been made to develop new methodologies and theories for high-dimensional regression. The paradigm in which $d$ is 1 or small and $p$ can increase with $n$ has received considerable attention, but the case where $d$ is very large and $p$ relatively small has been less studied. The model (2) in the latter setting has been applied to a number of research problems involving high-dimensional data such as DNA sequence data, gene expression microarray data and imaging data; see, for example, Wessel & Schork (2006) and Zapala & Schork (2006, 2012). Such studies typically generate huge amounts of data, i.e., responses, that, because of their expense and sophistication, are often collected on a relatively small number of individuals, and investigate how the data can be explained by a certain number of predictor variables such as the ages of individuals assayed, clinical diagnoses, strain memberships, cell line types or genotype information (Zapala & Schork, 2006). Owing to the inappropriateness of applying the standard MANOVA strategy and a lack of high-dimensional MANOVA theory, biological researchers often resort to some form of data reduction such as cluster analysis and factor analysis, which can suffer from many problems, as pointed out by Zapala & Schork (2012). Zapala & Schork (2006, 2012) incorporated a distance matrix to modify the standard MANOVA, but they commented that very little published material exists that can be used to guide a researcher as to which distance measure is the most appropriate for a given situation. Motivated by these real-world applications, we aim to develop a general methodology for high-dimensional MANOVA and lay a theoretical foundation for assessing statistical significance.

The testing problem (3) for model (2) is closely related to a group of high-dimensional hypothesis tests. The two-sample mean test, for testing $H_0 : \mu_1 = \mu_2$ where $\mu_1 \in \mathbb{R}^d$ and $\mu_2 \in \mathbb{R}^d$ are the mean vectors of two different populations, is a special case with $p = 2$, $B = (\mu_1, \mu_2)^{\mathrm{T}}$ and $\mathcal{C} = (1, -1)$. There is a large literature on implementting the Hotelling $T^2$-type statistic in high-dimensional situations where $d$ is large; see, for example, Bai & Saranadasa (1996), Chen & Qin (2010) and Srivastava et al. (2013), among many others. This approach can be generalized to test the equality of multiple mean vectors in high dimensions.

Notable works include Schott (2007), Cai & Xia (2014), Hu et al. (2017), Li et al. (2017), Zhang et al. (2017) and Zhou et al. (2017). In most existing work, the random samples are assumed to be Gaussian or to follow some linear structure such as that of Bai & Saranadasa (1996). In contrast, the testing problem we are interested in is much more general. First, all the aforementioned high-dimensional mean test problems can be fitted into our framework and, further, we can deal with the more general multivariate linear regression in the presence of an increasing number of predictor variables. Second, we do not assume Gaussianity or any particular structure of the error vectors $\{V_i\}_{i=1}^n$.

Throughout the paper, we assume $p < n$ and that the design matrix $X$ is of full column rank so that $X^{\mathrm{T}}X$ is invertible. The conventional MANOVA test statistic for (3) is

$$Q_n = |PY|_{\mathbb{F}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} P_{ij}\, Y_i^{\mathrm{T}} Y_j,$$

where $|\cdot|_{\mathbb{F}}$ stands for the Frobenius norm and

$$P = X(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}\{\mathcal{C}(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}\}^{-1}\mathcal{C}(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = (P_{ij})_{n\times n}$$

is the orthogonal projection matrix onto the column space of the matrix $X(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}$. We reject the null hypothesis $H_0$ if $Q_n$ is larger than some critical value. In the univariate case where $d = 1$, the asymptotic behaviour of $Q_n$ has been extensively studied; see Götze & Tikhomirov (1999, 2002) for detailed discussions. The validity of performing a test for (3) using $Q_n$ when $d$ is large has been an open question for a long time. The first goal of the present paper is to provide a solution to this problem by rigorously establishing a distributional approximation for the traditional MANOVA test statistic when $d$ is allowed to grow with $n$. Our key tool is the Gaussian approximation for degenerate $U$-type statistics: under fairly mild moment conditions, quadratic functionals of non-Gaussian random vectors can be approximated by those of Gaussian vectors with the same covariance structure. Chen (2018) established a Gaussian approximation result for high-dimensional nondegenerate $U$-statistics by Stein's method, which cannot be applied to the degenerate case considered here. From a technical point of view, we employ completely different arguments to bound the distance between the distribution functions of the test statistic and its Gaussian analogue.

The main contributions of this paper are three-fold. First, we develop a systematic theory for the conventional MANOVA test statistic $Q_n$ in the high-dimensional setting. More specifically, we establish a dichotomy result: $Q_n$ can be approximated either by a linear combination of independent chi-squared random variables or by a normal distribution under different conditions; see Theorem 1. While this reveals the interesting theoretical properties of the test statistics, it causes difficulties in applications as one may not know which asymptotic distribution to use in practice. To overcome this difficulty, as the second main contribution of our paper, we propose a new $U$-type test statistic. Using this modified test statistic, such a dichotomy does not appear; see Theorem 2 for the asymptotic result. Third, we propose a new estimator for the second spectral moment of the covariance matrix via a data-splitting technique. To the best of our knowledge, this is the first work dealing with an unbiased and ratio-consistent estimator in the multivariate linear regression model.

## 2. Theoretical results

### 2.1. *Notation and assumptions*

We now introduce some notation. Let $\mathbb{I}(\cdot)$ denote the indicator function. For random variables $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, the Kolmogorov distance is defined by $\rho(X, Y) = \sup_{z\in\mathbb{R}} |\mathrm{pr}(X \leqslant$

$z) - \mathrm{pr}(Y \leqslant z)|$. For $q > 0$, we write $\|X\|_q = \{E(|X|^q)\}^{1/q}$ if $E(|X|^q) < \infty$. For two matrices $A = (a_{ij})_{i\leqslant I, j\leqslant J}$ and $B = (b_{ij})_{i\leqslant I, j\leqslant J}$, $A \circ B = (a_{ij}b_{ij})_{i\leqslant I, j\leqslant J}$ denotes their Hardmard product. For any positive integer $m$, we use $I_m$ to denote the $m \times m$ identity matrix. For two sequences of positive numbers $(a_n)$ and $(b_n)$, we write $a_n \lesssim b_n$ if there exists some constant $C$ such that $a_n \leqslant Cb_n$ for all large $n$. We use $C, C_1, C_2, \ldots$ to denote positive constants whose values may differ at different places.

Let $\lambda_1(\Sigma) \geqslant \cdots \geqslant \lambda_d(\Sigma) \geqslant 0$ denote the eigenvalues of $\Sigma = \mathrm{cov}(V_1)$, and let $\varsigma = |\Sigma|_{\mathbb{F}} = \{\sum_{k=1}^d \lambda_k^2(\Sigma)\}^{1/2}$. For $q \geqslant 2$, we define

$$M_q = E\left(\left|\frac{V_1^{\mathrm{T}} V_2}{\varsigma}\right|^q\right) \quad \text{and} \quad L_q = E\left(\left|\frac{V_1^{\mathrm{T}} \Sigma V_1}{\varsigma^2}\right|^{q/2}\right).$$

*Assumption* 1. For the diagonal elements $P_{11}, \ldots, P_{nn}$ of the matrix $P$,

$$\frac{1}{m}\sum_{i=1}^n P_{ii}^2 \to 0$$

as $n \to \infty$.

*Remark* 1. Assumption 1 is quite natural and mild for testing (3). For instance, it automatically holds for the one-sample test of the mean vector as $m^{-1}\sum_{i=1}^n P_{ii}^2 = 1/n$. Additionally, in the context of the $K$-sample test, as discussed in § 3.1, Assumption 1 is satisfied as long as the minimum sample size goes to infinity. More generally, since $\sum_{i=1}^n P_{ii} = m$, a simple sufficient condition for Assumption 1 would be $\max_{1\leqslant i\leqslant n} P_{ii} \to 0$. Further discussion of this condition can be found in Remark 6 and Example 1.

### 2.2. *Asymptotic distribution of the conventional MANOVA test statistics*

Under the null hypothesis $\mathcal{C}B = 0$, $PXB = X(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}\{\mathcal{C}(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}\}^{-1}\mathcal{C}B = 0$ and hence $Q_n = |PXB + PV|_{\mathbb{F}}^2 \overset{H_0}{=} |PV|_{\mathbb{F}}^2$, which can be further decomposed as

$$Q_n \overset{H_0}{=} \sum_{i=1}^n\sum_{j=1}^n P_{ij}V_i^{\mathrm{T}}V_j = \sum_{i=1}^n P_{ii}V_i^{\mathrm{T}}V_i + \sum_{i=1}^n\sum_{j\neq i} P_{ij}V_i^{\mathrm{T}}V_j =: D_n + Q_n^{\star}. \tag{4}$$

Observe that $D_n$ is a weighted sum of independent and identically distributed random variables, and $Q_n^{\star}$ is a second-order nondegenerate $U$-statistic of high-dimensional random vectors. These two terms can be differently distributed in the high-dimensional setting. More specifically, since $D_n$ and $Q_n^{\star}$ are uncorrelated, we have $\mathrm{var}(Q_n) = \mathrm{var}(D_n) + \mathrm{var}(Q_n^{\star})$, where

$$\mathrm{var}(D_n) = \sum_{i=1}^n P_{ii}^2\|E_0(V_1^{\mathrm{T}}V_1)\|_2^2, \quad \mathrm{var}(Q_n^{\star}) = 2\left(m - \sum_{i=1}^n P_{ii}^2\right)\varsigma^2,$$

with $E_0(V_1^{\mathrm{T}}V_1) = V_1^{\mathrm{T}}V_1 - E(V_1^{\mathrm{T}}V_1)$. When the dimension $d$ increases with the sample size $n$, the magnitudes of $\mathrm{var}(D_n)$ and $\mathrm{var}(Q_n^{\star})$ can be quite different for non-Gaussian $\{V_i\}_{i=1}^n$; see Example 2. As a consequence, $Q_n$ can exhibit different asymptotic null distributions. More

precisely, to asymptotically quantify the discrepancy between $\mathrm{var}(D_n)$ and $\mathrm{var}(Q_n^\star)$, under Assumption 1 we define

$$\Lambda^2 = \frac{\sum_{i=1}^n P_{ii}^2 \|E_0(V_1^{\mathrm{T}} V_1)\|_2^2}{m\varsigma^2}.$$

Before presenting the distributional theory for $Q_n$, we first define its Gaussian analogue. Let $Z_1, \ldots, Z_n$ be independent and identically distributed $N(0, \Sigma)$ Gaussian random vectors and write $Z = (Z_1, \ldots, Z_n)^{\mathrm{T}}$. Then the Gaussian analogue of $Q_n$ is defined as the same quadratic functional of $\{Z_i\}_{i=1}^n$,

$$G_n = |PZ|_{\mathbb{F}}^2 = \sum_{i=1}^n \sum_{j=1}^n P_{ij} Z_i^{\mathrm{T}} Z_j. \tag{5}$$

THEOREM 1. *Let $q = 2 + \delta$, where $0 < \delta \leqslant 1$. Suppose that Assumption 1 holds and*

$$\Delta_q = \frac{\sum_{i=1}^n \sum_{j \neq i} |P_{ij}|^q}{m^{q/2}} M_q + \frac{\sum_{i=1}^n P_{ii}^{q/2}}{m^{q/2}} L_q \to 0. \tag{6}$$

(i) *Assume $\Lambda \to 0$. Then under (6) and the null hypothesis,*

$$\rho(Q_n, G_n) \leqslant C_1 \Lambda^{2/5} + C_q \Delta_q^{1/(2q+1)} + C_2 \left( \frac{1}{m} \sum_{i=1}^n P_{ii}^2 \right)^{1/5} \to 0.$$

(ii) *Assume $\Lambda \to \infty$ and that the Lindeberg condition holds for $W_i = E_0(P_{ii} V_i^{\mathrm{T}} V_i)/(\Lambda \varsigma \sqrt{m})$, i.e., $\sum_{i=1}^n E\{W_i^2 \mathbb{I}(|W_i| > \epsilon)\} \to 0$ for any $\epsilon > 0$. Then under the null hypothesis,*

$$\frac{Q_n - m\,\mathrm{tr}(\Sigma)}{\Lambda \varsigma \sqrt{m}} \Rightarrow N(0, 1).$$

*Remark* 2. Theorem 1 illustrates an interesting dichotomy: the conventional MANOVA test statistic $Q_n$ can have one of two different asymptotic null distributions, depending on the magnitude of the unknown quantity $\Lambda$. This dichotomy poses extra difficulties for using $Q_n$ to test (3) in practical situations, as we need to predetermine which asymptotic distribution to use. Any subjective choice may lead to an unreliable conclusion. To illustrate this, suppose $\Lambda \to 0$. For $\alpha \in (0, 1)$, let $G_n^{-1}(\alpha)$ denote the $(1-\alpha)$th quantile of $G_n$. Based on Theorem 1, an $\alpha$-level test for (3) is $\Phi_0 = \mathbb{I}\{Q_n > G_n^{-1}(\alpha)\}$. However, if one implements $\Phi_0$ in the case where $\Lambda \to \infty$, then the Type I error of $\Phi_0$ is such that $\mathrm{pr}(\Phi_0 = 1 \mid H_0) \to 1/2$, which implies that $\Phi_0$ in this scenario, i.e., $\Lambda \to \infty$, is no better than random guessing.

*Remark* 3. Recently much attention has been paid to studying the dichotomy and similar phase transition behaviour of the asymptotic distributions of classical tests in the high-dimensional setting. For instance, Xu et al. (2019) studied Pearson's chi-squared test in the scenario, where the number of cells can increase with the sample size and demonstrated that the corresponding asymptotic distribution can be either chi-squared or normal. He et al. (2021) derived the phase transition boundaries for several standard likelihood ratio

tests on multivariate mean and covariance structures of Gaussian random vectors. In addition to these tests, we suspect that similar phenomena can occur for many other traditional tests as the dimension increases with the sample size. More importantly, investigating the phase transition phenomena of classical tests, not only contributes to their theoretical development, but also provides motivation for proposing new test procedures and more advanced approximation distributional theory that are suitable for the high-dimensional scenario.

The following lemma establishes an upper bound for $\Delta_q$.

Lemma 1. *Assuming that $M_q < \infty$, we have*

$$\Delta_q < 2\left(\frac{1}{m}\max_{1\leqslant i\leqslant n} P_{ii}\right)^{\delta/2} M_q.$$

*Remark* 4. Condition (6) can be viewed as a Lyapunov-type condition for high-dimensional Gaussian approximation of $Q_n$. It is quite natural and does not directly impose any explicit restriction on the relation between the dimension $d$ and the sample size $n$. In particular, (6) can be dimension-free for some commonly used models; namely, (6) holds for an arbitrary dimension $d \geqslant 1$ as long as $n \to \infty$. For example, suppose that $\{V_i\}_{i=1}^n$ follow the linear process model

$$V_i = A\xi_i \quad (i = 1, \ldots, n), \tag{7}$$

where $A$ is a $d \times L$ matrix for some integer $L \geqslant 1$ and $\xi_i = (\xi_{i1}, \ldots, \xi_{iL})^{\mathrm{T}}$, with $\{\xi_{i\ell}\}_{i,\,\ell\in\mathbb{N}}$ being independent zero-mean random variables having $\mathrm{var}(\xi_{i\ell}) = 1$ for each $\ell \in \mathbb{N}$ and uniformly bounded $q$th moments $\max_{1\leqslant\ell\leqslant L} E(|\xi_{i\ell}|^q) \leqslant C < \infty$. Applying the Burkholder inequality leads to $M_q \leqslant (1+\delta)^q \max_{1\leqslant\ell\leqslant L} \|\xi_{i\ell}\|_q^{2q}$. Consequently, Lemma 1 reveals that a sufficient condition for $\Delta_q \to 0$ is

$$\frac{1}{m}\max_{1\leqslant i\leqslant n} P_{ii} \to 0. \tag{8}$$

Statement (8) depends only on the projection matrix $P$ and does not impose any restriction on the dimension $d$. Moreover, under Assumption 1, (8) is automatically satisfied in view of $\max_{1\leqslant i\leqslant n}(P_{ii}/m)^2 \leqslant m^{-2}\sum_{i=1}^n P_{ii}^2 \to 0$.

### 2.3. *Modified U-type test statistics*

The dichotomous nature of the asymptotic null distribution makes $Q_n$ unsuitable for testing (3) in the high-dimensional setting. This motivates us to propose a modified $U$-type test statistic of $Q_n$ for which such a dichotomy does not occur. Let $B_0 \in \mathbb{R}^{p\times d}$ denote the coefficient matrix of model (2) under the null hypothesis such that $\mathcal{C}B_0 = 0$ and $Y \overset{H_0}{=} XB_0 + V$. Motivated by Theorem 1, a natural candidate for the test statistic $Q_n$ would be

$$Q_{n,0} = Q_n - \sum_{k=1}^n P_{kk}(Y_k - B_0^{\mathrm{T}}X_k)^{\mathrm{T}}(Y_k - B_0^{\mathrm{T}}X_k), \tag{9}$$

which coincides with $Q_n^\star$ in (4) under the null hypothesis. However, $B_0$ is unknown in practice and hence $Q_{n,0}$ is infeasible. The primary goal of this section is to propose a consistent

empirical approximation $U_n$ for $Q_{n,0}$. In particular, motivated by the discussions in § 2.2, the modified test statistic $U_n$ should satisfy

$$U_n \overset{H_0}{=} \sum_{i=1}^{n} \sum_{j \neq i} K_{ij} V_i^{\mathrm{T}} V_j, \quad \frac{U_n - Q_{n,0}}{\{\mathrm{var}(Q_{n,0})\}^{1/2}} \overset{H_0}{=} o_{\mathrm{pr}}(1)$$

for some symmetric matrix $K = (K_{ij})_{n \times n}$. The latter ensures that $U_n$ is asymptotically equivalent to $Q_{n,0}$ in (9). Towards this end, let $\hat{B}_0$ be the least-squares estimator of $B$ under the constraint $\mathcal{C}B = 0$. Then $Y - X\hat{B}_0 = (I_n - P_0)Y$, where $P_0 = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} - P$ is the projection matrix of model (2) under the null hypothesis. In view of (9), the modified $U$-type test statistic is then defined by

$$\begin{aligned} U_n &= Q_n - \sum_{k=1}^{n} \theta_k (Y_k - \hat{B}_0^{\mathrm{T}} X_k)^{\mathrm{T}} (Y_k - \hat{B}_0^{\mathrm{T}} X_k) \\ &\overset{H_0}{=} \sum_{i=1}^{n} \left( P_{ii} - \sum_{k=1}^{n} \theta_k \bar{P}_{ik,0}^2 \right) V_i^{\mathrm{T}} V_i + \sum_{i=1}^{n} \sum_{j \neq i} \left( P_{ij} - \sum_{k=1}^{n} \theta_k \bar{P}_{ik,0} \bar{P}_{jk,0} \right) V_i^{\mathrm{T}} V_j \\ &= \sum_{i=1}^{n} \sum_{j \neq i} \left( P_{ij} - \sum_{k=1}^{n} \theta_k \bar{P}_{ik,0} \bar{P}_{jk,0} \right) V_i^{\mathrm{T}} V_j, \end{aligned} \tag{10}$$

where $\bar{P}_0 = I_n - P_0 = (\bar{P}_{ij,0})_{n \times n}$ and the last equality follows by taking $\theta_1, \ldots, \theta_n$ to be the solutions of the linear equations

$$\sum_{k=1}^{n} \bar{P}_{ik,0}^2 \theta_k = P_{ii} \quad (i = 1, \ldots, n). \tag{11}$$

Typically, the $\theta_k$ in (10) are not $P_{kk}$, as one would naturally like to use in view of (9). We can view (11) as a detailed balanced condition as it removes the diagonals in (10). Let $\theta = (\theta_1, \ldots, \theta_n)^{\mathrm{T}}$ and rewrite (11) in the more compact matrix form

$$(\bar{P}_0 \circ \bar{P}_0)\theta = (P_{11}, \ldots, P_{nn})^{\mathrm{T}}. \tag{12}$$

Let $P_\theta = P - \bar{P}_0 D_\theta \bar{P}_0 = (P_{ij,\theta})_{n \times n}$, where $D_\theta = \mathrm{diag}(\theta_1, \ldots, \theta_n)$ is a diagonal matrix. Then $P_{ii,\theta} = 0$ for all $i = 1, \ldots, n$ in view of (12) and

$$U_n \overset{H_0}{=} \mathrm{tr}(V^{\mathrm{T}} P_\theta V) = \sum_{i=1}^{n} \sum_{j \neq i} P_{ij,\theta} V_i^{\mathrm{T}} V_j.$$

Before proceeding, we introduce a sufficient condition for $U_n$ to exist and be well-defined.

LEMMA 2. *Assume there exists a positive constant $\varpi_0 < 1/2$ such that*

$$\max_{1 \leqslant i \leqslant n} P_{ii,0} \leqslant \varpi_0. \tag{13}$$

*Then the matrix* $\bar{P}_0 \circ \bar{P}_0$ *is strictly diagonally dominant and* $|P_\theta|_{\mathbb{F}}^2 = m - \sum_{i=1}^n \theta_i P_{ii}$. *Moreover, if* $\max_{1\leqslant i\leqslant n} P_{ii} \leqslant \varpi_1\zeta$ *for some positive constant* $\varpi_1 < 1/2$, *where* $\zeta = (1-2\varpi_0)(1-\varpi_0)$, *then* $\max_{1\leqslant i\leqslant n} |\theta_i| \leqslant \varpi_1 < 1/2$.

*Remark* 5. Condition (13) ensures that the matrix $\bar{P}_0 \circ \bar{P}_0$ is invertible. Consequently, the solution $\theta$ to (12) exists and is unique. Here $\theta$ is independent of the dimension $d$, and depends only on the projection matrices $P$ and $P_0$. Moreover, as shown in the proof of Lemma 2,

$$\sum_{i=1}^n \theta_i P_{ii} \leqslant \frac{1}{\zeta}\sum_{i=1}^n P_{ii}^2, \quad \max_{1\leqslant i\leqslant n} |\theta_i| \leqslant \frac{1}{\zeta}\max_{1\leqslant i\leqslant n} P_{ii},$$

which are essential for bounding from above the quantity $\Delta_{q,\theta}$ in Lemma 3 below. Consequently, under Assumption 1, supposing that $\sum_{i=1}^n P_{ii}^2 \leqslant m\zeta/2$ for sufficiently large $n$, we obtain

$$\mathrm{var}(U_n) = 2|P_\theta|_{\mathbb{F}}^2\varsigma^2 = 2\left(m - \sum_{i=1}^n \theta_i P_{ii}\right)\varsigma^2 > m\varsigma^2,$$

which ensures that the proposed test statistic $U_n$ is nondegenerate and well-defined.

*Remark* 6. Since $\mathrm{col}(X(X^{\mathrm{T}}X)^{-1}\mathcal{C}^{\mathrm{T}}) \subset \mathrm{col}(X)$, where $\mathrm{col}(\cdot)$ denotes the column space, $P_0 = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} - P$ defined above is also a projection matrix. Hence $\max\{P_{ii}, P_{ii,0}\} \leqslant X_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X_i$ uniformly for $i \in \{1,\ldots,n\}$, and a sufficient condition for Lemma 2 would be

$$\max_{1\leqslant i\leqslant n} X_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X_i \leqslant \min\{\varpi_0, (1-2\varpi_0)(1-\varpi_0)\varpi_1\}, \tag{14}$$

which is a fairly mild condition on the design matrix $X$. More specifically, for the linear regression model it is commonly assumed (Huber, 1973; Portnoy, 1985; Wu, 1986; Shao & Wu, 1987; Shao, 1988; Mammen, 1989; Navidi, 1989; Lahiri, 1992) that $\max_{1\leqslant i\leqslant n} X_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X_i \to 0$, which ensures a kind of 'robustness of design' (Huber, 1973). It also implies Assumption 1 in view of Remark 1 and can be viewed as an imbalance measure for model (2) (Shao & Wu, 1987).

*Example* 1. Suppose that $X_1,\ldots,X_n$ are independent Gaussian random vectors $N(0,\Gamma)$, where the covariance matrix $\Gamma \in \mathbb{R}^{p\times p}$ has minimal eigenvalue $\lambda_{\min}(\Gamma) > 0$. Then, with probability at least $1 - 2\exp(-n/2) - n^{-1}$,

$$\max_{1\leqslant i\leqslant n} X_i^{\mathrm{T}}(X^{\mathrm{T}}X)^{-1}X_i \leqslant \frac{9p + 18(2p\log n)^{1/2} + 36\log n}{n}. \tag{15}$$

Consequently, condition (14) holds with high probability as long as $p/n$ is sufficiently small.

PROPOSITION 1. *Under the conditions of Lemma 2, we have* $E(U_n) \geqslant 0$. *In particular,*

$$E(U_n) = 0 \iff \mathcal{C}B = 0.$$

### 2.4. *Asymptotic distribution of the modified test statistics*

The primary goal of this section is to establish a Gaussian approximation for the modified test statistic $U_n$. Following (5), the Gaussian analogue of $U_n$ is defined by

$$\mathcal{G}_n = \mathrm{tr}(Z^{\mathrm{T}} P_\theta Z) = \sum_{i=1}^{n} \sum_{j \neq i} P_{ij,\theta} Z_i^{\mathrm{T}} Z_j.$$

The following theorem establishes a non-asymptotic upper bound on the Kolmogorov distance between the distribution functions of $U_n$ and its Gaussian analogue $\mathcal{G}_n$. Compared with Theorem 1, it reveals that the modification of the test statistic $Q_n$ in (10) removes the dichotomous nature of its asymptotic null distribution.

THEOREM 2. *Let $q = 2 + \delta$, where $0 < \delta \leqslant 1$. Assume that (13) holds and that*

$$\Delta_{q,\theta} = \frac{\sum_{i=1}^{n} \sum_{j \neq i} |P_{ij,\theta}|^q}{m^{q/2}} M_q + \frac{\sum_{i=1}^{n} (\sum_{j \neq i} P_{ij,\theta}^2)^{q/2}}{m^{q/2}} L_q \to 0.$$

*Then under Assumption 1 and the null hypothesis,*

$$\rho(U_n, \mathcal{G}_n) \leqslant C_q \Delta_{q,\theta}^{1/(2q+1)} + C \left( \frac{1}{m} \sum_{i=1}^{n} P_{ii}^2 \right)^{1/5} \to 0.$$

Similar to Lemma 1, we establish an upper bound for $\Delta_{q,\theta}$ in the following lemma.

LEMMA 3. *Under the conditions of Lemma 2,*

$$\Delta_{q,\theta} \lesssim \left( \frac{1}{m} \max_{1 \leqslant i \leqslant n} P_{ii} \right)^{\delta/2} M_q.$$

For $\alpha \in (0, 1)$, Proposition 1 and Theorem 2 motivate an $\alpha$-level test for (3) as follows:

$$\Phi_\theta = \mathbb{I}\left( \frac{U_n}{\varsigma |P_\theta|_{\mathbb{F}} \sqrt{2}} > c_{1-\alpha} \right), \tag{16}$$

where $c_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standardized $\mathcal{G}_n / \{\mathrm{var}(\mathcal{G}_n)\}^{1/2}$.

*Remark* 7. The approximating distribution $\mathcal{G}_n$ may or may not be asymptotically normal. Let $\lambda_1(P_\theta), \ldots, \lambda_n(P_\theta)$ denote the eigenvalues of the symmetric matrix $P_\theta$. Being a quadratic functional of Gaussian random vectors $\{Z_i\}_{i=1}^n$, $\mathcal{G}_n$ is distributed as a linear combination of independent chi-squared random variables,

$$\mathcal{G}_n \overset{\mathcal{D}}{=} \sum_{k=1}^{d} \sum_{i=1}^{n} \lambda_k(\Sigma) \lambda_i(P_\theta) \eta_{ik}(1) = \sum_{k=1}^{d} \sum_{i=1}^{n} \lambda_k(\Sigma) \lambda_i(P_\theta) \{\eta_{ik}(1) - 1\},$$

where $\{\eta_{ik}(1)\}_{i,k \in \mathbb{N}}$ are independent $\chi_1^2$ random variables and the last equality follows from the fact that $\sum_{i=1}^n \lambda_i(P_\theta) = \sum_{i=1}^n P_{ii,\theta} = 0$. More specifically, the Lindeberg–Feller central limit theorem and Lemma 2 imply that $\mathcal{G}_n / \{\mathrm{var}(\mathcal{G}_n)\}^{1/2} \Rightarrow N(0, 1)$ if and only if

$$\frac{\lambda_1(\Sigma)}{\varsigma \sqrt{m}} \to 0. \tag{17}$$

Consequently, $c_{1-\alpha}$ in (16) is asymptotically equal to the standard normal quantiles whenever (17) holds.

When $m \to \infty$, condition (17) automatically holds for arbitrary dimensions $d \geqslant 1$ as $\lambda_1(\Sigma) \leqslant \varsigma$. Otherwise, (17) is equivalent to $\mathrm{tr}(\Sigma^4)/\varsigma^4 \to 0$, which is a common assumption for ensuring the asymptotic normality of high-dimensional quadratic statistics; see, for example, Bai & Saranadasa (1996), Chen & Qin (2010), Cai & Ma (2013), Yao et al. (2018) and Zhang et al. (2018), among others. In particular, it reveals that the asymptotic null distribution of $U_n$ can be nonnormal if (17) is violated. For example, let $Y_1, \ldots, Y_n \in \mathbb{R}^d$ be independent and identically distributed random vectors with mean vector $\mu_Y = E(Y_1)$, and consider testing whether $\mu_Y = 0$. Assume that $\Sigma = \mathrm{cov}(Y_1) = (\Sigma_{jk})_{d\times d}$ has entries $\Sigma_{jk} = \vartheta + (1-\vartheta)\,\mathbb{I}\{j = k\}$ for some constant $\vartheta \in (0,1)$. Then $\lambda_1(\Sigma)/(\varsigma\sqrt{m}) \to 1$ and it follows from Theorem 2 that

$$\frac{U_n}{\{\mathrm{var}(U_n)\}^{1/2}} = \frac{\sum_{i=1}^n \sum_{j\neq i} Y_i^{\mathrm{T}} Y_j}{\varsigma\{2n(n-1)\}^{1/2}} \overset{H_0}{\Rightarrow} \frac{\chi_1^2 - 1}{\sqrt{2}}.$$

The simulation study in § 5 shows that our Gaussian multiplier bootstrap approach has satisfactory performance regardless of whether $U_n$ is asymptotically normal or not.

## 3. Applications

### 3.1. *High-dimensional one-way MANOVA*

Let $\{\mathcal{Y}_{ij}\}_{j=1}^{n_i}$ for $i = 1, \ldots, K$ be $K \geqslant 2$ independent samples following the model

$$\mathcal{Y}_{ij} = \mu_i + \mathcal{V}_{ij} \quad (j = 1, \ldots, n_i;\ i = 1, \ldots, K),$$

where $\mu_1, \ldots, \mu_K \in \mathbb{R}^d$ are unknown mean vectors of interest, and $\{\mathcal{V}_{ij}\}_{j\in\mathbb{N}}$ are independent and identically distributed $d$-dimensional random vectors with $E(\mathcal{V}_{i1}) = 0$ and $\mathrm{cov}(\mathcal{V}_{i1}) = \Sigma$. We are interested in testing the equality of the $K$ mean vectors, i.e., testing the hypotheses

$$H_0 : \mu_1 = \cdots = \mu_K \quad \text{versus} \quad H_1 : \mu_i \neq \mu_l \text{ for some } 1 \leqslant i \neq l \leqslant K.$$

Following the construction of (10), we propose the $U$-type test statistic

$$U_{nK} = \sum_{i=1}^K P_{ii,K} \sum_{j=1}^{n_i} \sum_{k\neq j} \mathcal{Y}_{ij}^{\mathrm{T}} \mathcal{Y}_{ik} + \sum_{i=1}^K \sum_{l\neq i} P_{il,K} \sum_{j=1}^{n_i} \sum_{k=1}^{n_l} \mathcal{Y}_{ij}^{\mathrm{T}} \mathcal{Y}_{lk}, \tag{18}$$

where $n = \sum_{i=1}^K n_i$ is the total sample size and

$$P_{ii,K} = \frac{1}{n-2}\left(\frac{n}{n_i} - \frac{n+K-2}{n-1}\right), \quad P_{il,K} = \frac{1}{n-2}\left(\frac{1}{n_i} + \frac{1}{n_l} - \frac{n+K-2}{n-1}\right).$$

In the context of two-sample tests for mean vectors, where $K = 2$, $U_{nK}$ in (18) reduces to

$$U_{nK} = \frac{\sum_{i=1}^{n_1} \sum_{j\neq i} \sum_{k=1}^{n_2} \sum_{l\neq k} (\mathcal{Y}_{1i} - \mathcal{Y}_{2k})^{\mathrm{T}} (\mathcal{Y}_{1j} - \mathcal{Y}_{2l})}{(n-1)(n-2)n_1 n_2/n},$$

which coincides with the commonly used $U$-type test statistic (Chen & Qin, 2010).

For each $i \in \{1, \ldots, K\}$, let $\{\mathcal{Z}_{ij}\}_{j\in\mathbb{N}}$ be independent and identically distributed centred Gaussian random vectors with covariance matrix $\mathrm{cov}(\mathcal{Z}_{ij}) = \Sigma$. Following (5), the Gaussian analogue of $U_{nK}$ is defined by

$$\mathcal{G}_{nK} = \sum_{i=1}^{K} P_{ii,K} \sum_{j=1}^{n_i} \sum_{k\neq j} \mathcal{Z}_{ij}^{\mathrm{T}} \mathcal{Z}_{ik} + \sum_{i=1}^{K} \sum_{l\neq i} P_{il,K} \sum_{j=1}^{n_i} \sum_{k=1}^{n_l} \mathcal{Z}_{ij}^{\mathrm{T}} \mathcal{Z}_{lk}.$$

Let $n_{\min} = \min_{1\leqslant l\leqslant K} n_l$. Since $\max_{1\leqslant i\leqslant n} P_{ii} \leqslant n_{\min}^{-1}$, Assumption 1 holds as long as $n_{\min} \to \infty$. The following proposition establishes a non-asymptotic upper bound on the Kolmogorov distance between the distribution functions of $U_{nK}$ and $\mathcal{G}_{nK}$.

PROPOSITION 2. *Let $q = 2 + \delta$ for some $0 < \delta \leqslant 1$. Assume that $n_{\min} \to \infty$ and*

$$\tilde{M}_q = \max_{1\leqslant l, l'\leqslant K} E\left( \left| \frac{\mathcal{V}_{l1}^{\mathrm{T}} \mathcal{V}_{l'2}}{\varsigma} \right|^q \right) < \infty, \quad \varsigma = |\Sigma|_{\mathbb{F}}.$$

*Then under the null hypothesis,*

$$\rho(U_{nK}, \mathcal{G}_{nK}) \leqslant C_q \big( \tilde{M}_q n_{\min}^{-\delta/2} \big)^{1/(2q+1)} \to 0.$$

*Remark* 8. Both the dimension $d$ and the number of groups $K$ can grow with the total sample size $n$. In particular, as discussed in Remark 4, if all $K$ samples follow the linear process model in (7), then $\rho(U_{nK}, \mathcal{G}_{nK}) \to 0$ as long as $n_{\min} \to \infty$.

### 3.2. *High-dimensional nonparametric one-way MANOVA*

For each $i \in \{1, \ldots, K\}$, let $F_i$ denote the distribution function of $\mathcal{Y}_{i1}$. We consider the problem of testing whether these $K$ independent samples are equally distributed, i.e., testing the hypotheses

$$H_0 : F_1 = \cdots = F_K \quad \text{versus} \quad H_1 : F_i \neq F_l \text{ for some } 1 \leqslant i \neq l \leqslant K. \tag{19}$$

Being fundamental and important in statistical inference, (19) has been extensively studied; see, for example, Kruskal & Wallis (1952), Akritas & Arnold (1994), Brunner & Puri (2001), Rizzo & Székely (2010) and Thas (2010), among many others. However, all these works focus on the traditional low-dimensional scenario, and testing (19) for high-dimensional random vectors has been much less studied. In this section, we propose a new $U$-type test statistic for (19), following the intuition of (10), and establish the corresponding distributional theory. In particular, our asymptotic framework is fairly general, and allows both the dimension $d$ and the number of groups $K$ to grow with $n$.

To begin with, for each $i \in \{1, \ldots, K\}$ let $\phi_i(t) = E\{\exp(\iota t^{\mathrm{T}} \mathcal{Y}_{ij})\}$ denote the characteristic function of $\mathcal{Y}_{ij}$, where $\iota$ stands for the imaginary unit. Then testing (19) is equivalent to testing the hypotheses

$$H_0 : \phi_1 = \cdots = \phi_K \quad \text{versus} \quad H_1 : \phi_i \neq \phi_l \text{ for some } 1 \leqslant i \neq l \leqslant K. \tag{20}$$

Write $\mathcal{Y}_{ij}(t) = \exp(\iota t^{\mathrm{T}} \mathcal{Y}_{ij})$. Similar to (18), our test statistic for (20) is defined by

$$\tilde{U}_{nK} = \sum_{i=1}^{K} P_{ii,K} \sum_{j=1}^{n_i} \sum_{k\neq j} \int \mathcal{Y}_{ij}(t) \overline{\mathcal{Y}_{ik}(t)} w(t)\, \mathrm{d}t$$

$$+\sum_{i=1}^{K}\sum_{l\neq i}P_{il,K}\sum_{j=1}^{n_i}\sum_{k=1}^{n_l}\int \mathcal{Y}_{ij}(t)\overline{\mathcal{Y}_{lk}(t)}w(t)\,\mathrm{d}t,$$

where $w(t) \geqslant 0$ is a suitable weight function such that the integrals above are well-defined. Discussions of some commonly used weight functions are given in Remark 9.

Before proceeding, we define the Gaussian analogue of $\tilde{U}_{nK}$ under the null hypothesis that the $K$ samples are equally distributed. Define the covariance function of $\mathcal{Y}_{11}(t)$ as

$$\Sigma(t,s) = E\{\mathcal{Y}_{11}(t) - \phi_1(t)\}\overline{\{\mathcal{Y}_{11}(s) - \phi_1(s)\}} = \phi_1(t-s) - \phi_1(t)\phi_1(-s) \quad (t,s \in \mathbb{R}^d).$$

Throughout this section, by Mercer's theorem, we assume that this covariance function admits the eigendecomposition

$$\Sigma(t,s) = \sum_{m=1}^{\infty}\lambda_m\varphi_m(t)\overline{\varphi_m(s)} \quad (t,s \in \mathbb{R}^d),$$

where $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant 0$ are eigenvalues and $\varphi_1, \varphi_2, \ldots$ are the corresponding eigenfunctions. We now apply the Karhunen–Loève theorem. Let $\{Z_{ijk}\}_{i,j,k\in\mathbb{N}}$ be independent standard normal random variables and define Gaussian processes

$$\mathcal{Z}_{ij}(t) = \sum_{m=1}^{\infty}\lambda_m^{1/2}Z_{ijm}\varphi_m(t) \quad (t \in \mathbb{R}^d).$$

Then, following (5), the Gaussian analogue of $\tilde{U}_{nK}$ is defined by

$$\tilde{\mathcal{G}}_{nK} = \sum_{i=1}^{K}P_{ii,K}\sum_{j=1}^{n_i}\sum_{k\neq j}\int \mathcal{Z}_{ij}(t)\overline{\mathcal{Z}_{ik}(t)}w(t)\,\mathrm{d}t$$
$$+\sum_{i=1}^{K}\sum_{l\neq i}P_{il,K}\sum_{j=1}^{n_i}\sum_{k=1}^{n_l}\int \mathcal{Z}_{ij}(t)\overline{\mathcal{Z}_{lk}(t)}w(t)\,\mathrm{d}t.$$

Proposition 3. *Let $q = 2+\delta$ for some $0 < \delta \leqslant 1$. Assume that $n_{\min} \to \infty$ and*

$$\tilde{\mathcal{M}}_q = E\left[\left|\frac{\int E_0\{\mathcal{Y}_{11}(t)\}\overline{E_0\{\mathcal{Y}_{12}(t)\}}w(t)\,\mathrm{d}t}{\mathcal{F}}\right|^q\right] < \infty, \quad \mathcal{F}^2 = \sum_{m=1}^{\infty}\lambda_m^2.$$

*Then under the null hypothesis that these $K$ independent samples are equally distributed,*

$$\rho(\tilde{U}_{nK}, \tilde{\mathcal{G}}_{nK}) \leqslant C_q\big(\tilde{\mathcal{M}}_q n_{\min}^{-\delta/2}\big)^{1/(2q+1)} \to 0.$$

*Remark* 9. The proposed test statistic $\tilde{U}_{nK}$ contains a high-dimensional integral over $t \in \mathbb{R}^d$, which can be computationally intractable in practice. To make $\tilde{U}_{nK}$ well-defined and facilitate the computation, we shall choose a suitable weight function $w(t)$ such that $\tilde{U}_{nK}$ has a simple closed-form expression. In the literature, various kinds of weight functions have been proposed, such as the Gaussian kernel function (Gretton et al., 2012), the Laplace

kernel function (Gretton et al., 2012) and the energy kernel function (Székely et al., 2007; Rizzo & Székely, 2010). For instance, let $w(t)$ denote the density function of the random vector $\mathcal{X}\kappa/\sqrt{\eta}$ for some $\kappa > 0$, where $\mathcal{X} \sim N(0, I_d)$ and $\eta \sim \chi^2_1$ are independent or, equivalently, $\mathcal{X}\kappa/\sqrt{\eta}$ is a Cauchy random variable with location parameter 0 and scale parameter $\kappa$. Then it is straightforward to verify that

$$\int \mathcal{Y}_{ij}(t)\overline{\mathcal{Y}_{lk}(t)}w(t)\,\mathrm{d}t = \int \cos\{t^{\mathrm{T}}(Y_{ij} - Y_{lk})\}w(t)\,\mathrm{d}t = \exp(-\kappa|Y_{ij} - Y_{lk}|),$$

which is the same as the Laplace kernel function with bandwidth $1/\kappa$, where $|\cdot|$ stands for the Euclidean distance. A more general result can be derived using Bochner's theorem (see, e.g., Gretton et al., 2009, Theorem 3.1). Consequently, the proposed test statistic $\tilde{U}_{nK}$ reduces to

$$\tilde{U}_{nK} = \sum_{i=1}^{K} P_{ii,K} \sum_{j=1}^{N_i} \sum_{k\neq j} \exp(-\kappa|Y_{ij} - Y_{ik}|) + \sum_{i=1}^{K} \sum_{l\neq i} P_{il,K} \sum_{j=1}^{N_i} \sum_{k=1}^{N_l} \exp(-\kappa|Y_{ij} - Y_{lk}|),$$

which is fairly convenient to compute in practice. Moreover, a suitable choice of the weight function $w(t)$ will also facilitate analysis of the quantities $\mathcal{M}_q$ and $\mathcal{F}$.

## 4. Practical implementation

In this section, we propose an unbiased estimator for $\varsigma^2$, which is ratio-consistent under fairly mild moment conditions. To begin with, since $E(V_i^{\mathrm{T}}V_j)^2 = \varsigma^2$ for any $i \neq j$, a natural unbiased $U$-type estimator for $\varsigma^2$ based on $\{V_i\}_{i=1}^n$ would be

$$\hat{\varsigma}^2_o = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i} (V_i^{\mathrm{T}}V_j)^2. \tag{21}$$

Let $\bar{P}_1 = I_n - X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}} = (P_{ij,1})_{n\times n}$ and $\hat{V} = \bar{P}_1 Y = (\hat{V}_1, \ldots, \hat{V}_n)^{\mathrm{T}}$. Directly substituting the residual vectors $\{\hat{V}_i\}_{i=1}^n$ into (21) yields a feasible but generally biased estimator for $\varsigma^2$. More specifically, for any $i \neq j$,

$$\begin{aligned} E(\hat{V}_i^{\mathrm{T}}\hat{V}_j)^2 = {} & (\bar{P}_{ii,1}\bar{P}_{jj,1} + \bar{P}^2_{ij,1})\varsigma^2 + \bar{P}^2_{ij,1}E(V_1^{\mathrm{T}}V_1)(V_2^{\mathrm{T}}V_2) \\ & + \sum_{k=1}^{n} (\bar{P}_{ik,1}\bar{P}_{jk,1})^2\{\|E_0(V_1^{\mathrm{T}}V_1)\|_2^2 - 2\varsigma^2\}, \end{aligned}$$

which reveals that $(\hat{V}_i^{\mathrm{T}}\hat{V}_j)^2$ is no longer unbiased for $\varsigma^2$ even after proper scaling. This motivates us to propose a new unbiased estimator for $\varsigma^2$ via data-splitting, which excludes the bias terms $(V_i^{\mathrm{T}}V_i)^2$ and $(V_i^{\mathrm{T}}V_i)(V_j^{\mathrm{T}}V_j)$. Without loss of generality, we assume that the sample size $n$ is even in what follows.

*Step* 1. Randomly split $\{1, \ldots, n\}$ into two halves, $\mathcal{A}$ and $\mathcal{A}^{\mathrm{c}}$. Let $\mathcal{M}_{\mathcal{A}} = \{(X_i, Y_i),\ i \in \mathcal{A}\}$ and $\mathcal{M}_{\mathcal{A}^{\mathrm{c}}} = \{(X_i, Y_i),\ i \in \mathcal{A}^{\mathrm{c}}\}$.

*Step* 2. For both $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{A}^c}$, fit model (1) with the least-squares estimates and compute

$$\hat{\Sigma}_{\mathcal{A}} = \frac{1}{n/2-p}\hat{V}_{\mathcal{A}}^{\mathrm{T}}\hat{V}_{\mathcal{A}}, \quad \hat{\Sigma}_{\mathcal{A}^c} = \frac{1}{n/2-p}\hat{V}_{\mathcal{A}^c}^{\mathrm{T}}\hat{V}_{\mathcal{A}^c},$$

where $\hat{V}_{\mathcal{A}}$ and $\hat{V}_{\mathcal{A}^c}$ are the residual matrices of $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{A}^c}$, respectively.

*Step* 3. Compute the estimator $\hat{\varsigma}^2_{\mathcal{A}} = \mathrm{tr}(\hat{\Sigma}_{\mathcal{A}}\hat{\Sigma}_{\mathcal{A}^c})$.

Since $\hat{\Sigma}_{\mathcal{A}}$ and $\hat{\Sigma}_{\mathcal{A}^c}$ are independent and both are unbiased estimators of $\Sigma$, $\hat{\varsigma}^2_{\mathcal{A}}$ is unbiased for $\varsigma^2$ as $E(\hat{\varsigma}^2_{\mathcal{A}}) = \mathrm{tr}\{E(\hat{\Sigma}_{\mathcal{A}})E(\hat{\Sigma}_{\mathcal{A}^c})\} = \mathrm{tr}(\Sigma^2) = \varsigma^2$.

THEOREM 3. *Assume that $p/n < \varpi_2$ for some positive constant $\varpi_2 < 1/2$, and that the least-squares estimates are well-defined for both $\mathcal{M}_{\mathcal{A}}$ and $\mathcal{M}_{\mathcal{A}^c}$. Then*

$$E\left(\left|\frac{\hat{\varsigma}_{\mathcal{A}}}{\varsigma} - 1\right|^2\right) \lesssim \frac{M_4}{n^2} + \frac{p \times \mathrm{tr}(\Sigma^4)}{n^2\varsigma^4} + \frac{\|E_0(V_1^{\mathrm{T}}\Sigma V_1)\|_2^2}{n\varsigma^4}.$$

*Remark* 10. The proof of Theorem 3 is given in the Supplementary Material, where a more general upper bound on $E(|\hat{\varsigma}_{\mathcal{A}}/\varsigma - 1|^{\tau})$ is established for $1 < \tau \leqslant 2$. Theorem 3 reveals that $\hat{\varsigma}_{\mathcal{A}}$ is ratio-consistent under mild moment conditions. Suppose now that $\{V_i\}_{i\in\mathbb{N}}$ follow the linear process model (7) with $\max_{1\leqslant\ell\leqslant L} E(|\xi_{i\ell}|^4) \leqslant C < \infty$. Then $M_4$ is bounded and $\|E_0(V_1^{\mathrm{T}}\Sigma V_1)\|_2^2 \lesssim \mathrm{tr}(\Sigma^4)$. Consequently,

$$E\left(\left|\frac{\hat{\varsigma}_{\mathcal{A}}}{\varsigma} - 1\right|^2\right) \lesssim n^{-2} + \frac{\mathrm{tr}(\Sigma^4)}{n\varsigma^4}.$$

In this case, $\hat{\varsigma}_{\mathcal{A}}$ is ratio-consistent for arbitrary dimensions $d \geqslant 1$ as long as $n \to \infty$.

*Remark* 11. There are a total of $\binom{n}{n/2}$ different ways of splitting $\{1, \ldots, n\}$ into two halves. To reduce the influence of randomness of an arbitrary splitting, we can repeat the procedure independently multiple times and then take the average of the resulting estimators. We refer to Fan et al. (2012) for more discussion about data-splitting and repeated data-splitting.

*Remark* 12. Let $\hat{\Sigma} = (n-p)^{-1}\hat{V}^{\mathrm{T}}\hat{V}$. Observe that $E(\hat{V}_i^{\mathrm{T}}\hat{V}_j) = \bar{P}_{ij,1}\mathrm{tr}(\Sigma)$. We can estimate $\varsigma^2$ via

$$\hat{\varsigma}^2_S = \frac{\sum_{i,j=1}^n |\hat{V}_i^{\mathrm{T}}\hat{V}_j - \bar{P}_{ij,1}\mathrm{tr}(\hat{\Sigma})|^2}{(n-p+2)(n-p-1)} = \frac{(n-p)^2}{(n-p+2)(n-p-1)}\left[|\hat{\Sigma}|_{\mathbb{F}}^2 - \frac{\{\mathrm{tr}(\hat{\Sigma})\}^2}{n-p}\right],$$

which is the same as the estimator proposed in Srivastava & Fujikoshi (2006), where $\{V_i\}_{i=1}^n$ are assumed to be Gaussian random vectors. See also Bai & Saranadasa (1996). However, for non-Gaussian $\{V_i\}_{i=1}^n$ such that $\|E_0(V_1^{\mathrm{T}}V_1)\|_2^2 \neq 2\varsigma^2$, this estimator is generally biased as

$$E(\hat{\varsigma}^2_S) - \varsigma^2 = \frac{\sum_{i=1}^n \bar{P}_{ii,1}^2}{(n-p)(n-p+2)}\{\|E_0(V_1^{\mathrm{T}}V_1)\|_2^2 - 2\varsigma^2\}.$$

In particular, the bias of $\hat{\varsigma}^2_S$ can diverge when $\|E_0(V_1^{\mathrm{T}}V_1)\|_2^2$ is much larger than $\varsigma^2$. Below we provide an example that typifies the diverging bias.

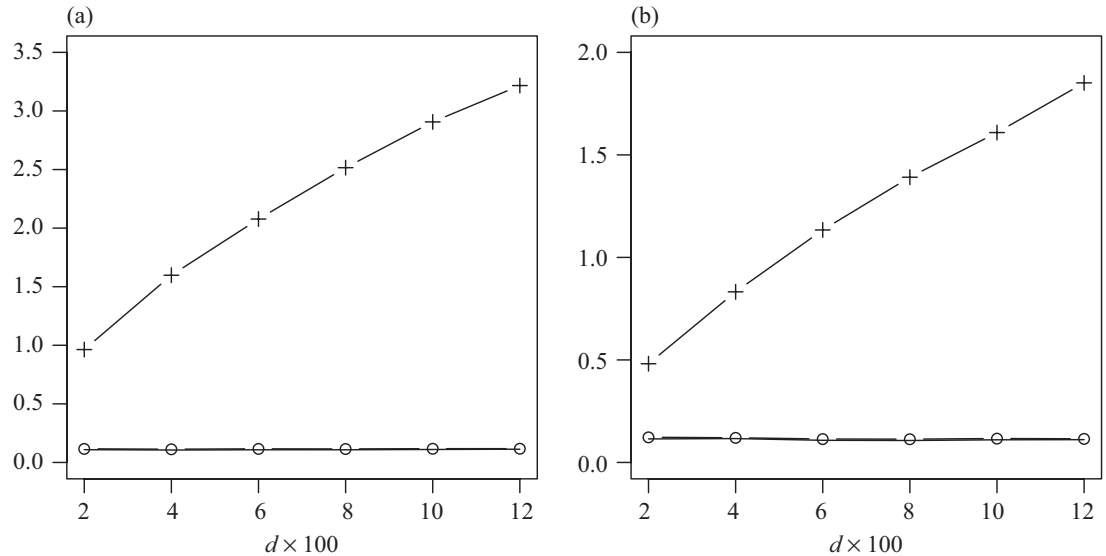


Fig. 1. Empirical averages of the values of $|\hat{\varsigma}/\varsigma - 1|$ for $\hat{\varsigma}_{\mathcal{A}}$ (circles), $\hat{\varsigma}_o$ (solid line) and $\hat{\varsigma}_S$ (crosses) with (a) $\vartheta = 0.3$ and (b) $\vartheta = 0.7$.

*Example* 2. Let $\{\xi_i\}_{i\in\mathbb{N}}$ and $\{\xi_i'\}_{i\in\mathbb{N}}$ be two sequences of independent Gaussian random vectors $N(0, \Sigma)$, where $\Sigma = (\Sigma_{ij})_{n\times n}$ has entries $\Sigma_{ij} = \vartheta^{|i-j|}$ for some $\vartheta \in (0, 1)$. Following Wang et al. (2015), we draw independent and identically distributed innovations $\{V_i\}_{i=1}^n$ from a scale mixture of two independent multivariate Gaussian distributions as follows:

$$V_i = \nu_i \times \xi_i + 3(1 - \nu_i) \times \xi_i' \quad (i = 1, \ldots, n),$$

where $\{\nu_i\}_{i\in\mathbb{N}}$ are independent Bernoulli random variables with $\mathrm{pr}(\nu_i = 1) = 0.9$. A simulation study is conducted in § 5 where we set $\vartheta = 0.3$ and 0.7. We report in Fig. 1 the average values of $|\hat{\varsigma}/\varsigma - 1|$ for $\hat{\varsigma}_{\mathcal{A}}$, $\hat{\varsigma}_o$ and $\hat{\varsigma}_S$, based on 1000 replications with the numerical set-up $(n, p, m) = (100, 20, 10)$ and $d = 200, 400, 800, 1000, 1200$. For both cases of $\vartheta$, $|\hat{\varsigma}_{\mathcal{A}}/\varsigma - 1|$ and $|\hat{\varsigma}_o/\varsigma - 1|$ are very close to 0, while $|\hat{\varsigma}_S/\varsigma - 1|$ is quite large. More precisely, we can derive that $\|E_0(V_1^{\mathrm{T}} V_1)\|_2^2 \asymp (18 + d)\varsigma^2$.

Substituting the ratio-consistent estimator $\hat{\varsigma}_{\mathcal{A}}^2$ into $\mathrm{var}(U_n) = 2|P_\theta|_{\mathbb{F}}^2\varsigma^2$ yields the central limit theorem $U_n/(\hat{\varsigma}_{\mathcal{A}}|P_\theta|_{\mathbb{F}}) \Rightarrow N(0, 2)$ under (17). Then, for $\alpha \in (0, 1)$, an asymptotic $\alpha$-level test is

$$\Phi_Z = \mathbb{I}\left(\frac{U_n}{\hat{\varsigma}_{\mathcal{A}}|P_\theta|_{\mathbb{F}}\sqrt{2}} > z_{1-\alpha}\right), \tag{22}$$

where $z_{1-\alpha}$ is the $(1 - \alpha)$th quantile of the standard normal distribution.

## 5. Simulation study

In this section, we conduct a Monte Carlo simulation study to assess the finite-sample performance of the proposed tests. In model (1), we write $X_i = (1, x_i^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^p$ to include an intercept. Here $x_1, \ldots, x_n \in \mathbb{R}^{p-1}$ are independent and identically distributed $N(0, I_{p-1})$

random vectors. Let $m < p$. For $k \in \{1, \ldots, p-m\}$, all entries of the coefficient vector $B_k$ are independent and identically distributed uniform random variables in the interval $(1, 2)$. After those $B_k$ are generated, we keep their values throughout the simulation. Our goal is to identify the zero $B_k$ entries by testing

$$H_0 : B_{p-m+1} = B_{p-m+2} = \cdots = B_p = 0.$$

In our simulation, we set $(p, m) = (20, 10)$, $n = 100, 200$ and $d = 400, 800, 1200$. We consider two different designs of the innovations $(V_i)$: the one introduced in Example 2 and the one in Example 3 below. In both examples, the parameter $\vartheta$ is set to 0.3 and 0.7.

*Example* 3. Let $\{\xi_{ij}\}_{i,j\in\mathbb{N}}$ be independent and identically distributed random variables with $E(\xi_{11}) = 0$ and $\mathrm{var}(\xi_{11}) = 1$. In particular, we consider two cases for $(\xi_{ij})$: drawn from the standardized $t_5$ distribution and from the standardized $\chi^2_5$ distribution. For some $\vartheta \in (0, 1)$, we generate

$$V_i = (1-\vartheta)^{1/2} \times \xi_i + \vartheta^{1/2} \times (\xi_{i0}, \xi_{i0}, \ldots, \xi_{i0})^{\mathrm{T}} \quad (i \in \mathbb{N}).$$

We apply a Gaussian multiplier bootstrap approach to implement our proposed test. The procedure is as follows.

*Step* 1. Compute the residual matrix $\hat{V} = (\hat{V}_1, \ldots, \hat{V}_n)^{\mathrm{T}} = \bar{P}_1 Y$. Generate independent and identically distributed $N(0, 1)$ random variables $\{\omega_{ij}\}_{i,j\in\mathbb{N}}$ and compute the bootstrap residuals $V^\star = (V_1^\star, \ldots, V_n^\star)^{\mathrm{T}}$, where

$$V_i^\star = \frac{1}{(n-p)^{1/2}} \sum_{j=1}^{n} \omega_{ij} \hat{V}_i \quad (i = 1, \ldots, n).$$

*Step* 2. Use $V^\star$ to compute $\hat{\varsigma}^\star_{\mathcal{A}}$ and the bootstrap test statistic $U_n^\star = \mathrm{tr}(V^{\star\mathrm{T}} P_\theta V^\star)$.

*Step* 3. Repeat the first two steps independently for $\mathcal{B}$ times and collect $U^\star_{nk}$ and $\hat{\varsigma}^\star_{\mathcal{A}k}$ $(k = 1, \ldots, \mathcal{B})$.

*Step* 4. Let $\hat{c}_{1-\alpha}$ be the $(1-\alpha)$th quantile of $\{U^\star_{nk}/(\hat{\varsigma}^\star_{\mathcal{A}k}|P_\theta|_{\mathbb{F}}\sqrt{2})\}_{k=1,\ldots,\mathcal{B}}$. Then our test is

$$\Phi_B = \mathbb{I}\left( \frac{U_n}{\hat{\varsigma}_{\mathcal{A}}|P_\theta|_{\mathbb{F}}\sqrt{2}} > \hat{c}_{1-\alpha} \right),$$

and we reject the null hypothesis whenever $\Phi_B = 1$.

Similar to $\mathcal{G}_n$, $U_n^\star$ is a quadratic functional of independent and identically distributed Gaussian random vectors conditional on $\{X, Y\}$, and is distributed as a linear combination of independent chi-squared random variables. To justify the validity of the proposed Gaussian multiplier bootstrap approach, it suffices to bound the distance between the distribution functions of these two quadratic functionals, which can be done by verifying the normalized consistency (Xu et al., 2015) of the corresponding covariance matrix. However, this can be highly nontrivial in the high-dimensional setting and is beyond the scope of the present paper, so we leave it for future work.

Table 1. *Empirical sizes for Example 2 with* $\alpha = 0.05$

| | | $\theta = 0.3$ | | | $\theta = 0.7$ | | |
|---|---|---|---|---|---|---|---|
| $n$ | $d$ | CLT | GMB | SK | CLT | GMB | SK |
| 100 | 400 | 0.057 | 0.047 | 0.041 | 0.059 | 0.051 | 0.036 |
| | 800 | 0.049 | 0.045 | 0.033 | 0.063 | 0.056 | 0.026 |
| | 1200 | 0.062 | 0.055 | 0.021 | 0.048 | 0.045 | 0.028 |
| 200 | 400 | 0.056 | 0.052 | 0.042 | 0.052 | 0.047 | 0.037 |
| | 800 | 0.052 | 0.049 | 0.037 | 0.053 | 0.050 | 0.033 |
| | 1200 | 0.045 | 0.044 | 0.029 | 0.050 | 0.046 | 0.035 |

CLT, test based on the central limit theorem; GMB, Gaussian multiplier bootstrap approach; SK, the test of Srivastava & Kubokawa (2013).

Table 2. *Empirical sizes for Example 3 with* $\alpha = 0.05$

| | | | $t_5$ | | | $\chi^2_5$ | | |
|---|---|---|---|---|---|---|---|---|
| $\theta$ | $n$ | $d$ | CLT | GMB | SK | CLT | GMB | SK |
| 0.3 | 100 | 400 | 0.068 | 0.058 | 0.023 | 0.083 | 0.065 | 0.036 |
| | | 800 | 0.082 | 0.066 | 0.023 | 0.074 | 0.058 | 0.016 |
| | | 1200 | 0.082 | 0.068 | 0.015 | 0.067 | 0.053 | 0.011 |
| | 200 | 400 | 0.073 | 0.059 | 0.022 | 0.067 | 0.054 | 0.018 |
| | | 800 | 0.071 | 0.057 | 0.012 | 0.074 | 0.058 | 0.014 |
| | | 1200 | 0.076 | 0.059 | 0.011 | 0.077 | 0.058 | 0.011 |
| 0.7 | 100 | 400 | 0.074 | 0.055 | 0.002 | 0.082 | 0.062 | 0.002 |
| | | 800 | 0.084 | 0.066 | 0.001 | 0.085 | 0.071 | 0.000 |
| | | 1200 | 0.073 | 0.057 | 0.000 | 0.076 | 0.062 | 0.001 |
| | 200 | 400 | 0.083 | 0.067 | 0.001 | 0.080 | 0.064 | 0.000 |
| | | 800 | 0.068 | 0.050 | 0.000 | 0.075 | 0.062 | 0.000 |
| | | 1200 | 0.070 | 0.051 | 0.001 | 0.074 | 0.056 | 0.000 |

In our simulation, we set the bootstrap size $\mathcal{B}$ to 1000. For comparison, we also perform the test suggested in (22) based on the central limit theorem and the one proposed by Srivastava & Kubokawa (2013). For each test, we report the empirical size based on 2000 replications, as displayed in Tables 1 and 2. The results suggest that our proposed test using the bootstrap procedure provides the best size accuracy in general, as the empirical sizes are close to the nominal level $\alpha$.

For Example 2, the tests using the central limit theorem and our Gaussian multiplier bootstrap method both have better performance than the test of Srivastava & Kubokawa (2013), since the latter is too conservative as $d$ is large. As expected from our theoretical results, the normal approximation can work reasonably well in this design.

For Example 3, the Gaussian multiplier bootstrap method outperforms the other two procedures in size accuracy for all cases. The test of Srivastava & Kubokawa (2013) suffers from size distortion. The test using the central limit theorem inflates the size more than does the Gaussian multiplier method, which can be explained by the fact that condition (18) does not hold and the test based on the central limit theorem fails for $U_n$. More specifically, for both $\theta = 0.3$ and $\theta = 0.7$, elementary calculations show that $\lambda_1(\Sigma)/\varsigma \to 1$. As a result, (17) is violated as $m = 10$; see also the comment at the end of § 2.3 for discussion of the non-normality of $U_n$. To gain insight, in Fig. 2 we display the density plots of $U_n/\{\mathrm{var}(U_n)\}^{1/2}$ for $n = 100$ as well as the density of $N(0, 1)$. As we can see from the plots, the distribution of $U_n/\{\mathrm{var}(U_n)\}^{1/2}$ is skewed to the right for all cases, which explains the inflated sizes of the test based on the central limit theorem.

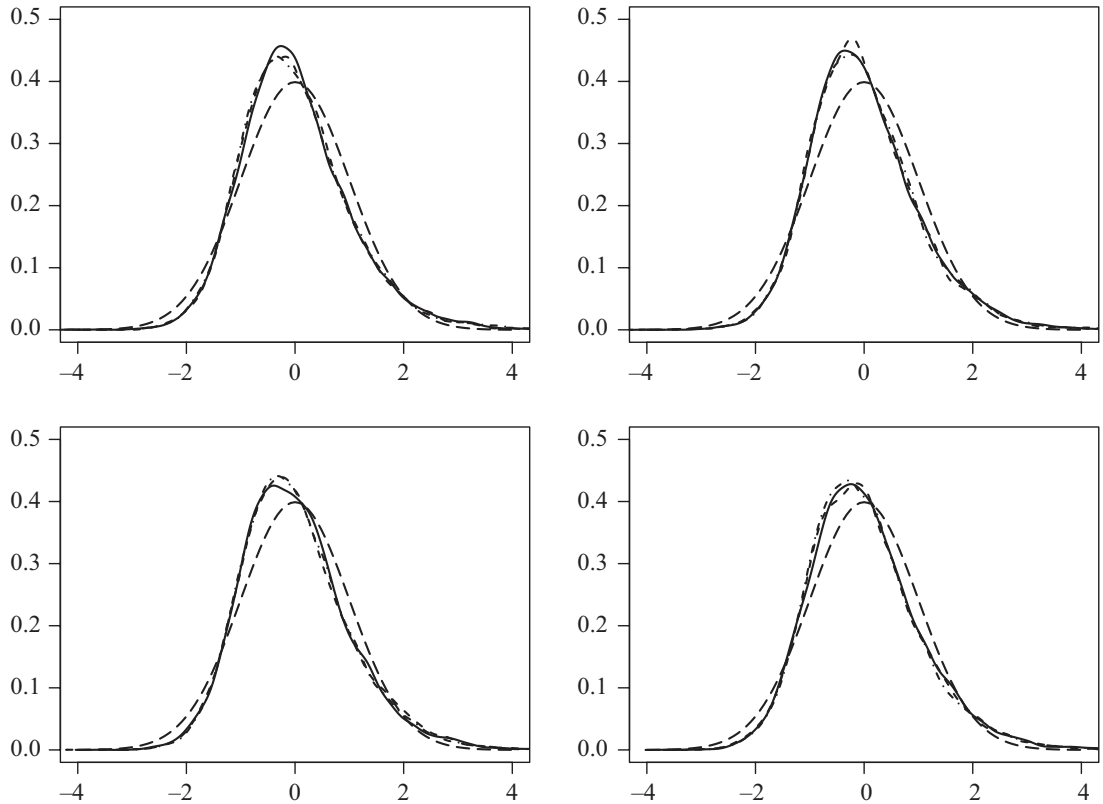


Fig. 2. Density plots of $U_n/\{\mathrm{var}(U_n)\}^{1/2}$ for $d = 400$ (solid), $d = 800$ (short dashed), $d = 1200$ (dash-dotted) and $N(0, 1)$ (long dashed).

More simulation studies on the power comparison for these three tests are reported in the Supplementary Material.

## 6. Data analysis

We apply the proposed method to two datasets. The first dataset comes from a study of the impact of the gut microbiome on the host serum metabolome and insulin sensitivity in non-diabetic Danish adults (Pedersen et al., 2016). It consists of measurements of 1201 metabolites, i.e., 325 serum polar metabolites and 876 serum molecular lipids, on 289 serum samples using mass spectrometry. The cleaned dataset was downloaded from `https://bitbucket.org/hellekp/clinical-micro-meta-integration` (Pedersen et al., 2018). We use this dataset to identify metabolites associated with insulin resistance. Insulin resistance was estimated by the homeostatic model assessment (Pedersen et al., 2016). Body mass index, BMI, is a confounder for this dataset since it is highly correlated with insulin resistance, with a Spearman's $\rho$ of 0.67, and is known to affect the serum metabolome. Two samples without insulin resistance measurements were excluded. For metabolites with zero measurements, the zeros were replaced by half of the minimal nonzero value. Log transformation was performed to make the data more symmetrically distributed before analysis. The $p$-values associated with the three methods are all very close to zero, indicating a strong dependence between metabolites and insulin resistance, see Table 3. We further perform a linear regression analysis on each metabolite using insulin resistance

Table 3. *The p-values of the three methods applied to the metabolomics and microbiome datasets*

| | Metabolomics | | | Microbiome | | |
|---|---|---|---|---|---|---|
| | CLT | GMB | SK | CLT | GMB | SK |
| *p*-value | 0.00 | 0.00 | 0.00 | $9.7 \times 10^{-6}$ | 0.002 | 0.13 |

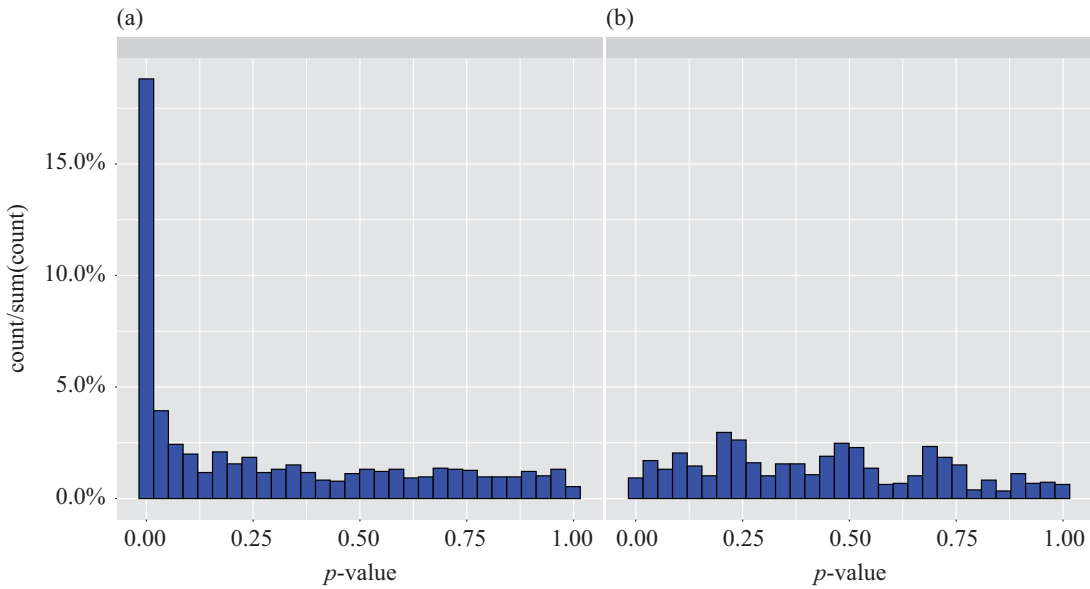


Fig. 3. Histograms of *p*-values for testing the association between each -omics feature and the variable of interest after adjusting for the confounder: (a) metabolomics dataset; (b) microbiome dataset.

and BMI as the covariates. Figure 3(a) presents the histogram of *p*-values on testing the significance of the coefficients associated with insulin resistance. We see a high peak close to zero, which provides strong evidence for the association between metabolites and insulin resistance. We further apply the Holm–Bonferroni procedure to the *p*-values to control the familywise error rate at the 5% level, resulting in 164 discoveries.

The second dataset we consider is from a study of the effects of smoking on the human upper respiratory tract (Charlson et al., 2010). The original dataset contains samples from throat and nose microbiomes and from both sides of the body. Here we focus on the throat microbiome on the left side of the body, which includes data from 60 subjects consisting of 32 nonsmokers and 28 smokers. More precisely, the dataset is presented as a $60 \times 856$ abundance table recording the frequencies of detected operational taxonomic units in the samples using the 16S metagenomics approach, together with a metadata table capturing sample-level information, including smoking status and sex. We transform the abundance of the operational taxonomic units using the centred log-ratio transformation after adding a pseudo-count of 0.5 to the zero counts. Our goal is to test the association of throat microbiomes with smoking status adjusting for sex. The proposed method using either the normal approximation or the bootstrap approximation detects a strong association between throat microbiomes and smoking status, see Table 3. In contrast, the method of Srivastava & Kubokawa (2013) fails to discover such an association.

We further perform an operational-taxonomic-unit-wise linear regression analysis using each operational taxonomic unit, after the centred log-ratio transformation as the response and smoking status and sex as covariates. Figure 3(b) presents the histogram of *p*-values

for testing the association between each operational taxonomic unit and smoking status, after adjusting for sex in each linear regression. Interestingly, adjusting the multiplicity using either the Holm–Bonferroni procedure or the Benjamini–Hochberg procedure at the 5% level gives zero discovery (Zhou et al., 2022). These results suggest that the association between individual operational taxonomic units and smoking status is weak. However, after aggregating the weak effects from all operational taxonomic units, the combined effect is strong enough to be detected by the proposed method.

## Supplementary material

The Supplementary Material contains proofs of (15) and (17) and a power comparison.

## References


Akritas, M. G. & Arnold, S. F. (1994). Fully nonparametric hypotheses for factorial designs I: Multivariate repeated measures designs. *J. Am. Statist. Assoc.* **89**, 336–43.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Hoboken, New Jersey: Wiley.

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

Brunner, E. & Puri, M. L. (2001). Nonparametric methods in factorial designs. *Statist. Papers* **42**, 1–52.

Cai, T. T. & Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli* **19**, 2359–88.

Cai, T. T. & Xia, Y. (2014). High-dimensional sparse MANOVA. *J. Mult. Anal.* **131**, 174–96.

Charlson, E. S., Chen, J., Custers-Allen, R., Bittinger, K., Li, H., Sinha, R., Hwang, J., Bushman, F. D. & Collman, R. G. (2010). Disordered microbial communities in the upper respiratory tract of cigarette smokers. *PLoS One* **5**, e15216.

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.

Chen, X. (2018). Gaussian and bootstrap approximations for high-dimensional U-statistics and their applications. *Ann. Statist.* **46**, 642–78.

Fan, J., Guo, S. & Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Statist. Soc.* B **74**, 37–65.

Götze, F. & Tikhomirov, A. (2002). Asymptotic distribution of quadratic forms and applications. *J. Theor. Prob.* **15**, 423–75.

Götze, F. & Tikhomirov, A. N. (1999). Asymptotic distribution of quadratic forms. *Ann. Prob.* **27**, 1072–98.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B. & Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.* **13**, 723–73.

Gretton, A., Fukumizu, K. & Sriperumbudur, B. K. (2009). Discussion of: Brownian distance covariance. *Ann. Appl. Statist.* **3**, 1285–94.

He, Y., Meng, B., Zeng, Z. & Xu, G. (2021). On the phase transition of Wilks' phenomenon. *Biometrika* **108**, 741–8.

Hu, J., Bai, Z., Wang, C. & Wang, W. (2017). On testing the equality of high dimensional mean vectors with unequal covariance matrices. *Ann. Inst. Statist. Math.* **69**, 365–87.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Kruskal, W. H. & Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Statist. Assoc.* **47**, 583–621.

Lahiri, S. N. (1992). Bootstrapping $M$-estimators of a multiple linear regression parameter. *Ann. Statist.* **20**, 1548–70.

Li, H., Hu, J., Bai, Z., Yin, Y. & Zou, K. (2017). Test on the linear combinations of mean vectors in high-dimensional data. *Test* **26**, 188–208.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382–400.

Navidi, W. (1989). Edgeworth expansions for bootstrapping regression models. *Ann. Statist.* **17**, 1472–8.

Pedersen, H. K., Forslund, S. K., Gudmundsdottir, V., Petersen, A. Ø., Hildebrand, F., Hyötyläinen, T., Nielsen, T., Hansen, T., Bork, P., Ehrlich, S. D. et al. (2018). A computational framework to integrate high-throughput '-omics' datasets for the identification of potential mechanistic links. *Nature Protocols* **13**, 2781–800.

Pedersen, H. K., Gudmundsdottir, V., Nielsen, H. B., Hyotylainen, T., Nielsen, T., Jensen, B. A., Forslund, K., Hildebrand, F., Prifti, E., Falony, G. et al. (2016). Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature* **535**, 376–81.

Portnoy, S. (1985). Asymptotic behavior of $M$ estimators of $p$ regression parameters when $p^2/n$ is large. II. Normal approximation. *Ann. Statist.* **13**, 1403–17.

Rizzo, M. L. & Székely, G. J. (2010). DISCO analysis: A nonparametric extension of analysis of variance. *Ann. Appl. Statist.* **4**, 1034–55.

Schott, J. R. (2007). Some high-dimensional tests for a one-way MANOVA. *J. Mult. Anal.* **98**, 1825–39.

Shao, J. (1988). On resampling methods for variance and bias estimation in linear models. *Ann. Statist.* **16**, 986–1008.

Shao, J. & Wu, C.-F. J. (1987). Heteroscedasticity-robustness of jackknife variance estimators in linear models. *Ann. Statist.* **15**, 1563–79.

Srivastava, M. S. & Fujikoshi, Y. (2006). Multivariate analysis of variance with fewer observations than the dimension. *J. Mult. Anal.* **97**, 1927–40.

Srivastava, M. S., Katayama, S. & Kano, Y. (2013). A two sample test in high dimensional data. *J. Mult. Anal.* **114**, 349–58.

Srivastava, M. S. & Kubokawa, T. (2013). Tests for multivariate analysis of variance in high dimension under non-normality. *J. Mult. Anal.* **115**, 204–16.

Székely, G. J., Rizzo, M. L. & Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35**, 2769–94.

Thas, O. (2010). *Comparing Distributions*. New York: Springer.

Wang, L., Peng, B. & Li, R. (2015). A high-dimensional nonparametric multivariate test for mean vector. *J. Am. Statist. Assoc.* **110**, 1658–69.

Wessel, J. & Schork, N. J. (2006). Generalized genomic distance–based regression methodology for multilocus association analysis. *Am. J. Hum. Genet.* **79**, 792–806.

Wu, C.-F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–350. With discussion and a rejoinder by the author.

Xu, M., Zhang, D. & Wu, W. B. (2015). $L^2$ asymptotics for high-dimensional data. *arXiv:* 1405.7244v3.

Xu, M., Zhang, D. & Wu, W. B. (2019). Pearson's chi-squared statistics: Approximation theory and beyond. *Biometrika* **106**, 716–23.

Yao, S., Zhang, X. & Shao, X. (2018). Testing mutual independence in high dimension via distance covariance. *J. R. Statist. Soc.* B **80**, 455–80.

Zapala, M. A. & Schork, N. J. (2006). Multivariate regression analysis of distance matrices for testing associations between gene expression patterns and related variables. *Proc. Nat. Acad. Sci.* **103**, 19430–5.

Zapala, M. A. & Schork, N. J. (2012). Statistical properties of multivariate distance matrix regression for high-dimensional data analysis. *Frontiers Genet.* **3**, 190.

Zhang, J.-T., Guo, J. & Zhou, B. (2017). Linear hypothesis testing in high-dimensional one-way MANOVA. *J. Mult. Anal.* **155**, 200–16.

Zhang, X., Yao, S. & Shao, X. (2018). Conditional mean and quantile dependence testing in high dimension. *Ann. Statist.* **46**, 219–46.

Zhou, B., Guo, J. & Zhang, J.-T. (2017). High-dimensional general linear hypothesis testing under heteroscedasticity. *J. Statist. Plan. Infer.* **188**, 36–54.

Zhou, H., He, K., Chen, J. & Zhang, X. (2022). LinDA: Linear models for differential abundance analysis of microbiome compositional data. *arXiv:* 2104.00242v3.

[*Received on* 14 *March* 2022. *Editorial decision on* 3 *November* 2022]