Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 1841–1859
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Diagnostic checking for multivariate regression models

Lixing Zhu[a,*], Ruoqing Zhu[b], Song Song[c]

[a] *Hong Kong Baptist University, Hong Kong, China*
[b] *Bowling Green State University, USA*
[c] *Humboldt University, Germany*

Received 23 December 2006
Available online 13 February 2008

Abstract

Diagnostic checking for multivariate parametric models is investigated in this article. A nonparametric Monte Carlo Test (NMCT) procedure is proposed. This Monte Carlo approximation is easy to implement and can automatically make any test procedure scale-invariant even when the test statistic is not scale-invariant. With it we do not need plug-in estimation of the asymptotic covariance matrix that is used to normalize test statistic and then the power performance can be enhanced. The consistency of NMCT approximation is proved. For comparison, we also extend the score type test to one-dimensional cases. NMCT can also be applied to diverse problems such as a classical problem for which we test whether or not certain covariables in linear model has significant impact for response. Although the Wilks lambda, a likelihood ratio test, is a proven powerful test, NMCT outperforms it especially in non-normal cases. Simulations are carried out and an application to a real data set is illustrated.

AMS 2000 subject classifications: 62H15; 62G09; 62G20

Keywords: Multivariate regression model; Goodness-of-fit; Wilks lambda; Score tests; Nonparametric Monte Carlo approximation

1. Introduction

Suppose that a response vector $\boldsymbol{Y} = (y_1, \ldots, y_q)^\tau$ depends on a vector $\boldsymbol{X} = (x_1, \ldots, x_p)^\tau$ of covariables, where τ denotes transposition. We may then decompose $\boldsymbol{Y}$ into a vector of functions

* Corresponding author.
E-mail address: lzhu@hkbu.edu.hk (L. Zhu).

0047-259X/$ - see front matter
doi:10.1016/j.jmva.2008.01.022

$\boldsymbol{m}(\boldsymbol{X}) = (m_1(\boldsymbol{X}), \ldots, m_q(\boldsymbol{X}))^\tau$ of $\boldsymbol{X}$ and a noise variable $\boldsymbol{\varepsilon}$, which is orthogonal to $\boldsymbol{X}$, i.e., for the conditional expectation of $\boldsymbol{\varepsilon}$ given $\boldsymbol{X}$, we have $\boldsymbol{E}(\boldsymbol{\varepsilon}|\boldsymbol{X}) = 0$. When $\boldsymbol{Y}$ is unknown, the optimal predictor of $\boldsymbol{Y}$ given $\boldsymbol{X} = \boldsymbol{x}$ equals $\boldsymbol{m}(\boldsymbol{x})$. Since in practice the regression function $\boldsymbol{m}$ is unknown, statistical inference about $\boldsymbol{m}$ is of importance. In a purely parametric framework, $\boldsymbol{m}$ is completely specified up to a parameter. For example, in linear regression, $\boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{\beta}^\tau \boldsymbol{x}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ is an unknown $p \times q$ matrix which needs to be estimated from the available data. More generally, we can study a nonlinear model with $\boldsymbol{m}(\boldsymbol{x}) = \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{x}) = (g_1(\beta_1, \boldsymbol{x}), \ldots, g_q(\beta_q, \boldsymbol{x}))^\tau$, where the vector of the link function $\boldsymbol{G}(\cdot)$ may be nonlinear but is specified.

When the dimension $q = 1$, the estimation and statistical inference have been studied extensively in the literature. In recent years, checking the adequacy of such parametric models becomes one of the central problems in regression analysis because any statistical analysis within the model, to avoid wrong conclusions, should be accompanied by a check of whether the model is valid or not. The literature is much elaborate. To review only a few contributions, Cox et al. [3] introduced tests of the null hypothesis that a regression function has a particular parametric structure. Azzalini et al. [1] considered nonparametric regression as an aid to model checking. Eubank and Spiegelman [7] considered spline approaches to testing the goodness-of-fit of a linear model. Simonoff and Tsai [17] proposed diagnostic methods for assessing the influence of individual data values on goodness-of-fit tests based on nonparametric regression. Gu [10] used spline methods in a diagnostic approach to model fitting. Eubank and La Riccia [6] derived properties of two-sided tests in nonparametric regression based on Fourier methods. Härdle and Mammen [11] considered comparisons between parametric and nonparametric fits and used the wild bootstrap for the computation of critical regions. Härdle et al. [12] investigated testing for parametric versus semiparametric modelling in Generalized Linear Models, again using the wild bootstrap. Stute, Thies and Zhu [20] proposed an innovation process approach. Two comprehensive reference books are Hart [13] and Zhu [24].

The above tests are almost in the class of locally smoothing methods. When the covariables are high-dimensional, the data sparseness in high-dimensional space causes a serious problem when nonparametric smoothing is used in the construction of tests. Another class consists of globally smoothing methods. Stute [18] proposed a nonparametric principal component analysis. Some optimally powerful tests can be constructed. Stute et al. [19] and Stute and Zhu [21] recommended an innovation process method. Fan and Huang [8] proposed adaptive Neyman tests. Stute et al. [19] also used the wild bootstrap.

It is worthwhile to note that the above methods focus on constructing omnibus tests. However, when we have prior knowledge about the alternative, more powerful test should be constructed invoking the prior information. Recently, Stute and Zhu [22] studied a score type test for Single-Index model. The score type test is of optimality for directional alternatives. The test statistic is the sum of weighted residuals. The optimality of the test can be achieved through selecting an appropriate weight function.

For the models with multivariate responses, in principle, Stute and Zhu's methodology can be extended to tackle the testing problem. We will define a score type test using similar idea. However, to define a scale-invariant test, an estimator of the limiting variance is often used to standardize the test. Note that such an estimator is model-dependent, and under alternative is consistent with larger value than the one under the null. As a result, the power is deteriorated. Take these problems into account, we propose a nonparametric Monte Carlo test (NMCT) procedure. This Monte Carlo procedure is used to approximate the null distribution of test statistic. With the help from it, we do not need to construct scale-invariant test because this test procedure can make any test scale-invariant. Particularly in our case, estimating the variance is unnecessary. In the simulations, we will make a comparison to the score type test. Furthermore, NMCT can fully apply the structure of the response/error, for instance the elliptical symmetry of the distribution, in the construction of NMCT. As we know, there are several proposals of Monte Carlo approximations available in the literature such as time-honored Bootstrap [5,2] is also a comprehensive reference. But these existing methods still involve the estimation for the asymptotic covariance. In these methods, the semistructure of the response/error cannot be used in the construction of the approximation.

Furthermore, NMCT can also be applied to some classical problems with multivariate linear models. To investigate which covariable(s) insignificantly affects the response, the likelihood ratio test called Wilks lambda is a standard test contained in the textbooks. The p-values can be determined by chi-square distribution, see e.g., [14]. When the underlying distribution of the error is normal, the Wilks lambda has been proved to be very powerful. However, it is not true when normality is violated. In this article, We will theoretically and empirically show how NMCT works. The limited simulations show that the power performance of NMCT is better, even in the normal case, than that of the Wilks lambda.

Therefore, NMCT shares the following desired features.

- As an alternative to the existing resampling methods, NMCT is especially suitable for the problems with semiparametric structured models where the errors are of semiparametric distributions such as elliptically symmetric distribution.
- NMCT is a self-scale-invariant. It does not involve the estimation of asymptotic covariance matrix which is model-dependent and often deteriorates the power performance when we use it in the test statistic.
- NMCT is a generic methodology which can be applied to other testing problems.

Since NMCT is based on a variant of score type test without scaling, we will study the score type test first so that we can similarly obtain the asymptotic properties of NMCT. It is worth mentioning that, in univariate response cases, although the score type test mainly deals with directional alternatives, it can be used to tackle composite alternatives through the residual plots to search for the weight function involved in the test. We will briefly discuss this issue for multivariate response cases in Sections 3 and 4. On the other hand, we also note that when the alternative is nonparametric, it is of importance to find how to make score type test possible to construct omnibus test. Technically, we could use the idea in [22] to do so when we use a set of weight functions. However, for multivariate cases, the structure of test statistic is rather complex and it is not easy to implement. This deserves further study.

The paper is organized as follows. The next section describes the construction of the score type test and its asymptotic behavior. Section 3 presents the algorithm of NMCT. The simulation study and an application are reported in Section 4. The technical proofs are presented in the Appendix.

2. Score type test

Suppose that $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ is a sample drawn from a population which follows the model as:

$$\boldsymbol{Y} = \boldsymbol{m}(\boldsymbol{X}) + \boldsymbol{\varepsilon}, \tag{2.1}$$

where $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_q)^{\mathrm{T}}$ is a q-dimensional error vector independent of $\boldsymbol{X}$. For model checking, we want to test the null hypothesis: for some matrix $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)$ almost surely

$$H_0: \quad \boldsymbol{m}(\cdot) = \boldsymbol{G}(\boldsymbol{\beta}, \cdot), \tag{2.2}$$

where for each i with $1 \le i \le q$, $m_i(\cdot) = g_i(\beta_i, \cdot)$. Let $\boldsymbol{e} = \boldsymbol{Y} - \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{X})$. Clearly, $H_0 \boldsymbol{e} = \boldsymbol{\varepsilon}$ and then $\boldsymbol{E}(\boldsymbol{e}|\boldsymbol{X}) = 0$. It implies that for any q-dimensional weight function $\boldsymbol{W}(\boldsymbol{\beta}, \cdot)$ of $\boldsymbol{X}$, $\boldsymbol{E}(\boldsymbol{e} \bullet \boldsymbol{W}(\boldsymbol{\beta}, \boldsymbol{X})) = 0$ where the dot product "$\bullet$" stands for the multiplication componentwise. From which we can define a score type test through an empirical version of $\boldsymbol{E}(\boldsymbol{e} \bullet \boldsymbol{W}(\boldsymbol{\beta}, \boldsymbol{X}))$. Let

$$\boldsymbol{T}_n = \frac{1}{n} \sum_{j=1}^{n} \hat{\boldsymbol{e}}_j \bullet \boldsymbol{W}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j), \tag{2.3}$$

where $\hat{\boldsymbol{e}}_j = \boldsymbol{y}_j - \boldsymbol{G}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j)$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}$. The resulting test statistic is a quadratic form $\mathcal{T}_n = \boldsymbol{T}_n^{\tau} \widehat{\boldsymbol{\Sigma}}^{-1} \boldsymbol{T}_n$ where "τ" stands for transposition and $\widehat{\boldsymbol{\Sigma}}$ is a consistent estimator of the covariance matrix of $\boldsymbol{T}_n$.

There are three quantities in the test statistic to be selected: two estimators $\hat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\Sigma}}$ and weight function $\boldsymbol{W}(\boldsymbol{\beta}, \cdot)$. The two estimators are for the consistency of the test statistic to obtain a tractable limit null distribution of $\mathcal{T}_n$. The selection for weight function is to enhance power performance of the test.

In this paper we use the least squares estimators $\hat{\beta}_i$ that are the solutions of the following equations:

$$\sum_{j=1}^{n} g_i'(\beta_i, \boldsymbol{x}_j)(\boldsymbol{y}_j^{(i)} - g_i(\beta_i, \boldsymbol{x}_j))^{\tau} = 0, \tag{2.4}$$

where g' is the $p \times 1$ derivative vector of g with respect to β_i provided that g_i are differentiable. As we know, each of $\hat{\beta}_i$ has an asymptotically linear representation. For model (2.1), $\boldsymbol{y}_j^{(i)} = m_i(x_j) + \boldsymbol{e}_j^{(i)}$, $j = 1, \ldots, n$. Denote $\boldsymbol{\eta} = \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{X}) - \boldsymbol{m}(\boldsymbol{X})$, and $\boldsymbol{\eta}_j = \boldsymbol{G}(\boldsymbol{\beta}, \boldsymbol{x}_j) - \boldsymbol{m}(\boldsymbol{x}_j)$. Then $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$ are $i.i.d.$ random variables. The asymptotically linear representations of $\hat{\beta}_i$ are as follows:

$$\hat{\beta}_i - \beta_i = \frac{1}{n} \sum_{j=1}^{n} S_{ni}^{-1} g_i'(\beta_i, \boldsymbol{x}_j) \boldsymbol{e}_j^{(i)} + \frac{1}{n} \sum_{j=1}^{n} S_{ni}^{-1} g_i'(\beta_i, \boldsymbol{x}_j) \boldsymbol{\eta}_j^{(i)} + o_p(1/\sqrt{n}), \tag{2.5}$$

where $S_{ni} = \frac{1}{n} \sum_{j=1}^{n} (g_i'(\beta_i, \boldsymbol{x}_j))(g_i'(\beta_i, \boldsymbol{x}_j))^{\tau}$.

Note that in probability, the first sum is of the rate $O(1/\sqrt{n})$ and under fixed alternatives, $S_{ni} \to S_i := \boldsymbol{E}\left((g_i'(\beta_i, \boldsymbol{X}))(g_i'(\beta_i, \boldsymbol{X}))^{\tau}\right)$, and the the second sum converges to $C_i := S_i^{-1} \boldsymbol{E}\left(g_i'(\beta_i, \boldsymbol{X}) \boldsymbol{\eta}^{(i)}\right)$. Thus, $\hat{\beta}_i$ converges to $\beta_i + C_i$, and $C_i \neq 0$ corresponds to the alternative H_1. In the following, we study the asymptotic behavior of the test statistics under both H_0 and H_1. To present the results, we give some conditions.

1. The second derivatives of g_i and the first derivative of $\boldsymbol{W}^{(i)}$ with respect to $\boldsymbol{x}$ and $\boldsymbol{\beta}$ are continuous and can be bounded by a function $M(\cdot)$ with $E(M(\boldsymbol{X}))^2 < \infty$.
2. The second moments of g_i, $\boldsymbol{W}^{(i)}$ and $\boldsymbol{e}^{(i)}$ are finite.
3. The asymptotic representation of $\hat{\beta}$ in (2.5) holds.

From the above conditions, we can easily obtain

$$\begin{aligned}\sqrt{n}T_{ni} &= \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{x}_j) - E[(\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{X}))(g_i'(\beta_i,\boldsymbol{X}))^{\tau}]S_i^{-1}(g_i'(\beta_i,\boldsymbol{x}_j))\right\}\boldsymbol{e}_j^{(i)}\\ &+\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left\{\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{x}_j) - E[(\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{X}))(g_i'(\beta_i,\boldsymbol{X}))^{\tau}]S_i^{-1}(g_i'(\beta_i,\boldsymbol{x}_j))\right\}\boldsymbol{\eta}_j^{(i)}\\ &+o_p(1)\\ &=: \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\boldsymbol{V}_j^{(i)}\boldsymbol{e}_j^{(i)} + \frac{1}{\sqrt{n}}\sum_{j=1}^{n}\boldsymbol{V}_j^{(i)}\boldsymbol{\eta}_j^{(i)} + o_p(1), \end{aligned} \tag{2.6}$$

where $\boldsymbol{V}_j^{(i)} = \left(\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{x}_j)\right) - E[(\boldsymbol{W}^{(i)}(\beta_i,\boldsymbol{X}))(g_i'(\beta_i,\boldsymbol{X}))^{\tau}]S_i^{-1}(g_i'(\beta_i,\boldsymbol{x}_j))$.

Theorem 2.1 states the asymptotic results of $\boldsymbol{T}_n = (T_{n1},\ldots,T_{nq})^{\tau}$.

Theorem 2.1. *Suppose that the above conditions hold.*

(1) *When H_0 holds, $\sqrt{n}T_{ni} \Longrightarrow T_i$ following normal distribution $N(0,\sigma_{ii})$ where the notation "$\Longrightarrow$" stands for weak convergence and σ_{ii} is the variance of $\boldsymbol{V}_j^{(i)}\boldsymbol{e}_j^{(i)}$. Therefore letting $\boldsymbol{T} = (T_1,\ldots,T_q)^{\tau}$, $\sqrt{n}\boldsymbol{T}_n \Longrightarrow \boldsymbol{T}$ following normal distribution $N(0,\boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma} = (\sigma_{lm})_{1\le l,m\le q}$ and σ_{lm} is the covariance between $\boldsymbol{V}_j^{(l)}\boldsymbol{e}_j^{(l)}$ and $\boldsymbol{V}_j^{(m)}\boldsymbol{e}_j^{(m)}$ for any pair of $1\le l,\ m\le q$. This results in that $\mathcal{T}_n$ is asymptotically chi-squared with degree of freedom q.*

(2) *When H_0 is false and for some i with $1\le i\le q$, if $\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\boldsymbol{V}_j^{(i)}\boldsymbol{\eta}_j^{(i)}\right]^2 \to \infty$, then $\mathcal{T}_n \to \infty$ in probability; and if $\left[\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\boldsymbol{V}_j^{(i)}\boldsymbol{\eta}_j^{(i)}\right] \to B_i$, a constant, then T_{ni} converges in distribution to $T_i + B_i$ where $B_i = E\left[\boldsymbol{V}^{(i)}\boldsymbol{\eta}^{(i)}\right]$. Let $\boldsymbol{T} = (T_1,\ldots,T_q)^{\tau}$ and $\boldsymbol{B} = (B_1,\ldots,B_q)^{\tau}$. $\mathcal{T}_n$ then converges in distribution to $(\boldsymbol{T}+\boldsymbol{B})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}+\boldsymbol{B})$ that is a non-central chi-squared random variable with the non-centrality $\boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$.*

This theorem shows that the test can detect the alternatives $n^{-1/2}$ distinct from the null hypothesis if the non-centrality $\boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ is not zero. In the following, we discuss the selection of $\boldsymbol{W}$.

When $q = 1$, the situation is reduced to the case similar to [22,25]. The distribution of $(\boldsymbol{T}+\boldsymbol{B})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}+\boldsymbol{B})$ is actually non-central chi-squared with the non-centrality $\boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$. When we consider the one-sided test, its power function is $\boldsymbol{G}(-c_{\alpha}/2+\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B})+\boldsymbol{G}(-c_{\alpha}/2-\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B})$ where c_{α} is the upper $(1-\alpha)$-quantile of the normal distribution. It is easy to prove that this function is a monotone function of $|\boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}|$, and then the power function is a monotone function of the non-centrality $\boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$. For the multivariate response case, we have the following lemma.

Lemma 2.1. *Under the conditions of Theorem 2.1, the power function relating to the distribution of $(\boldsymbol{T}+\boldsymbol{B})^{\tau}\boldsymbol{\Sigma}^{-1}(\boldsymbol{T}+\boldsymbol{B})$ is a monotone function of $\boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$.*

From Lemma 2.1, we can see that to enhance the power, we should select $\boldsymbol{W}$ to allow $\sum_{i=1}^{q} v_i^2 = \boldsymbol{B}^{\tau}\boldsymbol{\Sigma}^{-1}\boldsymbol{B}$ as large as possible.

Lemma 2.2. *Under the conditions of Theorem 2.1, the optimal choice of $\boldsymbol{W}$ satisfies the equation that $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V} = [\boldsymbol{E}(\boldsymbol{\eta}^2)]^{-1/2}\boldsymbol{\eta}$ where $[\boldsymbol{E}(\boldsymbol{\eta}^2)]$ is an diagonal matrix each element on* *the diagonal is* $\boldsymbol{E}(\boldsymbol{\eta}^{(i)})^2$ *where* $\boldsymbol{V} = (\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(q)})^\tau$ *and* $\boldsymbol{V}^{(i)}$ *depending on the weight* W *are defined in Theorem* 2.1.

Remark 2.1. Lemma 2.2 provides a way to search for an optimal weight through solving an equation. In a special case, the solution has a closed form. When $\boldsymbol{W}$ is orthogonal to g', $\boldsymbol{V}$ is actually equal to $\boldsymbol{W}$. If the components of $\boldsymbol{W}$ are orthogonal to one another, $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$ and has a common variance, σ^2, of all components $\boldsymbol{e}^{(i)}$, we have that $\boldsymbol{\Sigma}$ is also a diagonal matrix each element on the diagonal being $\sigma^2 E(\boldsymbol{W}^{(i)})$, $i = 1, \ldots, q$. Hence, $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{V} = \sigma^2[\boldsymbol{E}(\boldsymbol{W}^2)]^{-1/2}\boldsymbol{W} = [\boldsymbol{E}(\boldsymbol{\eta}^2)]^{-1/2}\boldsymbol{\eta}$. This means that $\boldsymbol{W}$ can be selected as $\boldsymbol{\eta}$ because σ^2 is a constant. Furthermore, $\boldsymbol{W} = \boldsymbol{\eta}$ should be a good weight function even when the above conditions on the model structure is violated because $\boldsymbol{W}$ exactly matches the directional departure $\boldsymbol{\eta}$ from the null model. The test should have good power performance. In univariate response case, a similar discussion can be seen in [22,25]. On the other hand, when we do not have much prior information on the alternatives, $\boldsymbol{\eta}$ is unknown and is not even estimable. Thus, this optimal weight cannot be used. To make use of the score test procedure, a more practically useful method is to choose it through residual plots of $\boldsymbol{Y}$ against $\boldsymbol{G}$. We will describe this residual-guided graphical method in Section 4. Clearly, it deserves further study on how to construct omnibus score type test. But this is beyond the scope of this paper.

3. Nonparametric Monte Carlo test

From Theorem 2.1 we can easily determine p-values through chi-square distributions. However, for $\mathcal{T}_n$ a deterioration for the power comes from a plug-in estimation for the covariance matrix $\boldsymbol{\Sigma} = \mathrm{Cov}(\boldsymbol{V} \bullet \boldsymbol{\varepsilon})$. This is because under the alternative $\boldsymbol{\varepsilon}$ is no longer centered and this covariance matrix will be larger than that under H_0. In this section, we propose a nonparametric Monte Carlo test (NMCT) procedure to approximate the null distribution of a test. An interesting point of this approximation is that although the test itself we will use is not scale-invariant because we do not use a plug-in estimator of the covariance matrix, NMCT can be scale-invariant. In comparison with the existing proposals for approximating the null distributions of the tests, for instance, classical bootstrap Efron [5], and their variants such as the wild bootstrap [11,19], we know that a bootstrap procedure also requests the estimation for Σ. More importantly, NMCT has a special usefulness in multivariate response cases because if the distribution of the error is of semiparametric structure, we can benefit from it to better mimic the null distribution, and if the error is fully nonparametrically distributed, NMCT is also consistent. To this end, we first give a definition about the independent decomposition for a random variable.

Definition 3.1. *A random vector* $\boldsymbol{X}$ *is said to be independently decomposable if* $\boldsymbol{X} = \boldsymbol{U} \bullet \boldsymbol{V}$ *in distribution,* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are independent and* $\boldsymbol{U} \bullet \boldsymbol{V}$ *is a dot product, that is* $\boldsymbol{U} \bullet \boldsymbol{V} = (U^{(1)}V^{(1)}, \ldots, U^{(d)}V^{(d)})$ *if both* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *are* d*-dimensional vectors, and* $\boldsymbol{U} \bullet \boldsymbol{V} = (U^{(1)}V, \ldots, U^{(d)}V)$ *if* V *is scalar, where the* $U^{(i)}$*'s are the components of* U*; similarly,* $\boldsymbol{U} \bullet \boldsymbol{V} = (UV^{(1)}, \ldots, UV^{(d)})$ *if* $\boldsymbol{U}$ *is scalar.*

There are several distribution families satisfying the independent decomposition including elliptically symmetric, reflectively symmetric, Liouville–Dirichlet and symmetric scale mixture distributions. These distributions are respectively the generalizations of normal, symmetric, Beta and stable distributions. See [9] and the references therein. To motivate our NMCT procedure, we first describe the algorithm with a simple case. When either $\boldsymbol{U}$ or $\boldsymbol{V}$ has an analytically tractable distribution, the above decomposition motivates the following NMCT procedure. Let

$\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be an *i.i.d.* sample of size n. A test statistic $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, say, can be rewritten as $T(\boldsymbol{u}_1 \bullet \boldsymbol{v}_1, \ldots, \boldsymbol{u}_n \cdot \boldsymbol{v}_n)$, if the $\boldsymbol{x}_i$'s are independently decomposable with $\boldsymbol{x}_i = \boldsymbol{u}_i \bullet \boldsymbol{v}_i$ under the null hypothesis. Then the unconditional distributions of $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ and $T(\boldsymbol{u}'_1 \bullet \boldsymbol{v}_1, \ldots, \boldsymbol{u}'_n \bullet \boldsymbol{v}_n)$ are identical, and due to the independent decomposition, when $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$ is given the conditional distribution of $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ is also identical to that of $T(\boldsymbol{u}'_1 \bullet \boldsymbol{v}_1, \ldots, \boldsymbol{u}'_n \bullet \boldsymbol{v}_n)$. These facts show that the expectation of tail probability of the conditional distribution over $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ is equal to tail probability of the unconditional distribution. Thus, the p-values based on the conditional distribution are reasonable estimators of the true p-values. Also from Theorem 3.1 below, we can see the consistency of such estimators as sample size goes to infinity. Therefore, we then propose a Monte Carlo procedure to simulate the conditional distribution of $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ when $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n)$ are given. We generate a reference set of values of the test statistic by sampling from $T(\boldsymbol{u}'_1 \bullet \boldsymbol{v}_1, \ldots, \boldsymbol{u}'_n \bullet \boldsymbol{v}_n)$, where $\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_n$ have the same distribution as that of $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$. That is, conditionally on $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n$, the Monte Carlo calculations can be performed based on $\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_n$, where $\boldsymbol{u}'_1, \ldots, \boldsymbol{u}'_n$ are drawn independently from the distribution of $\boldsymbol{U}$. The p-value of the test statistic T can be estimated as follows. Suppose that the null hypothesis will be rejected for large values of T; for two-sided tests, modifications are done easily. Let the values of T be T_0 for the original data set and $T_1, \ldots, T_m$ are obtained from the Monte Carlo procedure. The p-value is estimated as

$$\hat{p} = k/(m+1),$$

where k is the number of values in $T_0, T_1, \ldots, T_m$ that are larger than or equal to T_0. Therefore, for a given nominal level α, whenever $\hat{p} \leq \alpha$, the null hypothesis will be rejected. See for example [15,27,28].

3.1. NMCT in regression

When we apply NMCT to regression problems, some modification is needed because test statistic is based on the residuals that are not of a direct independent decomposition, and also we cannot simply simulate the residuals to approximate the null distribution because under the alternative, the conditional distribution of NMCT based on the simulated residuals is not an approximation to the null distribution, but an approximation to the distribution under the alternative. Therefore, we have to study the structure of the test first to see how NMCT should be constructed.

Recalling the format of $\mathcal{T}_n$ in Section 2, it is a standardized quadratic form of $\boldsymbol{T}_n$ with an estimator of $\boldsymbol{\Sigma}$ so that the test statistic is scale-invariant. However, to avoid the use of $\hat{\boldsymbol{\Sigma}}$, we use a test $\boldsymbol{T}_n^\tau \boldsymbol{T}_n$ with the help of NMCT to avoid the problem created by such a standardization. For the sake of notational simplicity, we only present an algorithm with elliptically symmetric distribution of the error, similarly with other classes of distributions.

- **Step 1**. Generate independent identically distributed random variables $\boldsymbol{u}_i = \boldsymbol{N}_i/\|\boldsymbol{N}_i\|, i = 1, \ldots, n$ where $\boldsymbol{N}_i$ has normal distribution $N(0, \boldsymbol{I}_q)$. Clearly, $\boldsymbol{u}_i$ is uniformly distributed on the sphere surface. Let $\boldsymbol{U}_n := \{\boldsymbol{u}_i, i = 1, \ldots, n\}$ and define the conditional counterpart of $\boldsymbol{T}_n$ as

$$\tilde{\boldsymbol{T}}_n(E_n) = \frac{1}{\sqrt{n}} \sum_{j=1}^{n} \hat{\boldsymbol{V}}_j \bullet \boldsymbol{u}_j \bullet \|\hat{\boldsymbol{e}}_j\|, \tag{3.1}$$

where $\hat{\boldsymbol{V}}_j = \left\{ \boldsymbol{W}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j) - \hat{\boldsymbol{E}}[(\boldsymbol{W}(\hat{\boldsymbol{\beta}}, \boldsymbol{X}))(g'(\hat{\boldsymbol{\beta}}, \boldsymbol{X}))^\tau] \hat{S}^{-1} (g'(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j)) \right\}$.

The resulting conditional counterpart of the test statistic $\mathcal{T}_n' = \boldsymbol{T}_n^\tau \boldsymbol{T}_n$ is

$$\mathcal{T}_n'(\boldsymbol{U}) = \left[\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n)\right]^\tau \left[\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n)\right]. \tag{3.2}$$

- **Step 2**. Generate m sets of $\boldsymbol{U}_n$, say $\boldsymbol{U}_n^{(i)}, i = 1, \ldots, m$ and get k values of $\mathcal{T}_n'(\boldsymbol{U}_n)$, say $(\mathcal{T}_n'(\boldsymbol{U}_n))^{(i)}, i = 1, \ldots, m$.
- **Step 3**. The p-value is estimated by $\hat{p} = k/(m+1)$ where k is the number of $(\mathcal{T}_n'(\boldsymbol{U}_n))^{(i)}$'s which are larger than or equal to $\mathcal{T}_n'$. Reject H_0 when $p \le \alpha$ for a designated level α.

Remark 3.1. From the above procedure, we can see that the test statistic $\mathcal{T}_n'$ does not involve $\boldsymbol{\Sigma}$, thus is different from $\mathcal{T}_n$. Also $\mathcal{T}_n'$ is not scale-invariant. However, in the Monte Carlo test procedure, we compute the conditional approximation $\mathcal{T}_n'(\boldsymbol{U}_n)$, and the p-values are computed by comparing the simulated values of the approximation with the original value of $\boldsymbol{T}_n^\tau \boldsymbol{T}_n$. Such a comparison makes any constant scalar to be eliminated and then makes no impact for the computation of p-values. The following result states the consistency of the approximation. On the other hand, note that when $\mathcal{T}_n$ is applied, the optimal weight is related to $\boldsymbol{\Sigma}$. Such an optimal weight is not used in $\mathcal{T}_n'$. However, we commented it in Remark 2.1, $\boldsymbol{W} = \boldsymbol{\eta}$ is actually a good weight and the selection by residual plot is also constructive. Thus, in our simulations, we do not use the optimal weight.

Theorem 3.1. *Assume that* $1/n \sum_{j=1}^n (\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)(\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)^\tau$ *converges to zero in probability and the conditions of Theorem 2.1 hold. Then we have that, for almost all sequences* $\{(\boldsymbol{x}_i, \boldsymbol{y}_i), i = 1, \ldots, n, \ldots,\}$, *the conditional distribution of* $\mathcal{T}_n'(\boldsymbol{U}_n)$ *converges to the limiting null distribution of* $\mathcal{T}_n$. *When* $1/n \sum_{j=1}^n (\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)(\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)^\tau$ *converges in probability to a constant matrix,* $\mathcal{T}_n'(\boldsymbol{U}_n)$ *converges in distribution to* $\mathcal{T}$ *which may have a different distribution from the limiting null distribution of* $\mathcal{T}_n$.

When the distribution of $\boldsymbol{\varepsilon}$ does not have a semiparametric structure, we can also construct a NMCT that is easy to implement. Other than **Step 1**, the others are the same. The modified **Step 1** is as follows.

- **Step** $1'$. Generate independent identically distributed random vectors $\boldsymbol{u}_i, i = 1, \ldots, n$ with bounded support and mean 0 and covariance matrix **1**. That is, all components are identical. Let $\boldsymbol{U}_n := \{\boldsymbol{u}_i, i = 1, \ldots, n\}$ and define the conditional counterpart of $\boldsymbol{T}_n$ as

$$\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n) = \frac{1}{\sqrt{n}} \sum_{j=1}^n \hat{\boldsymbol{V}}_j \bullet \boldsymbol{u}_j \bullet \hat{\boldsymbol{e}}_j, \tag{3.3}$$

where $\hat{\boldsymbol{V}}_j = \left\{\boldsymbol{W}(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j) - \hat{\boldsymbol{E}}[(\boldsymbol{W}(\hat{\boldsymbol{\beta}}, \boldsymbol{X}))(\boldsymbol{G}'(\hat{\boldsymbol{\beta}}, \boldsymbol{X}))^\tau]\hat{S}^{-1}(\boldsymbol{G}'(\hat{\boldsymbol{\beta}}, \boldsymbol{x}_j))\right\}$.

The resulting conditional counterpart of $\mathcal{T}_n'$ is

$$\mathcal{T}_n'(\boldsymbol{U}_n) = \left[\tilde{\boldsymbol{T}}_n(\boldsymbol{\varepsilon}_n)\right]^\tau \left[\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n)\right]. \tag{3.4}$$

Remark 3.2. In this algorithm, any kind of random vectors $\boldsymbol{u}_i$ satisfying the conditions in **Step** $1'$ can be used. The following theorem shows that the consistency of the conditional approximation holds. That is, the asymptotic validity is true. On the other hand, choosing an optimal distribution to generate data is of great interest. It deserves further research. However, since the distribution family in which the distribution satisfies the conditions is very large, the optimality issue is a very difficult problem, and is not even possible. Thus, it might be a way to do so when we restrict ourselves to a smaller family.

Remark 3.3. In univariate response cases, **Step 1** and **Step** $1'$ are similar to the wild bootstrap that was used by [11], and particularly to the one used by [19,26]. Specifically, **Step 1** for elliptically symmetric distribution is to generate $u_i = \pm 1$ that is a special case of **Step** $1'$. However, the major difference is as follows. With NMCT, we assign u_i in the summands when the test statistic can be asymptotically expressed as a linear statistic. This algorithm can ensure the consistency of NMCT. In contrast, for the wild bootstrap, the random variables are weighted on the residuals directly. When the underlying model is linear, these two algorithms are equivalent, however, for nonlinear models the consistency of the wild bootstrap test may not always hold. The readers can refer [23] for details. The special case is Zhu and Neuhaus [27].

Theorem 3.2. *Assume that* $1/n \sum_{j=1}^{n}(\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)(\boldsymbol{V}_j \bullet \boldsymbol{\eta}_j)^\tau$ *converges to zero in probability and the conditions in Theorem 2.1 hold. Then the conclusion of Theorem 3.1 holds.*

3.2. NMCT for regression parameters

Consider the linear model

$$\boldsymbol{Y} = \boldsymbol{\beta}^\tau \boldsymbol{X} + \boldsymbol{\varepsilon}, \tag{3.5}$$

where $\boldsymbol{\varepsilon}$ is independent of $\boldsymbol{X}$. To check whether some component of $\boldsymbol{X}$ has impact for $\boldsymbol{Y}$, we want to test the hypothesis

$$H_0: \quad \boldsymbol{\beta}_{(1)}^\tau = 0,$$

where $\boldsymbol{\beta}^\tau = (\boldsymbol{\beta}_{(1)}^\tau, \boldsymbol{\beta}_{(2)}^\tau)$, $\boldsymbol{\beta}_{(1)}^\tau$ is a $q \times l$ matrix and $\boldsymbol{\beta}_{(1)}^\tau$ is a $q \times (p-l)$ matrix. Let $\boldsymbol{x}^\tau = ((\boldsymbol{x}^{(1)})^\tau, (\boldsymbol{x}^{(2)})^\tau)$ with $(\boldsymbol{x}^{(1)})^\tau$ being a l-dimensional row vector and $(\boldsymbol{x}^{(1)})^\tau$ being a $(p-l)$-dimensional row vector. Under H_0, the model becomes

$$\boldsymbol{Y} = \boldsymbol{\beta}_{(2)}^\tau \boldsymbol{X}^{(2)} + \boldsymbol{\varepsilon}.$$

In any textbook of multivariate analysis, the likelihood ratio test: the Wilks lambda is a standard test, see, e.g. Johnson and Wichern [14]. When $\{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ is a sample, by least squares, we can separately obtain the least squares estimators $\hat{\boldsymbol{\beta}}^\tau$ and $\hat{\boldsymbol{\beta}}_{(2)}^\tau$ of $\boldsymbol{\beta}^\tau$ and $\boldsymbol{\beta}_{(2)}^\tau$ respectively. Hence two sums of squares and cross-products can be derived as

$$\hat{\boldsymbol{\Sigma}} = \sum_{j=1}^{n}(\boldsymbol{y}_j - \hat{\boldsymbol{\beta}}^\tau \boldsymbol{x}_j)(\boldsymbol{y}_j - \hat{\boldsymbol{\beta}}^\tau \boldsymbol{x}_j)^\tau;$$

$$\hat{\boldsymbol{\Sigma}}_2 = \sum_{j=1}^{n}(\boldsymbol{y}_j - \hat{\boldsymbol{\beta}}_{(2)}^\tau \boldsymbol{x}_j^{(2)})(\boldsymbol{y}_j - \hat{\boldsymbol{\beta}}_{(2)}^\tau \boldsymbol{x}_j^{(2)})^\tau.$$

A modified logarithm of the likelihood ratio test, popularly called Wilks lambda, is

$$\Lambda_n = -[n - p - 1 - 1/2(q - p + l + 1)] \ln \left(|\hat{\boldsymbol{\Sigma}}| / |\hat{\boldsymbol{\Sigma}}_2| \right). \tag{3.6}$$

Under H_0, this statistic converges to a chi-square distribution with $q(p-l)$ degrees of freedom.

Study the structure of the Wilks lambda first. Let $\mathcal{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$, $\mathcal{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, $\mathcal{X}_{(2)} = (\boldsymbol{x}_1^{(2)}, \ldots, \boldsymbol{x}_n^{(2)})$, $\mathcal{E} = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ be, respectively, the $q \times n$ response matrix, $p \times n$

and $(p-l)\times n$ covariate matrices and $q\times n$ error matrix. Note that $\hat{\boldsymbol\beta}=(\mathcal{X}\mathcal{X}^\tau)^{-1}\mathcal{X}\mathcal{Y}^\tau$ and $\hat{\boldsymbol\beta}_{(2)}=\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}\mathcal{Y}^\tau$. It is easy to obtain that

$$\mathcal{Y}-\mathcal{Y}\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}=\mathcal{E}\left[\boldsymbol I-\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}\right],$$
$$\mathcal{Y}-\mathcal{Y}\mathcal{X}_{(2)}^\tau\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}=\mathcal{E}\left[\boldsymbol I-\mathcal{X}_{(2)}^\tau\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}\right],$$

where $\boldsymbol I$ is an $n\times n$ identity matrix. From the definition of $\hat\Sigma$ and $\hat\Sigma_2$ above, we can rewrite them as

$$\begin{aligned}\hat{\boldsymbol\Sigma}&=\left[\mathcal{Y}-\mathcal{Y}\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}\right]\left[\mathcal{Y}-\mathcal{Y}\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}\right]^\tau\\&=\mathcal{E}\left[\boldsymbol I-\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}\right]\mathcal{E}^\tau;\\\hat{\boldsymbol\Sigma}_2&=\left[\mathcal{Y}-\mathcal{Y}\mathcal{X}_{(2)}^\tau\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}\right]\left[\mathcal{Y}-\mathcal{Y}\mathcal{X}_{(2)}^\tau\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}\right]^\tau\\&=\mathcal{E}\left[\boldsymbol I-\mathcal{X}_{(2)}^\tau\left(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau\right)^{-1}\mathcal{X}_{(2)}\right]\mathcal{E}^\tau.\end{aligned}$$

From these two formulae, we now define a NMCT. Like that in Section 3.1, we generate $q\times n$ random matrix $\mathcal{U}_n=(\boldsymbol u_1,\ldots,\boldsymbol u_n)$, and define

$$\hat{\boldsymbol\Sigma}(\mathcal{U}_n)=\left(\mathcal{U}_n\bullet\hat{\mathcal{E}}\right)\left[\boldsymbol I-\mathcal{X}^\tau\left(\mathcal{X}\mathcal{X}^\tau\right)^{-1}\mathcal{X}\right]\left(\mathcal{U}_n\bullet\hat{\mathcal{E}}\right)^\tau;$$
$$\hat{\boldsymbol\Sigma}_2(\mathcal{U}_n)=\left(\mathcal{U}_n\bullet\hat{\mathcal{E}}\right)\left[\boldsymbol I-\mathcal{X}_{(2)}^\tau(\mathcal{X}_{(2)}\mathcal{X}_{(2)}^\tau)^{-1}\mathcal{X}_{(2)}\right](\mathcal{U}_n\bullet\hat{\mathcal{E}})^\tau.$$

Repeat this step m times to generate m values of $\Lambda_n(\mathcal{U}_n)=-[n-p-1-1/2(q-p+l+1)]\ln\left(|\hat{\boldsymbol\Sigma}(\mathcal{U}_n)|/|\hat{\boldsymbol\Sigma}_2(\mathcal{U}_n)|\right)$, say $\Lambda_n(\mathcal{U}_n^{(1)}),\ldots,\Lambda_n(\mathcal{U}_n^{(m)})$; and count the number k of $\Lambda_n(\mathcal{U}_n^{(i)})$'s which are greater than or equal to Λ_n to obtain the estimated p-value $k/(m+1)$.

Similar to Theorem 3.1, we have the asymptotic equivalence between $\Lambda_n(\mathcal{U}_n))$ and Λ_n.

Theorem 3.3. *Assume that the fourth moment of $\boldsymbol X$ and $\boldsymbol Y$ exists. Then for almost all sequences $\{(\boldsymbol x_1,\boldsymbol y_1),\ldots,(\boldsymbol x_n,\boldsymbol y_n)\}$, the conditional distribution $\Lambda_n(\mathcal{U}_n)$ converges to the limit distribution of Λ_n.*

4. Simulations and application

In this section, we include three simulated examples and the application to a real dataset. The first two examples are for model checking and the third example is for diagnostic checking for regression parameters in the multivariate linear model. In all the cases, the sample sizes were $n=20,40,60$, the experiments were repeated 1000 times to compute the power of the tests and for NMCT, 1000 reference datasets were generated.

Example 1. The model is with continuous response, namely

$$\boldsymbol Y=(\boldsymbol\beta^\tau\boldsymbol X)+c\boldsymbol X^2+\boldsymbol\varepsilon, \tag{4.1}$$

where $\boldsymbol Y$ is q-dimensional and $\boldsymbol X$ p-dimensional, $\boldsymbol X$ and $\boldsymbol\varepsilon$ are independent, and $\boldsymbol X$ is multivariate normal $N(0,\boldsymbol I_p)$. To check the performance of NMCT procedure, we considered three

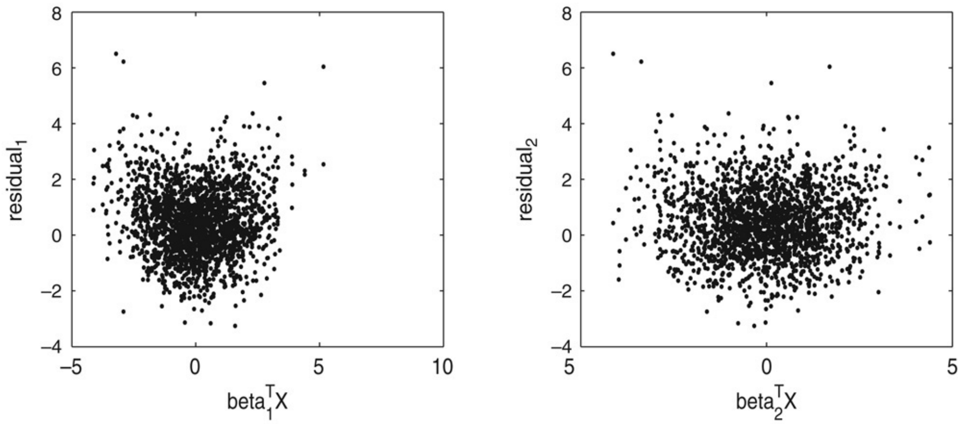

Fig. 1. The plots of residuals $Y_i - \beta_i^\tau X$ $i = 1, 2$ against the fitted linear models $\beta_i^\tau X$ with model (4.1) when $c = 0.5$.

Table 1
The empirical size of the tests in the examples at the normal level 0.05

Example 1	NMCT			Score test		
n	Normal	chi-square	Uniform	Normal	chi-square	Uniform
20	0.064	0.070	0.061	0.026	0.073	0.030
40	0.060	0.056	0.062	0.038	0.104	0.029
60	0.054	0.060	0.065	0.044	0.116	0.044
Example 2	NMCT			Score test		
n	Normal	chi-square	Uniform	Normal	chi-square	Uniform
20	0.067	0.065	0.075	0.0350	0.082	0.042
40	0.058	0.060	0.064	0.0390	0.078	0.060
60	0.051	0.056	0.052	0.0500	0.084	0.048
Example 3	NMCT			Wilks' lambda		
n	Normal	chi-square	Uniform	Normal	chi-square	Uniform
20	0.061	0.057	0.058	0.057	0.057	0.046
40	0.060	0.056	0.054	0.058	0.052	0.057
60	0.054	0.060	0.053	0.044	0.058	0.045

distributions of the error $\boldsymbol{\varepsilon}$: $N(0, I_q)$, normal; $U_q(-0.5, 0.5)$, uniform on the cube $(-0.5, 0.5)^q$, and $\chi_q^2(1)$ all components following chi-square with degree of freedom 1 respectively. The null hypothesis is $H_0 : \boldsymbol{m}(\boldsymbol{X}) = \boldsymbol{\beta}^\tau \boldsymbol{X}$. Therefore the null model holds if and only if $c = 0$. In the simulation, we considered $c = 0, 0.1, 0.2, \ldots, 1$ and $p = 3$ and $q = 2$ and the matrix $\boldsymbol{\beta} = [1, 0; 1, 1; 0, 1]$.

When we regard the alternatives as directional ones, the weight function can be selected as $\boldsymbol{X}^2$. As we discussed before, the residual plots of $\boldsymbol{\varepsilon}$ against $\boldsymbol{X}$ is also informative. We plotted all the components of $\boldsymbol{\varepsilon}$ against all the components of $\boldsymbol{X}$, and associated linear combinations $\beta_i^\tau \boldsymbol{X}$. In Fig. 1, we only report the plots of the residuals against $\beta_i^\tau \boldsymbol{X}$ with 300 generated data points. The plots indicate a pattern of quadratic curve and then suggest $\boldsymbol{W}(\boldsymbol{X}) = \boldsymbol{X}^2$ as a weight function.

Since this is the first research work with multivariate responses in this area, there are no other competitors in the literature, we compared the power performance of the score type test in

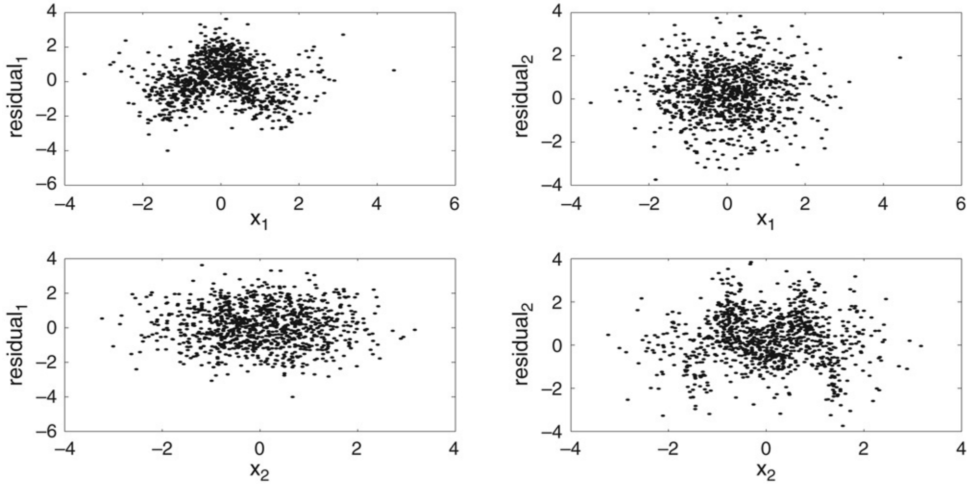

Fig. 2. The plots of residuals $Y_i - \beta_i^\tau X$ $i = 1, 2$ against each covariable x_i with model (4.2) when $c = 0.5$.

Section 2 and NMCT. From Table 1, we can see that although NMCT is a little bit conservative, it can better maintain the significance level while the score type test has difficulty to do so. From Fig. 3, we can clearly see that NMCT can gain the power much faster, while departing from the null, than the score type test does, especially when the sample size is $n = 20$. In all the cases with different error distributions, NMCT performs better than the limit distribution although when sample size is large, the powers with these two testing procedures are very close one another.

Example 2. The model is as

$$Y_1 = \beta_1^\tau \boldsymbol{X} - 2 + 2\cos(c\pi x_1) + \varepsilon_1$$
$$Y_2 = \beta_2^\tau \boldsymbol{X} + 2\sin(c\pi x_2^2) + \varepsilon_2, \tag{4.2}$$

where $\boldsymbol{Y} = (Y_1, Y_2)^\tau$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2)^\tau$ is independent of $\boldsymbol{X} = (x_1, x_2, x_3)^\tau$. $\boldsymbol{X}$ is normally distributed as $N(0, I_p)$. Let $p = 3$ and in this example, $q = 2$. Like that of Example 1, $c = 0$ corresponds to the hypothetical model, that is, the linear model. Again $\boldsymbol{\beta} = (\beta_1, \beta_2) = [1, 0; 1, 1; 0, 1]$, and $c = 0.1, 0.2, \ldots, 1$ were for power study. The three distributions of the error $\boldsymbol{\varepsilon}$ were the same as those in Example 1.

Clearly, we cannot consider this model with a directional alternative because c is inside the sine and cosine functions and with different c, the alternative regression function is rather different. Thus, we used the residual plots to provide idea for selecting the weight function. We used 300 generated data points from model (4.2) with $c = .5$ to obtain the residual plots presented in Fig. 2, we can clearly see the patterns of periodicity of the function. Hence, we used $\boldsymbol{W}(\boldsymbol{X}) = (\cos(\pi x_1), \sin(\pi x_2^2))^\tau$.

The results in Table 1 show that the score type test, especially in chi-square case, is rather conservative in both of Examples 1 and 2, and Fig. 4 suggests that even though it cannot gain higher power. Also when c is large the model becomes of high frequency, such an alternative is very difficult to detect. But compared with the score type test, NMCT also outperforms in all of the conducted cases.

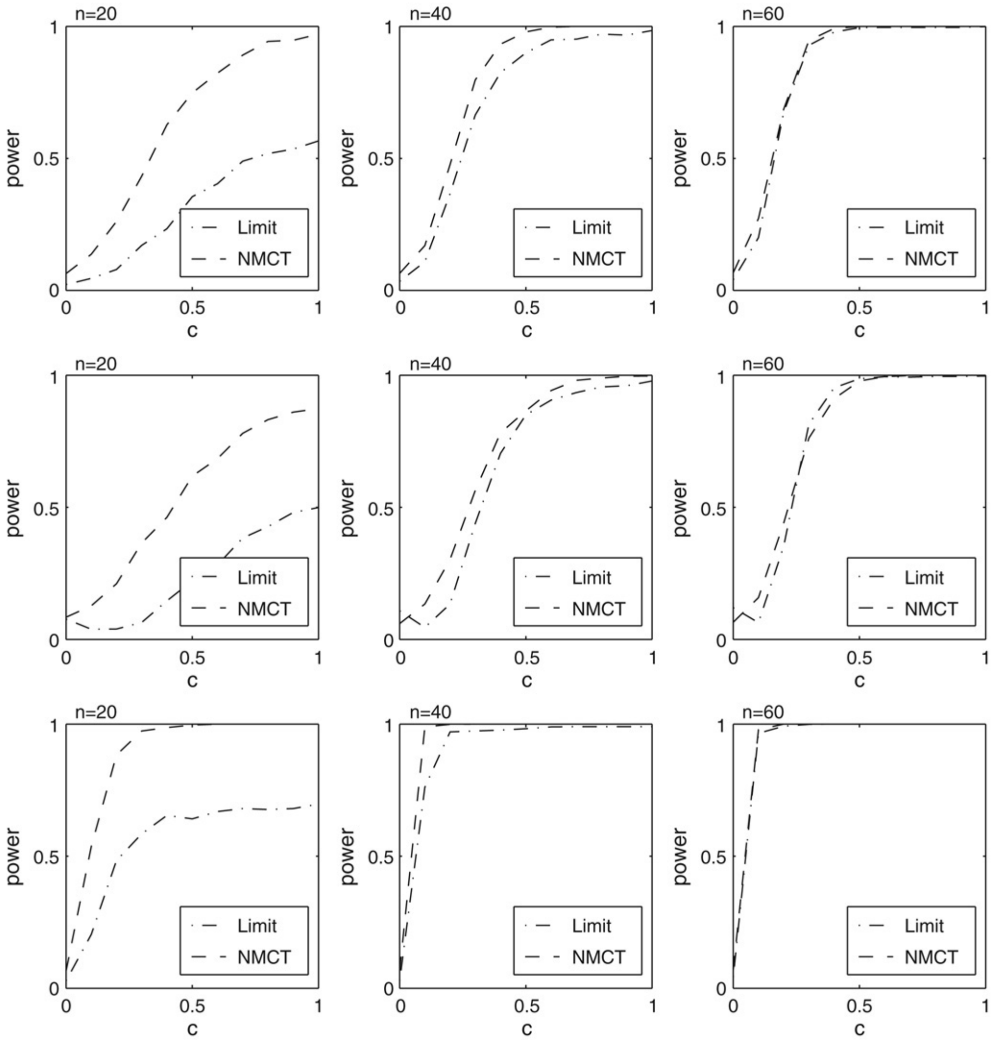

Fig. 3. The plots of the power against the departure with c for testing model (4.1). The first row is for normal error and the second row for chi-square error and the third row for uniform error.

Example 3. Consider the linear model as

$$\boldsymbol{Y} = c(\boldsymbol{\beta}_{(1)}\boldsymbol{X}^{(1)}) + (\boldsymbol{\beta}_{(2)}^{\tau}\boldsymbol{X}^{(2)}) + \boldsymbol{\varepsilon}, \tag{4.3}$$

where $\boldsymbol{Y}$ is q-dimensional, $\boldsymbol{X} = (\boldsymbol{X}^{(1)}, (\boldsymbol{X}^{(2)})^{\tau})^{\tau}$ where $\boldsymbol{X}^{(1)}$ is one-dimensional, $\boldsymbol{X}^{(2)}$ is $(p-1)$-dimensional independently of $\boldsymbol{\varepsilon}$, $\boldsymbol{X}$ is multivariate normal $N(0, \boldsymbol{I}_p)$. The three distributions of the error in Example 1 were considered. The hypothetical regression function is $\boldsymbol{\beta}_{(2)}^{\tau}\boldsymbol{X}^{(2)}$. Therefore the null model corresponds to $c = 0$. In the simulation, $\boldsymbol{\beta}_{(2)} = (1, 1; 0, 1)$ and $\boldsymbol{\beta}_1 = (1, 0)$. Again Table 1 provides the results that show a slightly more conservativeness of NMCT than the Wilks' lambda. Fig. 5 reports the power. We can find that NMCT outperforms the Wilks' lambda in all of the cases especially when sample size is small.

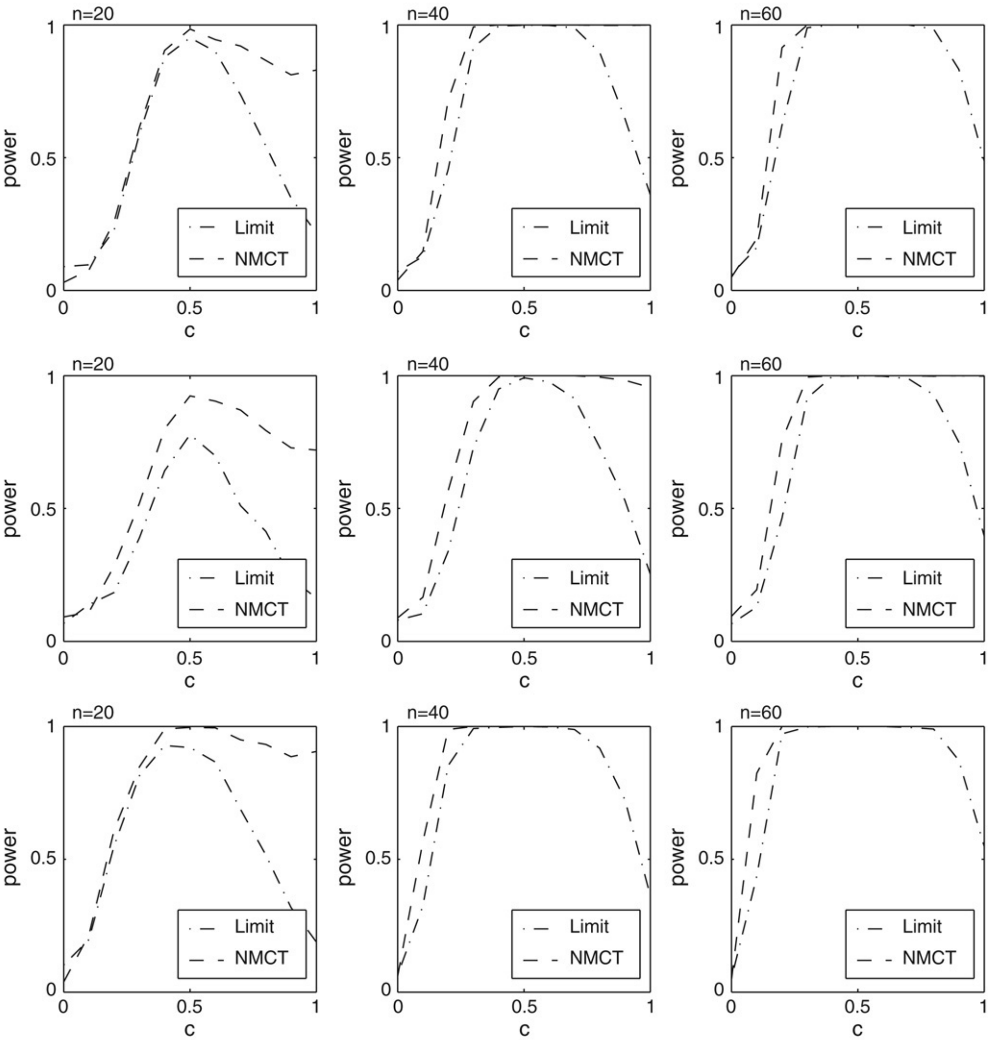

Fig. 4. The plots of the power against the departure with c for testing model (4.2). The first row is for normal error and the second row for chi-square error and the third row for uniform error.

We also did some simulations with other distributions such as Student $t(5)$. Since the results are very similar to those with normal distribution, we then did not report them in this paper.

Application

The 1984 Olympic records data on various track events were collected as reported by [14]. For a relevant dataset of Women's track records, [4] used principal component analysis to study the athletic excellence of a given nation and the relative strength of the nation at the various running distances. Zhu [23] studied the relation between the performance of a nation in long running distance and short running distance. For 55 countries winning times for women's running events at 100, 200, 400, 800, 1500, 3000 m and the Marathon distance were reported in, say, [14]. Now we want to know whether women of a nation whose performance is better in running long

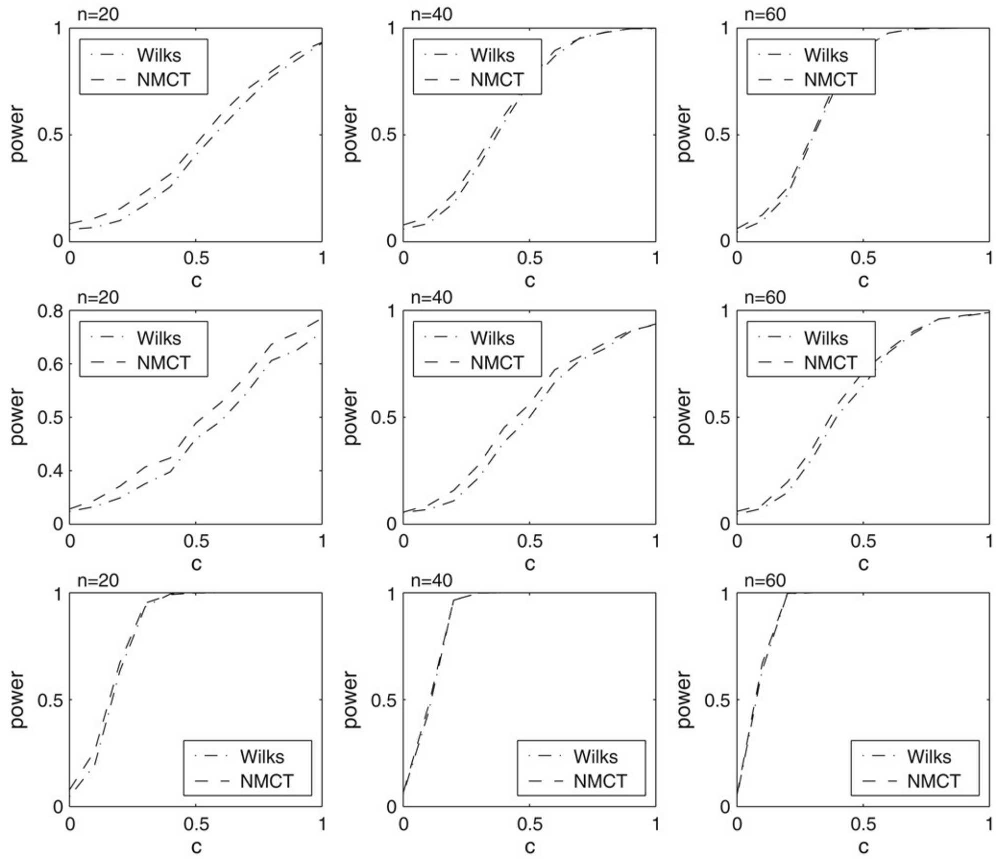

Fig. 5. The plots of the power against the departure with c for model (4.3). The first row is for normal error and the second row for chi-square error and the third row for uniform error.

distances may also have greater strength at short running distances. To make the analysis more reasonable, the winning time is transformed to speed. Let these speeds be $x_1, \ldots, x_7$. We regard 100, 200 and 400 m as short running distances, 1500 m and longer as long running distances. The hypothetical model is linear by considering the speed of the 100, 200 and 400 m running events (x_1, x_2, x_3) as the covariates and the speed of the 1500, 3000 m and the Marathon running events (Y_1, Y_2, Y_3) as covariates.

To test the linearity, we used the proposed test $\mathcal{T}_n$ in Section 2 and NMCT associated with $\mathcal{T}_n'$ in Section 3. For NMCT, we assumed two cases: the error follows an elliptically symmetric distribution and a general distribution respectively. Therefore, we used respective algorithms to construct NMCT statistics $\mathcal{T}_n'(U_n)$ and $\mathcal{T}_n'(E_n)$ as reported in Section 3. From Fig. 6, we found that the nonlinearity may be mainly from Y_3, the Marathon. There might be quadratic curves in the plots of Y_3 against X_i, $i = 1, 2, 3$. Hence, we chose X_3^2 as a weight function W. With these three tests, the p-values are, respectively, 0.09 with $\mathcal{T}_n$; 0.0001 with $\mathcal{T}_n'(U_n)$ under the elliptical distributional assumption and 0.03 with $\mathcal{T}_n'(U_n)$ under a general distribution assumption. Clearly, NMCT suggests a rejection for a linear model. Furthermore, the quadratic curves of Y_3 against X_i, $i = 1, 2, 3$ implies that the nation with either great strength or weak

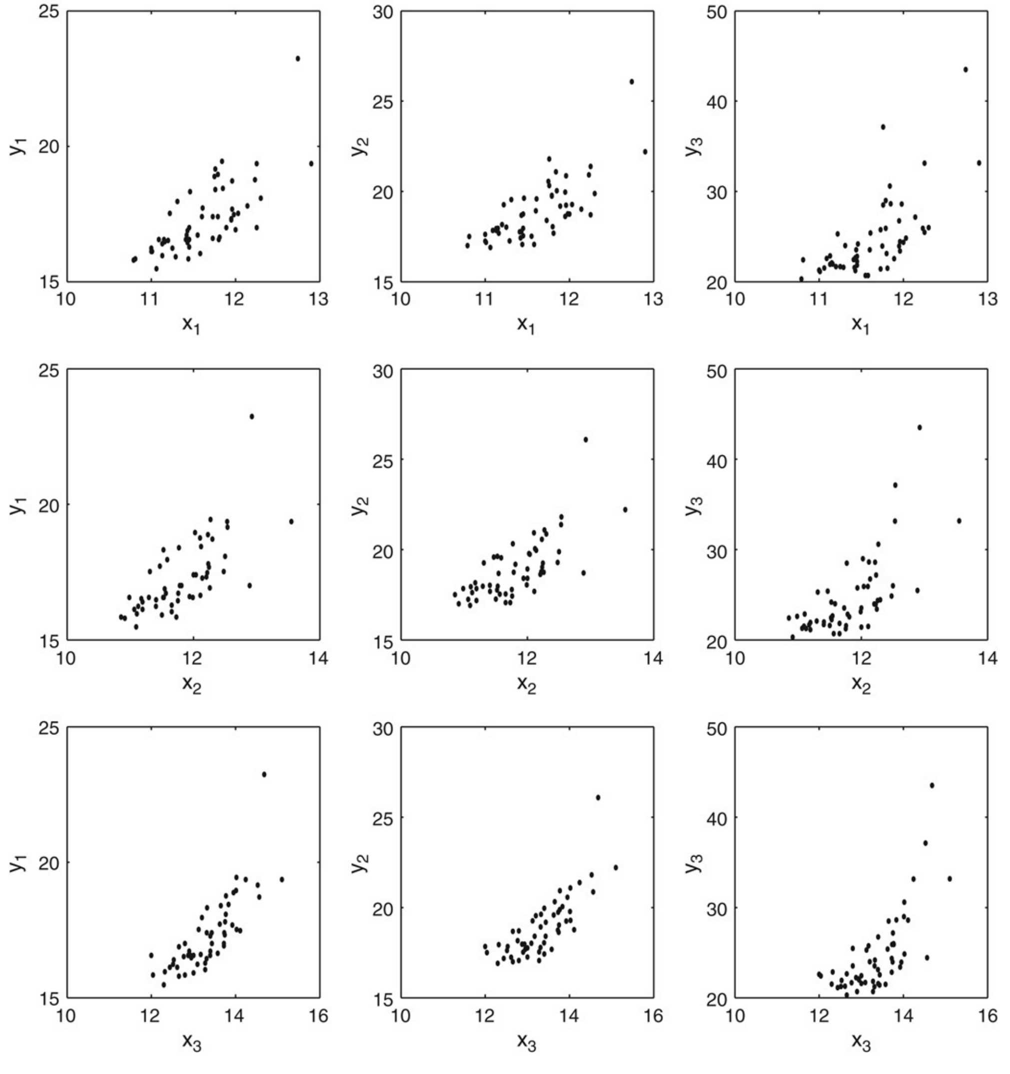

Fig. 6. The plots of the responses Y_1, Y_2, Y_3 against the covariates X_1, X_2, X_3 for the 1984 Olympic records data.

strength at short running distance may not have good performance in running the Marathon. We then fit a model linearly with X_1, X_2, X_3 and X_3^2. The p-values with the three tests are: 0.97 with $\mathcal{T}_n$; 0.34 with $\mathcal{T}_n'(U_n)$ under the elliptical distribution assumption and 0.99 with $\mathcal{T}_n'(U_n)$ under a general distributional assumption. These tests strongly suggested the tenability of the model with a two-order polynomial of X_3.

Let us turn to the classical testing problem with likelihood ratio test. First, we note that the speeds of 100 and 200 m are greatly correlated with the correlation coefficient 0.9528. Therefore, we transfer the variables by regressing X_2 on X_1 to obtain $\hat{X}_2 = 1.1492X_1$ and to get $\tilde{X}_1 = X_2 - 1.1492X_1$, $\tilde{X}_2 = X_2 + 1.1492X_1$. The new model is

$$\boldsymbol{Y} = \mathbf{a}\tilde{X}_1 + \mathbf{b}\tilde{X}_2 + \mathbf{d}X_3 + \mathbf{c}X_3^2 + \boldsymbol{\varepsilon}. \tag{4.4}$$

The purpose was to test whether $\tilde{X}_1$ has a significant impact for $\boldsymbol{Y}$: that is, the coefficient $a = 0$ or not. The p-values are: 0.08 for Wilks Lambda; 0.20 for NMCT with uniformly distributed weights and 0.38 for NMCT with normally distributed weights. All the three tests suggest that $\tilde{X}_1$ has less impact for $\boldsymbol{Y}$. Hence we can use a model as

$$\boldsymbol{Y} = \mathbf{b}\tilde{X}_2 + \mathbf{d}X_3 + \mathbf{c}X_3^2 + \boldsymbol{\varepsilon}.$$

to establish the relationship between $\boldsymbol{Y}$ and $\tilde{X}_2$, X_3, X_3^2.

Acknowledgments

The research was supported by a grant from the Research Grants Council of Hong Kong. The first author was in charge of the methodology and the second was in charge of the majority of the simulation results, and the third did part of the simulations. The authors thank the associate editor and referees for providing constructive comments and suggestions which led to an improvement of the presentation.

Appendix

Proof of Lemma 2.1. Note that $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{T} + \boldsymbol{S})$ is normally distributed as $N(\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}, \boldsymbol{I}_q)$ where $\boldsymbol{I}_q$ is a $q \times q$ identity matrix. Hence, the components, say, $u_i i = 1, \ldots, q$, of $\boldsymbol{\Sigma}^{-1/2}(\boldsymbol{T} + \boldsymbol{S})$ are independent normal with the mean v_i and variance 1 where v_i's are the components of $\boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}$. Therefore, $(\boldsymbol{T}+\boldsymbol{S})^\tau \boldsymbol{\Sigma}^{-1}(\boldsymbol{T}+\boldsymbol{S})$ can be written as the sum of independent non-central chi-squared variables $\sum_{i=1}^q (u_i + v_i)^2$, each has the non-centrality v_i^2. From the univariate response case, we know that $P\{(u_i + v_i)^2 \geq c\}$ for any $c > 0$ is smaller when $|v_i|$ gets larger. Consider $q = 2$. Note that the distribution of $(u_1 + v_1)^2 + (u_2 + v_2)^2$ is a convolution of two distributions each decreasing according to smaller value v_i respectively. First note that $P\{(u_1+v_1)^2+(u_2+v_2)^2 > c\} = \int (1 - F_{1,v_1}(c - x_2))\mathrm{d}F_{2,v_2}(x_2) = \int (1 - F_{2,v_2}(c - x_1))\mathrm{d}F_{1,v_1}(x_1)$. Then for any pairs (v_1, v_2) and (v_1', v_2') with $|v_i| \geq |v_i'|$, because of the independence between u_1 and u_2, we derive that

$$\begin{aligned}
P\{(u_1 + v_1)^2 + (u_2 + v_2)^2 > c\} &= \int (1 - F_{1,v_1}(c - x_2))\mathrm{d}F_{2,v_2}(x_2) \\
&\geq \int (1 - F_{1,v_1'}(c - x_2))\mathrm{d}F_{2,v_2}(x_2) = \int (1 - F_{2,v_2}(c - x_1))\mathrm{d}F_{1,v_1'}(x_1) \\
&\geq \int (1 - F_{2,v_2'}(c - x_2))\mathrm{d}F_{1,v_1'}(x_2) = P\{(u_1 + v_1')^2 + (u_2 + v_2')^2 > c\}.
\end{aligned}$$

When we use induction, the same can apply to prove the general case, we omit the details. □

Proof of Lemma 2.2. Let $\boldsymbol{V}' = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{V} =: ((\boldsymbol{V}')^{(1)}, \ldots, (\boldsymbol{V}')^{(q)})$. Since $S_i = E(\boldsymbol{V}^{(i)}\boldsymbol{\eta}^{(i)})$, then $\sum_{i=1}^q v_i^2 = \sum_{i=1}^q E[(\boldsymbol{V}')^{(i)}\boldsymbol{\eta}^{(i)}]^2$. Invoking the Cauchy–Schwarz inequality and the fact that $E[(\boldsymbol{V}')^{(i)}]^2 = 1$, we obtain that $\boldsymbol{B}^\tau \boldsymbol{\Sigma}^{-1}\boldsymbol{B} \leq \sum_{i=1}^q (E[\boldsymbol{\eta}^{(i)}])^2$ and the equality holds if and only if $(\boldsymbol{V}')^{(i)} = \boldsymbol{\eta}^{(i)}/\sqrt{(E(\boldsymbol{\eta}^{(i)})^2)}$. □

Proof of Theorem 3.1. First note that under the null,

$$\tilde{\boldsymbol{T}}_n = \frac{1}{\sqrt{n}} \sum_{j=1}^n \boldsymbol{V}_j \bullet \boldsymbol{e}_j + o_p(1). \tag{5.1}$$

It is easy to see that when the sequence of $\{(\boldsymbol{x}_i, \|\boldsymbol{\varepsilon}_i\|), i = 1, \ldots, n, \ldots, \}$ is given, $\frac{1}{\sqrt{n}}\sum_{j=1}^n \boldsymbol{V}_j \bullet \boldsymbol{e}_j$ has the same distribution as $\frac{1}{\sqrt{n}}\sum_{j=1}^n \boldsymbol{V}_j \bullet \boldsymbol{u}_j \bullet \|\boldsymbol{e}_j\|$ because $\boldsymbol{\varepsilon}_j/\|\boldsymbol{\varepsilon}_j\|$ is independent of $\|\boldsymbol{\varepsilon}_j\|$, and the distribution of $\boldsymbol{\varepsilon}_j/\|\boldsymbol{\varepsilon}_j\|$ is identical to that of $\boldsymbol{u}_j$ and so do the distributions of the associated unconditional counterparts. This implies that the limit distribution of $\tilde{\boldsymbol{T}}_n$ is the same as that of $\frac{1}{\sqrt{n}}\sum_{j=1}^n \boldsymbol{V}_j \bullet \boldsymbol{u}_j \bullet \|\boldsymbol{e}_j\|$. Note that this is in turn asymptotically equivalent to $\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n)$. The proof can be done as follows.

Note that in

$$\tilde{\boldsymbol{T}}_n(\boldsymbol{U}_n) = \frac{1}{\sqrt{n}}\sum_{j=1}^n \hat{\boldsymbol{V}}_j \bullet \boldsymbol{u}_j \bullet \|\hat{\boldsymbol{e}}_j\|, \tag{5.2}$$

$\boldsymbol{u}_j$ are independent of $\{(\boldsymbol{x}_j, \boldsymbol{y}_j),\ j = 1, \ldots, n\}$, and all estimators involved in $\hat{\boldsymbol{V}}_j$ and $\hat{\boldsymbol{e}}_j$ are consistent. Then we can easily derive that, by Taylor expansion,

$$\frac{1}{\sqrt{n}}\sum_{j=1}^n \left[\hat{\boldsymbol{V}}_j \bullet \boldsymbol{u}_j \bullet \|\hat{\boldsymbol{e}}_j\| - \boldsymbol{V}_j \bullet \boldsymbol{u}_j \bullet \|\boldsymbol{e}_j\|\right] = \frac{1}{\sqrt{n}}\sum_{j=1}^n \boldsymbol{u}_j \bullet \left[\hat{\boldsymbol{V}}_j \bullet \|\hat{\boldsymbol{e}}_j\| - \boldsymbol{V}_j \bullet \|\boldsymbol{e}_j\|\right]$$

in probability. The proof is finished. □

Proof of Theorem 3.2. To prove the result, we only need to prove two things: asymptotic normality of $\tilde{\mathbf{T}}_n(\boldsymbol{U}_n)$ and the identical of the covariance matrix to the limiting covariance of $\tilde{\boldsymbol{T}}_n$. Note that $\boldsymbol{u}_j$ are independent of $\{(\boldsymbol{x}_j, \boldsymbol{y}_j),\ j = 1, \ldots, n\}$. Then, when $\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}$ are given, the covariance matrix of $\tilde{\mathbf{T}}_n(\boldsymbol{U}_n)$ is $1/n\sum_{j=1}^n(\hat{\boldsymbol{V}}_j \bullet \hat{\boldsymbol{\varepsilon}}_j)(\hat{\boldsymbol{V}}_j \bullet \hat{\boldsymbol{\varepsilon}}_j)^\tau$. By the consistency of the estimators involved, Taylor expansion and the weak law of large numbers, it is easy to see that this sum converges to $\boldsymbol{E}[(\boldsymbol{V} \bullet \boldsymbol{\varepsilon})(\boldsymbol{V} \bullet \boldsymbol{\varepsilon})^\tau]$. This is just the limiting covariance of $\tilde{\boldsymbol{T}}_n$. As for the asymptotic normality, we only need to note that when $\{(\boldsymbol{x}_j, \boldsymbol{y}_j)\}$ are given, $\tilde{\mathbf{T}}_n(\boldsymbol{U}_n)$ is a sum of $i.i.d.$ random vectors. Combining central limit theorems (see [16]), we can verify that the limit is distributed as $N(0, \boldsymbol{E}[(\boldsymbol{V} \bullet \boldsymbol{\varepsilon})(\boldsymbol{V} \bullet \boldsymbol{\varepsilon})^\tau])$. This is identical to the limit distribution of $\tilde{\boldsymbol{T}}_n$. □

Proof of Theorem 3.3. The proof is almost the same as those for the previous theorems. We omit the details. □

References

[1] A. Azzalini, A.W. Bowman, W. Härdle, On the use of nonparametric regression for model checking, Biometrika 76 (1989) 1–11.

[2] R. Beran, G.R. Ducharme, Asymptotic Theory for Bootstrap Methods in Statistics, Universit de Montral, Centre de Recherches Mathmatiques, Montreal, QC, 1991 viii+81 pp.

[3] D.D. Cox, E. Koh, G. Wahba, B.S. Yandell, Testing the (parametric) null model hypothesis in (semiparametric) partial and generalized spline models, Ann. Statist. 16 (1988) 113–119.

[4] B. Dawkins, Multivariate analysis of national track records, Amer. Statist. 43 (1989) 110–112.

[5] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Statist. 7 (1979) 1–26.

[6] R.L. Eubank, V.N. LaRiccia, Testing for no effect in nonparametric regression, J. Statist. Plann. Inference 36 (1993) 1–14.

[7] R.L. Eubank, C.H. Spiegelman, Testing the goodness of fit of a linear model via nonparametric regression techniques, J. Amer. Statist. Assoc. 85 (1990) 387–392.

[8] J. Fan, L. Huang, Goodness-of-fit tests for parametric regression models, J. Amer. Statist. Assoc. 96 (2001) 674–688.

[9] K.T. Fang, S. Kotz, K.W. Ng, Symmetric Multivariate and Related Distributions, Chapman and Hall, London, 1990.

[10] C. Gu, Diagnostics for nonparametric regression models with additive terms, J. Amer. Statist. Assoc. 87 (1992) 1051–1058.

[11] W. Härdle, E. Mammen, Comparing non-parametric versus parametric regression fits, Ann. Statist. 21 (1993) 1926–1947.
[12] W. Härdle, E. Mammen, M. Müller, Testing parametric versus semiparametric modeling in generalized linear models, J. Amer. Statist. Assoc. 93 (1998) 1461–1474.
[13] J.D. Hart, Nonparametric Smoothing and Lack-of-fit Tests, Springer, New York, 1997.
[14] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall International, USA, 1998.
[15] G. Neuhaus, L.X. Zhu, Permutation tests for reflected symmetry, J. Multivariate Anal. 67 (1998) 129–53.
[16] D. Pollard, Convergence of Stochastic Processes, Springer, New York, 1984.
[17] J.S. Simonoff, C.L. Tsai, Assessing the influence of individual observations on a goodness-of-fit test based on nonparametric regression, Statist. Probab. Lett. 12 (1991) 9–17.
[18] W. Stute, Nonparametric model checks for regression, Ann. Statist. 25 (1997) 613–641.
[19] W. Stute, W. González Manteiga, M. Presedo Quindimil, Bootstrap approximations in model checks for regression, J. Amer. Statist. Assoc. 93 (1998) 141–149.
[20] W. Stute, S. Thies, L.X. Zhu, Model checks for regression: An innovation process approach, Ann. Statist. 26 (1998) 1916–1934.
[21] W. Stute, L.X. Zhu, Model checks for generalized linear models, Scand. J. Statist. 29 (2002) 535–546.
[22] W. Stute, L.X. Zhu, Nonparametric checks for single-index models, Ann. Statist. 33 (2005) 1048–1083.
[23] L.X. Zhu, Model checking of dimension-reduction type for regression, Statist. Sinica 13 (2003) 283–296.
[24] L.X. Zhu, Nonparametric Monte Carlo Tests and Their Applications, Springer, New York, 2005.
[25] L.X. Zhu, H.J. Cui, Tsting the adequacy for a general linear errors-in-varibles model, Statist. Sinica 15 (2005) 1049–1068.
[26] L.X. Zhu, K.T. Fang, I.M. Bhatti, On estimated projection pursuit type cramer-von mises statistics, J. Multivariate Anal. 63 (1997) 1–15.
[27] L.X. Zhu, G. Neuhaus, Nonparametric Monte Carlo test for multivariate distributions, Biometrika 87 (2000) 919–928.
[28] L.X. Zhu, K.W. Ng, Checking the adequacy of a partial linear model, Statist. Sinica 13 (2003) 763–781.