Statistics

A Journal of Theoretical and Applied Statistics

ISSN: 0233-1888 (Print) 1029-4910 (Online) Journal homepage: http://www.tandfonline.com/loi/gsta20

Testing linear hypotheses in high-dimensional regressions

Zhidong Bai , Dandan Jiang , Jian-feng Yao & Shurong Zheng

To cite this article: Zhidong Bai , Dandan Jiang , Jian-feng Yao & Shurong Zheng (2013) Testing linear hypotheses in high-dimensional regressions, Statistics, 47:6, 1207-1223, DOI: 10.1080/02331888.2012.708031

To link to this article: http://dx.doi.org/10.1080/02331888.2012.708031

Statistics, 2013
Vol. 47, No. 6, 1207–1223, http://dx.doi.org/10.1080/02331888.2012.708031

Testing linear hypotheses in high-dimensional regressions

Zhidong Bai[a], Dandan Jiang[b], Jian-feng Yao[c] and Shurong Zheng[a]*

[a] *KLASMOE and School of Mathematics and Statistics, Northeast Normal University, 5268 People's Road, 130024 Changchun, People's Republic of China;* [b] *Institute of Mathematics, Jilin University, 2699 Qianjin Street, 130021 Changchun, People's Republic of China;* [c] *Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam, Hong Kong*

(*Received 11 October 2010; final version received 23 February 2012*)

For a multivariate linear model, Wilk's likelihood ratio test (LRT) constitutes one of the cornerstone tools. However, the computation of its quantiles under the null or the alternative hypothesis requires complex analytic approximations, and more importantly, these distributional approximations are feasible only for moderate dimension of the dependent variable, say $p \leq 20$. On the other hand, assuming that the data dimension p as well as the number q of regression variables are fixed while the sample size n grows, several asymptotic approximations are proposed in the literature for Wilk's $\mathbf{\Lambda}$ including the widely used chi-square approximation. In this paper, we consider necessary modifications to Wilk's test in a high-dimensional context, specifically assuming a high data dimension p and a large sample size n. Based on recent random matrix theory, the correction we propose to Wilk's test is asymptotically Gaussian under the null hypothesis and simulations demonstrate that the corrected LRT has very satisfactory size and power, surely in the large p and large n context, but also for moderately large data dimensions such as $p = 30$ or $p = 50$. As a byproduct, we give a reason explaining why the standard chi-square approximation fails for high-dimensional data. We also introduce a new procedure for the classical multiple sample significance test in multivariate analysis of variance which is valid for high-dimensional data.

Keywords: high-dimensional data; multivariate regression; multivariate analysis of variance; Wilk's test; multiple sample significance test; random matrices

AMS 2000 Subject Classifications: Primary 62H15; secondary 62H10

1. Introduction

In more and more burgeoning science and technology fields and with the help of rapid development in information technology, a huge amount of data is collected where the number of variables is usually large. However, most of traditional statistical tools deeply depend on the assumption of a large sample size n compared with the number of variables p (data dimension). For high-dimensional data analysis, inevitably, these classical tools become inefficient or, even worse, inconsistent. For decades, statisticians devoted special efforts to seek for better approaches in such high-dimensional data case see Bai [1]. For the two-sample significance test problem in high dimensions, as early as in 1958, Dempster [2] proposed a so-called *non-exact test* (NET) as a remedy to the failure of Hotelling's T^2-test. A rigorous analysis of this NET arises much later in

*Corresponding author. Email: zhengsr@nenu.edu.cn

the paper by Bai and Saranadasa [3] using modern random matrix theory (RMT). These authors have found necessary correction for the T^2-test to cope with high-dimensional effects.

Recent work in high-dimensional statistics include Ledoit and Wolf [4], Srivastava [5] and Schott [6]. These authors propose several procedures in the high-dimensional setting for testing that (i) a covariance matrix is an identity matrix, proportional to an identity matrix (sphericity) and is a diagonal matrix or (ii) several covariance matrices are equal. These procedures have the following common feature: their construction involves some well-chosen distance function between the null and the alternative hypotheses and rely on the first two spectral moments, namely the statistics $\operatorname{tr} S_k$ and $\operatorname{tr} S_k^2$ from sample covariance matrices S_k. In a recent work by Bai *et al.* [7], we have considered likelihood-based tests about such high-dimensional covariance matrices where the failure of the classical likelihood ratio test (LRT) is explained using RMT. Necessary corrections to these LRTs are then introduced to achieve consistency.

This paper pursue the investigation of similar questions but for a multivariate regression model with high-dimensional data, that is, the dimensions of the dependent variable as well as the number of the regression variables are large compared with the sample size. More precisely, let a pth-dimensional regression model

$$\mathbf{x}_i = \mathbf{B}\mathbf{z}_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where ε_i is a sequence of i.i.d. zero-mean Gaussian noise $\mathcal{N}_p(0, \mathbf{\Sigma})$ with covariance matrices $\mathbf{\Sigma}$, $\mathbf{B}$ a $p \times q$ matrix of regression coefficients and $\mathbf{z}_i$ a sequence of known regression variables of dimension q. To simplify the presentation, we always assume that $n \geq p + q$ and that the rank of $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_n)$ equals q.

Let us define a block decomposition $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$ with q_1 and q_2 columns, respectively ($q = q_1 + q_2$). A general linear hypothesis is defined as

$$H_0 : \ \mathbf{B}_1 = \mathbf{B}_1^*, \tag{2}$$

where $\mathbf{B}_1^*$ is a given matrix. A well-studied example is the special case $\mathbf{B}_1^* = 0$ yielding a significance test for the first q_1 regression variables.

In the general case and under the alternative, the maximum-likelihood estimators of ($\mathbf{B}$, $\mathbf{\Sigma}$) are

$$\hat{\mathbf{B}} = \left(\sum_{i=1}^n \mathbf{x}_i\mathbf{z}_i'\right)\left(\sum_{i=1}^n \mathbf{z}_i\mathbf{z}_i'\right)^{-1} \tag{3}$$

and

$$\hat{\mathbf{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{x}_i - \hat{\mathbf{B}}\mathbf{z}_i)(\mathbf{x}_i - \hat{\mathbf{B}}\mathbf{z}_i)'. \tag{4}$$

The corresponding likelihood maximum equals

$$\mathcal{L}_1 = (2\pi)^{-(1/2)pn}|\hat{\mathbf{\Sigma}}|^{-(1/2)n}\,\mathrm{e}^{-(1/2)pn}.$$

On the other hand, under the null hypothesis, by using a partition $\mathbf{z}_i' = (\mathbf{z}_{i,1}', \mathbf{z}_{i,2}')$ on q_1 and q_2 variables, respectively, the maximum-likelihood estimators of ($\mathbf{B}_2$, $\mathbf{\Sigma}$) are

$$\hat{\mathbf{B}}_{20} = \left(\sum_{i=1}^n \mathbf{y}_i\mathbf{z}_{i,2}'\right)\left(\sum_{i=1}^n \mathbf{z}_{i,2}\mathbf{z}_{i,2}'\right)^{-1} \tag{5}$$

and

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n}\left(\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{B}}_{20}\mathbf{z}_{i,2})(\mathbf{y}_i - \hat{\mathbf{B}}_{20}\mathbf{z}_{i,2})'\right), \tag{6}$$

where $\mathbf{y}_i = \mathbf{x}_i - \mathbf{B}_1^* \mathbf{z}_{i,1}$. The associated likelihood maximum equals

$$\mathcal{L}_0 = (2\pi)^{-(1/2)pn} |\hat{\mathbf{\Sigma}}_0|^{-(1/2)n} \, \mathrm{e}^{-(1/2)pn}. \tag{7}$$

It follows that the likelihood ratio statistic for the test (2) equals

$$\frac{\mathcal{L}_0}{\mathcal{L}_1} = (\mathbf{\Lambda}_n)^{n/2}, \quad \mathbf{\Lambda}_n = \frac{|\hat{\mathbf{\Sigma}}|}{|\hat{\mathbf{\Sigma}}_0|}, \tag{8}$$

where $\mathbf{\Lambda}_n$ is the celebrated Wilk's $\mathbf{\Lambda}$ [8–10].

Let us define a similar block decomposition for the sum

$$\sum_{i=1}^{n} \mathbf{z}_i \mathbf{z}_i' = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

and the matrix

$$\mathbf{A}_{11:2} = \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}.$$

After some algebraic manipulations, we get [11, p. 302]

$$\mathbf{\Lambda}_n = \left| \mathbf{I} + \frac{q_1}{n-q} \mathbf{F} \right|^{-1}, \tag{9}$$

where

$$\mathbf{F} = \frac{n-q}{q_1} (n\hat{\mathbf{\Sigma}})^{-1} (\hat{\mathbf{B}}_1 - \mathbf{B}_1^*) \mathbf{A}_{11:2} (\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)', \tag{10}$$

and $\hat{\mathbf{B}}_1$ is a $p \times q_1$ matrix made of the first q_1 columns of $\hat{\mathbf{B}}$.

It is known that $n\hat{\mathbf{\Sigma}} \sim W_p(\mathbf{\Sigma}, n-q)$, a Wishart distribution. Moreover, under H_0,

$$(\hat{\mathbf{B}}_1 - \mathbf{B}_1^*) \mathbf{A}_{11:2} (\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)' \sim W_p(\mathbf{\Sigma}, q_1),$$

and this statistic is independent of $\hat{\mathbf{\Sigma}}$. Therefore, H_0 will be rejected if $\mathbf{\Lambda}_n < \lambda_0$ for some critical value λ_0, or equivalently, when the matrix $\mathbf{F}$ has some large enough eigenvalues.

Under the Gaussian assumptions made here, the exact distribution of $\mathbf{\Lambda}_n$ is known under the null hypothesis. However, in practice, it is usually a difficult task to compute the critical value λ_0 even for moderately large p and q. For example, Mathai [12] used complex analytical approximations and established tables for critical values with p and q smaller than 12.

On the other hand, in a large n asymptotic scheme, one assumes p and q are fixed and then the null distribution of $-n \log \mathbf{\Lambda}_n$ is approximated by a $\chi^2_{pq_1}$. Note that for this chi-squared approximation, one generally uses a rescaled LRT statistic

$$U_n = -k \log \mathbf{\Lambda}_n, \quad k = n - q - \tfrac{1}{2}(p - q_1 + 1). \tag{11}$$

This correction is known as Bartlett–Box correction (hereafter BBC) due to [13] and it is much less biased than the classical LRT $-n \log \mathbf{\Lambda}_n$ (see Section 3.3 for a detailed comparison).

However, for high-dimensional data where the dimensions p and q_1 are large compared with the sample size n, unfortunately the above $\chi^2_{pq_1}$ approximation becomes useless. As an example, even for moderate p, q and n with $y_n = p/(n-q)$ close to 1, the celebrated Marčenko–Pastur theorem tell us that the eigenvalues of $\hat{\mathbf{\Sigma}}$ tend to fill the whole interval $[(1-\sqrt{y_n})^2, (1+\sqrt{y_n})^2]$. Hence, a non-negligible proportion of these eigenvalues are close to zero. Consequently, any statistic based on the inverse $\hat{\mathbf{\Sigma}}^{-1}$ like $\mathbf{\Lambda}_n$ becomes unstable and non-robust.

In Section 3, by using modern RMT, we introduce a correction to Wilk's $\mathbf{\Lambda}$ to cope with the mentioned high-dimensional effects. The corrected LRT (CLRT) is asymptotically Gaussian and we will see that it has very satisfactory size and power, surely for the large p, q and n context, but also for moderate data dimensions such as $p = 30$ or $p = 50$.

Moreover, to assess the power of the CLRT, we examine two additional tests based on statistics of least-squares type as suggested in [3]. A quite intensive simulation experiment is then conducted to compare these different procedures for testing Equation (2).

Next in Section 4, we consider the classical multiple sample significance test problem but with high-dimensional data. As it is well known, this problem can be embedded into a special instance of the general linear hypothesis (2). Therefore, by an application of general results of Section 3, we obtain a valid LRT after necessary corrections.

All the proofs and technical derivations are postponed to Section 5.

2. A central limit theorem (CLT) for linear statistics of random Fisher matrices

We first recall a fundamental result from RMT for linear statistics of the so-called random Fisher matrices which will be used below. For any $p \times p$ square matrix M with real eigenvalues (λ_i^M), F^M denotes the empirical spectral distribution (ESD) of M, that is,

$$F^M(x) = \frac{1}{p}\sum_{i=1}^{p} \mathbf{1}_{\lambda_i^M \le x}, \quad x \in \mathbb{R}.$$

We will consider random matrices (M_n) whose ESD F^{M_n} converges, in a sense to be precise and when $n \to \infty$, to a limiting spectral distribution (LSD) F. Assume we have to estimate some parameter of F, say $\theta = \int f(x)\,\mathrm{d}F(x)$ for some function f; it is natural to use the empirical estimator

$$\hat{\theta}_n = \int f(x)\,\mathrm{d}F^{M_n}(x) = \frac{1}{p}\sum_{i=1}^{p} f(\lambda_i^{M_n}),$$

which is a so-called linear spectral statistic (LSS) of the random matrices M_n.

Let $\{\xi_{ki} \in \mathbb{C},\ i,k = 1,2,\ldots\}$ and $\{\eta_{kj} \in \mathbb{C}, j,k = 1,2,\ldots\}$ be two independent double arrays of i.i.d. complex variables with mean 0 and variance 1. Write $\xi_{\cdot i} = (\xi_{1i}, \xi_{2i}, \ldots, \xi_{pi})^{\mathrm{T}}$ and $\eta_{\cdot j} = (\eta_{1j}, \eta_{2j}, \ldots, \eta_{pj})^{\mathrm{T}}$. Also, for any positive integers n_1, n_2, the vectors $(\xi_{\cdot 1}, \ldots, \xi_{\cdot n_1})$ and $(\eta_{\cdot 1}, \ldots, \eta_{\cdot n_2})$ can be thought as independent samples of size n_1 and n_2, respectively, from some p-dimensional distributions. Let S_1 and S_2 be the associated sample covariance matrices, that is,

$$S_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \xi_{\cdot i}\xi_{\cdot i}^* \quad \text{and} \quad S_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} \eta_{\cdot j}\eta_{\cdot j}^*.$$

Then, the following so-called *F-matrix* generalizes the classical Fisher statistic to the present p-dimensional case,

$$V_n = S_1 S_2^{-1}, \tag{12}$$

where we assume that $n_2 > p$. Here, we use the notation $n = (n_1, n_2)$.

Let us also assume that

$$y_{n_1} = \frac{p}{n_1} \longrightarrow y_1 \in (0,1), \quad y_{n_2} = \frac{p}{n_2} \longrightarrow y_2 \in (0,1). \tag{13}$$

Under suitable moment conditions, the ESD F^{V_n} of V_n has an LSD F_{y_1,y_2} with the following density function [14, p. 72],

$$\ell(x) = \begin{cases} \dfrac{(1-y_2)\sqrt{(b-x)(x-a)}}{2\pi x(y_1+y_2x)}, & a \le x \le b, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where

$$a = \left(\frac{1-h}{1-y_2}\right)^2, \quad b = \left(\frac{1+h}{1-y_2}\right)^2, \quad h = \sqrt{y_1 + y_2 - y_1y_2}.$$

Let $\mathcal{U}$ be an open subset of the complex plane which contains the interval $[a,b]$ and $\mathcal{A}$ be the set of analytic functions $f : \mathcal{U} \mapsto \mathbb{C}$. Define the empirical process $G_n := \{G_n(f)\}$ indexed by $\mathcal{A}$

$$G_n(f) = p \cdot \int_{-\infty}^{+\infty} f(x)[F^{V_n} - F_{y_{n_1},y_{n_2}}](\mathrm{d}x), \quad f \in \mathcal{A}. \tag{15}$$

Here, $F_{y_{n_1},y_{n_2}}$ is the distribution in Equation (14) with indexes y_{n_k} (instead of y_k), $k = 1,2$.

Recently, Zheng [15] has established a general CLT for LSS of the large-dimensional F-matrix. The following theorem is a simplified one quoted from it. Throughout the paper, $\oint$ denotes a contour integral along a given contour.

THEOREM 2.1 *Let $f_1,\dots,f_k \in \mathcal{A}$ and assume the following:*

For each p, (ξ_{ij_1}) and (η_{ij_2}) variables are i.i.d., $1 \le i \le p$, $1 \le j_1 \le n_1$, $1 \le j_2 \le n_2$. $E\xi_{11} = E\eta_{11} = 0$, $E|\xi_{11}|^4 = E|\eta_{11}|^4 < \infty$, $y_{n_1} = p/n_1 \to y_1 \in (0,1)$, $y_{n_2} = p/n_2 \to y_2 \in (0,1)$.

(i) *Real case. Assume moreover (ξ_{ij}) and (η_{ij}) are real, $E|\xi_{11}|^2 = E|\eta_{11}|^2 = 1$, then the random vector $(G_n(f_1),\dots,G_n(f_k))$ weakly converges to a k-dimensional Gaussian vector with the mean vector*

$$\begin{aligned} m(f_j) &= \lim_{r\to 1_+} (16) + (17) + (18) \\ &= \frac{1}{4\pi \mathrm{i}} \oint_{|\zeta|=1} f_j(z(\zeta)) \left[\frac{1}{\zeta - 1/r} + \frac{1}{\zeta + 1/r} - \frac{2}{\zeta + y_2/h}\right] \mathrm{d}\zeta & (16) \\ &\quad + \frac{\beta \cdot y_1(1-y_2)^2}{2\pi \mathrm{i} \cdot h^2} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{1}{(\zeta + y_2/h)^3} \mathrm{d}\zeta & (17) \\ &\quad + \frac{\beta \cdot y_2(1-y_2)}{2\pi \mathrm{i} \cdot h} \oint_{|\zeta|=1} f_j(z(\zeta)) \frac{\zeta + 1/h}{(\zeta + y_2/h)^3} \mathrm{d}\zeta, \quad j = 1,\dots,k, & (18) \end{aligned}$$

where $z(\zeta) = (1-y_2)^{-2}[1+h^2+2h\mathcal{R}(\zeta)]$, $h = \sqrt{y_1+y_2-y_1y_2}$, $\beta = E|\xi_{11}|^4 - 3$ *and the covariance function*

$$\upsilon(f_j,f_\ell) = \lim_{r\to 1_+} (19) + (20)$$

$$= -\frac{1}{2\pi^2}\oint_{|\zeta_2|=1}\oint_{|\zeta_1|=1}\frac{f_j(z(\zeta_1))f_\ell(z(\zeta_2))}{(\zeta_1 - r\zeta_2)^2}\,\mathrm{d}\zeta_1\,\mathrm{d}\zeta_2, \tag{19}$$

$$-\frac{\beta\cdot(y_1+y_2)(1-y_2)^2}{4\pi^2h^2}\oint_{|\zeta_1|=1}\frac{f_j(z(\zeta_1))}{(\zeta_1+y_2/h)^2}\,\mathrm{d}\zeta_1\oint_{|\zeta_2|=1}\frac{f_\ell(z(\zeta_2))}{(\zeta_2+y_2/h)^2}\,\mathrm{d}\zeta_2$$

$$j,\ell \in \{1,\ldots,k\}. \tag{20}$$

(ii) *Complex case. Assume moreover* (ξ_{ij}) *and* (η_{ij}) *are complex,* $E(\xi_{11}^2) = E(\eta_{11}^2) = 0$, *then the conclusion of* (*i*) *also holds, except the means are* (17) + (18) *and the covariance function is* $\frac{1}{2}\lim_{r\to 1_+}$ (19) + (20) *with* $\beta = E|\xi_{11}|^4 - 2$.

We should point out that Zheng's CLT for F-matrices covers more general situations than those cited in Theorem 2.1. In particular, the fourth moments $E|\xi_{11}|^4$ and $E|\eta_{11}|^4$ can be different.

The following lemma will be used in Section 3 for an application of Theorem 2.1 (see Equations (26) and (27)). For a proof, see [7].

LEMMA 2.1 *For the function* $f(x) = \log(a+bx)$, $x\in\mathbb{R}$, $a,b>0$, *let* (c,d) *be the unique solution to the equations*

$$\begin{cases} c^2+d^2 = a+b\dfrac{(1+h^2)}{(1-y_2)^2}, \\ cd = \dfrac{bh}{(1-y_2)^2}, \\ 0<d<c. \end{cases} \tag{21}$$

Analogously, let γ *and* η *be the constants similar to* (c,d) *but for the function* $g(x) = \log(\alpha+\beta x)$, $\alpha>0$, $\beta>0$. *Then, the mean and covariance functions in Equations* (16) *and* (19) *equal*

$$m(f) = \frac{1}{2}\log\frac{(c^2-d^2)h^2}{(ch-y_2d)^2},$$
$$\upsilon(f,g) = 2\log\frac{c\gamma}{c\gamma - d\eta}.$$

3. Testing a general linear hypothesis in high-dimensional regressions

3.1. *A CLRT*

The construction of a correct scaling for the LRT statistic $\mathbf{\Lambda}_n$ of the test (2) will rely on the CLT 2.1. Recall that

$$\mathbf{\Lambda}_n = \left|\mathbf{I} + \frac{q_1}{n-q}\mathbf{F}\right|^{-1}, \quad \mathbf{F} = \frac{n-q}{q_1}(n\hat{\mathbf{\Sigma}})^{-1}(\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)\mathbf{A}_{11:2}(\hat{\mathbf{B}}_1-\mathbf{B}_1^*)'.$$

Under H_0, we have

$$n\hat{\mathbf{\Sigma}} \sim W_p(\mathbf{\Sigma}, n-q), \quad (\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)\mathbf{A}_{11:2}(\hat{\mathbf{B}}_1-\mathbf{B}_1^*)' \sim W_p(\mathbf{\Sigma}, q_1),$$

and they are independent. Consequently, $\mathbf{F}$ is exactly distributed as the F-matrix V_n defined in Equation (12), where in addition all the variables are Gaussian.

Our correction to the LRT statistic $\mathbf{\Lambda}_n$ is given in the following theorem.

THEOREM 3.1 *For the general linear hypothesis* (2) *in the regression model* (1)*, let* $\mathbf{\Lambda}_n$ *be Wilk's LRT statistic given in Equation* (9)*. Define also the function*

$$f(x) = \log\left(1 + \frac{y_{n_2}}{y_{n_1}}x\right)$$

and assume that

$$p \to \infty, \quad q_1 \to \infty, \quad n-q \to \infty, \quad y_{n_1} = \frac{p}{q_1} \to y_1 \in (0,1), \quad y_{n_2} = \frac{p}{n-q} \to y_2 \in (0,1). \tag{22}$$

Then, under the null hypothesis,

$$T_n = \upsilon(f)^{-1/2}[-\log \mathbf{\Lambda}_n - p \cdot F_{y_{n_1},y_{n_2}}(f) - m(f)] \Rightarrow \mathcal{N}(0,1), \tag{23}$$

where $m(f), \upsilon(f)$ *and* $F_{y_{n_1},y_{n_2}}(f)$ *are defined in Equations* (26)*,* (27) *and* (29)*, respectively.*

Before giving a proof, it is worth mentioning that at a first look, the asymptotic framework depicted in Equation (22) seems complicated. Indeed, this is a common set-up in RMT and simply requires that the degrees of freedom of the underlying Wishart matrices grow to infinity in a proportional way with the sample size.

Proof Since $\mathbf{F}$ can be represented by a Gaussian V_n, we have

$$\begin{aligned} -\log \mathbf{\Lambda}_n &= \log\left|I + \frac{q_1}{n-q}V_n\right| \\ &= \sum_{i=1}^{p} \log\left(1 + \frac{q_1}{n-q}\lambda_i^{V_n}\right) \\ &= p \cdot \int \log\left(1 + \frac{q_1}{n-q}x\right) \mathrm{d}F^{V_n}(x). \end{aligned}$$

Define $f(x) = \log(1 + (q_1/(n-q))x)$, by $y_{n_1} = p/q_1, y_{n_2} = p/(n-q)$, also it can be written as

$$f(x) = \log\left(1 + \frac{y_{n_2}}{y_{n_1}}x\right). \tag{24}$$

From

$$\begin{aligned} -\log \mathbf{\Lambda}_n &= p \cdot \int f(x)\,\mathrm{d}F^{V_n}(x) \\ &= p \cdot \int f(x)\,\mathrm{d}(F^{V_n}(x) - F_{y_{n_1},y_{n_2}}(x)) + p \cdot F_{y_{n_1},y_{n_2}}(f), \end{aligned}$$

where $F_{y_{n_1},y_{n_2}}(f) = \int f(x)\,\mathrm{d}F_{y_{n_1},y_{n_2}}(x)$ and $F_{y_{n_1},y_{n_2}}(x)$ is the limiting distribution which has a density in Equation (14) but with y_{n_k} instead of $y_k, k = 1, 2$. Then, we get

$$G_n(f) = -\log \mathbf{\Lambda}_n - p \cdot F_{y_{n_1},y_{n_2}}(f). \tag{25}$$

By Theorem 2.1, $G_n(f)$ weakly converges to a Gaussian vector with mean

$$m(f) = \frac{1}{2}\log\frac{(c^2-d^2)h^2}{(ch-y_2d)^2} \tag{26}$$

and variance

$$\upsilon(f) = 2\log\left(\frac{c^2}{c^2-d^2}\right) \tag{27}$$

for the real case, where

$$\begin{aligned} h &= \sqrt{y_1+y_2-y_1y_2}, \\ a_0, b_0 &= \frac{(1\mp h)^2}{(1-y_2)^2}, \\ c, d &= \frac{1}{2}\left[\sqrt{1+\frac{y_2}{y_1}b_0} \pm \sqrt{1+\frac{y_2}{y_1}a_0}\right], \quad c > d. \end{aligned}$$

This is calculated in Section 5 using Lemma 2.1. For the complex case, the mean $m(f)$ is zero and the variance is half of $\upsilon(f)$. In other words,

$$-\log \mathbf{\Lambda}_n - p\cdot F_{y_{n_1},y_{n_2}}(f) \Rightarrow N(m(f), \upsilon(f)). \tag{28}$$

Here,

$$F_{y_{n_1},y_{n_2}}(f) = \frac{y_{n_2}-1}{y_{n_2}}\log c_n + \frac{y_{n_1}-1}{y_{n_1}}\log(c_n-d_nh_n) + \frac{y_{n_1}+y_{n_2}}{y_{n_1}y_{n_2}}\log\left(\frac{c_nh_n-d_ny_{n_2}}{h_n}\right), \tag{29}$$

where

$$\begin{aligned} h_n &= \sqrt{y_{n_1}+y_{n_2}-y_{n_1}y_{n_2}}, \\ a_n, b_n &= \frac{(1\mp h_n)^2}{(1-y_{n_2})^2}, \\ c_n, d_n &= \frac{1}{2}\left[\sqrt{1+\frac{y_{n_2}}{y_{n_1}}b_n} \pm \sqrt{1+\frac{y_{n_2}}{y_{n_1}}a_n}\right], \quad c_n > d_n, \end{aligned}$$

is derived in Section 5 using the density function of $F_{y_{n_1},y_{n_2}}$. Then, we get, letting $q_1 \wedge (n-q_1) \to \infty$,

$$T_n = \upsilon(f)^{-1/2}[-\log\mathbf{\Lambda}_n - p\cdot F_{y_{n_1},y_{n_2}}(f) - m(f)] \Rightarrow N(0,1).$$

■

We call CLRT for testing Equation (2) the test based on the statistic T_n and its asymptotic distribution derived in the theorem above. Moreover, it is worth noting that in the above proof, we used the Gaussian assumption for entry variables to fit $\mathbf{F}$ to a Gaussian F-matrix. However, Theorem 2.1 does not need this Gaussian assumption. Therefore, we can expect (or conjecture) that the asymptotic distribution for T_n in Theorem 3.1, hence the CLRT, could be valid more generally. However, the kurtosis parameter β appeared in Theorem 2.1 is no more null and it will appears in the asymptotic parameters $m(f)$ and $\upsilon(f)$ above.

3.2. *Two least-squares-based procedures for testing Equation (2)*

To evaluate the CLRT, we consider two additional procedures based on least-squares-type statistics as suggested in [3]. We first need to find the asymptotic distributions of these statistics.

By Equation (3) and the partition of $\mathbf{B}$, we obtain

$$\hat{\mathbf{B}}_1 = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{z}'_{i,1} \mathbf{A}_{11:2}^{-1} - \sum_{i=1}^{n} \mathbf{x}_i \mathbf{z}'_{i,2} \mathbf{A}_{22}^{-1} \mathbf{A}_{21} \mathbf{A}_{11:2}^{-1}. \tag{30}$$

Let

$$M_{n,1} = \operatorname{tr}((\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)(\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)'), \tag{31}$$

$$M_{n,2} = \operatorname{tr}((\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)\mathbf{A}_{11:2}(\hat{\mathbf{B}}_1 - \mathbf{B}_1^*)'). \tag{32}$$

Because $\hat{\mathbf{B}}$ is an unbiased estimator of $\mathbf{B}$, then $E\hat{\mathbf{B}}_1 = \mathbf{B}_1^*$ under the null hypothesis. Thus,

$$EM_{n,1} = \operatorname{tr}(\boldsymbol{\Sigma}) \operatorname{tr}(\mathbf{A}_{11:2}^{-1}), \tag{33}$$

$$EM_{n,2} = q_1 \operatorname{tr}(\boldsymbol{\Sigma}), \tag{34}$$

$$\sigma_{n,1}^2 = \operatorname{Var}(M_{n,1}) = 2 \operatorname{tr}(\boldsymbol{\Sigma}^2) \operatorname{tr}(\mathbf{A}_{11:2}^{-2}) + \beta_x \beta_{z1}, \tag{35}$$

$$\sigma_{n,2}^2 = \operatorname{Var}(M_{n,2}) = 2q_1 \operatorname{tr}(\boldsymbol{\Sigma}^2) + \beta_x \beta_{z2}, \tag{36}$$

where

$$\beta_x = E(\varepsilon_1' \varepsilon_1)^2 - (\operatorname{tr}(\boldsymbol{\Sigma}))^2 - 2 \operatorname{tr}(\boldsymbol{\Sigma}^2),$$

$$\beta_{z1} = \sum_{i=1}^{n} [(\mathbf{z}'_{i,1} - \mathbf{z}'_{i,2} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}) \mathbf{A}_{11:2}^{-2} (\mathbf{z}_{i,1} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{z}_{i,2})]^2,$$

$$\beta_{z2} = \sum_{i=1}^{n} [(\mathbf{z}'_{i,1} - \mathbf{z}'_{i,2} \mathbf{A}_{22}^{-1} \mathbf{A}_{21}) \mathbf{A}_{11:2}^{-1} (\mathbf{z}_{i,1} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{z}_{i,2})]^2.$$

Define

$$\mathbf{Z}_i^{(k)} = \mathbf{A}_{11:2}^{-(3-k)/2} (\mathbf{z}_{i,1} - \mathbf{A}_{12} \mathbf{A}_{22}^{-1} \mathbf{z}_{i,2}), \quad k = 1, 2. \tag{37}$$

THEOREM 3.2 *Assuming that*

(1) $\min(q_1, p, n - q) \to \infty$;
(2) *as* $p \to \infty$, $\operatorname{tr} \boldsymbol{\Sigma}^2 = o((\operatorname{tr} \boldsymbol{\Sigma})^2)$;
(3) $\max_{1 \le i \le n} \mathbf{Z}_i^{(k)'} \mathbf{Z}_i^{(k)} = o([\operatorname{tr} \boldsymbol{A}_{11:2}^{-(2-k)}])$;
(4) (ε_i), $i = 1, \ldots, n$, *are i.i.d. zero-mean random vectors such that for any* $\eta > 0$, *there exists a* $K > 0$, *such that*

$$E(\varepsilon_1' \varepsilon_2)^2 \le K(\operatorname{tr} \boldsymbol{\Sigma}^2),$$

$$\max E(\varepsilon_1' \varepsilon_2)^2 I \left(|\varepsilon_1' \varepsilon_2| \ge \frac{\eta \sqrt{\operatorname{tr} \boldsymbol{A}_{11:2}^{-2(2-k)} \operatorname{tr} \boldsymbol{\Sigma}^2}}{|\mathbf{Z}_i^{(k)'} \mathbf{Z}_j^{(k)}|} \right) = o(\eta^2 (\operatorname{tr} \boldsymbol{\Sigma}^2)),$$

$$E(\varepsilon_1' \varepsilon_1 - \operatorname{tr} \boldsymbol{\Sigma})^2 \le K(\operatorname{tr} \boldsymbol{\Sigma}^2),$$

$$E(\varepsilon_1' \varepsilon_1 - \operatorname{tr} \boldsymbol{\Sigma})^2 I \left(\frac{|\varepsilon_1' \varepsilon_1 - \operatorname{tr} \boldsymbol{\Sigma}| \ge \eta \sqrt{\beta_{zk} \operatorname{tr} \boldsymbol{\Sigma}^2}}{|\mathbf{Z}_i^{(k)'} \mathbf{Z}_j^{(k)}|} \right) = o(\eta^2 (\operatorname{tr} \boldsymbol{\Sigma}^2)).$$

Then, for $k = 1, 2$ and under H_0 in Equation (2),

$$\Gamma_{n,k} := \frac{M_{n,k} - EM_{n,k}}{\sigma_{n,k}} \Rightarrow \mathcal{N}(0, 1).$$

Consequently, to test Equation (2), we can use any of the statistics $\Gamma_{n,1}$ and $\Gamma_{n,2}$. These tests will be referred below as ST1 and ST2.

3.3. *A simulation study for comparison of the tests*

We set up a simulation experiment to compare five procedures for testing Equation (2): the classical LRT with an asymptotic χ^2 approximation, the associated BBC recalled in Equation (11), our CLRT introduced in Section 3 and the two tests ST1 and ST2 based on least-squares-type statistics of Section 3.2. Denote the non-centre parameter as $\psi = c_0^2\psi_0$, where $\psi_0 = \mathrm{tr}((\mathbf{B}_1 - \mathbf{B}_1^*)'\mathbf{\Sigma}^{-1}(\mathbf{B}_1 - \mathbf{B}_1^*))$, and c_0 is a varying constant. Then, we consider the model (1) as the form $\mathbf{x}_i = c_0(\mathbf{B}_1 - \mathbf{B}_1^*)\mathbf{z}_i + \boldsymbol{\varepsilon}_i,\ i = 1, \ldots, n$. Assume that the elements of $(\mathbf{B}_1 - \mathbf{B}_1^*)$ follow the distribution $\mathcal{N}(1, 1)$. All the *i.i.d.* elements of $\mathbf{z}_i$ in the model are sampled from $\mathcal{N}(1, 0.5)$. The errors $\boldsymbol{\varepsilon}_i$ in Equation (1) have a multivariate normal distribution $\mathcal{N}_p(0, C)$ with

$$C = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \cdots & \cdots & & & \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}.$$

Therefore, ρ measures the degree of correlations between the p coordinates of the noise vectors. To understand the effect of these correlations on the test procedures, we consider two cases: $\rho = 0.9$ and $\rho = 0$.

For different values of (p, n, q, q_1), we compute the realized sizes (type-I errors) of the five tests based on 1000 independent replications. All the tests are defined with an nominal (and asymptotic) level $\alpha = 0.05$. The powers of the tests are evaluated under alternative hypotheses obtained by varying the parameter c_0.

Table 1 gives the sizes (line $c_0 = 0$, in bold) and the powers ($c_0 \neq 0$) for the case $\rho = 0$ and various choices of the dimensions (p, n, q, q_1). Table 2 displays analogous results for the case $\rho = 0.9$ where the coordinates of the noise sequence are highly correlated. The important conclusions from these tables are as follows.

Test size:

- The LRT and BBC correction are highly inconsistent: in all considered cases, the LRT and its BBC correction have a much higher size than the nominal value 5%. In particular, the LRT systematically rejects the null hypothesis, even for data dimension as small as $p = 10$, while the BBC correction is just less biased as expected.
- In the case where the coordinates of the noise are uncorrelated (Table 1), the three tests CLRT, ST1 and ST2, which are based on the RMT, achieve a correct level close to 5%.

  In contrary, when these correlations are high (Table 2), as the least-squares-type tests ST1 and ST2 heavily depend on an assumed non-correlation between these coordinates, these two tests become inconsistent.

The power function: In the case where the coordinates of the noise are uncorrelated (Table 1), while being all consistent, CLRT and ST2 outperform the test ST1.

When these coordinates are highly correlated (Table 2) and despite their inconsistency, the tests ST1 and ST2 are outperformed by the CLRT. For example, in the case $\rho = 0.9, n = 200, p = 30$, the highest power of ST1 and ST2 are only 0.283 and 0.115, respectively.

Table 1. Sizes ($c_0 = 0$) and powers ($c_0 \neq 0$) of the four methods, based on 1000 independent applications with real Gaussian variables.

$\rho = 0$	$(p, n, q, q_1) = (10, 100, 50, 30)$					$(p, n, q, q_1) = (20, 100, 60, 50)$				
Parameter c_0	LRT	CLRT	BBC	ST1	ST2	LRT	CLRT	BBC	ST1	ST2
0	**1**	**0.056**	**0.101**	**0.070**	**0.086**	**1**	**0.047**	**0.672**	**0.042**	**0.072**
0.01	1	0.064	0.113	0.071	0.096	1	0.084	0.741	0.044	0.129
0.02	1	0.083	0.150	0.080	0.136	1	0.203	0.879	0.050	0.395
0.03	1	0.150	0.224	0.098	0.222	1	0.381	0.963	0.063	0.851
0.04	1	0.247	0.342	0.125	0.387	1	0.583	0.992	0.091	0.998
0.05	1	0.382	0.500	0.156	0.588	1	0.784	0.999	0.127	1
0.06	1	0.574	0.676	0.200	0.792	1	0.914	1	0.173	1
0.07	1	0.747	0.829	0.279	0.932	1	0.979	1	0.257	1
0.08	1	0.885	0.925	0.375	0.988	1	0.996	1	0.374	1
0.09	1	0.953	0.980	0.496	0.997	1	0.999	1	0.526	1
0.10	1	0.986	0.990	0.624	1	1	1	1	0.681	1
$\rho = 0$	$(p, n, q, q_1) = (30, 200, 80, 60)$					$(p, n, q, q_1) = (50, 200, 80, 70)$				
Parameter c_0	LRT	CLRT	BBC	ST1	ST2	LRT	CLRT	BBC	ST1	ST2
0	**1**	**0.060**	**0.178**	**0.054**	**0.062**	**1**	**0.056**	**0.495**	**0.036**	**0.048**
0.003	1	0.062	0.190	0.055	0.065	1	0.063	0.551	0.040	0.065
0.006	1	0.078	0.221	0.060	0.083	1	0.099	0.668	0.042	0.135
0.009	1	0.106	0.276	0.068	0.123	1	0.210	0.797	0.048	0.372
0.012	1	0.164	0.357	0.071	0.229	1	0.363	0.908	0.060	0.734
0.015	1	0.232	0.462	0.082	0.352	1	0.560	0.972	0.073	0.974
0.018	1	0.348	0.584	0.097	0.501	1	0.742	0.991	0.103	0.999
0.021	1	0.483	0.725	0.131	0.715	1	0.871	0.998	0.152	1
0.024	1	0.616	0.831	0.182	0.874	1	0.939	0.999	0.207	1
0.027	1	0.771	0.911	0.241	0.970	1	0.984	1	0.304	1
0.03	1	0.872	0.954	0.325	0.993	1	0.995	1	0.414	1

Note: The parameter ρ in the covariance matrix of errors is equal to 0.

To summarize, among the five tests considered here, only the CLRT displays an overall consistency and a generally satisfactory power. In particular, this test is robust with regard to the correlations between the coordinates of the noise process.

Finally, Figures 1 and 2 give a dynamic view of these comparisons by varying the non-central parameter c_0 for the cases $\rho = 0$ and $\rho = 0.9$, respectively. Note that the left-first point of all the lines represent the realized sizes (type I errors) of the tests, and others are the powers.

4. A high-dimensional multiple sample significance test

In this section, we consider the following multiple sample significance test problem in an multivariate analysis of variance with high-dimensional data. For the two-sample case, this problem has been considered by Dempster [2] and Bai and Saranadasa [3]. Here, we treat the general multiple sample case. Consider q Gaussian populations $\mathcal{N}(\mu^{(i)}, \mathbf{\Sigma})$ of dimension p, $1 \le i \le q$, and for each population, assume that we have a sample of size n_i: $\{\mathbf{x}_k^{(i)},\ 1 \le k \le n_i\}$. We wish to test the hypothesis

$$H_0 : \boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}. \tag{38}$$

Here, high dimensional means that both the number q of the populations and the dimension p of the observation vectors are large with respect to the sample sizes (n_i)s.

Table 2. Sizes ($c_0 = 0$) and powers ($c_0 \neq 0$) of the four methods, based on 1000 independent applications with real Gaussian variables.

$\rho = 0.9$	$(p,n,q,q_1) = (10,100,50,30)$					$(p,n,q,q_1) = (20,100,60,50)$				
Parameter c_0	LRT	CLRT	BBC	ST1	ST2	LRT	CLRT	BBC	ST1	ST2
0	**1**	**0.056**	**0.089**	**0.105**	**0.119**	**1**	**0.055**	**0.681**	**0.087**	**0.155**
0.005	1	0.063	0.099	0.106	0.121	1	0.063	0.696	0.088	0.164
0.010	1	0.078	0.123	0.107	0.124	1	0.089	0.762	0.089	0.187
0.015	1	0.110	0.162	0.109	0.134	1	0.165	0.849	0.091	0.220
0.020	1	0.164	0.234	0.111	0.143	1	0.261	0.923	0.093	0.261
0.025	1	0.253	0.355	0.116	0.161	1	0.458	0.974	0.095	0.323
0.030	1	0.388	0.491	0.118	0.182	1	0.690	0.999	0.099	0.408
0.035	1	0.562	0.652	0.123	0.215	1	0.878	1	0.101	0.503
0.040	1	0.724	0.811	0.130	0.250	1	0.963	1	0.105	0.610
0.045	1	0.873	0.926	0.136	0.284	1	0.998	1	0.110	0.704
0.050	1	0.951	0.979	0.144	0.343	1	1	1	0.115	0.801
$\rho = 0.9$	$(p,n,q,q_1) = (30,200,80,60)$					$(p,n,q,q_1) = (50,200,80,70)$				
Parameter c_0	LRT	CLRT	BBC	ST1	ST2	LRT	CLRT	BBC	ST1	ST2
0	**1**	**0.054**	**0.181**	**0.089**	**0.105**	**1**	**0.059**	**0.520**	**0.098**	**0.100**
0.002	1	0.059	0.197	0.090	0.106	1	0.060	0.536	0.099	0.107
0.004	1	0.074	0.223	0.090	0.109	1	0.079	0.604	0.100	0.116
0.006	1	0.113	0.288	0.091	0.115	1	0.140	0.697	0.101	0.136
0.008	1	0.178	0.400	0.091	0.126	1	0.233	0.811	0.102	0.175
0.010	1	0.287	0.530	0.092	0.140	1	0.409	0.913	0.104	0.230
0.012	1	0.445	0.691	0.093	0.161	1	0.633	0.979	0.107	0.300
0.014	1	0.643	0.840	0.097	0.180	1	0.826	0.993	0.114	0.379
0.016	1	0.821	0.939	0.101	0.202	1	0.953	1	0.118	0.481
0.018	1	0.937	0.986	0.107	0.238	1	0.992	1	0.125	0.597
0.020	1	0.987	0.996	0.115	0.283	1	1	1	0.131	0.694

Note: The parameter ρ in the covariance matrix of errors is equal to 0.9.

Clearly, the observations can be put in the form

$$\mathbf{x}_k^{(i)} = \boldsymbol{\mu}^{(i)} + \varepsilon_k^{(i)}, \quad 1 \leq i \leq q, \ \ 1 \leq k \leq n_i, \tag{39}$$

where $\{\varepsilon_k^{(i)}\}$ is an array of i.i.d. random vectors distributed as $\mathcal{N}_p(0, \boldsymbol{\Sigma})$. We are going to embed the test (38) into a special instance of the regression test (2). To this end, let $\{e_i\}$ be the canonical base of $\mathbb{R}^p$ and we define the following regression vectors

$$\mathbf{z}_k^{(i)} = [e_i + e_q]\mathbf{1}_{\{i<q\}} + e_q\mathbf{1}_{\{i=q\}}, \quad 1 \leq i \leq q, \ \ 1 \leq k \leq n_i.$$

Define moreover the $p \times q$ matrix $\mathbf{B} = (\mathbf{B}_1, \mathbf{B}_2)$ with

$$\mathbf{B}_1 = (\boldsymbol{\mu}^{(1)} - \boldsymbol{\mu}^{(q)}, \ldots, \boldsymbol{\mu}^{(q-1)} - \boldsymbol{\mu}^{(q)}), \tag{40}$$

$$\mathbf{B}_2 = \boldsymbol{\mu}^{(q)}. \tag{41}$$

Note that the dimension q is split to $(q_1, q_2) = (q-1, 1)$ in the above decomposition.

Therefore, the observations follow a linear model

$$\mathbf{x}_k^{(i)} = \mathbf{B}\mathbf{z}_k^{(i)} + \varepsilon_k^{(i)}, \quad 1 \leq i \leq q, \ \ 1 \leq k \leq n_i. \tag{42}$$

The multiple sample test (38) is equivalent to the following regression test:

$$H_0: \ \mathbf{B}_1 = 0. \tag{43}$$

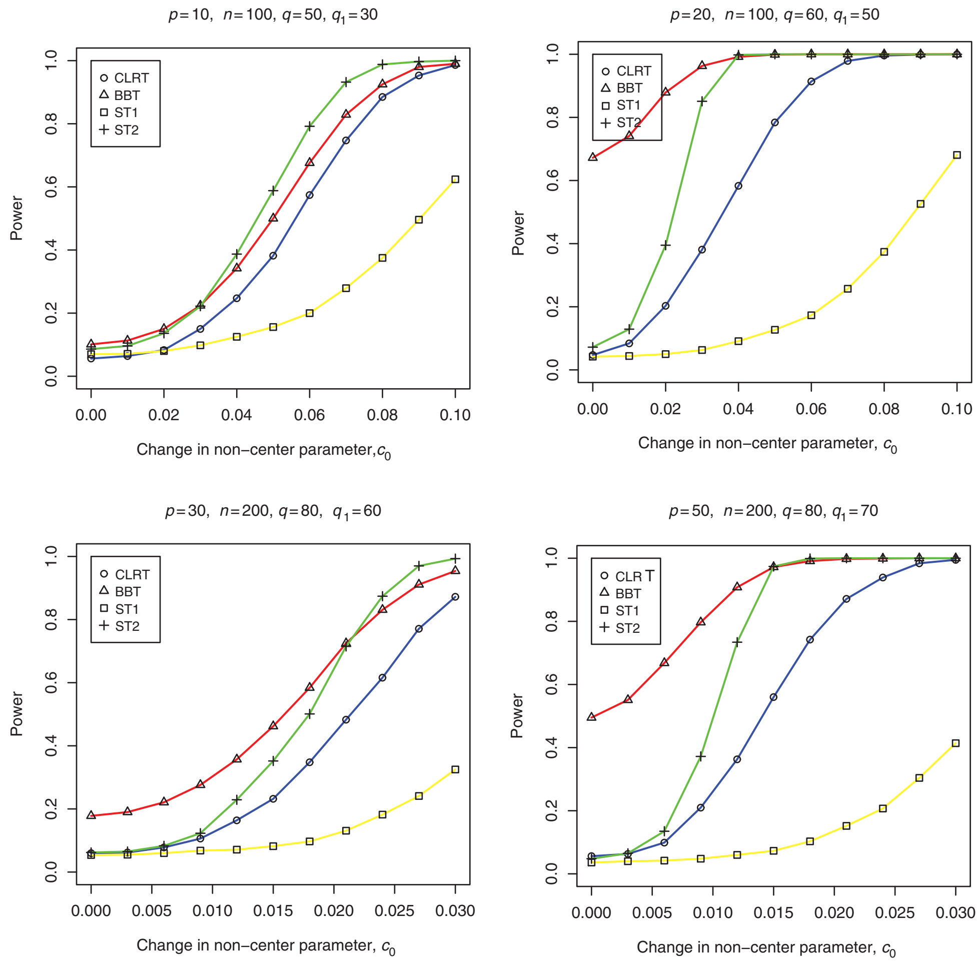

Figure 1. Sizes ($c_0 = 0$) and powers ($c_0 \neq 0$) of the four methods, which are the CLRT, BBC and two least-squares-type tests (ST1 and ST2), based on 1000 independent replications using Gaussian error variables from $\mathcal{N}(0, I)$. Top row: $(p, n, q, q_1) = (10, 100, 50, 30)$ and $(20, 100, 60, 50)$. Bottom row: $(p, n, q, q_1) = (30, 200, 80, 60)$ and $(50, 200, 80, 70)$.

In order to apply Theorem 3.1, we now identify the likelihood ratio statistic $\mathbf{\Lambda}_n$ defined in Equation (8). Here denote $n = \sum_{i=1}^{q} n_i$. Under the null hypothesis, the likelihood estimates of $(\mathbf{B}_2, \mathbf{\Sigma})$ are (see [11] for details of computation)

$$\hat{\mathbf{B}}_{20} = \bar{\mathbf{x}} = \frac{1}{n} \sum_{i,k} \mathbf{x}_k^{(i)}, \tag{44}$$

$$\hat{\mathbf{\Sigma}}_0 = \frac{1}{n} \sum_{i,k} (\mathbf{x}_k^{(i)} - \bar{\mathbf{x}})(\mathbf{x}_k^{(i)} - \bar{\mathbf{x}})'. \tag{45}$$

On the other hand, under the alternative hypothesis, the likelihood estimates of $(\boldsymbol{\mu}^{(i)}, \mathbf{\Sigma})$ are

$$\hat{\boldsymbol{\mu}}^{(i)} = \bar{\mathbf{x}}^{(i)} := \frac{1}{n_i} \sum_{k=1}^{n_i} \mathbf{x}_k^{(i)}, \quad 1 \leq i \leq q, \tag{46}$$

$$\hat{\mathbf{\Sigma}} = \frac{1}{n} \sum_{i,k} (\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k^{(i)})(\mathbf{x}_k^{(i)} - \bar{\mathbf{x}}_k^{(i)})'. \tag{47}$$

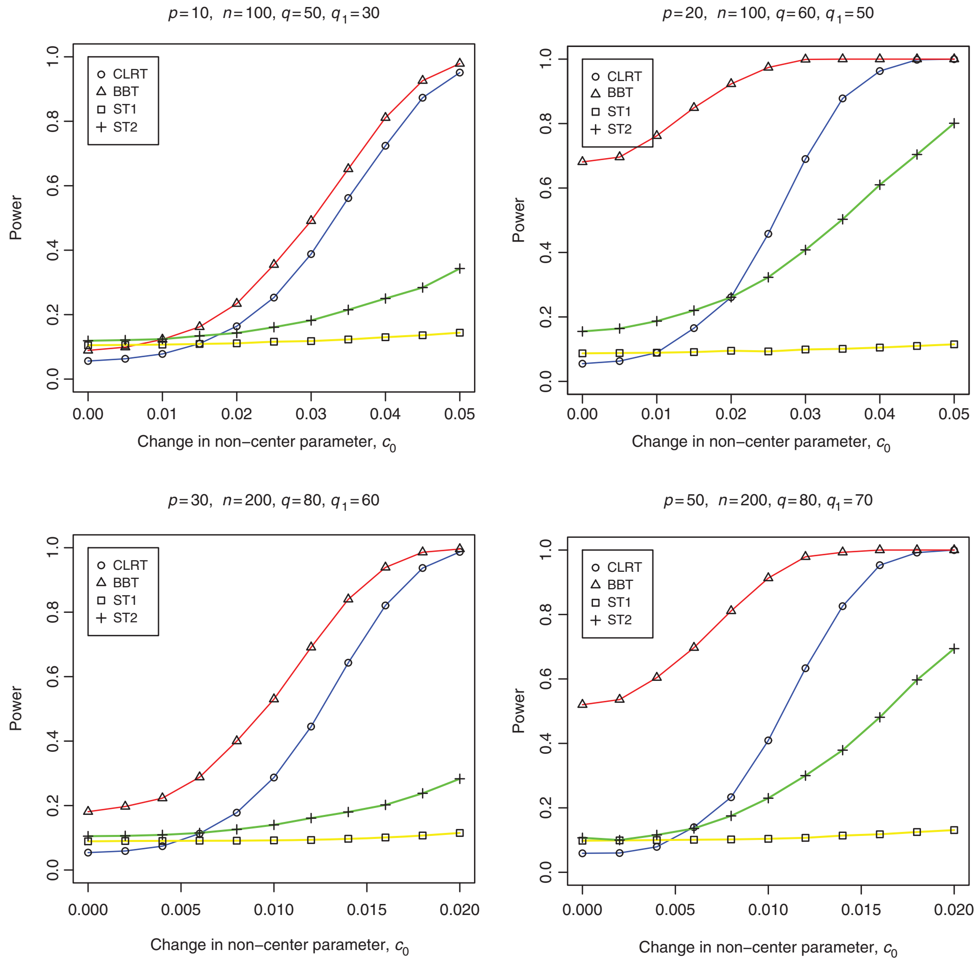

Figure 2. Sizes ($c_0 = 0$) and powers ($c_0 \neq 0$) of the four methods, which are the CLRT, BBC and two least-squares-type tests (ST1 and ST2), based on 1000 independent replications using Gaussian error variables from $\mathcal{N}(0, C)$ with the parameter $\rho = 0.9$. Top row: $(p, n, q, q_1) = (10, 100, 50, 30)$ and $(20, 100, 60, 50)$. Bottom row: $(p, n, q, q_1) = (30, 200, 80, 60)$ and $(50, 200, 80, 70)$.

The likelihood ratio statistic $\mathbf{\Lambda}_n = |\hat{\mathbf{\Sigma}}|/|\hat{\mathbf{\Sigma}}_0|$ readily follows.

By application of Theorem 3.1, we have the following.

PROPOSITION 4.1 *For the multiple sample significance test (38), assume that $q \to \infty$, $n_i \to \infty$, $1 \le i \le q$, $p \to \infty$ in such a manner that*

$$y_{n_1} := \frac{p}{q-1} \to y_1 \in (0,1), \quad y_{n_2} = \frac{p}{n-q} \to y_2 \in (0,1). \tag{48}$$

Then, for the same function f defined in Equation (24), we have

$$T_n^* = \upsilon(f)^{-1/2}[-\log \mathbf{\Lambda}_n - p \cdot F_{y_{n_1},y_{n_2}}(f) - m(f)] \Rightarrow \mathcal{N}(0,1),$$

where $\upsilon(f), m(f)$ and $F_{y_{n_1},y_{n_2}}(f)$ are defined in Equations (26), (27) and (29), respectively, with the values of $y_{n_1}, y_{n_2}, y_1, y_2$ defined in Equation (48).

It is worth noting here that the classical LRT for testing Equation (38 will rely on the following weak convergence theorem: under H_0 and assuming fixed p and q while letting $n_i \to \infty$,

$$-n \log \mathbf{\Lambda}_n \Rightarrow \chi^2_{p(q-1)}. \tag{49}$$

Inevitably, in a high-dimensional case, U_n will drift to infinity by Proposition 4.1. Consequently, this classical χ^2-approximation will lead to a test size much higher than a given nominal test level, exactly as for the general linear hypothesis considered in Section 3.

5. Proofs

5.1. *Proof of Equations (26) and (27)*

Because $\mathbf{x}_i$ are Gaussian variables, for the real case, $\beta = E|\xi|^4 - 3 = 0$, then Equations (17), (18) and (20) are all 0. Consider Equations (16) and (19), as $y_{n_k} \to y_k,\ k = 1, 2$, and during the process of Lemma 2.1 calculation, we will see that the constant and items approaching to zero do not effect on the circle integration results, and in practice, $y_{n_k} = y_k,\ k = 1, 2$. Therefore, we use

$$f(x) = \log\left(1 + \frac{y_2}{y_1}x\right)$$

instead of $f(x) = \log(1 + (y_{n_2}/y_{n_1})x)$. Because

$$c, d = \frac{1}{2}\left[\sqrt{1 + \frac{y_2}{y_1}b_0} \pm \sqrt{1 + \frac{y_2}{y_1}a_0}\right], \quad c > d,$$
$$a_0, b_0 = \frac{(1 + h)^2}{(1 - y_2)^2}$$

is the solution of Equation (21) with $a, \alpha = 1, b, \beta = y_2/y_1$, then, using Lemma 2.1, we have

$$m(f) = \frac{1}{2}\log\frac{(c^2 - d^2)h^2}{(ch - y_2 d)^2},$$
$$\upsilon(f) = 2\log\left(\frac{c^2}{c^2 - d^2}\right)$$

for the real case.

5.2. *Proof of $F_{y_{n_1}, y_{n_2}}(f)$ (Equation (29))*

For this computation, we drop the indexes n_1 and n_2 in the parameters y_{n_j} and compute the integral $F_{y_1,y_2}(f)$. Following a device designed in [15, Lemma A.2], let $\underline{m}(z)$ be the Stieltjes transform of the distribution function $\underline{F} := (1 - y_1)I_{(0,\infty)} + y_1 F_{y_1,y_2}$. For $r > 1$ but very close to 1 and $|\xi| = 1$, we use a change of variable $z = \phi(\xi)$ which is implicitly defined by the formula

$m_0(z) = -(1 + hr\xi)/(1 - y_2)$ and have the following relations

$$z = -\frac{m_0(z)(m_0(z) + 1 - y_1)}{(m_0(z) + 1/(1 - y_2))(1 - y_2)} \quad \text{and} \quad \underline{m}(z) = \frac{(1 - y_2)(m_0(z) + 1/(1 - y_2))}{m_0(z)(m_0(z) + 1)}.$$

Or equivalently,

$$z = \frac{1 + h^2 + hr^{-1}\bar{\xi} + hr\xi}{(1 - y_2)^2} \quad \text{and} \quad \underline{m}(z) = \frac{-(1 - y_2)^2\xi}{hr(\xi + 1/hr)(\xi + y_2/hr)}.$$

This shows that when ξ anticlockwise runs along the unit circle, z anticlockwise runs a contour which closely encloses the interval $[a, b]$ when r is close to 1, where $a = (1 - h)^2/(1 - y_2)^2$ and $b = (1 + h)^2/(1 - y_2)^2$. Therefore, we obtain

$$\begin{aligned}
F^{y_1,y_2}(f) &= \int_a^b f(x)\frac{(1 - y_2)\sqrt{(b - x)(x - a)}}{2\pi x(y_1 + y_2 x)}\,\mathrm{d}x = y_1^{-1}\int_a^b f(x)\,\mathrm{d}\underline{F}(x) \\
&= -\frac{1}{2\pi \mathrm{i} y_1}\oint_{\mathcal{C}} f(z)\underline{m}(z)\,\mathrm{d}z \quad \text{(any contour } \mathcal{C} \text{ enclosing the interval } [a, b]) \\
&= \frac{1}{2\pi \mathrm{i} y_1}\oint_{|\xi|=1} \log|c + \mathrm{d}\xi|^2 \frac{\xi^2 - 1}{\xi(\xi + 1/h)(\xi + y_2/h)}\,\mathrm{d}\xi \\
&= \frac{1}{2\pi \mathrm{i} y_1}\oint_{|\xi|=1} (\log(c + \mathrm{d}\xi) + \log(c + \mathrm{d}\xi^{-1}))\frac{\xi^2 - 1}{\xi(\xi + 1/h)(\xi + y_2/h)}\,\mathrm{d}\xi \\
&\quad \text{(making } \xi^{-1} \to \xi \text{ in the second integral)} \\
&= \frac{1}{2\pi \mathrm{i} y_1}\oint_{|\xi|=1} \log(c + \mathrm{d}\xi)\left(\frac{\xi^2 - 1}{\xi(\xi + 1/h)(\xi + y_2/h)} - \frac{h^2}{y_2}\frac{\xi^2 - 1}{\xi(\xi + h)(\xi + h/y_2)}\right)\mathrm{d}\xi \\
&= \frac{y_2 - 1}{y_2}\log(c) + \frac{y_1 - 1}{y_1}\log(c - dh) + \frac{y_1 + y_2}{y_1 y_2}\log\left(\frac{ch - dy_2}{h}\right),
\end{aligned}$$

where

$$c, d = \frac{1}{2}\left(\sqrt{1 + \frac{y_2}{y_1}b} \pm \sqrt{1 + \frac{y_2}{y_1}a}\right), \quad c > d.$$

Acknowledgements

The authors acknowledge the support from the following research grants: NSFC grant 11171057 (Z.D. Bai), NSFC grant 11101181 and RFDP grant 20110061120005 (D. Jiang), HKU Start-up Fund (J. Yao) and NSFC-11171058 and NECT-11-0616 (S. Zheng).

References

[1] Z.D. Bai, *Methodologies in spectral analysis of large dimensional random matrices*, A review, Statist. Sinica. 9 (1999), pp. 611–677.
[2] A.P. Dempster, *A high dimensional two sample significance test*, Ann. Math. Statist. 29 (1958), pp. 995–1010.
[3] Z.D. Bai and H. Saranadasa, *Effect of high dimension comparison of significance tests for a high dimensional two sample problem*, Statist. Sinica 6 (1996), pp. 311–329.
[4] O. Ledoit and M. Wolf, *Some hypothesis tests for the covariance matrix when the dimension is large compared to the sample size*, Ann. Statist. 30 (2002), pp. 1081–1102.
[5] M.S. Srivastava, *Some tests concerning the covariance matrix in high dimensional data*, J. Japan Statist. Soc. 35 (2005), pp. 251–272.
[6] J.R. Schott, *A test for the equality of covariance matrices when the dimension is large relative to the sample size*, Comput. Statist. Data Anal. 51 (2007), pp. 6535–6542.

[7] Z.D. Bai, D.D. Jiang, J.F. Yao, and S.R. Zheng, *Corrections to LRT on large dimensional covariance matrix by RMT*, Ann. Statist. 37(6B) (2009), pp. 3822–3840. A longer version with full proofs is available at arXiv:0902.0552[stat].
[8] S.S. Wilks, *Certain generalizations in the analysis of variance*, Biometrika 24 (1932), pp. 471–494.
[9] S.S. Wilks, *Moment-generating operators for determinants of product moments in samples from a normal system*, Ann. Math. 35 (1934), pp. 312–340.
[10] M.S. Bartlett, *The vector representation of a sample*, Proc. Cambridge Philos. Soc. 30 (1934), pp. 327–340.
[11] T.W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 3rd ed., John Wiley & Sons, Hoboken, NJ, 2003.
[12] A.M. Mathai, *On the distribution of the likelihood ratio criterion for testing linear hypotheses on regression coefficients*, Ann. Inst. Statist. Math. 23 (1971), pp. 181–197.
[13] G.E.P. Box, *A general distribution theory for a class of likelihood criteria*, Biometrika 36 (1949), pp. 317–346.
[14] Z.D. Bai and J.W. Silverstein, *Spectral Analysis of Large-Dimensional Random Matrices*, 1st ed., Science Press, Beijing, 2006.
[15] S. Zheng, *Central limit theorem for linear spectral statistics of large dimensional F-matrix*, Ann. Inst. H. Poincaré Probab. Statist. 48 (2012), pp. 444–476