Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Moderate-Dimensional Inferences on Quadratic Functionals in Ordinary Least Squares

Xiao Guo & Guang Cheng

To cite this article: Xiao Guo & Guang Cheng (2021): Moderate-Dimensional Inferences on Quadratic Functionals in Ordinary Least Squares, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1893177

To link to this article: https://doi.org/10.1080/01621459.2021.1893177

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–20: Theory and Methods
https://doi.org/10.1080/01621459.2021.1893177

Moderate-Dimensional Inferences on Quadratic Functionals in Ordinary Least Squares

Xiao Guo[a] and Guang Cheng[b]

[a]International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui, China; [b]Department of Statistics, Purdue University, West Lafayette, IN

ABSTRACT

Statistical inferences for quadratic functionals of linear regression parameter have found wide applications including signal detection, global testing, inferences of error variance and fraction of variance explained. Classical theory based on ordinary least squares estimator works perfectly in the low-dimensional regime, but fails when the parameter dimension p_n grows proportionally to the sample size n. In some cases, its performance is not satisfactory even when $n \geq 5p_n$. The main contribution of this article is to develop *dimension-adaptive* inferences for quadratic functionals when $\lim_{n\to\infty} p_n/n = \tau \in [0, 1)$. We propose a bias-and-variance-corrected test statistic and demonstrate that its theoretical validity (such as consistency and asymptotic normality) is adaptive to both low dimension with $\tau = 0$ and moderate dimension with $\tau \in (0, 1)$. Our general theory holds, in particular, without Gaussian design/error or structural parameter assumption, and applies to a broad class of quadratic functionals covering all aforementioned applications. As a by-product, we find that the classical fixed-dimensional results continue to hold *if and only if* the signal-to-noise ratio is large enough, say when p_n diverges but slower than n. Extensive numerical results demonstrate the satisfactory performance of the proposed methodology even when $p_n \geq 0.9n$ in some extreme cases. The mathematical arguments are based on the random matrix theory and leave-one-observation-out method.

ARTICLE HISTORY
Received August 2019
Accepted June 2020

KEYWORDS
Fraction of variance explained; Linear regression model; Moderate dimension; Quadratic functional; Signal-to-noise ratio

1. Introduction

The linear regression model is one of the most widely used statistical tools to discover the relation between a continuous response and a class of explanatory variables in different scientific areas. Specifically, we consider

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}_0 + \epsilon_i, \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_0 = (\beta_{0,1}, \ldots, \beta_{0,p_n})^T \in \mathbb{R}^{p_n}$ is an unknown vector of parameters, and $\{\epsilon_i\}_{i=1}^n$ are iid errors independent of $\{\boldsymbol{X}_i\}_{i=1}^n$ with $\mathrm{E}(\epsilon_i) = 0$ and $\mathrm{var}(\epsilon_i) = \sigma_\epsilon^2$. We assume $\{Y_i, \boldsymbol{X}_i\}_{i=1}^n$ are iid observations with $\mathrm{E}(\boldsymbol{X}_i) = \boldsymbol{0}_{p_n}$ and $\mathrm{cov}(\boldsymbol{X}_i) = \Sigma$, without imposing any specific distributional assumption on either $\boldsymbol{X}_i$ or ϵ_i throughout this article. Denoting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $X = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^T$, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$, (1) can be re-expressed as

$$\boldsymbol{Y} = X\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}.$$

For fixed dimension, statistical estimation and inference for $\boldsymbol{\beta}_0$ and σ_ϵ^2 have been well studied based on the ordinary least squares (OLS) estimator,

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y}.$$

In the modern high-dimensional regime, the parameter dimension p_n is allowed to be much larger than n, for example, $(\log p_n)/n = o(1)$, but in most cases the number of nonzero elements in $\boldsymbol{\beta}_0$ is a vanishing fraction of n. Such a sparsity condition is commonly assumed in the high-dimensional literature, for example, Meinshausen and Yu (2009), van de Geer (2008), and Zhang and Huang (2008) on oracle inequality and parameter estimation; Tibshirani (1996), Fan and Lv (2008), and Meinshausen and Bühlmann (2006) on variable selection, and Javanmard and Montanari (2014), van de Geer et al. (2014), and Zhang and Zhang (2014) on statistical inference. However, in reality, p_n may be moderately large, that is, of the same magnitude as n, and $\boldsymbol{\beta}_0$ is not necessarily sparse. One example is the genomic study, where the number of *significantly identified* genes with association in *trans*, that is, $p_n = 108$, is moderately large compared with $n = 270$; see Stranger et al. (2007).

For moderate dimension with $\lim_{n\to\infty} p_n/n = \tau \in (0, 1)$, which is of major concern in this article, some classical statistical inference procedures developed for fixed-dimensional data are no longer valid. For example, when p_n is fixed, we can test

$$H_0 : ||\boldsymbol{\beta}_0||_2 = c_0 \quad \text{versus} \quad H_1 : ||\boldsymbol{\beta}_0||_2 \neq c_0, \tag{2}$$

for a known constant $c_0 \geq 0$, by calculating the Z-score

$$\mathbb{Z}_0 = \frac{||\hat{\boldsymbol{\beta}}||_2^2 - c_0^2}{\hat{\zeta}_0}, \tag{3}$$

where

$$\hat{\zeta}_0^2 = 4\hat{\sigma}_\epsilon^2 \hat{\boldsymbol{\beta}}^T (X^T X)^{-1} \hat{\boldsymbol{\beta}} \quad \text{and} \quad \hat{\sigma}_\epsilon^2 = \frac{||\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}||_2^2}{n - p_n}. \tag{4}$$

Under the null hypothesis, $\mathbb{Z}_0 \xrightarrow{\mathcal{D}} N(0, 1)$; see Theorem 4. Hence, the p-value for testing (2) is $2\Phi(-|\mathbb{z}_0|)$, where $\mathbb{z}_0$ is a

CONTACT Guang Cheng chengg@purdue.edu Department of Statistics, Purdue University, West Lafayette, IN 47906.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

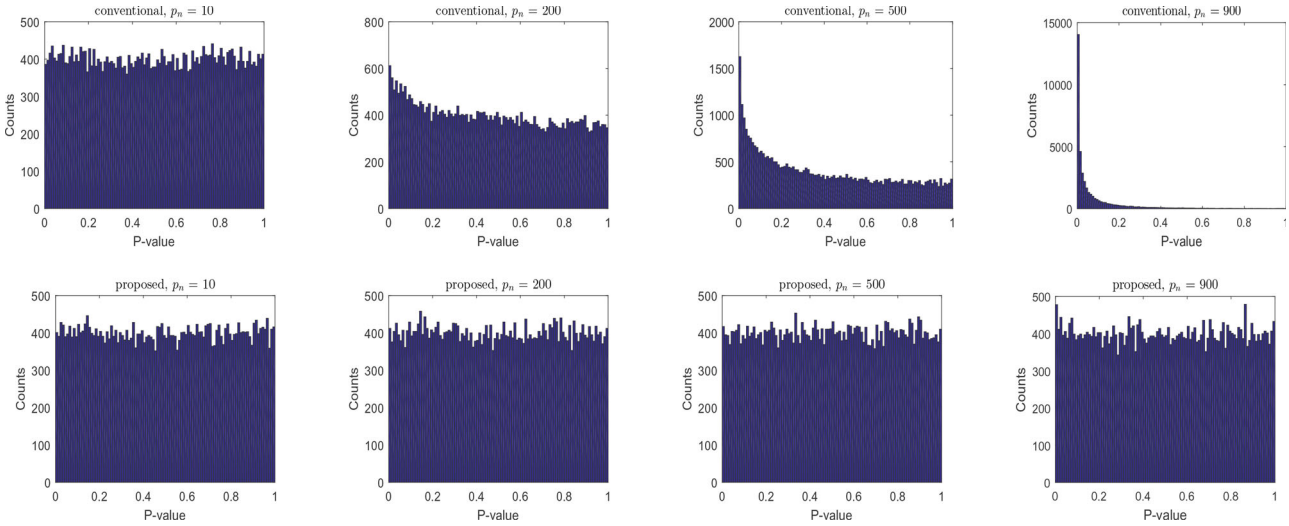

Figure 1. p-values of $\mathbb{Z}_0$ (top panels) and $\mathbb{Z}_n$ (bottom panels). The panels from left to right are for $p_n = 10/200/500/900$.

realization of $\mathbb{Z}_0$ and $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution.

We next examine the empirical performance of the conventional Z-test by setting $n = 1000$ with $p_n = 10$ for fixed dimension and $p_n = 200$, 500, and 900 for moderate dimension. Consider $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$ and $\epsilon_i \overset{\text{iid}}{\sim} N(0,1)$, where $\mathbf{I}_{p_n}$ denotes the $p_n \times p_n$ identity matrix. The true parameter $\beta_{0,j}$'s were generated independently from Unif(0, 1), and 40,000 replications were conducted in each setup. The plots of the p-values under the valid null hypothesis are given in the top panels of Figure 1. The uniform distribution of the p-values when $p_n = 10$ is consistent with the classical fixed-dimensional theory. But for $p_n = 200$, 500, and 900, p-values are relatively concentrated around 0. We further test the uniform distribution of the p-values by the formal Kolmogorov–Smirnov (KS) test (Kolmogorov 1933; Smirnov 1939), and find that the p-values for $p_n = 10$, 200, 500, and 900 are 0.2518, 8.05×10^{-68}, 0, and 0, respectively. Hence, the naive Z-score does not work under moderate dimension, say even when $n \geq 5p_n$.

The main focus of this article is on the moderate-dimensional inference without imposing any type of structural conditions on $\boldsymbol{\beta}_0$ and Σ, while our results are also adaptive to the low-dimensional case with $\tau = 0$.[1] Specifically, we conduct statistical inferences for a class of quadratic functionals such as $||\boldsymbol{\beta}_0||_2^2$ and σ_ϵ^2, which cover a wide range of applications including signal detection and global testing. A related line of work is the study of the signal strength $\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0$ by Dicker (2014) and Janson, Barber, and Candès (2017). However, their procedures crucially rely on the fact that $Y_i \sim N(0, \boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0 + \sigma_\epsilon^2)$, and their theoretical results hold only when $\boldsymbol{X}_i$ and ϵ_i are both Gaussian. Hence, their results are not readily carried over into our case, for example, two-sample inference. Additionally, different tools such as leave-one-observation-out method (El Karoui 2013, 2018) are used in our article. Please see more discussions in the end of Section 3.4. As a side remark, we point out that the classical fixed-dimensional inference may still be applied to the low-dimensional regime *if and only if* the signal-to-noise ratio SNR $:= \text{var}(\boldsymbol{X}_i^T \boldsymbol{\beta}_0)/\text{var}(\epsilon_i) = \boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0 / \sigma_\epsilon^2$ is large. However, the strength of the SNR cannot be directly examined in practice. Hence, the adaptiveness of our proposed method (without relying on SNR) is practically important. In case of interest, readers may refer to Figure B1 in Appendix B for the precise relation between τ and SNR.

Our primary contribution is to propose a bias-corrected estimator $||\hat{\boldsymbol{\beta}}||_2^2$ for $||\boldsymbol{\beta}_0||_2^2$ in (10), based on which a bias-and-variance-corrected test statistic $\mathbb{Z}_n$ is developed in (12). The bottom panels of Figure 1 plot the p-values of $\mathbb{Z}_n$ for $p_n = 10$, 200, 500, and 900. The p-values of the KS test for the uniformity are 0.4755, 0.1175, 0.8972 and 0.2672 correspondingly. Figure 2 plots the amount of empirical corrections of bias and variance needed in $||\hat{\boldsymbol{\beta}}||_2^2$ under the same setting. It reveals that the bias correction tends to $-\infty$ as $\tau \to 1$, while the variance correction diverges to ∞. The right panel of Figure 2 plots the relative difference between $\mathbb{Z}_n$ and $\mathbb{Z}_0$ versus τ. As τ deviates from zero, the amount of correction rapidly increases to its largest value, and then decreases and stabilizes around 1. As an immediate application, global testing

$$H_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{\text{null}} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_0^{\text{null}}, \tag{5}$$

can also be performed with a bias-and-variance-corrected version of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^{\text{null}}||_2^2$ as the test statistic. Please see Portnoy (1985), Arias-Castro, Candès, and Plan (2011), Zhong and Chen (2011), and Zhang and Cheng (2017) for low- and high-dimensional results, respectively.

Our general moderate-dimensional theory can also be applied to other statistical inference problems. For example, we can detect the existence of signal by setting $c_0 = 0$ in (2). By formulating a sequence of alternatives $H_{1n} : ||\boldsymbol{\beta}_0||_2^2 = \delta_n$, we further show that $\delta_n^* := \sigma_\epsilon^2 \sqrt{p_n} n^{-1}$ is the smallest separation rate such that successful detection of H_{1n} is still possible, which matches with the minimax detection rate in Ingster, Tsybakov, and Verzelen (2010). As far as we are aware, the existing results concerned with detection boundary only focus on either Gaussian mean models with $p_n = n$ (e.g., Donoho and Jin 2004; Cai, Jin, and Low 2007; Hall and Jin 2010), or high-dimensional

[1] We call it low-dimensional regime when $p_n \to \infty$ but $p_n/n \to 0$. Hence, both fixed- and low-dimensional regimes correspond to that $\tau = 0$.

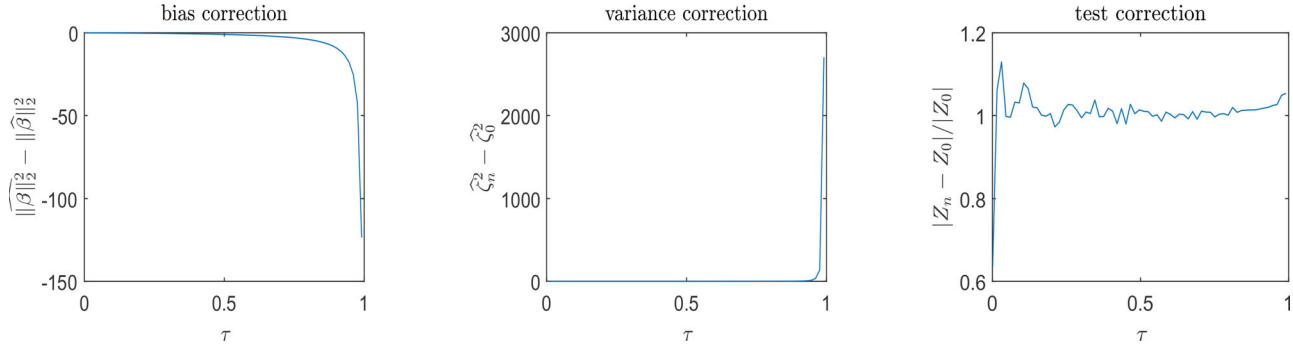

Figure 2. Amount of empirical corrections of bias (left panel) and variance (middle panel) versus τ for $||\hat{\boldsymbol{\beta}}||_2^2$ compared with $||\boldsymbol{\beta}||_2^2$. The right panel plots $|\mathbb{Z}_n - \mathbb{Z}_0|/|\mathbb{Z}_0|$ versus τ.

data (e.g., Ingster, Tsybakov, and Verzelen 2010; Arias-Castro, Candès, and Plan 2011).

New results of inference on the error variance will also be established for moderate dimension. We still use the estimator $\hat{\sigma}_\epsilon^2$ defined in (4) for low-dimensional data, but modify its asymptotic variance as $\zeta_\epsilon^2 = \{\nu_4 + \sigma_\epsilon^4(3\tau - 1)/(1 - \tau)\}/n$ with $\nu_4 = \mathrm{E}(\epsilon_i^4)$ to derive that

$$\frac{\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2}{\zeta_\epsilon} \xrightarrow{\mathcal{D}} N(0, 1).$$

One related result is concerned with the fraction of variance explained (and also SNR), defined as

$$\rho_0 := \frac{\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0}{\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0 + \sigma_\epsilon^2} = \frac{\mathrm{SNR}}{\mathrm{SNR} + 1}, \tag{6}$$

which describes the proportion of the variance in the dependent variable that is predictable from the independent variable. The high-dimensional estimation of σ_ϵ^2, ρ_0 and SNR can be found in Sun and Zhang (2012), Fan, Guo, and Hao (2012), and Verzelen and Gassiat (2018).

Our results can be naturally extended to two-sample inference. Here, we give two examples in Section S.4 in the supplementary materials. Let $\boldsymbol{\gamma}_0 \in \mathbb{R}^{p_n}$ be the regression parameter in another linear regression model independent of (1). The first issue is to test the equality of $\boldsymbol{\gamma}_0$ and $\boldsymbol{\beta}_0$, while the second is concerned with the co-heritability, defined as

$$\theta_0 = \frac{\boldsymbol{\gamma}_0^T \boldsymbol{\beta}_0}{||\boldsymbol{\gamma}_0||_2 ||\boldsymbol{\beta}_0||_2}. \tag{7}$$

The measure θ_0 is an important concept that characterizes the genetic associations within pairs of quantitative traits, whose high-dimensional estimation has recently been studied in Guo et al. (2016). Besides, an immediate application of our arguments is the inference for the linear functionals as discussed in Appendix A.

As a summary, a list of hypotheses in consideration together with potential applications is given below:

- Testing the quadratic functional: hypotheses in (2);
- Signal detection: hypotheses in (2) with $c_0 = 0$;
- Global testing: hypotheses in (5);
- Inference for the error variance σ_ϵ^2 using Proposition 1;
- Testing the fraction of variance explained (or SNR): hypotheses in (6);
- Inference for the signal strength $\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0$ using (16);
- Two-sample inferences: hypotheses in (S.4.2) and (S.4.3).

Our asymptotic normality result relies on the application of the martingale difference central limit theorem (CLT) Heyde and Brown (1970) to linear-quadratic forms, that is, $\boldsymbol{\epsilon}^T A_n \boldsymbol{\epsilon} + \boldsymbol{b}_n^T \boldsymbol{\epsilon}$, where $A_n \in \mathbb{R}^{n \times n}$ ($\boldsymbol{b}_n \in \mathbb{R}^n$) is some random matrix (vector) independent of $\boldsymbol{\epsilon}$. Although CLT has been studied for quadratic and linear-quadratic forms, to the best of our knowledge, those results cannot be directly applied to our problem. For example, Dicker and Erdogdu (2017) provided the concentration bounds and finite sample multivariate normal approximation for quadratic forms, but these results are not applicable to the linear-quadratic forms. de Jong (1987) developed CLT for "clean" quadratic forms requiring zero elements on the diagonal (see Definition 2.1 therein), which however is not satisfied by the linear-quadratic form in our article. A more related example is the CLT for the linear-quadratic form in Kelejian and Prucha (2001) with A_n and $\boldsymbol{b}_n$ being deterministic. An important assumption in Kelejian and Prucha (2001) is $||A_n||_1 \leq C < \infty$ which is violated in our work (see Section S.1 in the supplementary materials for detailed explanations). Besides, two technical tools have been used in our article: random matrix theory (Bai and Silverstein 2010) and leave-one-observation-out method (El Karoui 2013, 2018). The former contributes to bounding the eigenvalues of $X^T X/n$ from 0 and ∞ as in Lemma 1, while the latter is employed here to demonstrate the consistency of terms like $\mathrm{tr}\{(X^T X)^{-1}\}$ as in Lemma 2. Note that no sparsity assumption on Σ is needed in our technical analysis. It is worth pointing out that the theoretical results above are adaptive to the low-dimensional regime, which makes our proposed method concretely applicable in practice.

In the end, we conduct a real data analysis on the relationship between gene expression and single nucleotide polymorphism (SNP) with $n = 377$ and p_n ranging from 33 to 87. Specifically, confidence intervals for the fraction of variance explained, that is, ρ_0, are constructed using our proposed method, the conventional method and the one in Dicker (2014) based on the method of moment. We find that the conventional method may falsely discover nonzero ρ_0 for some genes due to the moderate dimension and insufficient SNR, and that our confidence interval is mostly narrower than that by Dicker (2014).

1.1. Related Works

Some earlier studies, for example, Portnoy (1984, 1985), focused on the quadratic functional $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T X^T X(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, under the low-dimensional regime, that is, $\tau = 0$. In the moderate-dimensional regime, El Karoui (2013, 2018), El Karoui et al. (2013), and Donoho and Montanari (2016) studied the consistency of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2$ for a general M-estimator $\hat{\boldsymbol{\beta}}$. As far as we are aware, these techniques and results for consistency are not ready for deriving the asymptotic distributions of the quadratic functionals, which is the main contribution of our work. Another line of research is the element-wise inference (Bai et al. 2013; Dobriban and Su 2018; Lei, Bickel, and El Karoui 2018; Sur, Chen, and Candès 2019) whose strategies for analyzing single-element estimation error cannot be easily adapted for the analysis of aggregated estimation errors, for example, quadratic functionals. To elucidate the difference between the two types of inferences, we plot $\sqrt{n}(\hat{\beta}_j^2 - \beta_{0,j}^2)$ versus j and $||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2$ in Figure A1. In the high-dimensional regime, a more recent result is Cai and Guo (2018) who studied the point and interval estimations of $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_q^2$ with $1 \le q \le 2$.

The rest of the article is organized as follows. Section 2 develops the bias-and-variance-corrected inference for $||\boldsymbol{\beta}_0||_2^2$ and demonstrates that the conventional procedure works if and only if the SNR is large. Section 3 consists of important applications of inferences for the quadratic functionals, including signal detection, global testing, inferences for the error variance and fraction of variance explained. Simulations are conducted in Section 4 and a real data analysis is performed in Section 5. The proofs of some main theoretical results are included in the Appendix while the remaining proofs are relegated to the supplementary materials.

1.2. Notation

Let $\lfloor \cdot \rfloor$ be the floor function. For any set G, denote by $\bar{G}$ the complement of G. Let $\mathrm{I}(\cdot)$ be the indicator function. Denote by $\mathbf{I}_m$ the $m \times m$ identity matrix and by $\mathbf{e}_{j,m}$ $(j = 1, \ldots, m)$ the jth column of $\mathbf{I}_m$. Let $\mathbf{0}_m \in \mathbb{R}^m$ and $\mathbf{1}_m \in \mathbb{R}^m$ be the vectors of zeros and ones, respectively. For a vector $\boldsymbol{v} = (v_1, \ldots, v_m)^T$, the L_1, L_2 and L_∞ norms are $||\boldsymbol{v}||_1 = \sum_{i=1}^m |v_i|$, $||\boldsymbol{v}||_2 = (\sum_{i=1}^m v_i^2)^{1/2}$ and $||\boldsymbol{v}||_\infty = \max_{i \le m} |v_i|$, respectively. For an $m \times m$ matrix $A = \{a_{ij}\}_{1 \le i,j \le m}$, denote by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ the maximum and minimum eigenvalues of A, respectively. Let $|A|$ be the determinant of A. The L_1, L_2 and L_∞ norms of A are defined as $||A||_1 = \max_{1 \le j \le m} \sum_{i=1}^m |a_{ij}|$, $||A||_2 = \{\lambda_{\max}(A'A)\}^{1/2}$ and $||A||_\infty = \max_{1 \le i \le m} \sum_{j=1}^m |a_{ij}|$, respectively. For sequences $\{a_n\}_{n \ge 1}$ and $\{b_n\}_{n \ge 1}$, we write $a_n \lesssim b_n$ $(a_n \gtrsim b_n)$ if there exists a constant $C > 0$ independent with n such that $|a_n| \le C|b_n|$ $(|a_n| \ge C|b_n|)$. Denote $a_n = \Omega(b_n)$ if $a_n = O(b_n)$ and $b_n = O(a_n)$. If $\{U_n\}_{n \ge 1}$ and $\{V_n\}_{n \ge 1}$ are random sequences, then $U_n = \Omega_{\mathrm{P}}(V_n)$ denotes that $U_n = O_{\mathrm{P}}(V_n)$ and $V_n = O_{\mathrm{P}}(U_n)$. Notation "$S_1 \iff S_2$" means that statements S_1 and S_2 are equivalent, while "$S_1 \implies S_2$" denotes that S_1 implies S_2. In the following, C and c are generic finite constants which may vary from place to place and do not depend on sample size n.

2. Statistical Inference for Quadratic Functionals

This section establishes the dimension-adaptive inference for $||\boldsymbol{\beta}_0||_2^2$, which is the main theoretical result of this article. As a by-product, we discover that the classical (fixed-dimensional) statistical inference procedure continues to work in the low-dimensional regime if and only if the signal-to-noise ratio is large. As far as we are aware, this finding is new for quadratic functional $||\boldsymbol{\beta}_0||_2^2$.

We start with an examination of the plug-in estimator $||\hat{\boldsymbol{\beta}}||_2^2$ for $||\boldsymbol{\beta}_0||_2^2$. The estimation error $||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2$ can be expressed as a linear-quadratic form, that is, sum of a quadratic term and a linear term, as follows

$$||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2 = \boldsymbol{\epsilon}^T X(X^T X)^{-2} X^T \boldsymbol{\epsilon} + 2\boldsymbol{\beta}_0^T (X^T X)^{-1} X^T \boldsymbol{\epsilon}. \quad (8)$$

The linear term has zero mean and hence the bias of $||\hat{\boldsymbol{\beta}}||_2^2$ is

$$\begin{aligned} \mathrm{E}(||\hat{\boldsymbol{\beta}}||_2^2) - ||\boldsymbol{\beta}_0||_2^2 &= \mathrm{E}\{\boldsymbol{\epsilon}^T X(X^T X)^{-2} X^T \boldsymbol{\epsilon}\} \\ &= \mathrm{Etr}\{(X^T X)^{-1}\}\sigma_\epsilon^2 > 0. \end{aligned} \quad (9)$$

For a special case that $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$, $(X^T X)^{-1}$ follows the inverse Wishart distribution and hence

$$\mathrm{Etr}\{(X^T X)^{-1}\} = p_n/(n - p_n - 1) \to \tau/(1 - \tau), \quad \text{as } n \to \infty.$$

For low dimension with $\tau = 0$, $||\hat{\boldsymbol{\beta}}||_2^2$ is asymptotically unbiased and its asymptotic distribution can be established based on the dominating linear term as in Theorem 4 to be introduced later. However, when $\tau > 0$, the bias (9) is nonignorable, leading to failure of the conventional low-dimensional results.

The analysis above suggests a bias-corrected estimator for $||\boldsymbol{\beta}_0||_2^2$:

$$||\hat{\boldsymbol{\beta}}||_2^2 = ||\hat{\boldsymbol{\beta}}||_2^2 - \mathrm{tr}\{(X^T X)^{-1}\}\hat{\sigma}_\epsilon^2, \quad (10)$$

where $\hat{\sigma}_\epsilon^2$ is defined in (4). Since X and $\boldsymbol{\epsilon}$ are independent, $||\hat{\boldsymbol{\beta}}||_2^2$ is unbiased for $||\boldsymbol{\beta}_0||_2^2$. Before presenting the asymptotic properties of $||\hat{\boldsymbol{\beta}}||_2^2$, we first provide our assumptions below.

Condition A.

- A1. Assume $\{\boldsymbol{X}_i\}_{i=1}^n$ are iid, $\boldsymbol{X}_i = \Sigma^{1/2}\boldsymbol{Z}_i$ where $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{ip_n})^T$, $\{z_{ij}\}_{j=1}^{p_n}$ are independent for each $i \le n$, $\mathrm{E}(z_{ij}) = 0$, $\mathrm{E}(z_{ij}^2) = 1$ and there exists a constant $c^* > 0$ such that for any $n \ge 1$, $i \le n$, $j \le p_n$, and $t > 0$, $\mathrm{P}(|z_{ij}| \ge t) \le 2\exp(-c^* t^2)$.
- A2. Suppose $\{\epsilon_i\}_{i=1}^n$ are iid and independent of $\{\boldsymbol{X}_i\}_{i=1}^n$, $\mathrm{E}(\epsilon_i) = 0$, $\mathrm{E}(\epsilon_i^2) = \sigma_\epsilon^2 \ge c > 0$, and $\mathrm{E}(\epsilon_i^8) = O(\sigma_\epsilon^8)$.
- A3. There exist constants c and C, such that $0 < c < \lambda_{\min}(\Sigma) \le \lambda_{\max}(\Sigma) < C < \infty$.
- A4. There exists a constant C, such that $||\boldsymbol{\beta}_0||_\infty \le C < \infty$.

Conditions A1 and A2 only require sub-Gaussian tail for $\boldsymbol{X}_i$ and moment conditions on $\boldsymbol{\epsilon}$, rather than impose any specific distributional restriction. The independence between $\{\epsilon_i\}_{i=1}^n$ and $\{\boldsymbol{X}_i\}_{i=1}^n$ is crucial for applying the martingale difference CLT, and is a standard assumption for inference of the quadratic functionals in, for example, Dicker (2014) and Janson, Barber, and Candès (2017). The error variance σ_ϵ^2 could either be bounded or diverging with n. Under Condition A4, $||\boldsymbol{\beta}_0||_2 = O(\sqrt{p_n})$,

and will reach $\Omega(\sqrt{p_n})$ when $\boldsymbol{\beta}_0$ is not sparse. Throughout this article, both $\boldsymbol{\beta}_0$ and σ_ϵ^2 are allowed to vary with n, except when p_n is fixed.

Under Condition A, $\hat{\boldsymbol{\beta}}$ is well defined, that is, the $p_n \times p_n$ matrix $(X^TX)^{-1}$ exists with probability tending to 1. Lemma 1 shows that the eigenvalues of X^TX/n are bounded away from 0 and ∞ with probability tending to 1 based on the random matrix theory in Bai and Silverstein (2010). The proof is given in Appendix C.

Lemma 1. If $\tau \in [0, 1)$ and Conditions A1 and A3 hold, then for any $\ell \in \mathbb{N}$, we have

$$\mathrm{P}(||X^TX/n||_2 \geq x_1) = o(n^{-\ell}),$$
$$\mathrm{P}(||(X^TX/n)^{-1}||_2 \geq x_2^{-1}) = o(n^{-\ell}),$$

where $x_1 = 4(1+\sqrt{\tau})^2||\Sigma||_2$ and $x_2 = (1-\sqrt{\tau})^2/(4||\Sigma^{-1}||_2)$.

Define event $K = H \cap J$, where H and J denote the events $||(X^TX/n)^{-1}||_2 < x_2^{-1}$ and $||X^TX/n||_2 < x_1$, respectively. Event K is introduced to truncate the eigenvalues of X^TX/n. Constants x_1 and x_2^{-1} may not be the smallest for our analysis, and can be replaced by any constants larger than them. From Lemma 1, for any $\ell \in \mathbb{N}$, we have $\mathrm{P}(\bar{K}) = o(n^{-\ell})$.

We now present our main result: the asymptotic normality and ratio consistency of $||\hat{\boldsymbol{\beta}}||_2^2$.

Theorem 1. (a) Assume $\tau \in [0, 1)$ and Condition A for (1). If either of the following conditions hold: (1) $\lim_{n\to\infty} p_n = \infty$; (2) p_n is fixed and $\boldsymbol{\beta}_0 \neq \mathbf{0}_{p_n}$, then,

$$\zeta_n^{-1}(||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2) \xrightarrow{\mathcal{D}} N(0, 1),$$

where

$$\zeta_n^2 = 4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0 + 2\sigma_\epsilon^4\mathrm{Etr}\{(X^TX)^{-2}\mathrm{I}(K)\} + 2\sigma_\epsilon^4[\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}]^2/(n-p_n).$$

(b) Additionally, if $p_n^{1/2}/n = o(\mathrm{SNR})$, then

$$\frac{||\hat{\boldsymbol{\beta}}||_2^2}{||\boldsymbol{\beta}_0||_2^2} \xrightarrow{\mathcal{P}} 1. \tag{11}$$

The proof of Theorem 1 relies on the martingale difference CLT Heyde and Brown (1970) and is provided in Appendix C.

A few remarks are in order: (i) After bias correction, $||\hat{\boldsymbol{\beta}}||_2^2$ is asymptotically normal under fixed, low or moderate dimension. Hence, the proposed method is adaptive to dimension and generally applicable for $p_n < n$ in practice. (ii) As $||\boldsymbol{\beta}_0||_2$ may vary with n, the ratio consistency of $||\hat{\boldsymbol{\beta}}||_2^2$ in (11) is not automatically implied by the asymptotic normality but requires an additional assumption $p_n^{1/2}/n = o(\mathrm{SNR})$. (iii) Random design is assumed in Theorem 1, but the result also holds for fixed design, that is, conditioning on X. (The SNR is not well defined for fixed design, and needs to be replaced by $||\boldsymbol{\beta}_0||_2^2/\sigma_\epsilon^2$ in the condition of part (b).) As discussed in the end of the proof of Theorem 1, there exists a set $\mathcal{X}_n \subseteq \mathbb{R}^{n\times p_n}$ as in (C.13) satisfying $\mathrm{P}(X \in \mathcal{X}_n) \to 1$, such that for any $x \in \mathcal{X}_n$, $t \in \mathbb{R}$ and $\varepsilon > 0$, $\mathrm{P}(\zeta_n^{-1}(||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2) \leq t|X = x) \to \Phi(t)$ and $\mathrm{P}(|||\hat{\boldsymbol{\beta}}||_2^2/||\boldsymbol{\beta}_0||_2^2 - 1| \geq \varepsilon|X = x) \to 0$. In the following, all theoretical results (theorems, corollaries, and propositions) are applicable to fixed design unless otherwise specified.

Remark 1. In Theorem 1, we assume homoscedasticity for the error. Here, we consider heteroscedasticity, that is, ϵ_i are independent with different variances $\sigma_i^2 = \mathrm{E}(\epsilon_i^2)$ for $i = 1, \ldots, n$. We first introduce a general result. For $k \in \mathbb{N}$, denoting $D = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2)$, then

$$\mathrm{E}\{\boldsymbol{\epsilon}^TX(X^TX)^{-k}X^T\boldsymbol{\epsilon}\} = \sum_{i=1}^n \mathrm{E}\{\boldsymbol{X}_i^T(X^TX)^{-k}\boldsymbol{X}_i\}\sigma_i^2$$
$$= \mathrm{Etr}\{X(X^TX)^{-k}X^T/n\}\mathrm{tr}(D) = \mathrm{Etr}\{(X^TX)^{-k+1}\}\mathrm{tr}(D)/n,$$

since $\mathrm{E}\{\boldsymbol{X}_i^T(X^TX)^{-k}\boldsymbol{X}_i\}$ are identical and equal $\mathrm{Etr}\{X(X^TX)^{-k}X^T/n\}$ for $i = 1, \ldots, n$. From (8), the bias of $||\hat{\boldsymbol{\beta}}||_2^2$ becomes $\mathrm{E}\{\boldsymbol{\epsilon}^TX(X^TX)^{-2}X^T\boldsymbol{\epsilon}\} = \mathrm{Etr}\{(X^TX)^{-1}\}\mathrm{tr}(D)/n$. In Lemma 2, $\mathrm{tr}\{(X^TX)^{-1}\}$ is ratio consistent for $\mathrm{Etr}\{(X^TX)^{-1}\}$ given event K, while $\hat{\sigma}_\epsilon^2$ is unbiased for $\mathrm{tr}(D)/n$ because $\mathrm{E}(\hat{\sigma}_\epsilon^2) = \mathrm{E}[\boldsymbol{\epsilon}^T\{\mathbf{I}_n - X(X^TX)^{-1}X^T\}\boldsymbol{\epsilon}/(n-p_n)] = \{\mathrm{tr}(D) - p_n/n\mathrm{tr}(D)\}/(n-p_n) = \mathrm{tr}(D)/n$. Hence, $||\hat{\boldsymbol{\beta}}||_2^2$ is still a bias-corrected estimator for $||\boldsymbol{\beta}_0||_2^2$ under heteroscedasticity. It will be an interesting future work to derive the asymptotic distribution of $||\hat{\boldsymbol{\beta}}||_2^2$ under heteroscedasticity, and derive consistent estimators for the parameters in the limiting distribution.

Remark 2. For ease of presenting the proofs, we assume $\mathrm{E}(\boldsymbol{X}_i) = \mathbf{0}_{p_n}$ in Condition A1 and hence $\mathrm{E}(Y_i) = 0$. For the general form of the linear model

$$Y_i = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\beta}_0 + \epsilon_i, \quad \text{for } i = 1, \ldots, n,$$

where α_0 is the intercept and $\mathrm{E}(\boldsymbol{X}_i) = \boldsymbol{\mu}$, our method is still applicable to the centralized data $\{Y_i - \bar{Y}, \boldsymbol{X}_i - \bar{\boldsymbol{X}}\}_{i=1}^n$ with $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ and $\bar{\boldsymbol{X}} = n^{-1}\sum_{i=1}^n \boldsymbol{X}_i$. Specifically, Theorem 1 together with the theorems, corollaries, and propositions below still holds after data centralization. Please see Section S.2 in the supplementary materials for a brief explanation.

Remark 3. From (4), the variance term of the conventional inference procedure $\hat{\zeta}_0^2 = 4\hat{\sigma}_\epsilon^2\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}$ is ratio consistent for $\zeta_0^2 = 4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0$ (see Theorem 4 to be introduced later). Hence, the removal of bias in $||\hat{\boldsymbol{\beta}}||_2^2$ leads to a larger variance ζ_n^2 than ζ_0^2. To see that more clearly, we consider a special case that $\boldsymbol{X}_i \overset{\mathrm{iid}}{\sim} N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$. In this case, $\mathrm{E}\{(X^TX)^{-1}\} = \mathbf{I}_{p_n}/(n-p_n-1)$ and $n\mathrm{Etr}\{(X^TX)^{-2}\} \to \tau/(1-\tau)^3$ based on Letac and Massam (2004). From (9), we know that the amount of theoretical correction of bias for $||\hat{\boldsymbol{\beta}}||_2^2$ compared with $||\hat{\boldsymbol{\beta}}||_2^2$ is $-\tau\sigma_\epsilon^2/(1-\tau)$. Also,

$$\zeta_n^2 = \zeta_0^2 + \frac{2\sigma_\epsilon^4}{n}\frac{\tau(1+\tau)}{(1-\tau)^3}\{1 + o(1)\}$$

for $\tau \in (0, 1)$. Both bias and variance corrections deviate from zero significantly as $\tau \to 1$; see Figure 3 for $n = 100$ and $\sigma_\epsilon^2 = 1$. The patterns in Figure 3 are consistent with the empirical ones observed in Figure 2.

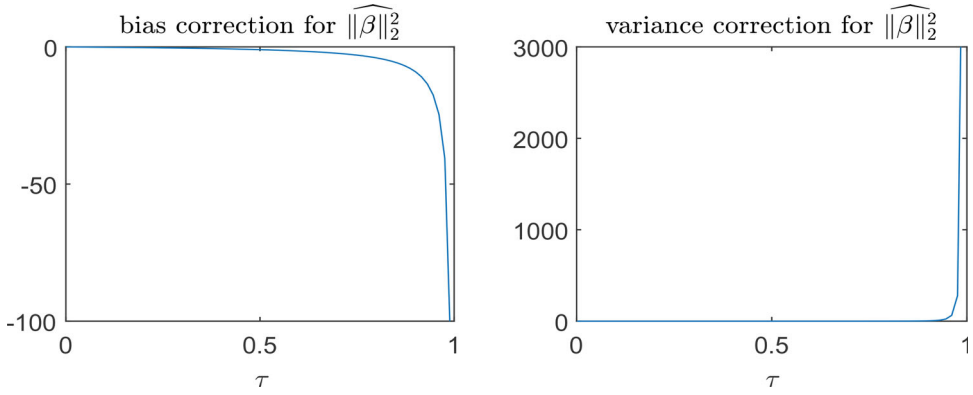

Figure 3. Amount of theoretical corrections of bias (left panel) and variance (right panel) versus τ for $||\hat{\boldsymbol{\beta}}||_2^2$ compared with $||\hat{\boldsymbol{\beta}}||_2^2$.

Remark 4. We now discuss that $||\hat{\boldsymbol{\beta}}||_2^2$ is not a uniformly minimum variance unbiased estimator (UMVUE). Denote by $T(\boldsymbol{Y}, X)$ a generic unbiased estimator of $||\boldsymbol{\beta}_0||_2^2$. If we assume $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\mathbf{0}_{p_n}, \Sigma)$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{I}_n)$, then the joint probability density function of $(Y_i, \boldsymbol{X}_i)$ is $f(y_i, \boldsymbol{x}_i) = \{(2\pi)^{p_n+1}\sigma_\epsilon^2|\Sigma|\}^{-1/2} \exp\{-(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta}_0)^2/(2\sigma_\epsilon^2) - \boldsymbol{x}_i^T\Sigma^{-1}\boldsymbol{x}_i/2\}$ implying that the Fisher information matrix with the full data $(\boldsymbol{Y}, X)$ for $\boldsymbol{\beta}_0$ is $I(\boldsymbol{\beta}_0) = n\Sigma/\sigma_\epsilon^2$. Using the Cramér–Rao lower bound (see, e.g., Shao 2003, Theorem 3.3), we have $\text{var}\{T(\boldsymbol{Y}, X)\} \geq 4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\Sigma^{-1}\boldsymbol{\beta}_0/n$. From the fact that $\text{E}\{(X^TX)^{-1}\} = \Sigma^{-1}/(n - p_n - 1)$, we know $\zeta_n^2 > 4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\Sigma^{-1}\boldsymbol{\beta}_0/n$ and hence $||\hat{\boldsymbol{\beta}}||_2^2$ may not be a UMVUE of $||\boldsymbol{\beta}_0||_2^2$.

As discussed in Remarks 3 and 4, the bias correction for $||\hat{\boldsymbol{\beta}}||_2^2$ leads to larger asymptotic variance and $||\hat{\boldsymbol{\beta}}||_2^2$ is not a UMVUE for $||\boldsymbol{\beta}_0||_2^2$. However, we can show that $||\hat{\boldsymbol{\beta}}||_2^2$ achieves the optimal rate of convergence in terms of the quadratic loss.

Theorem 2. Assume model (1) with $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\mathbf{0}_{p_n}, \Sigma)$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \sigma_\epsilon^2\mathbf{I}_n)$, $\sigma_\epsilon^2 = O(n)$, that $\{\epsilon_i\}_{i=1}^n$ are independent of $\{\boldsymbol{X}_i\}_{i=1}^n$, and Conditions A3 and A4 hold. For any estimator T of $||\boldsymbol{\beta}_0||_2^2$, we have

$$\inf_T \sup_{\boldsymbol{\beta} \in \mathcal{G}_{\boldsymbol{\beta}_0}(c)} \text{E}_{(\boldsymbol{Y},X)|(\boldsymbol{\beta},\sigma_\epsilon^2,\Sigma)}(T - ||\boldsymbol{\beta}||_2^2)^2 = \Omega(\zeta_n^2),$$

where $\mathcal{G}_{\boldsymbol{\beta}_0}(c) = \{\boldsymbol{\beta} \in \mathbb{R}^{p_n} : ||\boldsymbol{\beta}||_\infty \leq C < \infty, ||\boldsymbol{\beta}||_2 \leq c(||\boldsymbol{\beta}_0||_2 + \sigma_\epsilon\sqrt{p_n}/\sqrt{n})\}$, $c > 1$ is a generic constant and $\text{E}_{(\boldsymbol{Y},X)|(\boldsymbol{\beta},\sigma_\epsilon^2,\Sigma)}(\cdot)$ denotes taking expectation with respect to $(\boldsymbol{Y}, X)$ given parameters $(\boldsymbol{\beta}, \sigma_\epsilon^2, \Sigma)$.

Theorem 2 implies that, $||\hat{\boldsymbol{\beta}}||_2^2$ achieves the optimal convergence rate over all $\boldsymbol{\beta} \in \mathcal{G}_{\boldsymbol{\beta}_0}(c)$ under the quadratic loss. Since ζ_n^2 involves the true parameter $\boldsymbol{\beta}_0$, the set of parameters $\mathcal{G}_{\boldsymbol{\beta}_0}(c)$ also depends on $\boldsymbol{\beta}_0$, which covers a wide range of p_n-dimensional vectors including $\boldsymbol{\beta}_0$.

To estimate the variance term ζ_n^2, we need to estimate the following four terms: σ_ϵ^2, $\text{II} := \boldsymbol{\beta}_0^T\text{E}\{(X^TX)^{-1}\text{I}(K)\}\boldsymbol{\beta}_0$, $\text{III} := \text{Etr}\{(X^TX)^{-2}\text{I}(K)\}$ and $\text{IV} := \text{Etr}\{(X^TX)^{-1}\text{I}(K)\}$. The error variance σ_ϵ^2 can be consistently estimated by $\hat{\sigma}_\epsilon^2$ as in Proposition 1 to be introduced in Section 3.3. For the other three terms, we need to utilize the following general result.

Lemma 2. Assume $\tau \in [0, 1)$ and Conditions A1 and A3 for (1). For any $k \in \mathbb{N}$,

$$\text{var}[\text{tr}\{(X^TX/n)^{-k}\}\text{I}(K)] = o(p_n^2),$$
$$\text{var}\{\boldsymbol{\beta}_0^T(X^TX/n)^{-k}\boldsymbol{\beta}_0\text{I}(K)\} = o(||\boldsymbol{\beta}_0||_2^4).$$

The key strategy to prove Lemma 2 is the leave-one-observation-out method. See Appendix C for the detailed proof. According to Lemma 2, $\text{tr}\{(X^TX)^{-2}\}$ and $\text{tr}\{(X^TX)^{-1}\}$ are ratio consistent for terms III and IV, respectively. It's not necessary to include $\text{I}(K)$ in the estimators, since $\text{I}(K) \xrightarrow{\mathcal{P}} 1$ due to $\text{P}(K) \to 1$. Lemma 2 further induces Lemma S.12 in the supplementary materials that,

$$\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}} - \hat{\sigma}_\epsilon^2\text{tr}\{(X^TX)^{-2}\} - \text{II} = o_\text{P}(\zeta_n^2/\sigma_\epsilon^2).$$

Subsequently, the plug-in estimator of ζ_n^2 is

$$\hat{\zeta}_n^2 = 4\hat{\sigma}_\epsilon^2\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}} - 2\hat{\sigma}_\epsilon^4\text{tr}\{(X^TX)^{-2}\} + 2\hat{\sigma}_\epsilon^4[\text{tr}\{(X^TX)^{-1}\}]^2/(n - p_n).$$

We summarize the above discussion into Theorem 3.

Theorem 3. Under the conditions in part (a) of Theorem 1, we have

$$\hat{\zeta}_n^2/\zeta_n^2 \xrightarrow{\mathcal{P}} 1.$$

The proof of Theorem 3 is provided in Appendix C.

We are now ready to test the hypothesis in (2) by proposing the following test statistic

$$\mathbb{Z}_n = \frac{||\hat{\boldsymbol{\beta}}||_2^2 - c_0^2}{\hat{\zeta}_n}. \tag{12}$$

Theorems 1 and 3 directly imply that the null limiting distribution of $\mathbb{Z}_n$ is standard normal, the p-value for testing (2) is $2\Phi(-|\mathbb{z}_n|)$, where $\mathbb{z}_n$ is a realization of $\mathbb{Z}_n$, and the asymptotic power function is $1 - \Phi\{-(c_1^2 - c_0^2)/\hat{\zeta}_n + \Phi^{-1}(1 - \alpha/2)\} + \Phi\{-(c_1^2 - c_0^2)/\hat{\zeta}_n - \Phi^{-1}(1 - \alpha/2)\}$ under the fixed alternative $H_1 : ||\boldsymbol{\beta}_0||_2 = c_1 \neq c_0$.

In the end, we point out that as long as the SNR is large enough, the conventional estimator $||\hat{\boldsymbol{\beta}}||_2^2$ is still ratio consistent and asymptotically normal. However, the strength of SNR is

usually unknown in practice. Hence, this result highlights the importance of our proposed *adaptive* method that works for both moderate and low dimensions, regardless of weak or strong signals.

Theorem 4. Assume $\tau \in [0, 1)$ and Condition A for (1). Then,

$$\frac{||\hat{\boldsymbol{\beta}}||_2^2}{||\boldsymbol{\beta}_0||_2^2} \xrightarrow{\mathcal{P}} 1 \Longleftrightarrow p_n/n = o(\text{SNR}) \Longleftrightarrow \frac{\hat{\zeta}_0^2}{\zeta_0^2} \xrightarrow{\mathcal{P}} 1,$$

$$\frac{||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2}{\zeta_0} \xrightarrow{\mathcal{D}} N(0,1) \Longleftrightarrow p_n^2/n = o(\text{SNR}) \Longrightarrow \tau = 0,$$

where $\zeta_0^2 = 4\sigma_\epsilon^2 \boldsymbol{\beta}_0^T \mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0$ and $\hat{\zeta}_0^2$ is defined in (4).

Theorem 4 verifies the asymptotic normality of $\mathbb{Z}_0$ in (3) under valid null hypothesis. The proof is provided in the supplementary materials.

3. Applications

This section consists of four important applications of our general theory: signal detection, global testing, inferences for the error variance, and fraction of variance explained. The results on two-sample inference are postponed to the supplementary materials.

3.1. Detection Boundary for $||\boldsymbol{\beta}_0||_2^2$

Hypothesis (2) can be used to perform signal detection by setting $c_0 = 0$. In this problem, the detection boundary is often of interest, which is the smallest separation rate between the null and a sequence of contiguous alternatives H_{1n} indexed by $\delta_n \to 0$, that is,

$$H_{1n} : ||\boldsymbol{\beta}_0||_2^2 = \delta_n,$$

such that successful detection is still possible. From Theorem 1, we propose the following test statistic for hypothesis (2) with $c_0 = 0$

$$\mathbb{Z}_n^* = \frac{||\hat{\boldsymbol{\beta}}||_2^2}{\hat{\zeta}_*},$$

where $\hat{\zeta}_*^2 = 2\hat{\sigma}_\epsilon^4 \mathrm{tr}\{(X^TX)^{-2}\} + 2\hat{\sigma}_\epsilon^4[\mathrm{tr}\{(X^TX)^{-1}\}]^2/(n - p_n)$. The difference between $\mathbb{Z}_n^*$ and $\mathbb{Z}_n$ lies in the variance term $\hat{\zeta}_*^2$. Using Lemma 2, $\hat{\zeta}_*^2$ is ratio consistent for $\zeta_*^2 = 2\sigma_\epsilon^4 \mathrm{Etr}\{(X^TX)^{-2}\mathrm{I}(K)\} + 2\sigma_\epsilon^4[\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}]^2/(n - p_n)$ which equals ζ_n^2 when $\boldsymbol{\beta}_0 = \mathbf{0}_{p_n}$. In other words, $\hat{\zeta}_*^2$ is a refined estimator of ζ_n^2 under the null hypothesis (2) with $c_0 = 0$. Then, the asymptotic standard normality of $\mathbb{Z}_n^*$ under the null follows directly from Theorem 1 for diverging p_n. Corollary 1 presents the detection boundary using $\mathbb{Z}_n^*$.

Corollary 1. Assume that $\tau \in [0, 1)$, $\lim_{n\to\infty} p_n = \infty$, Condition A holds for (1). If $\delta_n = \Omega(\sigma_\epsilon^2 p_n^{1/2}/n)$, then

$$\mathbb{Z}_n^* - \hat{\zeta}_*^{-1}\delta_n \xrightarrow{\mathcal{D}} N(0,1),$$

where $\hat{\zeta}_*^{-1}\delta_n = \Omega_{\mathrm{P}}(1)$. If $\delta_n = o(\sigma_\epsilon^2 p_n^{1/2}/n)$, then

$$\mathbb{Z}_n^* \xrightarrow{\mathcal{D}} N(0,1).$$

Therefore, the detection boundary is $\sigma_\epsilon^2 p_n^{1/2}/n$, which matches with the minimax detection rate in Ingster, Tsybakov, and Verzelen (2010) (see (1.2) therein). It is worth mentioning that Corollary 1 requires diverging p_n.

3.2. Global Inference for $\boldsymbol{\beta}_0$

This section is concerned with the global hypothesis (5)

$$H_0 : \boldsymbol{\beta}_0 = \boldsymbol{\beta}_0^{\text{null}} \quad \text{versus} \quad H_1 : \boldsymbol{\beta}_0 \neq \boldsymbol{\beta}_0^{\text{null}},$$

by proposing a bias-and-variance-corrected test statistic based on $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^{\text{null}}||_2^2$ as follows

$$\mathbb{G}_n = \frac{||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^{\text{null}}||_2^2 - \mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2}{\hat{\zeta}_*}. \tag{13}$$

The construction of $\mathbb{G}_n$ is based on the fact that the distribution of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0^{\text{null}}$ under H_0 in (5) is the same as that of $\hat{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}_0 = \mathbf{0}_{p_n}$. Thus, the amount of bias correction $-\mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2$ and the variance term $\hat{\zeta}_*^2$ for $\mathbb{G}_n$ are the same as those for $\mathbb{Z}_n^*$. From the asymptotic results of $\mathbb{Z}_n^*$, $\mathbb{G}_n$ is also asymptotically standard normal under the null for diverging p_n, and the smallest separation rate for $H_{1n} : ||\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^{\text{null}}||_2^2 = \delta_n$ is $\delta_n^* = \sigma_\epsilon^2 p_n^{1/2}/n$, the same as that identified by Corollary 1.

From (13), we can construct $1 - \alpha$ confidence regions for $\boldsymbol{\beta}_0$ using one-sided and two-sided strategies as

$$\begin{aligned}
\mathrm{CR}_1 &= \{\boldsymbol{\beta} : ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 - \mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2 \leq \Phi^{-1}(1-\alpha)\hat{\zeta}_*\};\\
\mathrm{CR}_2 &= \{\boldsymbol{\beta} : |||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_2^2 - \mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2| \\
&\qquad \leq \Phi^{-1}(1-\alpha/2)\hat{\zeta}_*\}.
\end{aligned} \tag{14}$$

Confidence region for high-dimensional sparse $\boldsymbol{\beta}_0$ was studied in Cai and Guo (2020) and Nickl and van de Geer (2013).

3.3. Inference for σ_ϵ^2

This section is concerned with moderate-dimensional inference for the error variance σ_ϵ^2.

Proposition 1. Assume $\tau \in [0, 1)$ and Conditions A1–A3 for (1). Then

$$\frac{\hat{\sigma}_\epsilon^2}{\sigma_\epsilon^2} \xrightarrow{\mathcal{P}} 1 \quad \text{and} \quad \frac{\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2}{\zeta_\epsilon} \xrightarrow{\mathcal{D}} N(0,1),$$

where $\zeta_\epsilon^2 = n^{-1}\{\nu_4 + \sigma_\epsilon^4(3\tau - 1)/(1 - \tau)\}$ and $\nu_4 = \mathrm{E}(\epsilon_i^4)$.

Our result is adaptive to data dimension by incorporating τ in the variance term ζ_ϵ^2 for both fixed and diverging p_n. Specifically, ζ_ϵ^2 increases with τ. For a special case that $\epsilon_i \sim N(0, \sigma_\epsilon^2)$, $\zeta_\epsilon^2 = 2\sigma_\epsilon^4/\{n(1 - \tau)\}$.

To estimate ζ_ϵ^2, it suffices to provide a ratio consistent estimator for ν_4. We first examine a straightforward estimator $1/n\sum_{i=1}^n \hat{\epsilon}_i^4$ where $(\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n)^T = \boldsymbol{Y} - X\hat{\boldsymbol{\beta}}$. However, as in the proof of Lemma S.15 in the supplementary materials,

$1/n\mathrm{E}(\sum_{i=1}^n \hat{\epsilon}_i^4) = (1-\tau)^4\nu_4 + 3\sigma_\epsilon^4\tau(1-\tau)^2(2-\tau) + o(\sigma_\epsilon^4)$. Although the naive estimator is biased, it induces an estimator for ν_4 after centering and rescaling

$$\hat{\nu}_4 = (1-p_n/n)^{-4} \times \Big\{1/n\sum_{i=1}^n \hat{\epsilon}_i^4 - 3\hat{\sigma}_\epsilon^4(p_n/n)(1-p_n/n)^2(2-p_n/n)\Big\}.$$

Lemma S.15 demonstrates that $\hat{\nu}_4$ is ratio consistent for ν_4. Hence, the plug-in estimator $\hat{\zeta}_\epsilon^2 = n^{-1}\{\hat{\nu}_4 + \hat{\sigma}_\epsilon^4(3p_n/n - 1)/(1-p_n/n)\}$ is ratio consistent for ζ_ϵ^2. From Proposition 1, we have $(\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)/\hat{\zeta}_\epsilon \xrightarrow{\mathcal{D}} N(0,1)$, which can be used to conduct inference for σ_ϵ^2.

3.4. Inference for ρ_0

Consider the hypotheses

$$H_0 : \rho_0 \geq \rho_0^{\mathrm{null}} \quad \text{versus} \quad H_1 : \rho_0 < \rho_0^{\mathrm{null}}, \tag{15}$$

where $0 < \rho_0^{\mathrm{null}} < 1$ is a given constant. Recalling the definition of ρ_0 in (6), its conventional plug-in estimator can be obtained by replacing $\eta_0 := \boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$ and σ_ϵ^2 with $\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}}$ and $\hat{\sigma}_\epsilon^2$, respectively. The asymptotic normality of this estimator is studied under $\tau = 0$ in Theorem S.1 in the supplementary materials followed by the low-dimensional inference for ρ_0.

However, in the moderate-dimensional regime, the bias of $\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}}$ for η_0 is nonignorable, that is,

$$\begin{aligned}
&\mathrm{E}\{\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}}\} - \eta_0 \\
&= \mathrm{E}\{\boldsymbol{\beta}_0^T(X^TX/n)\boldsymbol{\beta}_0\} - \eta_0 + 2n^{-1}\mathrm{E}(\boldsymbol{\beta}_0^TX^T\boldsymbol{\epsilon}) \\
&\quad + n^{-1}\mathrm{E}\{\boldsymbol{\epsilon}^TX(X^TX)^{-1}X^T\boldsymbol{\epsilon}\} \\
&= n^{-1}\mathrm{E}\{\boldsymbol{\epsilon}^TX(X^TX)^{-1}X^T\boldsymbol{\epsilon}\} \\
&= \sigma_\epsilon^2 p_n/n \to \sigma_\epsilon^2\tau > 0.
\end{aligned}$$

Consequently, we propose an unbiased estimator for η_0 as

$$\hat{\eta} = \hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}} - \hat{\sigma}_\epsilon^2 p_n/n.$$

Hence, a new plug-in estimator for ρ_0 is

$$\hat{\rho} = \frac{\hat{\eta}}{\hat{\eta} + \hat{\sigma}_\epsilon^2} = \frac{\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}} - \hat{\sigma}_\epsilon^2 p_n/n}{\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}} + \hat{\sigma}_\epsilon^2(1-p_n/n)}$$

with the following asymptotic distribution.

Theorem 5. Assume $\tau \in [0,1)$, $\rho_0 \in [C_1, C_2]$ for some constants $0 < C_1 \leq C_2 < 1$ and Condition A holds for (1). Then, $\hat{\rho} - \rho_0 = o_{\mathrm{P}}(1)$ and

$$\frac{\hat{\rho} - \rho_0}{\sigma_{\hat{\rho}}} \xrightarrow{\mathcal{D}} N(0,1),$$

where $\sigma_{\hat{\rho}}^2 = n^{-1}(\eta_0 + \sigma_\epsilon^2)^{-4}[2\sigma_\epsilon^8\tau/(1-\tau) - \{2 + 4\tau/(\tau - 1)\}\sigma_\epsilon^6\eta_0 + \sigma_\epsilon^4\{\mathrm{E}(Y_1^4) - \nu_4 + \eta_0^2(4\tau - 2)/(1-\tau)\} + \eta_0^2\nu_4]$.

Simple calculation implies that $\hat{\rho} = 1 - (1-p_n/n)^{-1}||\boldsymbol{Y} - X\hat{\boldsymbol{\beta}}||_2^2/||\boldsymbol{Y}||_2^2 = 1 - (1-p_n/n)^{-1}(1-R^2)$, where R^2 is the coefficient of determination. Therefore, if $\tau = 0$, then R^2 is asymptotically unbiased for ρ_0, but when $\tau > 0$, a rescaled R^2, that is, $\hat{\rho}$, is required for the inference of ρ_0. The definition of SNR is not applicable to fixed design, and hence the results of Theorem 5 and Proposition 2 are not available for fixed design.

For the variance term $\sigma_{\hat{\rho}}^2$, the plug-in estimator $\hat{\sigma}_{\hat{\rho}}^2$ is obtained by replacing $\mathrm{E}(Y_1^4)$, η_0, σ_ϵ^2, ν_4, and τ in $\sigma_{\hat{\rho}}^2$ with $n^{-1}\sum_{i=1}^n Y_i^4$, $\hat{\eta}$, $\hat{\sigma}_\epsilon^2$, $\hat{\nu}_4$, and p_n/n, respectively, and its consistency is demonstrated below.

Proposition 2. Assume the conditions in Theorem 5. Then, $\sigma_{\hat{\rho}}^2 = \Omega(1/n)$ and

$$\hat{\sigma}_{\hat{\rho}}^2 - \sigma_{\hat{\rho}}^2 = o_{\mathrm{P}}(1/n).$$

Hence, (15) can be tested by $\hat{\sigma}_{\hat{\rho}}^{-1}(\hat{\rho} - \rho_0^{\mathrm{null}})$, whose null limiting distribution is standard normal. Also the smallest separation rate for contiguous alternative is $n^{-1/2}$.

For the inference of η_0, the proof of Theorem 5 immediately implies that

$$\sigma_{\hat{\eta}}^{-1}(\hat{\eta} - \eta_0) \xrightarrow{\mathcal{D}} N(0,1), \tag{16}$$

with $\sigma_{\hat{\eta}}^2 = n^{-1}\{\mathrm{E}(Y_1^4) - \nu_4 - 2\sigma_\epsilon^2\eta_0 - \eta_0^2 + 2\sigma_\epsilon^4 p_n/(n-p_n)\}$ which is consistently estimated by the plug-in estimator following the proof of Proposition 2.

In the end, we comment on related works concerned with signal strength (i.e., Dicker 2014; Dicker and Erdogdu 2016, 2017; Janson, Barber, and Candès 2017; Verzelen and Gassiat 2018). The first three works, that is, Dicker and Erdogdu (2016), Janson, Barber, and Candès (2017), and Dicker (2014), conduct statistical inference for moderate-dimensional fixed effect models. However, our OLS-based methods are essentially different from their methods in the following aspects: parameters of interest and weak assumptions.

First, the parameter of interest in the aforementioned three works is $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$, and their procedures crucially rely on the fact that $Y_i \sim N(0, \boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0 + \sigma_\epsilon^2)$ and that $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$ is a part of the variance term. Therefore, their results are not readily translated into inference for our parameter of interest, that is, $||\boldsymbol{\beta}_0||_2^2$, unless Σ is identity. In contrast, our strategy depends on the OLS estimator. This flexibility also allows us to conduct inference for $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$ based on a bias-corrected version of $\hat{\boldsymbol{\beta}}^T(X^TX/n)\hat{\boldsymbol{\beta}}$ as in (16), and even two-sample inferences, for example, the coheritability (7), to which it is unclear how their methods can be applied.

Second, some assumptions of our article are weaker due to the use of different technical tools. Specifically, the proofs as well as the development of the estimation and inference procedures in the aforementioned three works rely heavily on the Gaussian assumption of the design matrix X and error $\boldsymbol{\epsilon}$. For example, among other implications, the Gaussian design is important in deriving the invariant distribution of X under orthogonal transformations in Dicker and Erdogdu (2016), the Haar distribution of the right-singular vectors from the singular value decomposition of X in Janson, Barber, and Candès (2017)

and the Wishart distribution of X^TX in Dicker (2014). However, our OLS-based result is derived using the martingale difference CLT without requiring any specific distributional assumption of X or $\boldsymbol{\epsilon}$. Also, our results can be easily extended to fixed design. Besides, the three works above need $\Sigma = \mathbf{I}_{p_n}$ to conduct inference for $||\boldsymbol{\beta}_0||_2^2$. Even for the inference of $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$, they still require strong conditions on Σ, for example, known or consistently estimable Σ. And, some further sparsity assumptions need to be imposed if Σ will be estimated. Our inference methods for $||\boldsymbol{\beta}_0||_2^2$ and $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$ neither need known or a consistent estimator of Σ nor require any sparsity assumption on Σ. From the simulations in the end of Section 4, our method performs better than or at least as well as those in Dicker (2014) and Dicker and Erdogdu (2016).

As for Dicker and Erdogdu (2017) and Verzelen and Gassiat (2018), the former conducted inference for the variance of the regression parameter by considering the "random effect" model conditioning on the design matrix, and hence is different from the setup of fixed effect model in our article; the latter derives the minimax estimators of ρ_0 under Gaussian design and error, but they did not derive the asymptotic distribution of the estimators and hence their results cannot be applied to the inference for ρ_0.

4. Simulations

Numerical studies are conducted to support the proposed statistical inference procedures. Set $n \in \{400, 800\}$ and $p_n = 4, \lfloor n/6 \rfloor, n/4, n/2.5$ corresponding to fixed dimension ($p_n = 4$), low dimension ($p_n = \lfloor n/6 \rfloor$) and moderate dimension ($p_n = n/4, n/2.5$), unless otherwise specified. In the simulations, we consider a general form of the linear model $Y_i = 1 + \boldsymbol{X}_i^T\boldsymbol{\beta}_0 + \epsilon_i$ with $\mathrm{E}(\boldsymbol{X}_i) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_{p_n})^T$ generated by $\{\mu_i\}_{i=1}^{p_n} \overset{\text{iid}}{\sim} \text{Unif}[1, 2]$. Both $\boldsymbol{Y}$ and $\{\boldsymbol{X}_i\}_{i=1}^n$ are centralized before conducting inference for the quadratic functionals. To generate data, we consider the following four cases representing various situations in reality:

I. Gaussian design with $\Sigma = \mathbf{I}_{p_n}$: $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \mathbf{I}_{p_n})$ for $i = 1, \ldots, n$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \mathbf{I}_n)$ and $\boldsymbol{\beta}_0 = \tilde{\boldsymbol{\beta}}/||\tilde{\boldsymbol{\beta}}||_2$ with $\{\tilde{\beta}_j\}_{j=1}^{p_n} \overset{\text{iid}}{\sim} \text{Unif}[1, 2]$;

II. Gaussian design with general Σ: $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \Sigma)$ for $i = 1, \ldots, n$ and $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \mathbf{I}_n)$, where $\Sigma = \Sigma^{*T}\Sigma^*/\lambda_{\max}(\Sigma^{*T}\Sigma^*) + \text{diag}(d_1, \ldots, d_{p_n})$, $\Sigma_{ij}^* \overset{\text{iid}}{\sim} \text{Unif}[-0.5, 0.5]$ for $1 \le i, j \le p_n$ and $\{d_i\}_{i=1}^{p_n} \overset{\text{iid}}{\sim} \text{Unif}[0.4, 1]$. Here $\boldsymbol{\beta}_0 = c_{p_n}\boldsymbol{\beta}^*$ where $\boldsymbol{\beta}^*$ is the normalized eigenvector of Σ corresponding to the smallest eigenvalue and $c_{p_n} = 1$ for $p_n = 4$, $c_{p_n} = 2$ for $p_n = \lfloor n/6 \rfloor, n/4$ and $c_{p_n} = 5$ for $p_n = n/2.5$;

III. t-distributed design: $X_{ij} - \mu_j \overset{\text{iid}}{\sim} t_5/\sqrt{5/3}$ for $i = 1, \ldots, n; j = 1, \ldots, p_n$, $\epsilon_i \overset{\text{iid}}{\sim} t_{16}/\sqrt{8/7}$ and $\boldsymbol{\beta}_0 = (1, 1, 1, 0, \ldots, 0)^T$;

IV. fixed design: X is identical for all replications and generated by $\boldsymbol{X}_i \overset{\text{iid}}{\sim} N(\boldsymbol{\mu}, \mathbf{I}_{p_n})$, while $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \mathbf{I}_n)$ are independently generated for each replication, and $\boldsymbol{\beta}_0$ is the normalized eigenvector of $\sum_{i=1}^n (\boldsymbol{X}_i - \bar{\boldsymbol{X}})(\boldsymbol{X}_i - \bar{\boldsymbol{X}})^T$ corresponding to the largest eigenvalue.

In Case II, Σ is not necessarily sparse and allows different diagonal elements. Case III is for non-Gaussian design and Case IV corresponds to fixed design matrix. In what follows, QQ plots (for the 1st to 99th percentiles) under the valid null hypotheses and confidence intervals were obtained with 1000 replications, while the power function was computed using 500 replications for each setup.

First, consider hypothesis (2) with $c_0 = 0$. Data are generated by Cases I–IV with $\boldsymbol{\beta}_0 = \mathbf{0}_{p_n}$. From Figure 4, under low and moderate dimensions, $\mathbb{Z}_n^*$ follows standard normal distribution under the null hypothesis. The fixed-dimensional results are not reported as the signal detection is only conducted for diverging p_n. The empirical power of $\mathbb{Z}_n^*$ is given in Figure 5 by varying $\boldsymbol{\beta}_0 = \mathbf{1}_{p_n}\delta\sigma_\epsilon/(n^{1/2}p_n^{1/4})$ with $\delta = 0, 0.5, 1, 1.5, \ldots, 6$. This choice of alternative values is supported by the derived detection boundary $\delta_n^* = \sigma_\epsilon^2 p_n^{1/2}/n$ for signal detection. From Figure 5, we can tell that the empirical rejection rate grows from the nominal level to one as δ increases from zero.

We also check the coverage probability of the two-sided (CR_2) and one-sided (CR_1) confidence regions of $\boldsymbol{\beta}_0$ based on (14), with 1000 replications at $\alpha = 0.05$. Table 1 reveals that both CR_1 and CR_2 are satisfactory while the latter slightly outperforms the former. The coverage probabilities of CR_2 are around 0.95 while those of CR_1 are generally below 0.95. Hence, we suggest to use CR_2 in practice. Note that our proposed method particularly works for diverging p_n, but when p_n is fixed, the finite-sample performance is still satisfactory.

Testing error variance:

$$H_0 : \sigma_\epsilon^2 = 1 \quad \text{versus} \quad H_1 : \sigma_\epsilon^2 \neq 1 \tag{17}$$

is performed by test statistic $(\hat{\sigma}_\epsilon^2 - 1)/\hat{\zeta}_\epsilon$. Figure 6 provides the QQ plots of the test statistic under the null hypothesis. Clearly, the proposed test statistic well adapts to fixed-, low-, and moderate-dimensional regimes. The empirical powers under $\sigma_\epsilon^2 - 1 = \delta/n^{1/2}$ are provided in Figure 7 with $\delta = -10, -8, \ldots, 0, \ldots, 8, 10$. Again, the power behaviors are satisfactory.

We compare the performance of the conventional (in Section S.3) and proposed test statistics for testing $H_0 : \rho_0 \ge \rho_0^{\text{null}}$, that is, (15). Figures 8 and 9 provide the QQ plots of the conventional and proposed test statistics, respectively. We find that both the conventional and proposed tests perform well for the fixed dimension. Under low and moderate dimensions, the conventional method fails but the proposed test continues to perform satisfactorily.

In the end, we consider the performance of the confidence intervals for three parameters $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$, σ_ϵ^2 and ρ_0 in Tables 2–4, respectively. The averaged coverage probability and length of the confidence intervals are calculated using the proposed method, the MLE method in Dicker and Erdogdu (2016), the method of moment for $\Sigma = \mathbf{I}_{p_n}$ (MM_1) in Dicker (2014) (see Corollary 1 therein) and its alternative version for unknown Σ (MM_2) (see Proposition 2 therein). The MLE and MM_1 methods are applied by assuming that $\Sigma = \mathbf{I}_{p_n}$, that is, Σ is correctly identified in Cases I and III but misidentified in Cases II and IV. Since the EigenPrism method in Janson, Barber, and Candès (2017) is particularly proposed for $p_n > n$, it is not compared with our methods. For all three parameters, in Case I with Gaussian

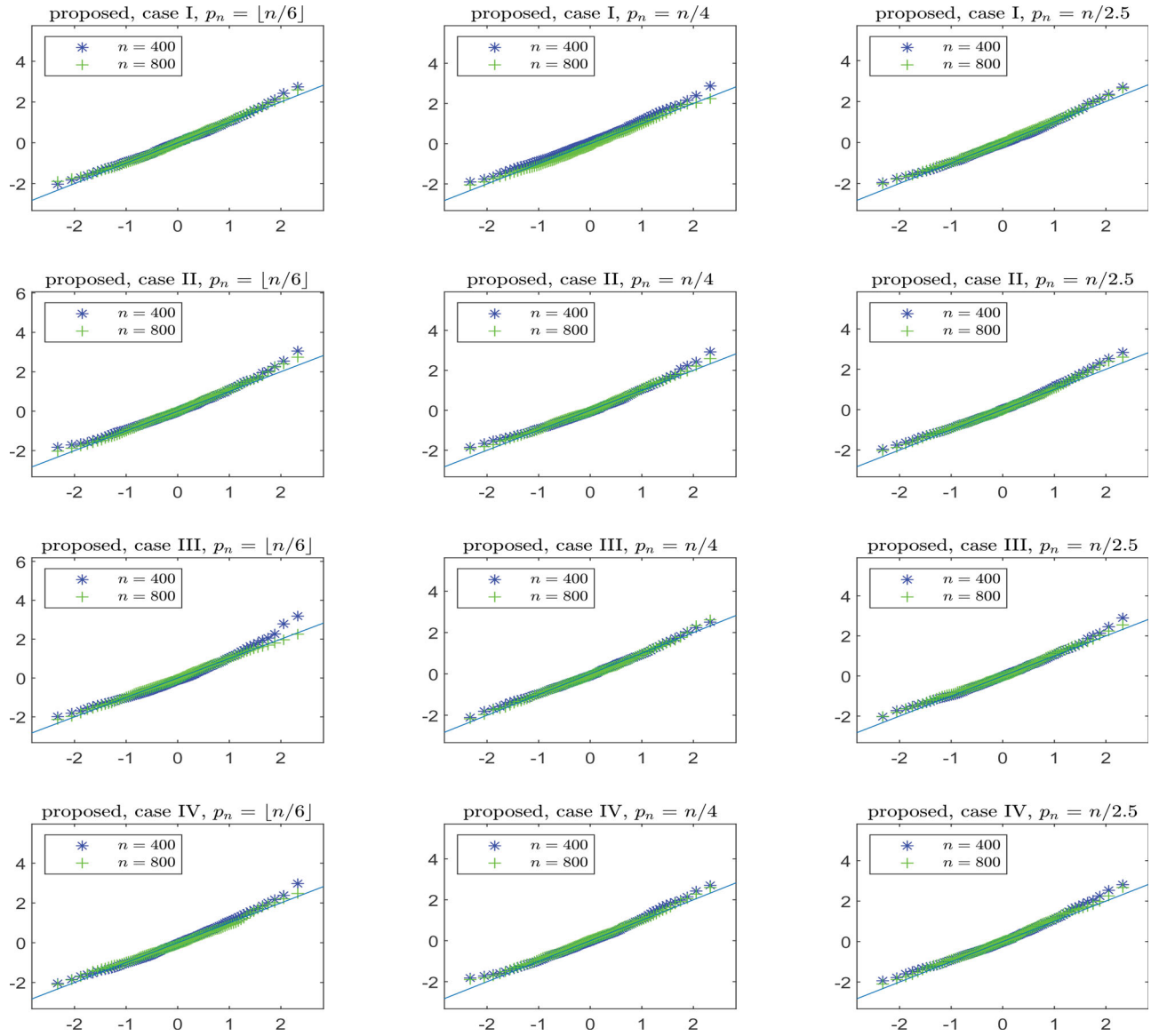

Figure 4. QQ plots for testing H_0 in (2) with $c_0 = 0$ using $\mathbb{Z}_n^*$. Panels from top to bottom are for Cases I–IV, respectively, while panels from left to right are for $p_n = \lfloor n/6 \rfloor, n/4, n/2.5$, respectively.

design and $\Sigma = \mathbf{I}_{p_n}$, all methods are satisfactory with coverage probability close to the nominal level 95%. For Cases II–IV, our method still performs well with the coverage probability close to 95% for all three parameters. But, due to the misidentified Σ or non-Gaussian design, the coverage probabilities of the MLE method are away from 95% for $\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0$ and ρ_0 in Case II and for σ_ϵ^2 in Cases III and IV, while the MM_1 and MM_2 methods result in invalid confidence intervals in most situations. Since $\boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0$ and ρ_0 are not defined for fixed design, the corresponding results for Case IV are not reported.

5. Real Data

We study a dataset from the International HapMap Project to investigate the relationship between gene expression and single nucleotide polymorphism (SNP). In Stranger et al. (2007), it is revealed that the expression levels of certain genes are associated with its nearby SNPs. Specifically, they identified 803 genes that were significantly associated with certain SNPs located within 1-Mbp of the gene midpoint using 30 Caucasian trios of northern and western European origin (CEU), 45 unrelated Chinese individuals from Beijing University (CHB), 45 unrelated Japanese individuals from Tokyo (JPT), and 30 Yoruba trios from Ibadan, Nigeria (YRI). We select 9 genes among these 803 genes and investigate the relationship between each gene and its nearby SNPs from n = 377 individuals (80 individuals in CHB population, 82 from JPT, 107 from CEU, and 108 from YRI). We use the gene expression dataset from *https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-264/* and *https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-198/*. The SNP data were obtained from the International HapMap Project (*https://www.ncbi.nlm.nih.gov/variation/news/NCBI_retiring_HapMap/*), release 28.

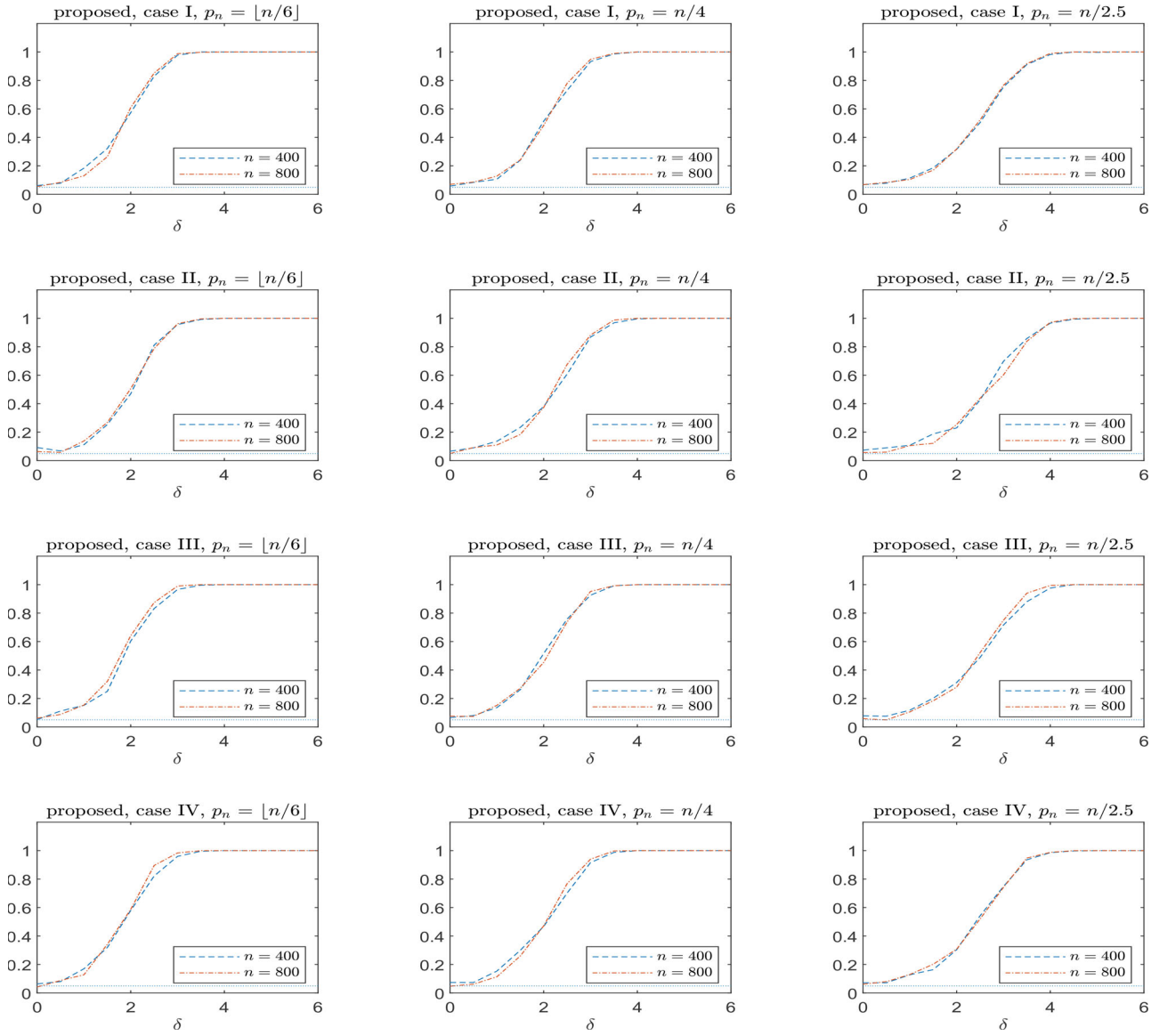

Figure 5. Empirical rejection rates versus δ for testing H_1 in (2) using $\mathbb{Z}_n^*$. Panels from top to bottom are for Cases I–IV, respectively, while panels from left to right are for $p_n = \lfloor n/6 \rfloor, n/4, n/2.5$, respectively. The dotted line indicates the true significance level $\alpha = 0.05$.

Table 1. Coverage probability of 95% confidence regions for β_0.

(n, p_n)	Case I	Case II	Case III	Case IV
		CR_2		
(400, 4)	0.943	0.946	0.950	0.932
(400, 66)	0.945	0.946	0.941	0.945
(400, 100)	0.949	0.948	0.951	0.953
(400, 160)	0.950	0.940	0.946	0.942
(800, 4)	0.944	0.960	0.964	0.948
(800, 133)	0.958	0.945	0.959	0.951
(800, 200)	0.944	0.961	0.947	0.957
(800, 320)	0.952	0.943	0.957	0.954
		CR_1		
(400, 4)	0.913	0.921	0.936	0.908
(400, 66)	0.925	0.931	0.920	0.928
(400, 100)	0.933	0.933	0.944	0.916
(400, 160)	0.941	0.923	0.936	0.911
(800, 4)	0.929	0.938	0.942	0.923
(800, 133)	0.951	0.935	0.953	0.937
(800, 200)	0.936	0.940	0.953	0.941
(800, 320)	0.922	0.928	0.942	0.934

For each selected gene, we only focus on the SNPs that are significantly associated with this gene. A list of these significant SNPs for each identified gene is provided in Supplementary Table S2 of Stranger et al. (2007). Furthermore, among these significant SNPs, we only choose those with minor allele frequency greater than 5% and missingness no larger than 20%. For the selected SNPs included in the analysis, we impute the missing values by the marginal mean. The minor allele counts are assigned as the numerical values for the SNPs.

For each gene, we aim to regress the gene expression levels on the minor allele counts of its related SNPs. We first center the gene expression levels and SNP minor allele counts, and hence each variable has mean 0. Denote by Y_{ik} the centered expression level for the kth gene ($k = 1, \ldots, 9$) and ith individual ($i = 1, \ldots, n = 377$), and by X_{ijk} the centered minor allele count for the jth SNP ($j = 1, \ldots, p_k$) corresponding to the kth gene and ith individual. For the kth gene, if the design matrix $X_k = \{X_{ijk}\}_{i=1,\ldots,n, j=1,\ldots,p_k}$ does not have full column rank, then we will randomly delete one column which is linearly correlated with

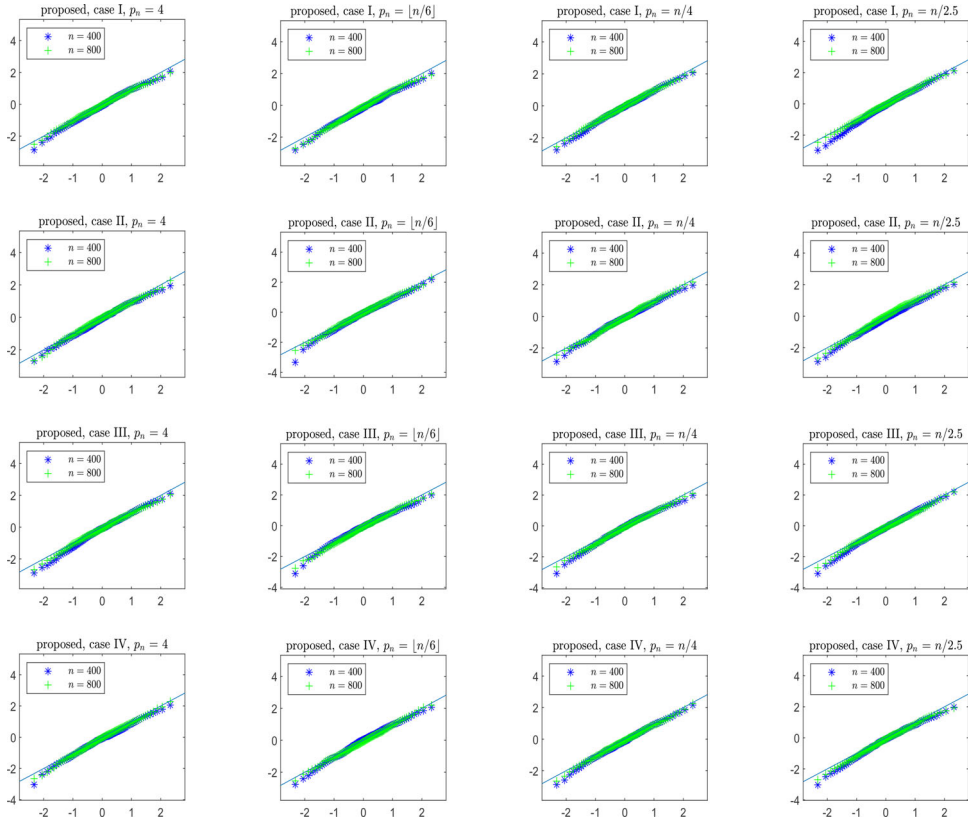

Figure 6. QQ plots for testing H_0 in (17) using the proposed test statistics. Panels from top to bottom are for Cases I–IV, respectively, while panels from left to right are for $p_n = 4, \lfloor n/6 \rfloor, n/4, n/2.5$, respectively.

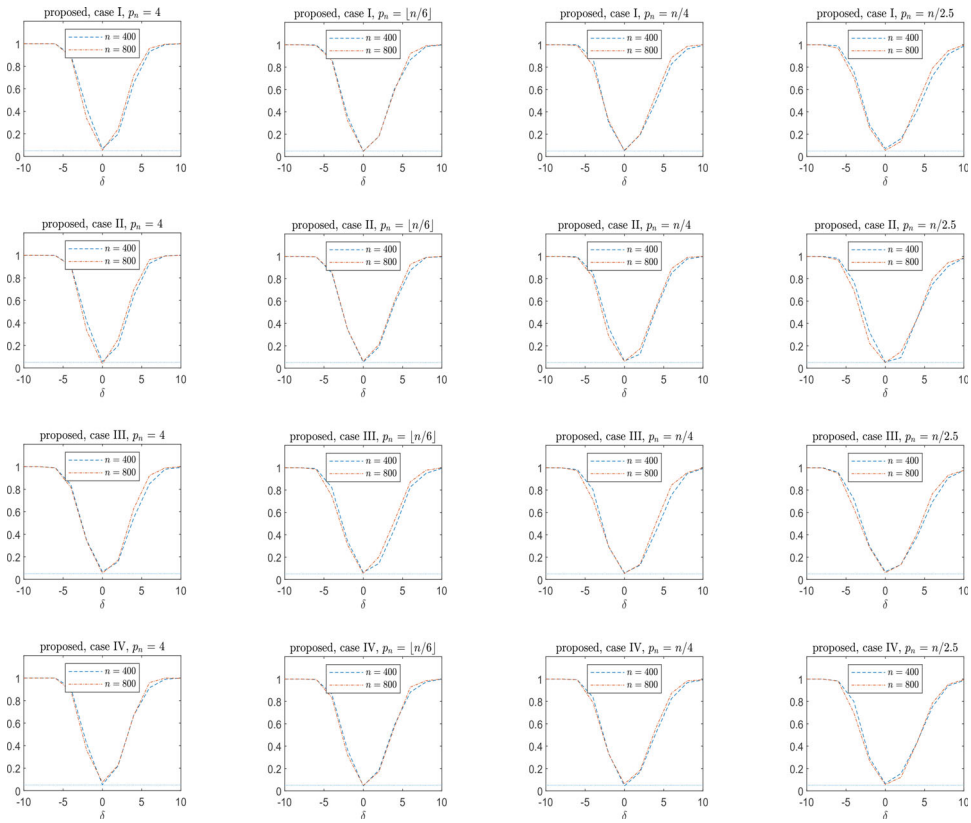

Figure 7. Empirical rejection rates versus δ for testing H_1 in (17) using the proposed test statistics. Panels from top to bottom are for Cases I–IV, respectively, while panels from left to right are for $p_n = 4, \lfloor n/6 \rfloor, n/4, n/2.5$, respectively. The dotted line indicates the true significance level $\alpha = 0.05$.

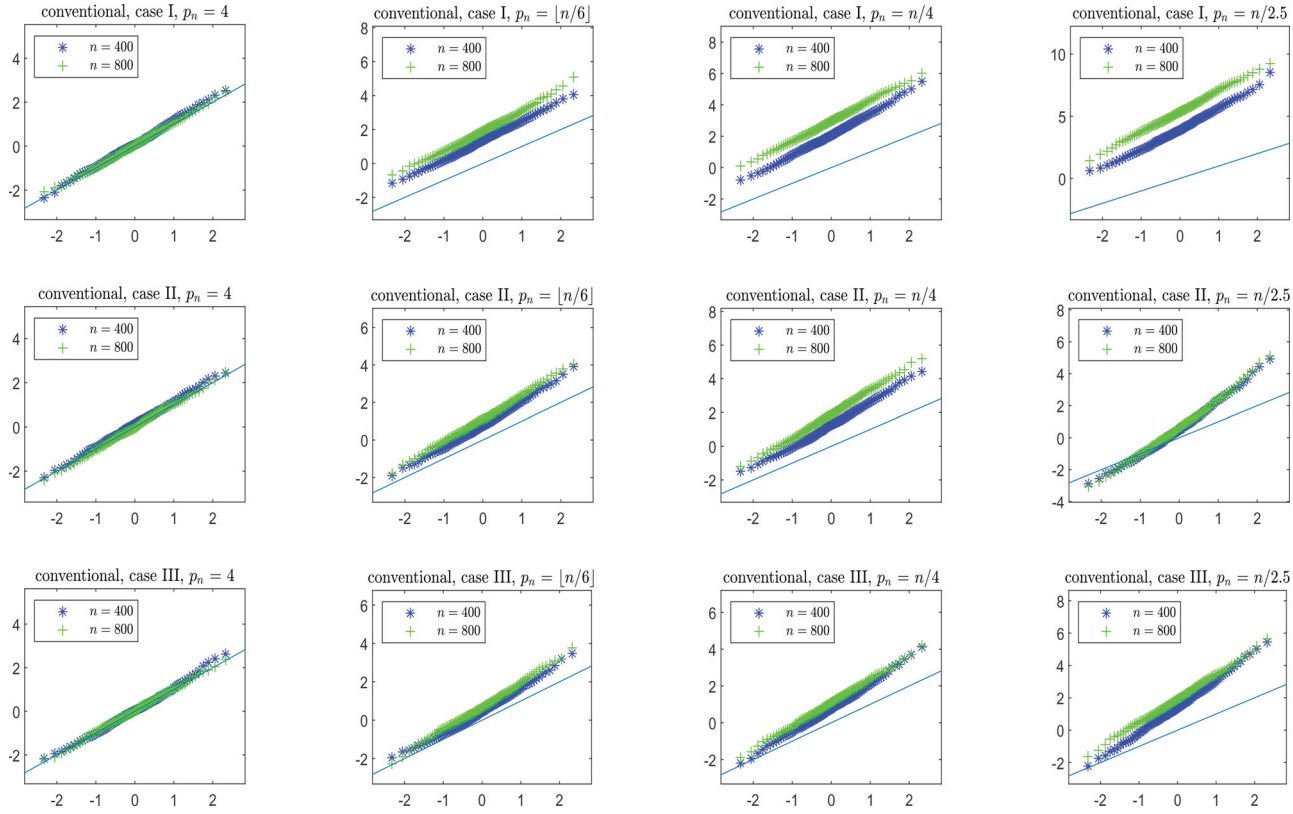

Figure 8. QQ plots for testing H_0 in (15) using the conventional method. Panels from top to bottom are for Cases I–III, respectively, while panels from left to right are for $p_n = 4, \lfloor n/6 \rfloor, n/4, n/2.5$, respectively.

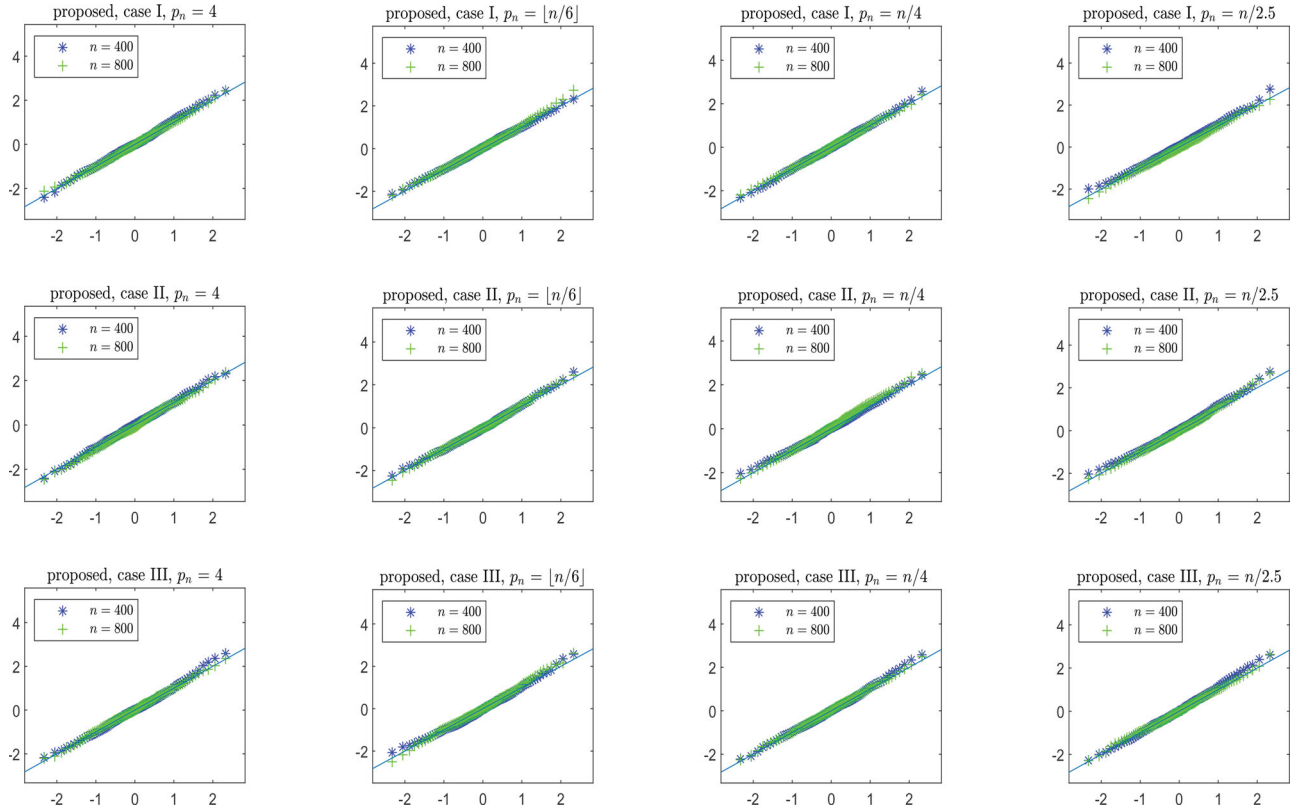

Figure 9. QQ plots for testing H_0 in (15) using the proposed method. Panels from top to bottom are for Cases I–III, respectively, while panels from left to right are for $p_n = 4, \lfloor n/6 \rfloor, n/4, n/2.5$, respectively.

Table 2. Coverage probabilities (Cov) and length (Len) of 95% confidence intervals for $\boldsymbol{\beta}_0^T\Sigma\boldsymbol{\beta}_0$.

		Proposed		MLE		MM_1		MM_2	
n	p_n	Cov	Len	Cov	Len	Cov	Len	Cov	Len
					Case I				
400	4	0.941	0.478	0.940	0.395	0.943	0.681	0.959	0.682
	66	0.939	0.489	0.947	0.440	0.949	0.712	0.951	0.712
	100	0.946	0.504	0.950	0.468	0.933	0.729	0.934	0.730
	160	0.957	0.529	0.948	0.514	0.960	0.762	0.960	0.763
800	4	0.950	0.338	0.949	0.278	0.939	0.479	0.953	0.479
	133	0.940	0.350	0.940	0.312	0.951	0.506	0.954	0.506
	200	0.947	0.356	0.954	0.332	0.955	0.517	0.952	0.517
	320	0.948	0.373	0.948	0.363	0.932	0.536	0.933	0.536
					Case II				
400	4	0.946	0.317	0.015	0.456	0.070	0.275	0.024	0.275
	66	0.942	0.767	0.000	0.887	0.008	0.808	0.004	0.811
	100	0.954	0.812	0.000	0.895	0.023	0.923	0.005	0.888
	160	0.927	3.625	0.000	2.658	0.000	4.690	0.000	4.689
800	4	0.938	0.284	0.117	0.263	0.405	0.327	0.002	0.279
	133	0.928	0.546	0.000	0.640	0.000	0.568	0.000	0.579
	200	0.932	0.556	0.000	0.646	0.000	0.600	0.000	0.599
	320	0.946	2.563	0.000	1.890	0.000	3.283	0.000	3.298
					Case III				
400	4	0.938	1.265	0.945	0.686	0.871	1.804	0.936	1.796
	66	0.938	1.260	0.948	0.756	0.872	1.852	0.868	1.844
	100	0.944	1.285	0.951	0.796	0.870	1.888	0.876	1.880
	160	0.933	1.290	0.937	0.875	0.867	1.943	0.860	1.935
800	4	0.931	0.912	0.939	0.483	0.848	1.274	0.913	1.270
	133	0.933	0.914	0.954	0.532	0.850	1.315	0.850	1.312
	200	0.947	0.919	0.955	0.561	0.870	1.334	0.868	1.331
	320	0.942	0.925	0.937	0.617	0.862	1.364	0.855	1.361

Table 3. Coverage probabilities (Cov) and length (Len) of 95% confidence intervals for σ_ϵ^2.

		Proposed		MLE		MM_1		MM_2	
n	p_n	Cov	Len	Cov	Len	Cov	Len	Cov	Len
					Case I				
400	4	0.940	0.276	0.944	0.278	0.953	0.398	0.973	0.398
	66	0.938	0.299	0.944	0.301	0.951	0.452	0.953	0.453
	100	0.938	0.318	0.938	0.319	0.951	0.483	0.953	0.483
	160	0.913	0.354	0.921	0.350	0.967	0.529	0.966	0.529
800	4	0.952	0.196	0.956	0.196	0.962	0.279	0.978	0.279
	133	0.941	0.213	0.941	0.214	0.953	0.321	0.956	0.321
	200	0.946	0.226	0.946	0.226	0.948	0.341	0.949	0.341
	320	0.956	0.252	0.959	0.249	0.951	0.372	0.953	0.372
					Case II				
400	4	0.952	0.276	0.956	0.278	0.144	0.357	0.088	0.379
	66	0.936	0.299	0.946	0.305	0.000	0.698	0.000	0.721
	100	0.951	0.317	0.949	0.326	0.000	0.759	0.000	0.786
	160	0.935	0.352	0.938	0.361	0.000	3.489	0.000	3.574
800	4	0.945	0.196	0.944	0.196	0.209	0.259	0.001	0.285
	133	0.951	0.214	0.954	0.216	0.000	0.495	0.000	0.510
	200	0.942	0.224	0.936	0.231	0.000	0.525	0.000	0.538
	320	0.948	0.253	0.937	0.258	0.000	2.456	0.000	2.513
					Case III				
400	4	0.926	0.304	0.906	0.277	0.883	0.868	0.968	0.894
	66	0.942	0.328	0.925	0.303	0.879	0.958	0.889	1.005
	100	0.940	0.345	0.919	0.319	0.878	1.004	0.896	1.060
	160	0.921	0.377	0.912	0.355	0.892	1.076	0.891	1.147
800	4	0.943	0.218	0.920	0.196	0.853	0.622	0.957	0.627
	133	0.938	0.233	0.920	0.214	0.859	0.697	0.869	0.707
	200	0.944	0.245	0.929	0.226	0.890	0.734	0.892	0.742
	320	0.936	0.271	0.927	0.251	0.892	0.785	0.895	0.800
					Case IV				
400	4	0.945	0.277	0.951	0.277	0.978	0.407	0.877	0.411
	66	0.941	0.301	0.940	0.300	0.000	0.004	0.000	1.141
	100	0.936	0.315	0.912	0.309	0.000	0.001	0.000	1.212
	160	0.931	0.351	0.812	0.326	0.000	0.000	0.000	1.773
800	4	0.932	0.195	0.935	0.196	0.960	0.286	0.951	0.285
	133	0.946	0.214	0.936	0.212	0.000	0.000	0.000	0.748
	200	0.945	0.225	0.900	0.219	0.000	0.000	0.000	0.924
	320	0.948	0.251	0.725	0.231	0.000	0.000	0.000	1.383

the others and repeat this procedure until the design matrix is of full column rank. With a slight abuse of notation, denote by p_k the number of eventually selected SNPs for the kth gene. The list of the selected genes and the corresponding p_k are provided in Table 5. Our strategy for selecting the 9 genes in this study is that their corresponding p_k is large enough, that is, at least 33, such that τ ranges from 0.088 to 0.231.

The linear model for regressing $\boldsymbol{Y}_k = (Y_{1k}, \ldots, Y_{nk})^T$ on X_k is fitted for each gene and the confidence intervals for ρ_0 are given in Table 5 using the conventional method in Section S.3 in the supplementary materials, our proposed method and the MM_2 method in Dicker (2014) (see Proposition 2 therein) proposed for general covariance matrix Σ. The MLE method in Dicker and Erdogdu (2016) and MM_1 in Dicker (2014) are not designed for general Σ, and hence are not compared here. We observe that the upper bounds of the confidence intervals of ρ_0 by our method are bounded away from 1, which indicates that the SNRs are not exploding. Specifically, using (6), we can calculate the confidence intervals for SNR directly by those of ρ_0, and we find that the largest value of the upper bounds of the confidence intervals for SNR is 2.9526. Also, the values of p_k^2/n range from 2.8886 to 20.0769, and hence, the strong signal condition (as in Theorem 4) that $p_k^2/n = o(\text{SNR})$ is not satisfied by this data. From Table 5, for most genes, the confidence intervals of ρ_0 do not cover 0, indicating that these selected SNPs are indeed significantly associated with the genes, which supports the findings in Stranger et al. (2007). However, for gene AKAP10, 0 is covered by the confidence interval using our proposed method but excluded by the conventional method. This discrepancy may be due to the failure of the conventional method under moderate dimension and insufficient SNR. More importantly, we can see that the confidence intervals for ρ_0 using our method are narrower than those using MM_2 in Dicker (2014) for most genes, which means that our method is more accurate in the moderate-dimensional case with $\tau \in [0, 1)$.

Table 4. Coverage probabilities (Cov) and length (Len) of 95% confidence intervals for ρ_0.

		Proposed		MLE		MM_1		MM_2	
n	p_n	Cov	Len	Cov	Len	Cov	Len	Cov	Len
					Case I				
400	4	0.944	0.138	0.940	0.121	0.949	0.241	0.970	0.241
	66	0.955	0.150	0.958	0.142	0.945	0.264	0.950	0.264
	100	0.942	0.160	0.950	0.154	0.938	0.275	0.941	0.276
	160	0.959	0.178	0.948	0.176	0.966	0.296	0.966	0.296
800	4	0.955	0.098	0.942	0.085	0.951	0.170	0.970	0.170
	133	0.945	0.107	0.939	0.100	0.964	0.187	0.963	0.187
	200	0.961	0.113	0.951	0.109	0.951	0.195	0.952	0.195
	320	0.948	0.127	0.947	0.125	0.937	0.209	0.942	0.209
					Case II				
400	4	0.935	0.150	0.026	0.135	0.008	0.159	0.000	0.164
	66	0.951	0.115	0.001	0.069	0.000	0.230	0.000	0.234
	100	0.951	0.117	0.023	0.077	0.000	0.251	0.000	0.248
	160	0.947	0.033	0.000	0.015	0.000	0.287	0.000	0.291
800	4	0.936	0.103	0.203	0.083	0.211	0.146	0.000	0.134
	133	0.938	0.082	0.000	0.049	0.000	0.162	0.000	0.167
	200	0.939	0.085	0.001	0.055	0.000	0.173	0.000	0.175
	320	0.933	0.024	0.000	0.011	0.000	0.203	0.000	0.206
					Case III				
400	4	0.938	0.098	0.919	0.068	0.876	0.268	0.960	0.270
	66	0.954	0.102	0.934	0.076	0.868	0.289	0.873	0.285
	100	0.940	0.106	0.931	0.083	0.868	0.300	0.881	0.294
	160	0.930	0.113	0.914	0.095	0.874	0.319	0.874	0.307
800	4	0.946	0.070	0.939	0.048	0.842	0.190	0.950	0.191
	133	0.930	0.073	0.937	0.054	0.848	0.205	0.857	0.205
	200	0.948	0.076	0.939	0.058	0.872	0.213	0.870	0.213
	320	0.954	0.081	0.941	0.067	0.880	0.226	0.885	0.224

Table 5. 90% confidence intervals of ρ_0 for gene data.

Gene	Probe	p_k	Conventional	Proposed	MM_2 (Dicker 2014)
AKAP10	ILMN_1718808	33	[0.071, 0.161]	[0.000, 0.092]	[0.000, 0.040]
CPNE1	ILMN_1670841	35	[0.293, 0.524]	[0.244, 0.504]	[0.259, 0.474]
NUDT13	ILMN_1680420	59	[0.366, 0.467]	[0.294, 0.422]	[0.274, 0.482]
PIGN	ILMN_1691112	36	[0.225, 0.325]	[0.162, 0.280]	[0.183, 0.376]
PKHD1L1	ILMN_1717886	87	[0.680, 0.747]	[0.640, 0.747]	[0.831, 1.000]
SPG7	ILMN_1675583	38	[0.265, 0.375]	[0.212, 0.328]	[0.208, 0.406]
ST7L	ILMN_1659926	40	[0.548, 0.637]	[0.521, 0.627]	[0.522, 0.796]
TGM5	ILMN_1699925	39	[0.298, 0.406]	[0.243, 0.368]	[0.232, 0.441]
TSGA10	ILMN_1674645	44	[0.512, 0.613]	[0.479, 0.599]	[0.496, 0.761]

Appendix A: Inference for Single Element and Linear Functional of β_0

We provide a brief discussion of the element-wise inference for $\beta_{0,j}$ and $\beta_{0,j}^2$ $(j = 1, \ldots, p_n)$. The estimator for $\beta_{0,j}$ is

$$\hat{\beta}_j = \mathbf{e}_{j,p_n}^T \hat{\boldsymbol{\beta}} = \beta_{0,j} + \mathbf{e}_{j,p_n}^T (X^T X)^{-1} X^T \boldsymbol{\epsilon}.$$

If $\sigma_\epsilon^2 = o(n)$, then $\sigma_{\hat{\beta}_j}^{-1}(\hat{\beta}_j - \beta_{0,j}) \xrightarrow{\mathcal{D}} N(0,1)$, where $\sigma_{\hat{\beta}_j}^2 = \sigma_\epsilon^2 \mathbf{e}_{j,p_n}^T \mathrm{E}\{(X^T X)^{-1} \mathrm{I}(K)\} \mathbf{e}_{j,p_n} = \Omega(\sigma_\epsilon^2/n)$. If $\sigma_\epsilon^2 = \Omega(1)$ and $\boldsymbol{X}_i \sim N(\mathbf{0}_{p_n}, \mathbf{I}_{p_n})$, the bias of $\hat{\beta}_j^2$ is

$$\mathrm{E}(\hat{\beta}_j^2) - \beta_{0,j}^2 = \sigma_\epsilon^2 \mathbf{e}_{j,p_n}^T \mathrm{E}\{(X^T X)^{-1}\} \mathbf{e}_{j,p_n} = \Omega(1/n).$$

Therefore, the bias of $\hat{\beta}_j^2$ is ignorable if $n^{-1} = o(\beta_{0,j}^2)$. Inference for $\beta_{0,j}^2$ can be conducted using $(2\beta_{0,j}\sigma_{\hat{\beta}_j})^{-1}(\hat{\beta}_j^2 - \beta_{0,j}^2) \xrightarrow{\mathcal{D}} N(0,1)$. The $\sqrt{n}(\hat{\beta}_j^2 - \beta_{0,j}^2)$ versus j and the bias for $||\hat{\boldsymbol{\beta}}||_2^2$ are plotted in Figure A1.

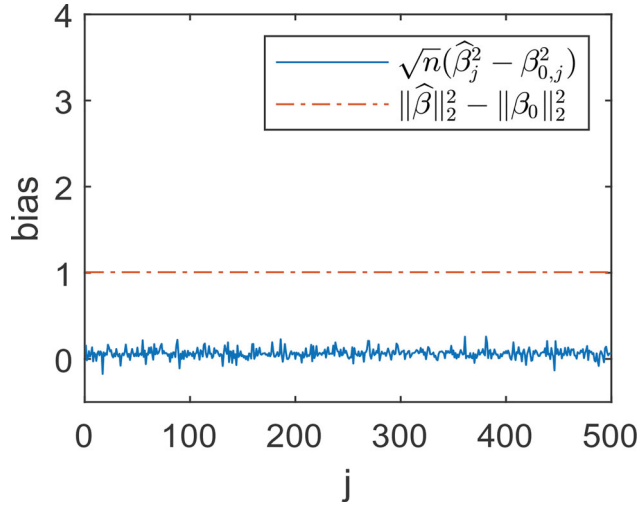

Figure A1. Plots of $\sqrt{n}(\hat{\beta}_j^2 - \beta_{0,j}^2)$ versus j (solid line) and bias for $||\hat{\boldsymbol{\beta}}||_2^2$ (dash-dotted line) under the same setting as in Figure 1 with $p_n = 500$.

We then discuss the inference for linear functionals $\boldsymbol{c}^T\boldsymbol{\beta}_0$ where $\boldsymbol{c} \in \mathbb{R}^{p_n}$ is deterministic. The estimator for $\boldsymbol{c}^T\boldsymbol{\beta}_0$ is $\boldsymbol{c}^T\hat{\boldsymbol{\beta}} = \boldsymbol{c}^T\boldsymbol{\beta}_0 + \boldsymbol{c}^T(X^TX)^{-1}X^T\boldsymbol{\epsilon}$. Following the proof of Theorem 1, its limiting distribution is $\sigma_L^{-1}(\boldsymbol{c}^T\hat{\boldsymbol{\beta}} - \boldsymbol{c}^T\boldsymbol{\beta}_0) \xrightarrow{\mathcal{D}} N(0,1)$ where $\sigma_L^2 = \sigma_\epsilon^2 \boldsymbol{c}^T \mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{c}$ with a ratio consistent estimator $\hat{\sigma}_L^2 = \hat{\sigma}_\epsilon^2 \boldsymbol{c}^T(X^TX)^{-1}\boldsymbol{c}$.

Appendix B: Relation Between τ and SNR

According to Theorem 4, define

- strong SNR: $p_n^2/n = o(\mathrm{SNR})$;
- weak SNR: $\mathrm{SNR} \lesssim p_n^2/n$.

Figure B1 describes the precise relation between τ and the signal strength under mild conditions. In particular, $\tau = 0/\tau > 0$ may imply strong/weak signals unless we allow $||\boldsymbol{\beta}_0||_2$ or σ_ϵ^2 to diminish.

Appendix C: Proofs of Main Theoretical Results

This section includes the proofs of Lemmas 1 and 2 and Theorems 1 and 3. In all the proofs, we only consider the case that σ_ϵ^2 is fixed. The results for diverging σ_ϵ^2 can be simply obtained by replacing Y_i, $\boldsymbol{\epsilon}$ and $\boldsymbol{\beta}_0$ with Y_i/σ_ϵ, $\boldsymbol{\epsilon}/\sigma_\epsilon$ and $\boldsymbol{\beta}_0/\sigma_\epsilon$, respectively, in the proofs.

We introduce some notations and equations. Let $X_{(i)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_{i-1}, \boldsymbol{X}_{i+1}, \ldots, \boldsymbol{X}_n)^T$ for $i = 1, \ldots, n$, that is, the design matrix without the ith observation. Similarly, $X_{(i,j)}$ denotes the design matrix without the ith and jth observations for $1 \le i \ne j \le n$. From the Sherman–Morrison formula (Sherman and Morrison 1950),

$$(X^TX)^{-1} = (X_{(1)}^TX_{(1)} + \boldsymbol{X}_1\boldsymbol{X}_1^T)^{-1} = (X_{(1)}^TX_{(1)})^{-1} - \frac{(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}}{1 + \boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1}, \tag{C.1}$$

and hence,

$$(X^TX)^{-2} = (X_{(1)}^TX_{(1)})^{-2} - (X_{(1)}^TX_{(1)})^{-2}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}/\{1 + \boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\} - (X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-2}/ \tag{C.2}$$

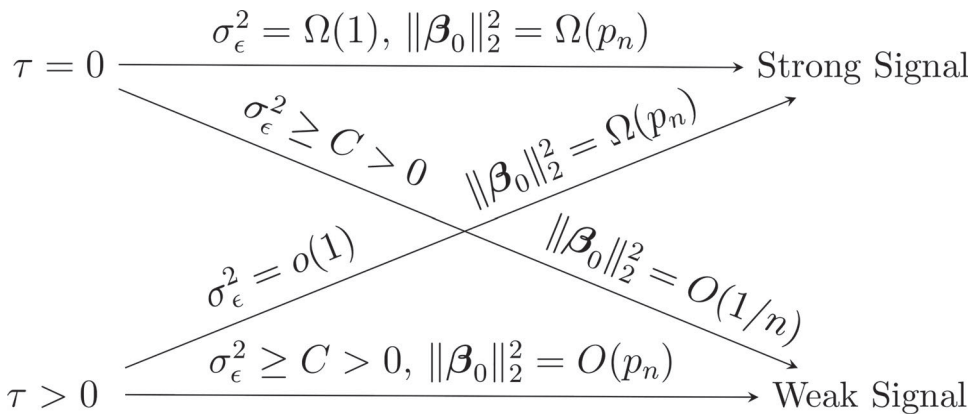

Figure B1. Relation between $\tau = 0/\tau > 0$ and strong/weak signal.

$$\{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1\} \tag{C.3}$$
$$+ \{(X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1 \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1}\}^2 / \{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1\}^2. \tag{C.4}$$

Therefore,

$$(X^T X)^{-1} \boldsymbol{X}_1 = \frac{(X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1}{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1}, \tag{C.5}$$

$$(X^T X)^{-2} \boldsymbol{X}_1 = \frac{(X_{(1)}^T X_{(1)})^{-2} \boldsymbol{X}_1}{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1} - \frac{(X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1 \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-2} \boldsymbol{X}_1}{\{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1\}^2}, \tag{C.6}$$

$$\boldsymbol{X}_1^T (X^T X)^{-2} \boldsymbol{X}_1 = \frac{\boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-2} \boldsymbol{X}_1}{\{1 + \boldsymbol{X}_1^T (X_{(1)}^T X_{(1)})^{-1} \boldsymbol{X}_1\}^2}. \tag{C.7}$$

The following are the proofs of the main results in this article.

Proof of Lemma 1. For $\boldsymbol{Z}_i = (z_{i1}, \ldots, z_{ip_n})^T$ defined in Condition A1, let $z_{ij}^* = z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n}) - \mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}$, $\tilde{z}_{ij} = z_{ij} - z_{ij}^* = z_{ij}\mathrm{I}(|z_{ij}| > \sqrt{n}/\sqrt{\log n}) + \mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}$, $\boldsymbol{Z}_i^* = (z_{i1}^*, \ldots, z_{ip_n}^*)^T$, $\tilde{\boldsymbol{Z}}_i = (\tilde{z}_{i1}, \ldots, \tilde{z}_{ip_n})^T$, $Z^* = (\boldsymbol{Z}_1^*, \ldots, \boldsymbol{Z}_n^*)^T = (z_{ij}^*)_{i \leq n, j \leq p_n}$ and $\tilde{Z} = (\tilde{\boldsymbol{Z}}_1, \ldots, \tilde{\boldsymbol{Z}}_n)^T = (\tilde{z}_{ij})_{i \leq n, j \leq p_n}$.

Then, $\mathrm{E}(z_{ij}^*) = 0$ and by Cauchy–Schwarz inequality and Chebyshev's inequality,

$$\begin{aligned}
&1 - \mathrm{E}(z_{ij}^{*2}) \\
&= 1 - \mathrm{E}\{z_{ij}^2 \mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\} + [\mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}]^2 \\
&= 1 - 1 + \mathrm{E}\{z_{ij}^2 \mathrm{I}(|z_{ij}| > \sqrt{n}/\sqrt{\log n})\} \\
&\quad + [\mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| > \sqrt{n}/\sqrt{\log n})\}]^2 \\
&\leq 2\mathrm{E}\{z_{ij}^2 \mathrm{I}(|z_{ij}| > \sqrt{n}/\sqrt{\log n})\} \leq 2\{\mathrm{E}(z_{ij}^4)\mathrm{P}(|z_{ij}| > \sqrt{n}/\sqrt{\log n})\}^{1/2} \\
&\lesssim \{\mathrm{P}(|z_{ij}| > \sqrt{n}/\sqrt{\log n})\}^{1/2} \leq \{\mathrm{E}(z_{ij}^4)/(\sqrt{n}/\sqrt{\log n})^4\}^{1/2} \\
&\lesssim (\log n)/n,
\end{aligned}$$

which implies that $\max_{j \leq p_n} \sum_{i=1}^n |1 - \mathrm{E}(z_{ij}^{*2})| \lesssim \log n = o(n)$. Also,

$$\begin{aligned}
&\sup_{i \leq n, j \leq p_n, n \geq 1} \mathrm{E}(z_{ij}^{*4}) \\
&\lesssim \sup_{i \leq n, j \leq p_n, n \geq 1} \big(\mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}^4 \\
&\qquad + [\mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}]^4\big) \\
&\lesssim \sup_{i \leq n, j \leq p_n, n \geq 1} \mathrm{E}\{z_{ij}\mathrm{I}(|z_{ij}| \leq \sqrt{n}/\sqrt{\log n})\}^4 + C \\
&\leq \sup_{i \leq n, j \leq p_n, n \geq 1} \mathrm{E}(z_{ij}^4) + C \leq 2C < \infty.
\end{aligned}$$

It is easy to see $|z_{ij}^*| \leq 2\sqrt{n}/\sqrt{\log n}$. From Theorem 9.13 of Bai and Silverstein (2010), for any $s_1 > (1 + \sqrt{\tau})^2$, $s_2 < (1 - \sqrt{\tau})^2$ and any $\ell > 0$, we have

$$\begin{aligned}
&\mathrm{P}(||Z^{*T} Z^*/n||_2 > s_1) = o(n^{-\ell}), \\
&\mathrm{P}(||(Z^{*T} Z^*/n)^{-1}||_2 > 1/s_2) = o(n^{-\ell}).
\end{aligned}$$

Since $Z = Z^* + \tilde{Z}$, we have $Z^T Z/n = Z^{*T} Z^*/n + \tilde{Z}^T Z^*/n + Z^{*T} \tilde{Z}/n + \tilde{Z}^T \tilde{Z}/n$. We know that

$$||\tilde{Z}^T \tilde{Z}/n||_2 \leq ||\tilde{Z}^T \tilde{Z}/n||_1 = \max_{i \leq p_n} \sum_{j=1}^{p_n} \Big| \sum_{t=1}^n \tilde{z}_{ti} \tilde{z}_{tj}/n \Big|. \tag{C.8}$$

Note $\mathrm{E}(\tilde{z}_{ti}) = 0$, $\mathrm{E}(\tilde{z}_{ti}\tilde{z}_{tj}) = 0$ for $i \neq j$, and, for any integer $k > 0$, from Cauchy–Schwarz inequality,

$$\begin{aligned}
\mathrm{E}|\tilde{z}_{ti}|^k &= \mathrm{E}|z_{ti}\mathrm{I}(|z_{ti}| > \sqrt{n}/\sqrt{\log n}) + \mathrm{E}\{z_{ti}\mathrm{I}(|z_{ti}| \leq \sqrt{n}/\sqrt{\log n})\}|^k \\
&= \mathrm{E}|z_{ti}\mathrm{I}(|z_{ti}| > \sqrt{n}/\sqrt{\log n}) - \mathrm{E}\{z_{ti}\mathrm{I}(|z_{ti}| > \sqrt{n}/\sqrt{\log n})\}|^k \\
&\lesssim \mathrm{E}|z_{ti}\mathrm{I}(|z_{ti}| > \sqrt{n}/\sqrt{\log n})|^k \\
&\leq \{\mathrm{E}|z_{ti}|^{2k}\mathrm{P}(|z_{ti}| > \sqrt{n}/\sqrt{\log n})\}^{1/2} \\
&\lesssim \{\mathrm{P}(|z_{ti}| > \sqrt{n}/\sqrt{\log n})\}^{1/2}.
\end{aligned}$$

Therefore, for any integer $\ell > 0$, taking $x = 1/\sqrt{n}$, from (C.8) and Markov's inequality,

$$\begin{aligned}
&\mathrm{P}(||\tilde{Z}^T \tilde{Z}/n||_2 > x) \\
&\leq \mathrm{P}\Big(\max_{i \leq p_n} \sum_{j=1}^{p_n} \Big| \sum_{t=1}^n \tilde{z}_{ti}\tilde{z}_{tj}/n \Big| > x\Big) \leq p_n \mathrm{P}\Big(\sum_{j=1}^{p_n} \Big| \sum_{t=1}^n \tilde{z}_{ti}\tilde{z}_{tj}/n \Big| > x\Big) \\
&\leq p_n^2 \mathrm{P}\Big(\Big| \sum_{t=1}^n \tilde{z}_{ti}\tilde{z}_{tj}/n \Big| > x/p_n\Big) \\
&\leq p_n^2 \mathrm{P}\Big(\Big| \sum_{t=1}^n \tilde{z}_{ti}^2/n \Big| \Big| \sum_{t=1}^n \tilde{z}_{tj}^2/n \Big| > x^2/p_n^2\Big) \\
&\lesssim p_n^2 \mathrm{P}\Big(\Big| \sum_{t=1}^n \tilde{z}_{ti}^2/n \Big| > x/p_n\Big) \leq p_n^2 \mathrm{E}\Big(\Big| \sum_{t=1}^n \tilde{z}_{ti}^2/n \Big|^{2\ell}\Big) \Big/ (x/p_n)^{2\ell}
\end{aligned}$$

$$
\begin{aligned}
&= p_n^{2\ell+2}x^{-2\ell}n^{-2\ell}\mathrm{E}\Big(\Big|\sum_{t=1}^{n}\tilde{z}_{ti}^2\Big|^{2\ell}\Big)\\
&\lesssim p_n^{2\ell+2}x^{-2\ell}n^{-2\ell}n^{2\ell}\mathrm{E}|\tilde{z}_{ti}^2|^{2\ell} \lesssim p_n^{2\ell+2}x^{-2\ell}\mathrm{E}|\tilde{z}_{ti}|^{4\ell}\\
&\lesssim p_n^{2\ell+2}x^{-2\ell}\{\mathrm{P}(|z_{ti}| > \sqrt{n}/\sqrt{\log n})\}^{1/2}\\
&\lesssim p_n^{2\ell+2}x^{-2\ell}\{\mathrm{E}|z_{ti}|^{28\ell}/(\sqrt{n}/\sqrt{\log n})^{28\ell}\}^{1/2}\\
&\lesssim p_n^{2\ell+2}x^{-2\ell}(\sqrt{\log n}/\sqrt{n})^{14\ell}\\
&\le (\log n)^{7\ell}n^{-5\ell+2}x^{-2\ell} = (\log n)^{7\ell}n^{-4\ell+2} = o(n^{-\ell}).
\end{aligned}
$$

Then, for n large enough,

$$
\begin{aligned}
&\mathrm{P}(||\tilde{Z}^TZ^*/n||_2 > 1/\log n) \le \mathrm{P}(||\tilde{Z}^T||_2||Z^*||_2/n > 1/\log n)\\
&= \mathrm{P}(||\tilde{Z}^T\tilde{Z}/n||_2||Z^{*T}Z^*/n||_2 > 1/(\log n)^2)\\
&\le \mathrm{P}(||\tilde{Z}^T\tilde{Z}/n||_2 > n^{-1/4}/\log n) + \mathrm{P}(||Z^{*T}Z^*/n||_2 > n^{1/4}/\log n)\\
&= o(n^{-\ell}).
\end{aligned}
$$

Therefore, taking $\mu_1 = 4(1+\sqrt{\tau})^2$ and $\mu_2 = (1-\sqrt{\tau})^2/4$, we have

$$
\begin{aligned}
&\mathrm{P}(||Z^TZ/n||_2 \ge \mu_1)\\
&\le \mathrm{P}(||Z^{*T}Z^*/n||_2 > \mu_1/2) + \mathrm{P}(||\tilde{Z}^TZ^*/n||_2 > \mu_1/8)\\
&\quad +\mathrm{P}(||Z^{*T}\tilde{Z}/n||_2 > \mu_1/8) + \mathrm{P}(||\tilde{Z}^T\tilde{Z}/n||_2 > \mu_1/8) = o(n^{-\ell}),
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathrm{P}(||(Z^TZ/n)^{-1}||_2 \ge 1/\mu_2) = \mathrm{P}(\lambda_{\min}(Z^TZ/n) \le \mu_2)\\
&\le \mathrm{P}(\lambda_{\min}(Z^{*T}Z^*/n) - ||\tilde{Z}^TZ^*/n||_2 - ||Z^{*T}\tilde{Z}/n||_2\\
&\quad -||\tilde{Z}^T\tilde{Z}/n||_2 \le \mu_2)\\
&\le \mathrm{P}(\lambda_{\min}(Z^{*T}Z^*/n) < 2\mu_2) + \mathrm{P}(||\tilde{Z}^TZ^*/n||_2 \ge \mu_2/4)\\
&\quad +\mathrm{P}(||Z^{*T}\tilde{Z}/n||_2 \ge \mu_2/4) + \mathrm{P}(||\tilde{Z}^T\tilde{Z}/n||_2 \ge \mu_2/4)\\
&= \mathrm{P}(||(Z^{*T}Z^*/n)^{-1}||_2 > 1/(2\mu_2)) + o(n^{-\ell}) = o(n^{-\ell}).
\end{aligned}
$$

Then, taking $x_1 = ||\Sigma||_2\mu_1$ and $x_2 = \mu_2/||\Sigma^{-1}||_2$, we have

$$
\begin{aligned}
\mathrm{P}(||X^TX/n||_2 \ge x_1) &\le \mathrm{P}(||\Sigma||_2||Z^TZ/n||_2 \ge x_1)\\
&= \mathrm{P}(||Z^TZ/n||_2 \ge x_1/||\Sigma||_2) = o(n^{-\ell}),\\
\mathrm{P}(||(X^TX/n)^{-1}||_2 \ge x_2^{-1}) &\le \mathrm{P}(||\Sigma^{-1}||_2||(Z^TZ/n)^{-1}||_2 \ge x_2^{-1})\\
&= \mathrm{P}(||(Z^TZ/n)^{-1}||_2 \ge (x_2||\Sigma^{-1}||_2)^{-1})\\
&= o(n^{-\ell}).
\end{aligned}
$$

□

Proof of Theorem 1. We first consider the situation that $\lim_{n\to\infty} p_n = \infty$ for part (a). From Lemma 2 that $\mathrm{tr}\{(X^TX)^{-1}\} - \mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\} = o_{\mathrm{P}}(p_n/n)$. Under event K, the eigenvalues of X^TX/n are bounded away from 0 and infinity. Hence, $\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\} = \Omega(p_n/n)$. Therefore, we have $\mathrm{tr}\{(X^TX)^{-1}\} = \mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}\{1 + o_{\mathrm{P}}(1)\}$. From

$$
\begin{aligned}
&||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2 - \mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2\\
&= ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2^2 + 2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) - \mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2\\
&= [||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0||_2^2 - \mathrm{tr}\{(X^TX)^{-1}\}\sigma_\epsilon^2] - \mathrm{tr}\{(X^TX)^{-1}\}(\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2)\\
&\quad +2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\\
&\equiv \mathrm{I}_1 - \mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}\{1 + o_{\mathrm{P}}(1)\}\mathrm{I}_2 + 2\mathrm{I}_3,
\end{aligned}
$$

we first demonstrate the asymptotic normality of $\zeta_n^{-1}(c_1\mathrm{I}_1 + c_2\mathrm{I}_2 + c_3\mathrm{I}_3)\mathrm{I}(K)$ for any constants $c_1 = \Omega(1)$, $c_2 = \Omega(p_n/n)$ and $c_3 = \Omega(1)$.

For notational simplicity, denote $M_1 = X(X^TX)^{-2}X^T$, $M_2 = \{\mathbf{I}_n - X(X^TX)^{-1}X^T\}/(n-p_n)$ and $\boldsymbol{\nu}^T = \boldsymbol{\beta}_0^T(X^TX)^{-1}X^T$. Then,

$$
\begin{aligned}
\mathrm{I}_1 &= \boldsymbol{\epsilon}^TM_1\boldsymbol{\epsilon} - \mathrm{tr}\{(X^TX)^{-1}\}\sigma_\epsilon^2\\
&= 2\sum_{1\le i<j\le n} M_1(i,j)\epsilon_i\epsilon_j + \sum_{j=1}^{n} M_1(j,j)\epsilon_j^2 - \mathrm{tr}\{(X^TX)^{-1}\}\sigma_\epsilon^2\\
&= 2\sum_{1\le i<j\le n} M_1(i,j)\epsilon_i\epsilon_j + \sum_{j=1}^{n} M_1(j,j)(\epsilon_j^2 - \sigma_\epsilon^2),
\end{aligned}
$$

$$
\mathrm{I}_2 = \boldsymbol{\epsilon}^TM_2\boldsymbol{\epsilon} - \sigma_\epsilon^2 = 2\sum_{1\le i<j\le n} M_2(i,j)\epsilon_i\epsilon_j + \sum_{j=1}^{n} M_2(j,j)(\epsilon_j^2 - \sigma_\epsilon^2),
$$

$$
\mathrm{I}_3 = \boldsymbol{\nu}^T\boldsymbol{\epsilon} = \sum_{j=1}^{n}\nu_j\epsilon_j,
$$

where $M_k(i,j)$ is the (i,j)th element of M_k ($k = 1, 2$) and $\boldsymbol{\nu} = (\nu_1,\ldots,\nu_n)^T$. Hence,

$$
\begin{aligned}
&c_1\mathrm{I}_1 + c_2\mathrm{I}_2 + c_3\mathrm{I}_3\\
&= 2\sum_{1\le i<j\le n}\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\epsilon_j\\
&\quad +\sum_{j=1}^{n}\{c_1M_1(j,j) + c_2M_2(j,j)\}(\epsilon_j^2 - \sigma_\epsilon^2) + c_3\sum_{j=1}^{n}\nu_j\epsilon_j\\
&= \sum_{j=1}^{n}\Big[\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\epsilon_j\\
&\quad +\{c_1M_1(j,j) + c_2M_2(j,j)\}(\epsilon_j^2 - \sigma_\epsilon^2) + c_3\nu_j\epsilon_j\Big]\\
&\equiv \sum_{j=1}^{n} U_j.
\end{aligned}
$$

Note that $U_j\mathrm{I}(K)$, $j = 1, 2, \ldots$, is a martingale difference, with

$$
\mathrm{E}(U_j\mathrm{I}(K)|X,\epsilon_1,\ldots,\epsilon_{j-1}) = 0
$$

and

$$
\begin{aligned}
&\sum_{j=1}^{n}\mathrm{E}[\{U_j\mathrm{I}(K)\}^2|X,\epsilon_1,\ldots,\epsilon_{j-1}]\\
&= \mathrm{I}(K)\sum_{j=1}^{n}\mathrm{E}\Big(\Big[\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\epsilon_j\\
&\quad +\{c_1M_1(j,j) + c_2M_2(j,j)\}(\epsilon_j^2 - \sigma_\epsilon^2) + c_3\nu_j\epsilon_j\Big]^2\Big|X,\epsilon_1,\ldots,\epsilon_{j-1}\Big)\\
&= \mathrm{I}(K)\sum_{j=1}^{n}\mathrm{E}\Big(\Big[\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\Big]^2\epsilon_j^2\\
&\quad +\{c_1M_1(j,j) + c_2M_2(j,j)\}^2(\epsilon_j^2 - \sigma_\epsilon^2)^2 + c_3^2\nu_j^2\epsilon_j^2\\
&\quad +2\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\epsilon_j\{c_1M_1(j,j) + c_2M_2(j,j)\}\\
&\quad \times(\epsilon_j^2 - \sigma_\epsilon^2)\\
&\quad +2\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\epsilon_jc_3\nu_j\epsilon_j\\
&\quad +2\{c_1M_1(j,j) + c_2M_2(j,j)\}(\epsilon_j^2 - \sigma_\epsilon^2)c_3\nu_j\epsilon_j\Big|X,\epsilon_1,\ldots,\epsilon_{j-1}\Big)\\
&= \mathrm{I}(K)\sum_{j=1}^{n}\Big(\Big[\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\epsilon_i\Big]^2\sigma_\epsilon^2\\
&\quad +\{c_1M_1(j,j) + c_2M_2(j,j)\}^2\mathrm{var}(\epsilon_j^2) + c_3^2\nu_j^2\sigma_\epsilon^2\\
&\quad +2\sum_{1\le i<j} 2\{c_1M_1(i,j) + c_2M_2(i,j)\}\{c_1M_1(j,j) + c_2M_2(j,j)\}\mathrm{E}(\epsilon_j^3)\epsilon_i
\end{aligned}
$$

$$+2\sum_{1\le i<j}2\{c_1M_1(i,j)+c_2M_2(i,j)\}\epsilon_i c_3 v_j\sigma_\epsilon^2$$
$$+2\{c_1M_1(j,j)+c_2M_2(j,j)\}\mathrm{E}(\epsilon_j^3)c_3v_j\Big)$$
$$\equiv \mathrm{I}(K)\sum_{j=1}^{n}(\mathrm{II}_{1,j}+\mathrm{II}_{2,j}+\mathrm{II}_{3,j}+\mathrm{II}_{4,j}+\mathrm{II}_{5,j}+\mathrm{II}_{6,j}).$$

Denote $t_n = ||\boldsymbol{\beta}_0||_2/\sqrt{n}+\sqrt{p_n}/n$. Lemmas S.6–S.11 imply

$$\mathrm{var}\Big\{\sum_{j=1}^{n}\mathrm{II}_{k,j}\mathrm{I}(K)\Big\} = o(t_n^4), \quad \text{for } k=1,\dots,6. \tag{C.9}$$

Lemma S.5 indicates

$$\sum_{j=1}^{n}\mathrm{E}\{U_j\mathrm{I}(K)\}^4 = o(t_n^4). \tag{C.10}$$

From Lemmas S.12–S.14, we have

$$\sum_{j=1}^{n}\sum_{k=1}^{3}\mathrm{E}\{\mathrm{II}_{k,j}\mathrm{I}(K)\} = O(t_n^2). \tag{C.11}$$

Lemmas S.9–S.11 imply that

$$\sum_{j=1}^{n}\sum_{k=4}^{6}\mathrm{E}\{\mathrm{II}_{k,j}\mathrm{I}(K)\} = o(t_n^2). \tag{C.12}$$

Checking conditions (2) and (4) with $\delta = 1$ in the theorem of Heyde and Brown (1970), from (C.9)–(C.12), taking $c_1 = 1$, $c_2 = -\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}$, and $c_3 = 2$,

$$\zeta_n^{-1}(c_1\mathrm{I}_1+c_2\mathrm{I}_2+c_3\mathrm{I}_3)\mathrm{I}(K)\xrightarrow{\mathcal{D}}N(0,1),$$

where, from Lemmas S.12–S.14,

$$\begin{aligned}\zeta_n^2 &= 4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0+2\sigma_\epsilon^4\mathrm{Etr}\{(X^TX)^{-2}\mathrm{I}(K)\}\\ &\quad+\frac{2\sigma_\epsilon^4}{n-p_n}[\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}]^2\\ &= \Omega(\sigma_\epsilon^2||\boldsymbol{\beta}_0||_2^2/n+\sigma_\epsilon^4p_n/n^2+\sigma_\epsilon^4p_n^2/n^3)\\ &= \Omega(\sigma_\epsilon^2||\boldsymbol{\beta}_0||_2^2/n+\sigma_\epsilon^4p_n/n^2) = \Omega(t_n^2).\end{aligned}$$

For part (b), if $p_n^{1/2}/n = o(\mathrm{SNR})$, then $\zeta_n^2 = o(||\boldsymbol{\beta}_0||_2^4)$. Then, $||\hat{\boldsymbol{\beta}}||_2^2 - ||\boldsymbol{\beta}_0||_2^2 = O_P(\zeta_n) = o_P(||\boldsymbol{\beta}_0||_2^2)$.

Last, we consider the case for fixed p_n. Since $\boldsymbol{\beta}_0$ and σ_ϵ^2 do not change with n and $\boldsymbol{\beta}_0 \neq \mathbf{0}_{p_n}$, we have $n^{-1} = o(\mathrm{SNR})$. Note

$$\begin{aligned}&||\hat{\boldsymbol{\beta}}||_2^2-||\boldsymbol{\beta}_0||_2^2-\mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2\\ &= ||\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0||_2^2+2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2\\ &= \boldsymbol{\epsilon}^TX(X^TX)^{-2}X^T\boldsymbol{\epsilon}+2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)-\mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2.\end{aligned}$$

Following the proof for I_3, we have $2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0)/[4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0]^{1/2}\xrightarrow{\mathcal{D}}N(0,1)$ and hence $2\boldsymbol{\beta}_0^T(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0) = \Omega_P(\sigma_\epsilon||\boldsymbol{\beta}_0||_2/\sqrt{n})$. From $\mathrm{E}\{\boldsymbol{\epsilon}^TX(X^TX)^{-2}X^T\boldsymbol{\epsilon}\mathrm{I}(K)\} = \Omega(\sigma_\epsilon^2p_n/n)$, $\mathrm{tr}\{(X^TX)^{-1}\}\hat{\sigma}_\epsilon^2 = \sigma_\epsilon^2O_P(p_n/n)$ and $n^{-1} = o(\mathrm{SNR})$, we have the following results $\zeta_n^2 = \Omega[4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0]$ and $(||\hat{\boldsymbol{\beta}}||_2^2-||\boldsymbol{\beta}_0||_2^2)/[4\sigma_\epsilon^2\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0]^{1/2}\xrightarrow{\mathcal{D}}N(0,1)$. Hence, $(||\hat{\boldsymbol{\beta}}||_2^2-||\boldsymbol{\beta}_0||_2^2)/\zeta_n\xrightarrow{\mathcal{D}}N(0,1)$. The proof for part(b) is similar to that for diverging p_n.

In the end, we will discuss the extension to fixed design. From the proofs for (C.9)–(C.12) with $c_1 = 1$, $c_2 = -\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}$ and $c_3 = 2$, there exists a sequence of positive real numbers $\{\omega_n\}_{n\ge1}$ with $\omega_n = o(1)$ and constants $0 < C_1 < C_2 < \infty$, such that $\mathrm{P}(X\in\mathcal{X}_n)\to 1$ where $\mathcal{X}_n \subseteq \mathbb{R}^{n\times p_n}$ is a collection of all $x\in\mathbb{R}^{n\times p_n}$ satisfying

$$\sum_{k=1}^{6}\mathrm{var}\Big\{\sum_{j=1}^{n}\mathrm{II}_{k,j}\mathrm{I}(K)\Big|X=x\Big\}\le\omega_nt_n^4,$$
$$\sum_{j=1}^{n}\mathrm{E}[\{U_j\mathrm{I}(K)\}^4|X=x]\le\omega_nt_n^4,$$
$$\sum_{j=1}^{n}\sum_{k=1}^{3}\mathrm{E}\{\mathrm{II}_{k,j}\mathrm{I}(K)|X=x\}\in[C_1t_n^2,C_2t_n^2],$$
$$\sum_{j=1}^{n}\sum_{k=4}^{6}\mathrm{E}\{\mathrm{II}_{k,j}\mathrm{I}(K)|X=x\}\le\omega_nt_n^2$$
$$\Big|4\sigma_\epsilon^2\boldsymbol{\beta}_0^T(x^Tx)^{-1}\boldsymbol{\beta}_0+2\sigma_\epsilon^4\mathrm{tr}\{(x^Tx)^{-2}\}+\frac{2\sigma_\epsilon^4}{n-p_n}[\mathrm{tr}\{(x^Tx)^{-1}\}]^2-\zeta_n^2\Big|\le\omega_nt_n^2, \tag{C.13}$$

while the last equation above is due to Lemma 2. Then, using the martingale difference CLT in Heyde and Brown (1970), the asymptotic standard normality holds for $(||\hat{\boldsymbol{\beta}}||_2^2-||\boldsymbol{\beta}_0||_2^2)/\zeta_n$ conditioning on $X = x$ for any $x\in\mathcal{X}_n$. The consistency result in part (b) can be derived using similar arguments. □

Proof of Lemma 2. We provide the proof given event H. The results given event K can be similarly derived.

From Efron–Stein inequality in Efron and Stein (1981), if W is a function of n independent random variables and $W_{(i)}$ is any function of all those random variables except the ith, then

$$\mathrm{var}(W)\le\sum_{i=1}^{n}\mathrm{var}(W-W_{(i)})\le\sum_{i=1}^{n}\mathrm{E}(W-W_{(i)})^2. \tag{C.14}$$

First, we use (C.14) with

$$W = n^k/p_n\mathrm{tr}\{(X^TX)^{-k}\}\mathrm{I}(H),$$
$$W_{(i)} = n^k/p_n\mathrm{tr}\{(X_{(i)}^TX_{(i)})^{-k}\}\mathrm{I}(H_{(i)}),$$

where $H_{(i)}$ denotes the event that $||(X_{(i)}^TX_{(i)}/n)^{-1}||_2\le 1/x_2$. Note

$$\begin{aligned}&\sum_{i=1}^{n}\mathrm{E}(W-W_{(i)})^2 = n\mathrm{E}(W-W_{(i)})^2\\ &\lesssim n\mathrm{E}[n^k/p_n\mathrm{tr}\{(X^TX)^{-k}\}\{\mathrm{I}(H)-\mathrm{I}(H_{(i)})\}]^2\\ &\quad+n\mathrm{E}[n^k/p_n\mathrm{tr}\{(X^TX)^{-k}\}\mathrm{I}(H_{(i)})-n^k/p_n\mathrm{tr}\{(X_{(i)}^TX_{(i)})^{-k}\}\mathrm{I}(H_{(i)})]^2\\ &= \mathrm{I}+\mathrm{II}.\end{aligned}$$

Since $X^TX\succeq X_{(i)}^TX_{(i)}$, we know $||(X^TX)^{-1}||_2\le||(X_{(i)}^TX_{(i)})^{-1}||_2$ and hence $H\supseteq H_{(i)}$. Then, $\mathrm{I}(H)-\mathrm{I}(H_{(i)}) = \mathrm{I}(H\cap\bar{H}_{(i)}) = \mathrm{I}(H)\mathrm{I}(\bar{H}_{(i)})$. From Lemma 1,

$$\mathrm{I}\le n(n^k/p_n)^2(p_nn^{-k})^2\mathrm{P}(\bar{H}_{(i)}) = O(1/n).$$

Next, given $H_{(1)}$, we will show that

$$n^{2k+1}/p_n^2\mathrm{E}[\mathrm{tr}\{(X^TX)^{-k}\}-\mathrm{tr}\{(X_{(1)}^TX_{(1)})^{-k}\}]^2 = O(1/n).$$

From (C.1), we have

$$(X^TX)^{-k} = (X_{(1)}^TX_{(1)})^{-k}+\Delta,$$

where Δ is a sum of 2^k-1 terms, each of which can be expressed as $A_1\times A_2\times\cdots\times A_k$ with $A_i=(X_{(1)}^TX_{(1)})^{-1}$ or $A_i=B$ $(i=1,\ldots,k)$ where

$$B=-(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}/\{1+\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\},$$

and at least one of $A_1,\ldots,A_k$ is B. It suffices to show that for each of the 2^k-1 terms in Δ, $\mathrm{E}\{\mathrm{tr}(A_1A_2\cdots A_k)\}^2=O(p_n^2n^{-2k-2})$. Without loss of generality, if $A_1=B$, then from Lemmas 1 and S.1, given event $H_{(1)}$,

$$\begin{aligned}\mathrm{E}\{\mathrm{tr}(A_1A_2\cdots A_k)\}^2&\le\mathrm{E}\{\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}A_2\cdots A_k(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\}^2\\&=O(p_n^2n^{-2k-2}).\end{aligned}$$

Next, we prove the second result of this lemma. Without loss of generality, assume $||\boldsymbol{\beta}_0||_2=1$, and we will use (C.14) again with $W=n^k\boldsymbol{\beta}_0^T(X^TX)^{-k}\boldsymbol{\beta}_0\mathrm{I}(H)$ and $W_{(i)}=n^k\boldsymbol{\beta}_0^T(X_{(i)}^TX_{(i)})^{-k}\boldsymbol{\beta}_0\mathrm{I}(H_{(i)})$ to show that, for each of the 2^k-1 terms in Δ,

$$n^{2k+1}\mathrm{E}\{\boldsymbol{\beta}_0^TA_1A_2\cdots A_k\boldsymbol{\beta}_0\mathrm{I}(H_{(i)})\}^2=O(1/n).$$

We only give the proof of a special case that $A_1=A_2=B$, and $A_3=\cdots=A_k=(X_{(1)}^TX_{(1)})^{-1}$. From Lemma S.1,

$$\begin{aligned}&\mathrm{E}\{\boldsymbol{\beta}_0^TA_1A_2\cdots A_k\boldsymbol{\beta}_0\mathrm{I}(H_{(1)})\}^2\\&\le\mathrm{E}\{\boldsymbol{\beta}_0^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-2}\boldsymbol{X}_1\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-1}\\&\quad\times(X_{(1)}^TX_{(1)})^{-k+2}\boldsymbol{\beta}_0\mathrm{I}(H_{(1)})\}^2\\&=\mathrm{E}\{(\boldsymbol{\beta}_0^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1)^2(\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-2}\boldsymbol{X}_1)^2\\&\quad\times(\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-k+1}\boldsymbol{\beta}_0)^2\mathrm{I}(H_{(1)})\}\\&\le[\mathrm{E}\{(\boldsymbol{\beta}_0^T(X_{(1)}^TX_{(1)})^{-1}\boldsymbol{X}_1)^4\mathrm{I}(H_{(1)})\}]^{1/2}\\&\quad\times[\mathrm{E}\{(\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-2}\boldsymbol{X}_1)^8\mathrm{I}(H_{(1)})\}]^{1/4}\\&\quad\cdot[\mathrm{E}\{(\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-k+1}\boldsymbol{\beta}_0)^8\mathrm{I}(H_{(1)})\}]^{1/4}\\&\lesssim\{\mathrm{E}||\boldsymbol{\beta}_0^T(X_{(1)}^TX_{(1)})^{-1}\mathrm{I}(H_{(1)})||_2^4\}^{1/2}\\&\quad\times[\mathrm{E}\{\boldsymbol{X}_1^T(X_{(1)}^TX_{(1)})^{-2}\boldsymbol{X}_1\mathrm{I}(H_{(1)})\}^8]^{1/4}\\&\quad\cdot\{\mathrm{E}||(X_{(1)}^TX_{(1)})^{-k+1}\boldsymbol{\beta}_0\mathrm{I}(H_{(1)})||_2^8\}^{1/4}\\&\lesssim n^{-2}n^{-2}n^{-2k+2}=O(n^{-2k-2}).\end{aligned}$$

The proofs for the other terms are similar. We complete the proof. □

Proof of Theorem 3. First, we consider $p_n\to\infty$. Following the proof of Theorem 1, if $c_1=1$, $c_2=-\mathrm{Etr}\{(X^TX)^{-1}\mathrm{I}(K)\}$, $c_3=2$, we have $\hat{\zeta}_n^2-\zeta_n^2=o_{\mathrm{P}}(t_n^2)$ using the results in Lemmas S.12–S.14 and Proposition 1, where t_n is defined in the proof of Theorem 1.

For fixed p_n, we first show that $\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}/B\xrightarrow{\mathcal{P}}1$, where $B=\boldsymbol{\beta}_0^T\mathrm{E}\{(X^TX)^{-1}\mathrm{I}(K)\}\boldsymbol{\beta}_0=\Omega(||\boldsymbol{\beta}_0||_2^2/n)$. Note

$$\begin{aligned}\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}&=\boldsymbol{\beta}_0^T(X^TX)^{-1}\boldsymbol{\beta}_0+2\boldsymbol{\beta}_0^T(X^TX)^{-2}X^T\boldsymbol{\epsilon}\\&\quad+\boldsymbol{\epsilon}^TX(X^TX)^{-3}X^T\boldsymbol{\epsilon}.\end{aligned}$$

From Lemma 2, $\boldsymbol{\beta}_0^T(X^TX)^{-1}\boldsymbol{\beta}_0/B\xrightarrow{\mathcal{P}}1$. Since $2\boldsymbol{\beta}_0^T(X^TX)^{-2}X^T\boldsymbol{\epsilon}=O_{\mathrm{P}}(\sigma_\epsilon||\boldsymbol{\beta}_0||_2n^{-3/2})=o_{\mathrm{P}}(B)$ and $\boldsymbol{\epsilon}^TX(X^TX)^{-3}X^T\boldsymbol{\epsilon}=O_{\mathrm{P}}(\sigma_\epsilon^2n^{-2})=o_{\mathrm{P}}(B)$, we claim that $\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}/B\xrightarrow{\mathcal{P}}1$.

Recall $\hat{\zeta}_n^2=4\hat{\sigma}_\epsilon^2\hat{\boldsymbol{\beta}}^T(X^TX)^{-1}\hat{\boldsymbol{\beta}}-2\hat{\sigma}_\epsilon^4\mathrm{tr}\{(X^TX)^{-2}\}+2\hat{\sigma}_\epsilon^4[\mathrm{tr}\{(X^TX)^{-1}\}]^2/(n-p_n)$. Then $2\hat{\sigma}_\epsilon^4\mathrm{tr}\{(X^TX)^{-2}\}=O_{\mathrm{P}}(\sigma_\epsilon^4n^{-2})=o_{\mathrm{P}}(\sigma_\epsilon^2B)$ and $2\hat{\sigma}_\epsilon^4[\mathrm{tr}\{(X^TX)^{-1}\}]^2/(n-p_n)=O_{\mathrm{P}}(\sigma_\epsilon^4n^{-3})=o_{\mathrm{P}}(\sigma_\epsilon^2B)$. Therefore, from Proposition 1, we have $\hat{\zeta}_n^2/(4\sigma_\epsilon^2B)\xrightarrow{\mathcal{P}}1$.

Following the proof of Theorem 1, we have $\zeta_n^2/(4\sigma_\epsilon^2B)\xrightarrow{\mathcal{P}}1$, which implies that $\hat{\zeta}_n^2/\zeta_n^2\xrightarrow{\mathcal{P}}1$. We complete the proof. □

Supplementary Materials

The supplementary material includes a discussion of the conditions in Kelejian and Prucha (2001), extension of our results to centralized data, conventional inference for the fraction of variance explained, two-sample inferences, and the proofs for the rest of the main theoretical results as well as the technical lemmas.

Acknowledgments

We thank Mr. Ching-Wei Cheng for implementing our codes in Purdue supercomputer.

Funding

Xiao Guo's research was sponsored by the National Natural Science Foundation of China grants 12071452, 11601500, 11671374, and 11771418, and the Fundamental Research Funds for the Central Universities. Guang Cheng's research was sponsored by NSF DMS-1712907, DMS-1811812, DMS-1821183, and Office of Naval Research (ONR N00014-18-2759).

References

Arias-Castro, E., Candès, E. J., and Plan, Y. (2011), "Global Testing Under Sparse Alternatives: ANOVA, Multiple Comparisons and the Higher Criticism," *The Annals of Statistics*, 39, 2533–2556. [2,3]

Bai, Z., Jiang, D., Yao, J., and Zheng, S. (2013), "Testing Linear Hypotheses in High-Dimensional Regressions," *Statistics*, 47, 1207–1223. [4]

Bai, Z., and Silverstein, J. W. (2010), *Spectral Analysis of Large Dimensional Random Matrices*, New York: Springer. [3,5,16]

Cai, T. T., and Guo, Z. (2018), "Accuracy Assessment for High-Dimensional Linear Regression," *The Annals of Statistics*, 46, 1807–1836. [4]

——— (2020), "Semisupervised Inference for Explained Variance in High Dimensional Linear Regression and Its Applications," *Journal of the Royal Statistical Society*, Series B, 82, 391–419. [7]

Cai, T. T., Jin, J., and Low, M. G. (2007), "Estimation and Confidence Sets for Sparse Normal Mixtures," *The Annals of Statistics*, 35, 2421–2449. [2]

de Jong, P. (1987), "A Central Limit Theorem for Generalized Quadratic Forms," *Probability Theory and Related Fields*, 75, 261–277. [3]

Dicker, L. H. (2014), "Variance Estimation in High-Dimensional Linear Models," *Biometrika*, 101, 269–284. [2,3,4,8,9,14,15]

Dicker, L. H., and Erdogdu, M. A. (2016), "Maximum Likelihood for Variance Estimation in High-Dimensional Linear Models," in *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (Vol. 51), pp. 159–167. [8,9,14]

——— (2017), "Flexible Results for Quadratic Forms With Applications to Variance Components Estimation," *The Annals of Statistics*, 45, 386–414. [3,8,9]

Dobriban, E., and Su, W. J. (2018), "Robust Inference Under Heteroskedasticity via the Hadamard Estimator," arXiv no. 1807.00347. [4]

Donoho, D., and Jin, J. (2004), "Higher Criticism for Detecting Sparse Heterogeneous Mixtures," *The Annals of Statistics*, 32, 962–994. [2]

Donoho, D., and Montanari, A. (2016), "High Dimensional Robust M-Estimation: Asymptotic Variance via Approximate Message Passing," *Probability Theory and Related Fields*, 166, 935–969. [4]

Efron, B., and Stein, C. (1981), "The Jackknife Estimate of Variance," *The Annals of Statistics*, 9, 586–596. [18]

El Karoui, N. (2013), "Asymptotic Behavior of Unregularized and Ridge-Regularized High-Dimensional Robust Regression Estimators: Rigorous Results," arXiv no. 1311.2445. [2,3,4]

——— (2018), "On the Impact of Predictor Geometry on the Performance on High-Dimensional Ridge-Regularized Generalized Robust Regression Estimators," *Probability Theory and Related Fields*, 170, 95–175. [2,3,4]

El Karoui, N., Bean, D., Bickel, P. J., Lim, C., and Yu, B. (2013), "On Robust Regression With High-Dimensional Predictors," *Proceedings of the National Academy of Sciences of the United States of America*, 110, 14557–14562. [4]

Fan, J., Guo, S., and Hao, N. (2012), "Variance Estimation Using Refitted Cross-Validation in Ultrahigh Dimensional Regression," *Journal of the Royal Statistical Society*, Series B, 74, 37–65. [3]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1]

Guo, Z., Wang, W., Cai, T. T., and Li, H. (2016), "Optimal Estimation of Co-Heritability in High-Dimensional Linear Models," arXiv no. 1605.07244. [3]

Hall, P., and Jin, J. (2010), "Innovated Higher Criticism for Detecting Sparse Signals in Correlated Noise," *The Annals of Statistics*, 38, 1686–1732. [2]

Heyde, C. C., and Brown, B. M. (1970), "On the Departure From Normality of a Certain Class of Martingales," *The Annals of Mathematical Statistics*, 41, 2161–2165. [3,5,18]

Ingster, Y. I., Tsybakov, A. B., and Verzelen, N. (2010), "Detection Boundary in Sparse Regression," *Electronic Journal of Statistics*, 4, 1476–1526. [2,3,7]

Janson, L., Barber, R. F., and Candès, E. (2017), "EigenPrism: Inference for High Dimensional Signal-to-Noise Ratios," *Journal of the Royal Statistical Society*, Series B, 79, 1037–1065. [2,4,8,9]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1]

Kelejian, H. H., and Prucha, I. R. (2001), "On the Asymptotic Distribution of the Moran I Test Statistic With Applications," *Journal of Econometrics*, 104, 219–257. [3,19]

Kolmogorov, A. N. (1933), "Sulla Determinazione Empirica di una Legge di Distribuzione," *Giornale dell'Instituto Italiano degli Attuari*, 4, 83–91. [2]

Lei, L., Bickel, P. J., and El Karoui, N. (2018), "Asymptotics for High Dimensional Regression M-Estimates: Fixed Design Results," *Probability Theory and Related Fields*, 172, 983–1079. [4]

Letac, G., and Massam, H. (2004), "All Invariant Moments of the Wishart Distribution," *Scandinavian Journal of Statistics*, 31, 295–318. [5]

Meinshausen, N., and Bühlmann, P. (2006), "High-Dimensional Graphs and Variable Selection With the Lasso," *The Annals of Statistics*, 34, 1436–1462. [1]

Meinshausen, N., and Yu, B. (2009), "Lasso-Type Recovery of Sparse Representations for High-Dimensional Data," *The Annals of Statistics*, 37, 246–270. [1]

Nickl, R., and van de Geer, S. (2013), "Confidence Sets in Sparse Regression," *The Annals of Statistics*, 41, 2852–2876. [7]

Portnoy, S. (1984), "Asymptotic Behavior of M-Estimators of p Regression Parameters When p^2/n Is Large. I. Consistency," *The Annals of Statistics*, 12, 1298–1309. [4]

——— (1985), "Asymptotic Behavior of M Estimators of p Regression Parameters When p^2/n Is Large. II. Normal Approximation," *The Annals of Statistics*, 13, 1403–1417. [2,4]

Shao, J. (2003), *Mathematical Statistics*, New York: Springer. [6]

Sherman, J., and Morrison, W. J. (1950), "Adjustment of an Inverse Matrix Corresponding to a Change in One Element of a Given Matrix," *The Annals of Mathematical Statistics*, 21, 124–127. [15]

Smirnov, N. V. (1939), "Estimate of Deviation Between Empirical Distribution Functions in Two Independent Samples," *Bulletin Moscow University*, 2, 3–16. [2]

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavaré, S., Deloukas, P., and Dermitzakis, E. T. (2007), "Population Genomics of Human Gene Expression," *Nature Genetics*, 39, 1217–1224. [1,10,11,14]

Sun, T., and Zhang, C. H. (2012), "Scaled Sparse Linear Regression," *Biometrika*, 99, 879–898. [3]

Sur, P., Chen, Y., and Candès, E. J. (2019), "The Likelihood Ratio Test in High-Dimensional Logistic Regression Is Asymptotically a Rescaled Chi-Square," *Probability Theory and Related Fields*, 175, 487–558. [4]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1]

van de Geer, S. (2008), "High-Dimensional Generalized Linear Models and the Lasso," *The Annals of Statistics*, 36, 614–645. [1]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *The Annals of Statistics*, 42, 1166–1202. [1]

Verzelen, N., and Gassiat, E. (2018), "Adaptive Estimation of High-Dimensional Signal-to-Noise Ratios," *Bernoulli*, 24, 3683–3710. [3,8,9]

Zhang, C. H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [1]

Zhang, C. H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1]

Zhang, X., and Cheng, G. (2017), "Simultaneous Inference for High-Dimensional Linear Models," *Journal of the American Statistical Association*, 112, 757–768. [2]

Zhong, P., and Chen, S. (2011), "Tests for High-Dimensional Regression Coefficients With Factorial Designs," *Journal of the American Statistical Association*, 106, 260–274. [2]