Statistica Sinica **34** (2024), 2265-2275
doi:https://doi.org/10.5705/ss.202022.0068

CLT FOR HIGH-DIMENSIONAL R^2 STATISTICS UNDER A GENERAL INDEPENDENT COMPONENTS MODEL

Weiming Li and Shizhe Hong*

Shanghai University of Finance and Economics

Abstract: This study establishes a central limit theorem (CLT) for R^2 statistics in a moderately high-dimensional asymptotic framework. The underlying population accommodates a general independent components model, by which our result unifies two existing CLTs. Beyond this, the new CLT characterizes the effect of kurtosis of the latent independent components on the fluctuation of R^2 statistics. As an application, a novel confidence interval is constructed for the coefficient of multiple correlation in a high-dimensional linear regression.

Key words and phrases: High dimension, independent components model, multiple correlation coefficient.

1. Introduction

The coefficient of multiple correlation ρ_p measures the linear dependence between a scalar random variable y and a set of variables $x_1, \ldots, x_p$. It maximizes the Pearson correlation between y and any linear combination of $\mathbf{x} = (x_1, \ldots, x_p)'$, that is,

$$\rho_p = \rho(y, \mathbf{x}) \triangleq \max_{\alpha \in \mathbb{R}^p} \mathrm{Cor}\,(y, \alpha'\mathbf{x})\,; \tag{1.1}$$

see Anderson (2003).

The R^2 statistic, or squared sample multiple correlation coefficient, is by definition the moment estimator of ρ_p^2. Under Gaussian distributions, its exact distribution is derived by Fisher (1928). Additional discussions on this sampling distribution can be found in Wilks (1932), Gurland (1968), Lee (1971), Williams (1978), and Nandi and Choudhury (2005). Under general populations, numerous works have examined the asymptotic behavior of R^2 in a low-dimensional asymptotic regime, where the dimension p of the observations is fixed, while the sample size n tends to infinity; see for instance, Muirhead (1982), Ali and Nagar (2002), Anderson (2003), and Ogasawara (2006).

When the dimension p is non-negligible with respect to the sample size n, the distribution of R^2 deviates from its predicted limit in low-dimensional situations. First, consider a moderately high-dimensional framework, that is,

$$n \to \infty, \quad p = p_n \to \infty, \quad \frac{p}{n} \to c \in (0, 1), \tag{1.2}$$

*Corresponding author.

which is commonly used in the literature on *random matrix theory*, and is referred to as the *Marčenko–Pastur* (MP) asymptotic regime (Marčenko and Pastur (1967)). In this regime, Zheng et al. (2014) prove that the R^2 statistic converges to $c+(1-c)\rho^2$, almost surely, where ρ^2 denotes the limit of ρ_p^2 as $p \to \infty$. Moreover, under a specific *independent components* (IC) model (Bai and Silverstein (2004)), the R^2 statistic is asymptotically Gaussian, with a limiting variance determined jointly by the limit ρ^2 and the ratio c. Similar results are reported by Guo and Cheng (2022), who studied the R^2 statistic in a high-dimensional linear regression. However, note that the models considered in Zheng et al. (2014) and Guo and Cheng (2022), as well as their corresponding results, overlap, but not entirely. Therefore, we need to study the R^2 statistic under more general situations and provide a unified limiting theory.

The main contribution of this study is a unified *central limit theorem* (CLT) for the R^2 statistic, established under a general IC model (Bai and Silverstein (2010)) in the MP asymptotic regime (1.2). Our results show that the R^2 statistic converges in distribution to a Gaussian variable, the variance of which is a function of the limiting ratio c, the whole dependence structure of $(y, x_1, \ldots, x_p)$, and the fourth moments of their latent independent components. By specifying the structure of the dependence and/or the fourth moments, our CLT reduces to those in Zheng et al. (2014) and Guo and Cheng (2022). In general cases, the CLT represents the moment contribution of the latent components to the fluctuation of R^2. As an application, we develop a novel interval estimation procedure for the multiple correlation coefficient in a high-dimensional linear regression.

The rest of the paper is organized as follows. Section 2 details our model assumptions and presents the new CLT for R^2 statistics. Section 3 proposes our interval estimation of multiple correlation coefficients, which is then applied to an empirical analysis of a breast cancer data set. Technical proofs are relegated to online Supplementary Material.

2. Main Results

2.1. Multiple correlation coefficient and the R^2 statistic

Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be a sequence of independent and identically distributed (i.i.d.) observations from a population $\mathbf{z} = (y, x_1, \ldots, x_p)' \in \mathbb{R}^{p+1}$, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The sample mean and sample covariance matrix are $\bar{\mathbf{z}} = \sum_{j=1}^n \mathbf{z}_j/n$ and

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n-1}\sum_{j=1}^{n}(\mathbf{z}_j - \bar{\mathbf{z}})(\mathbf{z}_j - \bar{\mathbf{z}})', \tag{2.1}$$

respectively. Partitioning the population $\mathbf{z}$ into y and $\mathbf{x} = (x_1, \ldots, x_p)'$, the covariance matrices $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{\Sigma}}$ have partitions

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{yy} & \boldsymbol{\sigma}'_{xy} \\ \boldsymbol{\sigma}_{xy} & \boldsymbol{\Sigma}_{xx} \end{pmatrix} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} \hat{\sigma}_{yy} & \hat{\boldsymbol{\sigma}}'_{xy} \\ \hat{\boldsymbol{\sigma}}_{xy} & \hat{\boldsymbol{\Sigma}}_{xx} \end{pmatrix}, \tag{2.2}$$

respectively. By solving the optimization problem in (1.1), the squared multiple correlation coefficient ρ_p^2 and its moment estimator, the R^2 statistic, are given by

$$\rho_p^2 = \frac{\boldsymbol{\sigma}'_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\boldsymbol{\sigma}_{xy}}{\sigma_{yy}} \quad \text{and} \quad R^2 = \frac{\hat{\boldsymbol{\sigma}}'_{xy}\hat{\boldsymbol{\Sigma}}_{xx}^{-1}\hat{\boldsymbol{\sigma}}_{xy}}{\hat{\sigma}_{yy}}, \tag{2.3}$$

respectively (Anderson (2003)).

2.2. CLT for the R^2 statistic

Our study of the R^2 statistic is under a general IC model (Bai and Silverstein (2010)). It assumes that the population $\mathbf{z}$ has a stochastic representation

$$\mathbf{z} = \boldsymbol{\mu} + \mathbf{A}\mathbf{w} = \boldsymbol{\mu} + \begin{pmatrix} \mathbf{a}'_1 \\ \mathbf{A}_2 \end{pmatrix} \mathbf{w}, \tag{2.4}$$

where $\boldsymbol{\mu} \in \mathbb{R}^{p+1}$ denotes the unknown mean vector, $\mathbf{w} = (w_1, \ldots, w_m)' \in \mathbb{R}^m$ ($m \geq p+1$) is a vector of independent random variables representing the m latent components, and $\mathbf{A} \in \mathbb{R}^{(p+1)\times m}$ is a deterministic transformation matrix with $\operatorname{rank}(\mathbf{A}) = p+1$ and $\mathbf{A}\mathbf{A}' = \boldsymbol{\Sigma}$. Here, the transformation matrix $\mathbf{A}$ is partitioned into $\mathbf{a}_1 \in \mathbb{R}^m$ and $\mathbf{A}_2 \in \mathbb{R}^{p\times m}$, according to the partition of the population $\mathbf{z}$.

Our main assumptions on this model are listed below.

Assumption 1. *The dimensions (p, m, n) tend to infinity in a related way, such that*

$$p = p_n \to \infty, \quad m = m_n \to \infty, \quad c_n \triangleq \frac{p}{n} \to c \in (0,1), \quad \limsup_{n\to\infty} \frac{m}{n} < 1.$$

Assumption 2. *The latent independent variables (w_i) satisfy*

$$\mathbb{E}(w_i) = 0, \quad \mathbb{E}(w_i^2) = 1, \quad \mathbb{E}(w_i^4) = \tau_i, \quad \sup_{i\geq 1} \mathbb{E}|w_i|^6 < \infty,$$

and $\sup_{i\geq 1} \mathbb{E}|w_i|^6 I_{(|w_i|\geqslant \delta n^{1/3})} \to 0$, for any fixed $\delta > 0$.

Assumption 3. *As $(p, m) \to \infty$, the multiple correlation coefficient $\rho_p \to \rho \in [0, 1)$ and the limits of the following quantities exist:*

$$\frac{1}{\sigma_{yy}^2} \sum_{i=1}^{m} (\tau_i - 3)[\mathbf{a}'_1\mathbf{e}_i]^{5-k}[\boldsymbol{\sigma}'_{xy}\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{A}_2\mathbf{e}_i]^{k-1} \to \zeta_k, \quad k = 1, \ldots, 5,$$

where $\{\sigma_{yy}, \boldsymbol{\sigma}_{xy}, \boldsymbol{\Sigma}_{xx}\}$ are defined in (2.2), and $\mathbf{e}_i$ denotes an $m\times 1$ column vector, with its ith coordinate equals to one, and all others equal to zero.

Remark 1. The IC model (2.4) generalizes the one studied in Zheng et al. (2014). Their model assumes that the latent variables (w_i) are i.i.d. and have common finite fourth moment. Moreover, for the first row $\mathbf{a}_1'$ of the matrix $\mathbf{A}$, after normalization, its ℓ_∞-norm should converge to zero, that is, $||\mathbf{a}_1/\sqrt{\mathbf{a}_1'\mathbf{a}_1}||_\infty = o(1)$, which implies that $\max_k \mathrm{Cor}(y, w_k) \to 0$. This condition is now removed from our model and, as a price, we need the condition of a finite sixth moment; see Assumption 2. An alternative condition on their model is $\mathbb{E}(w_i^4) = 3$, under which our Assumption 3 holds automatically with $\zeta_k = 0$, for $k = 1, \ldots, 5$. In general cases, the five quantities $\{\zeta_k\}$ may contribute to the fluctuation of R^2.

Remark 2. The IC model (2.4) includes the linear regression model as a special case. Consider the following linear model:

$$y = \beta_0 + \boldsymbol{\beta}'\mathbf{x} + \epsilon, \tag{2.5}$$

where $y \in \mathbb{R}$ is the response variable, $\beta_0 \in \mathbb{R}$ is the intercept, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of regression coefficients, $\mathbf{x} \in \mathbb{R}^p$ is the vector of explanatory variables with zero means, and $\epsilon = \sigma_{\epsilon,p}\epsilon_p$, independent of $\mathbf{x}$, denotes the error term, with mean zero and variance $\sigma_{\epsilon,p}^2$ (note that ϵ_p is a standardized variable). Suppose that $\mathbf{x}$ has the following IC representation:

$$\mathbf{x} = \mathbf{A}_\mathbf{x}\boldsymbol{\xi},$$

with $\mathbf{A}_\mathbf{x} \in \mathbb{R}^{p\times m_x}$ a row full-rank transformation matrix and $\boldsymbol{\xi} \in \mathbb{R}^{m_x}$ a vector of independent components. Then, the joint vector of y and $\mathbf{x}$ in the linear model can be written as

$$\begin{pmatrix} y \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}'\mathbf{A}_\mathbf{x} & \sigma_{\epsilon,p} \\ \mathbf{A}_\mathbf{x} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \boldsymbol{\xi} \\ \epsilon_p \end{pmatrix}, \tag{2.6}$$

which is a special case of the IC model (2.4) with the correspondence

$$\boldsymbol{\mu} = \begin{pmatrix} \beta_0 \\ \mathbf{0} \end{pmatrix}, \mathbf{a}_1 = \begin{pmatrix} \mathbf{A}_\mathbf{x}'\boldsymbol{\beta} \\ \sigma_{\epsilon,p} \end{pmatrix}, \mathbf{A}_2 = \begin{pmatrix} \mathbf{A}_\mathbf{x} & \mathbf{0} \end{pmatrix}, \mathbf{w} = \begin{pmatrix} \boldsymbol{\xi} \\ \epsilon_p \end{pmatrix}, \tag{2.7}$$

and $m = m_x + 1$.

Theorem 1. *Suppose that Assumptions* 1 *to* 3 *hold. Then,*

$$\sqrt{n}\{R^2 - c_n - (1 - c_n)\rho_p^2\} \to N\{0, \sigma^2(c, \rho^2)\} \tag{2.8}$$

in distribution. The variance function is $\sigma^2(c, x) = \sigma_1^2(c, x) + \sigma_2(c, x)$*, with*

$$\begin{aligned}\sigma_1^2(c, x) = {} & 2\{c + (1 - c)x\}^2 \\ & + 4\{(1 - c)x^2 - 2(1 - c)x - c\}\left\{c + (1 - c)x - \frac{1}{2}\right\},\end{aligned}$$

$$\sigma_2(c,x) = h_1 - 2\{c+(1-c)x\}h_2 + \{c+(1-c)x\}^2\zeta_1,$$

where $h_1 = c^2\zeta_1 + 4c(1-c)\zeta_2 + 2(2-3c)(1-c)\zeta_3 - 4(1-c)^2\zeta_4 + (1-c)^2\zeta_5$ *and* $h_2 = c\zeta_1 + 2(1-c)\zeta_2 - (1-c)\zeta_3$.

Theorem 1 establishes a new CLT for the R^2 statistic under the IC model (2.4). Its limiting variance $\sigma^2(c,\rho^2)$ is represented as the sum of $\sigma_1^2(c,\rho^2)$ and $\sigma_2(c,\rho^2)$. In particular, the second part $\sigma_2(c,\rho^2)$ consists of all quantities involving the five auxiliary parameters $\{\zeta_k\}$ defined in Assumption 3. Thus, this part characterizes how the fourth moments $\{\tau_i\}$ of the latent independent components $\{w_i\}$ contribute to the fluctuation of R^2.

When the coefficient of multiple correlation has the limit $\rho = 0$, we have $\zeta_3 = \zeta_4 = \zeta_5 = 0$, because

$$\frac{1}{\sigma_{yy}}\sum_{i=1}^{m}(\boldsymbol{\sigma}_{xy}'\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{A}_2\mathbf{e}_i)^2 = \rho_p^2 \to 0,$$

which gives $\sigma_1^2(c,0) = 2c(1-c)$ and $\sigma_2(c,0) = 0$. It follows immediately that

$$\sqrt{n}(R^2 - c_n) \to N\{0, 2c(1-c)\}$$

in distribution, which coincides with the result in Zheng et al. (2014). This conclusion does not depend on the distributions of the latent independent components, and thus can facilitate the testing procedure for $H_0 : \rho = 0$.

Under the linear model (2.5), $\sigma_2(c,\rho^2)$ can be simplified to

$$\sigma_2(c,\rho^2) = (1-c)^2(1-\rho^2)^2\left\{\tau_y - 3 + (2\rho^2-1)(\tau_\epsilon - 3)\right\},$$

where τ_y and τ_ϵ are the (limiting) kurtosis parameters of the response variable y and the error term ϵ_p, respectively. This shows that the overall contribution of the fourth moments of $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_{m_x})'$ to the variance of R^2 can be quantified by the kurtosis of the response y. This result coincides with Theorem 5 in Guo and Cheng (2022). Note that their CLT is established under a concentration condition on $\boldsymbol{\xi}$, that is, for some $\alpha > 0$,

$$\max_i \mathrm{P}\left(|\xi_i| \geq t\right) \leq 2\exp\left(-\alpha t^2\right), \quad \forall t \geq 0, \tag{2.9}$$

whereas ours is established under finite sixth moments, a weaker condition.

As an illustration, we numerically examine the fluctuation of R^2 under the following model.

Model 1. Take $\mathbf{A} = \mathbf{I}_{p+1}$, except with the (2,1)th entry equal to $q \in \{0,1\}$, set $\boldsymbol{\mu} = \mathbf{0}$, and let the components of $\mathbf{w}$ be i.i.d. standardized Gamma(1,2) random variables.

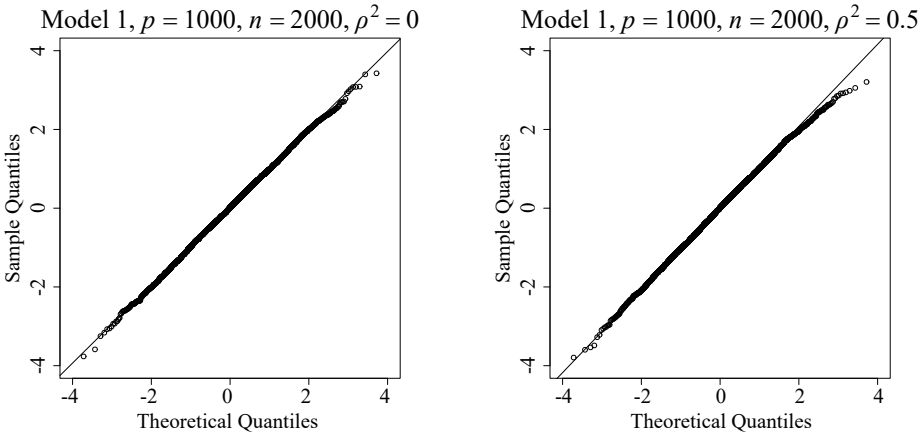

Figure 1. Normal Q–Q plots for the normalized R^2 from 5,000 independent replications, with $\rho^2 = 0$ (left panel) and $\rho^2 = 0.5$ (right panel).

The dimensional setting is $(p, n, c) = (1000, 2000, 0.5)$. Under this model, for $q = 0$, we have $\rho^2 = 0$ and $\sigma^2(c, \rho^2) = 0.5$, with $\sigma_2(c, \rho^2) = 0$; for $q = 1$, we have $\rho^2 = 0.5$ and $\sigma^2(c, \rho^2) = 0.4375$, with $\sigma_2(c, \rho^2) = 0.1875$. Normal Q–Q plots for the normalized R^2 from 5,000 independent replications are displayed in Figure 1, which confirms its asymptotic standard normality.

3. Interval Estimation of ρ^2 in a Linear Regression

3.1. Confidence interval for ρ^2

This section considers the interval estimation of the squared multiple correlation coefficient ρ^2 in the linear regression (2.5). Using the CLT developed in Section 2, it is sufficient to present a reasonable estimate of the limiting variance $\sigma^2(c, \rho^2)$, which involves three unknown parameters, namely, ρ^2, τ_y, and τ_ϵ.

Let $(y_1, \mathbf{x}_1'), \ldots, (y_n, \mathbf{x}_n')$ be a sequence of i.i.d. observations from the regression model. Then, the moment estimates of the three parameters are, respectively,

$$\hat{\rho}^2 = R^{*2} \triangleq \frac{R^2 - c_n}{1 - c_n}, \quad \hat{\tau}_y = \frac{(1/n)\sum_{j=1}^n (y_j - \bar{y})^4}{\left\{(1/n)\sum_{j=1}^n (y_j - \bar{y})^2\right\}^2},$$

and

$$\hat{\tau}_\epsilon = \frac{1}{(1 - c_n)^4}\left[\frac{1}{n}\sum_{j=1}^n \left\{\frac{\hat{\epsilon}_j^2}{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/(n-p)}\right\}^2 - 3c_n(1 - c_n)^2(2 - c_n)\right],$$

where $\bar{y}$ is the sample mean of $\{y_j\}$, and $\hat{\boldsymbol{\epsilon}}$ denotes the residual vector of the regression. Note that the consistency of $\hat{\rho}^2$ and $\hat{\tau}_y$ is obvious, and that of $\hat{\tau}_\epsilon$ is verified in Guo and Cheng (2022) under the concentration condition (2.9),

which can be relaxed to our moment conditions. Therefore, a plug-in estimator of $\sigma^2(c,\rho^2)$ is given by

$$\hat{\sigma}^2 = \sigma_1^2(c_n,\hat{\rho}^2) + (1-c_n)^2(1-\hat{\rho}^2)^2\left\{\hat{\tau}_y - 3 + (2\hat{\rho}^2-1)(\hat{\tau}_\epsilon - 3)\right\}.$$

However, as attested by our simulations, for moderately large p and n, the estimator $\hat{\sigma}^2$ sometimes takes negative values due to the fluctuations of $\hat{\rho}^2$ and $\hat{\tau}_\epsilon$, especially when c_n is large and ρ^2 is small. To cope with this irrational situation, we find a lower bound for $\sigma^2(c,\rho^2)$, that is,

$$\sigma^2(c,\rho^2) \geq \sigma_1^2(c,\rho^2) - 4(1-c)^2(1-\rho^2)^2\rho^4 > 0,\ \forall \rho^2 \in [0,1), \tag{3.1}$$

and propose a truncated estimator of $\sigma^2(c,\rho^2)$ as

$$\begin{aligned}\hat{\sigma}_t^2 &= \sigma_1^2(c_n,R_t^{*2})\\ &\quad +(1-c_n)^2(1-R_t^{*2})^2\max\left\{-4R_t^{*4},\hat{\tau}_y - 3 + (2R_t^{*2}-1)(\hat{\tau}_\epsilon-3)\right\},\end{aligned}$$

where $R_t^{*2} = \max\{R^{*2},0\}$ is the truncated statistic of R^{*2}.

Theorem 2. *Under the assumptions of Theorem* 1, $\hat{\sigma}_t^2$ *converges to* $\sigma^2(c,\rho^2)$ *in probability.*

Based on Theorems 1 and 2, the $(1-\alpha)100\%$ confidence interval for ρ^2 in a linear regression can be constructed as

$$\mathcal{C}(R^{*2}) \triangleq \left\{\rho^2 : R^{*2} - \frac{z_{\alpha/2}\hat{\sigma}_t}{\sqrt{n}(1-c_n)} \leq \rho^2 \leq R^{*2} + \frac{z_{\alpha/2}\hat{\sigma}_t}{\sqrt{n}(1-c_n)}\right\} \cap [0,1],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$-quantile of the standard normal distribution. If $\mathcal{C}(R^{*2})$ is a null set, the confidence interval is deemed to be nonexistent; the probability of this event is $o(1)$ for any $\rho^2 \in (0,1)$.

3.2. Simulations

We numerically evaluate the performance of our confidence interval for ρ^2, referred to as CI_{new}, and compare it with that of the original estimator $\hat{\sigma}^2$ (Guo and Cheng (2022)), referred to as CI_{gc}.

The model settings are as follows:

1. We set $\beta_0 = 0$ and $\mathbf{A}_\mathbf{x} = \mathbf{I}_p$ in the model, because the R^2 statistic is invariant under any invertible affine transformation on the explanatory vector $\mathbf{x}$.

2. Two distributional settings for $\boldsymbol{\xi}$ and ϵ are considered:

   Case 1. The first $[p/2]$ components of $\boldsymbol{\xi}$ are generated from a standardized Gamma(1,2) distribution and the rest are from a Unif$(-\sqrt{3},\sqrt{3})$ distribution, and ϵ follows N(0,1).

Table 1. Coverage (%) and average length of the 95% confidence intervals for ρ^2 under Case 1.

c	p		(ρ^2, k)			
			$(0.2, \frac{3p}{4})$	$(0.8, \frac{3p}{4})$	(0.2,2)	(0.8,2)
0.2	200	CI_{new}	95.10(0.1134)	94.54(0.0478)	94.42(0.1177)	94.58(0.0582)
		CI_{gc}	94.78(0.1127)	94.54(0.0478)	94.40(0.1173)	94.58(0.0582)
	300	CI_{new}	95.34(0.0924)	94.90(0.0390)	94.78(0.0963)	95.16(0.0478)
		CI_{gc}	95.20(0.0921)	94.90(0.0390)	94.68(0.0962)	95.16(0.0478)
	500	CI_{new}	95.06(0.0714)	94.74(0.0302)	94.92(0.0746)	95.14(0.0370)
		CI_{gc}	95.02(0.0713)	94.74(0.0302)	94.90(0.0746)	95.14(0.0370)
0.5	200	CI_{new}	95.12(0.2606)	94.50(0.0891)	95.02(0.2645)	94.56(0.1032)
		CI_{gc}	93.72(0.2523*)	94.50(0.0891)	93.76(0.2572*)	94.56(0.1032)
	300	CI_{new}	95.42(0.2168)	94.38(0.0728)	95.22(0.2165)	94.64(0.0846)
		CI_{gc}	94.36(0.2117*)	94.38(0.0728)	94.40(0.2108*)	94.64(0.0846)
	500	CI_{new}	95.00(0.1682)	94.58(0.0565)	94.70(0.1705)	94.38(0.0654)
		CI_{gc}	94.48(0.1653)	94.58(0.0565)	94.32(0.1684)	94.38(0.0654)
0.8	200	CI_{new}	92.44(0.4822)	91.50(0.1660)	92.38(0.4847)	91.70(0.1761)
		CI_{gc}	85.20(0.4384*)	89.84(0.1594*)	86.26(0.4426*)	90.80(0.1729)
	300	CI_{new}	93.64(0.4196)	92.64(0.1362)	93.76(0.4181)	91.88(0.1444)
		CI_{gc}	88.64(0.3873*)	91.52(0.1327)	88.60(0.3846*)	91.30(0.1430)
	500	CI_{new}	94.62(0.3472)	93.98(0.1049)	94.54(0.3485)	93.32(0.1118)
		CI_{gc}	91.38(0.3260*)	93.32(0.1034)	91.16(0.3269*)	93.20(0.1115)

Case 2. The first $[p/2]$ components of $\boldsymbol{\xi}$ are generated from a standardized Poisson(1) distribution and the rest are from N(0,1), and ϵ follows a t(9) distribution.

3. For the regression coefficient vector $\boldsymbol{\beta}$, we fix its ℓ_2-norm $\|\boldsymbol{\beta}\|$, and let its first to kth elements be equal to $\|\boldsymbol{\beta}\|/\sqrt{k}$, and the rest be zero. Here, we set $||\boldsymbol{\beta}|| = 0.5$ or 2, corresponding to a small or large ρ^2, respectively. Note that ρ^2 is equal to 0.2 or 0.8 under Case 1, and 7/43 or 28/37 under Case 2. The parameter k is set to 2 or $[3p/4]$, representing a sparse and dense regression, respectively.

4. The dimensional settings are $p = 200, 300, 500$ and $c_n = 0.2, 0.5, 0.8$.

The empirical coverage rates and average lengths of CI_{new} and CI_{gc} from 5,000 independent replications are collected in Tables 1 and 2. Starred results imply that there are some negative estimates $\hat{\sigma}^2$ among the 5,000 replications. When this occurs, we set the length of the corresponding CI_{gc} to zero and judge that this interval does not cover ρ^2.

The results in Tables 1 and 2 show that when the ratio c_n is small and ρ^2 is large, the two interval estimates are comparable, with similar average lengths, and their coverage rates are all close to the nominal level 0.95. However, for

Table 2. Coverage (%) and average length of the 95% confidence intervals for ρ^2 under Case 2.

c	p		(ρ^2, k)			
			$(\frac{7}{43}, \frac{3p}{4})$	$(\frac{28}{37}, \frac{3p}{4})$	$(\frac{7}{43}, 2)$	$(\frac{28}{37}, 2)$
0.2	200	CI_{new}	94.84(0.1152)	94.94(0.0613)	94.40(0.1155)	94.80(0.0635)
		CI_{gc}	92.44(0.1108*)	94.94(0.0613)	92.76(0.1112*)	94.80(0.0635)
	300	CI_{new}	94.88(0.0934)	95.08(0.0504)	95.08(0.0938)	95.06(0.0520)
		CI_{gc}	93.22(0.0907*)	95.08(0.0504)	93.84(0.0914*)	95.06(0.0520)
	500	CI_{new}	95.24(0.0720)	95.08(0.0391)	95.46(0.0724)	94.68(0.0404)
		CI_{gc}	94.24(0.0707*)	95.08(0.0391)	94.54(0.0714*)	94.68(0.0404)
0.5	200	CI_{new}	95.30(0.2617)	93.44(0.1126)	95.42(0.2620)	93.92(0.1148)
		CI_{gc}	90.82(0.2451*)	93.44(0.1126)	91.14(0.2448*)	93.92(0.1148)
	300	CI_{new}	95.08(0.2214)	94.26(0.0922)	95.46(0.2212)	94.48(0.0944)
		CI_{gc}	92.36(0.2092*)	94.26(0.0922)	91.94(0.2097*)	94.48(0.0944)
	500	CI_{new}	95.26(0.1739)	94.84(0.0717)	95.52(0.1747)	94.54(0.0733)
		CI_{gc}	92.92(0.1667*)	94.84(0.0717)	93.62(0.1676*)	94.54(0.0733)
0.8	200	CI_{new}	92.64(0.4796)	91.50(0.2039)	92.64(0.4764)	91.72(0.2046)
		CI_{gc}	84.52(0.4347*)	90.28(0.1979*)	83.32(0.4283*)	90.70(0.1995*)
	300	CI_{new}	93.42(0.4111)	92.68(0.1662)	94.02(0.4103)	91.88(0.1677)
		CI_{gc}	86.76(0.3758*)	92.06(0.1633)	86.70(0.3717*)	91.02(0.1653)
	500	CI_{new}	93.94(0.3414)	93.68(0.1294)	94.08(0.3412)	93.58(0.1304)
		CI_{gc}	88.80(0.3153*)	93.34(0.1285)	89.06(0.3162*)	93.36(0.1296)

large c_n, their coverage rates tend to be biased downward, and the biases become small as the dimensions increase. In particular, when c_n is large and ρ^2 is small, CI_{new} outperforms CI_{gc}, with more accurate coverage rates. This demonstrates the necessity and validity of using the truncated estimate of the limiting variance $\sigma^2(c, \rho^2)$.

3.3. An empirical study

We study a breast cancer data set collected by Yau et al. (2010) that can be downloaded from the UCSC Xena platform (`http://xena.ucsc.edu`). This data set consists of measurements on 9,168 gene expression levels of $n = 228$ cancer patients and their (uncensored) distant metastasis-free survival times T. Our interest is the extent to which a linear function of a set of gene expressions can explain the variation of the survival time.

Motivated by the accelerated failure time model (Kalbfleisch and Prentice (2002)), we regress the logarithm of the survival time $\log(T)$ on p gene expression levels selected to have the largest marginal correlations with the response. Three 95% confidence intervals, IC_{new}, IC_{gc}, and IC_{zheng} (Zheng et al. (2014)), for the squared coefficient of multiple correlation ρ^2 are reported in Table 3, where the dimension p varies from 100 to 160. The results illustrate that, compared with

Table 3. 95% confidence intervals of ρ^2 for breast cancer data.

p	CI_{new}	CI_{gc}	CI_{zheng}
100	(0.1816,0.4679)	(0.1816,0.4679)	(0.1765,0.4730)
110	(0.1728,0.4749)	(0.1728,0.4749)	(0.1679,0.4799)
120	(0.1530,0.4733)	(0.1530,0.4733)	(0.1470,0.4794)
130	(0.1495,0.4852)	(0.1495,0.4852)	(0.1418,0.4928)
140	(0.1371,0.4944)	(0.1445,0.4869)	(0.1285,0.5030)
150	(0.1064,0.4986)	(0.1138,0.4911)	(0.0989,0.5061)
160	(0.1119,0.5286)	(0.1308,0.5096)	(0.1043,0.5361)

CI_{zheng}, both CI_{new} and CI_{gc} suggest a slightly narrower confidence interval in all cases under study. In addition, when p is large ($p \geq 140$), CI_{new} indicates the need to truncate the estimate of $\sigma^2(c, \rho^2)$, which results in a wider confidence interval than that of CI_{gc}.

Supplementary Material

The online Supplementary Material includes proofs of Theorems 1 and 2.

Acknowledgments

The authors would like to thank the Editor, Associate Editor, and anonymous reviewers for their valuable and insightful comments.

References

Ali, H. and Nagar, D. K. (2002). Null distribution of multiple correlation coefficient under mixture normal model. *International Journal of Mathematics and Mathematical Sciences* **30**, 249–255.

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd Edition. Wiley, New York.

Bai, Z. D. and Silverstein, J. W. (2004). CLT for linear spectral statistics of large-dimensional sample covariance matrices. *The Annals of Probability* **32**, 553–605.

Bai, Z. D. and Silverstein, J. W. (2010). *Spectral Analysis of Large Dimensional Random Matrices*. 2nd Edition. Springer, New York.

Fisher, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* **121**, 654–673.

Guo, X. and Cheng, G. (2022). Moderate-dimensional inferences on quadratic functionals in ordinary least squares. *Journal of The American Statistical Assocation* **117**, 1931–1950.

Gurland, J. (1968). A relatively simple form of the distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society. Series B (Methodological)* **30**, 276–283.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd Edition. Wiley.

Lee, Y. (1971). Some results on the sampling distribution of the multiple correlation coefficient. *Journal of the Royal Statistical Society. Series B (Methodological)* **33**, 117–130.

Marčenko, V. and Pastur, L. (1967). The distribution of eigenvalues in certain sets of random matrices. *Math. USSR-Sbornik* **1**, 457–483.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. Wiley, New York.

Nandi, S. and Choudhury, S. (2005). Series representation of non null distribution of the square of sample multiple correlation coefficient by use of the mellin integral transform. *Communications in Statistics - Theory and Methods* **34**, 679–685.

Ogasawara, H. (2006). Asymptotic expansion and conditional robustness for the sample multiple correlation coefficient under nonnormality. *Communications in Statistics - Simulation and Computation* **35**, 177–199.

Wilks, S. S. (1932). On the sampling distribution of the multiple correlation coefficient. *The Annals of Mathematical Statistics* **3**, 196–203.

Williams, E. J. (1978). A simple derivation of the distribution of the multiple correlation coefficient. *Communications in Statistics - Theory and Methods* **7**, 1413–1420.

Yau, C., Esserman, L., Moore, D., Waldman, F., Sninsky, J. and Benz, C. (2010). A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Research: BCR* **12**, R85.

Zheng, S. R., Jiang, D. D., Bai, Z. D. and He, X. M. (2014). Inference on multiple correlation coefficients with moderately high dimensional data. *Biometrika* **101**, 748–754.

Weiming Li

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: li.weiming@shufe.edu.cn

Shizhe Hong

School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai 200433, China.

E-mail: hong.shizhe@163.sufe.edu.cn

(Received February 2022; accepted April 2023)