The Annals of Statistics
2018, Vol. 46, No. 3, 989–1017
https://doi.org/10.1214/17-AOS1575
© Institute of Mathematical Statistics, 2018

ARE DISCOVERIES SPURIOUS? DISTRIBUTIONS OF MAXIMUM SPURIOUS CORRELATIONS AND THEIR APPLICATIONS

BY JIANQING FAN[*,†,1], QI-MAN SHAO[‡,2] AND WEN-XIN ZHOU[†,§,3]

Fudan University[*], *Princeton University*[†], *Chinese University of Hong Kong*[‡]
and University of California, San Diego[§]

Over the last two decades, many exciting variable selection methods have been developed for finding a small group of covariates that are associated with the response from a large pool. Can the discoveries from these data mining approaches be spurious due to high dimensionality and limited sample size? Can our fundamental assumptions about the exogeneity of the covariates needed for such variable selection be validated with the data? To answer these questions, we need to derive the distributions of the maximum spurious correlations given a certain number of predictors, namely, the distribution of the correlation of a response variable Y with the best s linear combinations of p covariates $\mathbf{X}$, even when $\mathbf{X}$ and Y are independent. When the covariance matrix of $\mathbf{X}$ possesses the restricted eigenvalue property, we derive such distributions for both a finite s and a diverging s, using Gaussian approximation and empirical process techniques. However, such a distribution depends on the unknown covariance matrix of $\mathbf{X}$. Hence, we use the multiplier bootstrap procedure to approximate the unknown distributions and establish the consistency of such a simple bootstrap approach. The results are further extended to the situation where the residuals are from regularized fits. Our approach is then used to construct the upper confidence limit for the maximum spurious correlation and to test the exogeneity of the covariates. The former provides a baseline for guarding against false discoveries and the latter tests whether our fundamental assumptions for high-dimensional model selection are statistically valid. Our techniques and results are illustrated with both numerical examples and real data analysis.

1. Introduction.

Information technology has forever changed the data collection process. Massive amounts of very high dimensional or unstructured data are continuously produced and stored at an affordable cost. Massive and complex data and high dimensionality characterize contemporary statistical problems in many emerging fields of science and engineering. Various statistical and machine learning methods and algorithms have been proposed to find a small group of covariate

Received October 2016; revised April 2017.

[1] Supported in part by NSF Grants DMS-1206464, DMS-1406266, and NIH Grant R01-GM072611-10.

[2] Supported in part by Hong Kong Research Grants Council GRF-403513 and GRF-14302515.

[3] Supported by NIH Grant R01-GM072611-10.

MSC2010 subject classifications. Primary 62H10, 62H20; secondary 62E17, 62F03.

Key words and phrases. High dimension, spurious correlation, bootstrap, false discovery.

variables that are associated with given responses such as biological and clinical outcomes. These methods have been very successfully applied to genomics, genetics, neuroscience, economics and finance. For an overview of high-dimensional statistical theory and methods, see the review article by Fan and Lv (2010) and monographs by Dudoit and van der Laan (2008), Hastie, Tibshirani and Friedman (2009), Efron (2010) and Bühlmann and van de Geer (2011).

Underlying machine learning, data mining, and high-dimensional statistical techniques, there are many model assumptions and even heuristic arguments. For example, the LASSO [Tibshirani (1996)] and the SCAD [Fan and Li (2001)] are based on an exogeneity assumption, meaning that all of the covariates and the residual of the true model are uncorrelated. However, it is nearly impossible that such a random variable, which is the part of the response variable that cannot be explained by a small group of covariates and lives in a low-dimensional space spanned by the response and the small group of variables, is uncorrelated with any of the tens of thousands of coviariates. Indeed, Fan and Liao (2014) and Fan, Han and Liu (2014) provide evidence that such an ideal assumption might not be valid, although it is a necessary condition for model selection consistency. Even under the exogenous assumption, conditions such as the restricted eigenvalue condition [Bickel, Ritov and Tsybakov (2009)] and homogeneity [Fan, Han and Liu (2014)] are needed to ensure model selection consistency or oracle properties. Despite their critical importance, these conditions have rarely been verified in practice. Their violations can lead to false scientific discoveries. A simpler question is then, for a given data set, do data mining techniques produce results that are better than spurious correlation? The answer depends on not only the correlation between the fitted and observed values, but also on the sample size, the number of variables selected and the total number of variables.

To better appreciate the above two questions, let us consider an example. We take the gene expression data on 90 Asians (45 Japanese and 45 Han Chinese) from the international "HapMap" project [Thorisson et al. (2005)]. The normalized gene expression data are generated with an Illumina Sentrix Human-6 Expression Bead Chip [Stranger et al. (2007)] and are available on ftp://ftp.sanger.ac.uk/pub/genevar/. We take the expressions of gene *CHRNA6*, a cholinergic receptor, nicotinic, alpha 6, as the response Y and the remaining expressions of probes as covariates $\mathbf{X}$ with dimension $p = 47{,}292$. We first fit an ℓ_1-penalized least-squares regression (LASSO) on the data with a tuning parameter automatically selected via tenfold cross validation (25 genes are selected). The correlation between the LASSO-fitted value and the response is 0.8991. Next, we refit an ordinary least-squares regression on the selected model to calculate the fitted response and residual vector. The sample correlation between the post-LASSO fit and observed responses is 0.9214, a remarkable fit! But is it any better than the spurious correlation? The model diagnostic plot, which depicts the empirical distribution of the correlations between each covariate X_j and the residual $\widehat{\varepsilon}$ after the LASSO

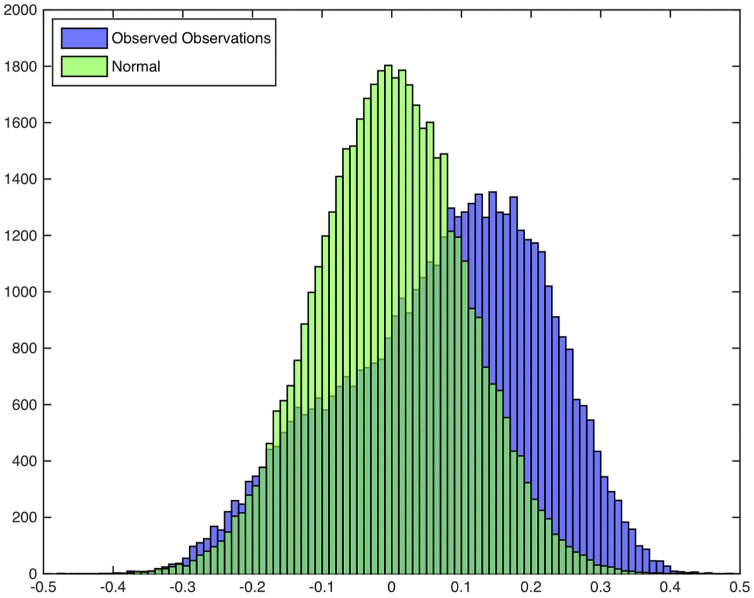

FIG. 1. *Histogram of the sample correlations between the residuals and each covariate (blue) and histogram of $N(0, 1/\sqrt{n})$ random variables (green).*

fit, is given in Figure 1. Does the exogenous assumption that $\mathbb{E}(\varepsilon X_j) = 0$ for all $j = 1, \dots, p$ hold?

To answer the above two important questions, we need to derive the distributions of the maximum spurious correlations. Let $\mathbf{X}$ be the p-dimensional random vector of the covariates and $\mathbf{X}_S$ be a subset of covariates indexed by S. Let $\widehat{\operatorname{corr}}_n(\varepsilon, \boldsymbol{\alpha}_S^{\mathrm{T}}\mathbf{X}_S)$ be the sample correlation between the random noise ε (independent of $\mathbf{X}$) and $\boldsymbol{\alpha}_S^{\mathrm{T}}\mathbf{X}_S$ based on a sample of size n, where $\boldsymbol{\alpha}_S$ is a constant vector. Then the maximum spurious correlation is defined as

$$\widehat{R}_n(s, p) = \max_{|S|=s} \max_{\boldsymbol{\alpha}_S} \widehat{\operatorname{corr}}_n(\varepsilon, \boldsymbol{\alpha}_S^{\mathrm{T}}\mathbf{X}_S), \tag{1.1}$$

when $\mathbf{X}$ and ε are independent, where the maximization is taken over all $\binom{p}{s}$ subsets of size s and all of the linear combinations of the selected s covariates. Next, let $(Y_i, \mathbf{X}_i), \dots, (Y_n, \mathbf{X}_n)$ be independent and identically distributed (i.i.d.) observations from the linear model $Y = \mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}^* + \varepsilon$. Assume that s covariates are selected by a certain variable selection method for some $1 \le s \ll \min(p, n)$. If the correlation between the fitted response and observed response is no more than the 90th or the 95th percentile of $\widehat{R}_n(s, p)$, it is hard to claim that the fitted value is impressive or even genuine. In this case, the finding is hardly more impressive than the best fit using data that consist of independent response and explanatory variables, 90% or 95% of the time. To simplify and unify the terminology, we call this result the spurious discovery throughout this paper.

For the aforementioned gene expression data, as 25 probes are selected, the observed correlation of 0.9214 between the fitted value and the response should be compared with the distribution of $\widehat{R}_n(25, p)$. Further, a simple method to test the null hypothesis

$$\mathbb{E}(\varepsilon X_j) = 0 \qquad \text{for all } j = 1, \dots, p, \tag{1.2}$$

is to compare the maximum absolute correlation in Figure 1 with the distribution of $\widehat{R}_n(1, p)$; see additional details in Section 5.3.

The importance of such spurious correlation was recognized by Cai and Jiang (2011), Fan, Guo and Hao (2012) and Cai, Fan and Jiang (2013). When the data are independently and normally distributed, they derive the distribution of $\widehat{R}_n(1, p)$, which is equivalent to the distribution of the minimum angle to the north pole among p random points uniformly distributed on the $(n+1)$-dimensional sphere. Fan, Guo and Hao (2012) conducted simulations to demonstrate that the spurious correlation can be very high when p is large and grows quickly with s. To demonstrate this effect and to examine the impact of correlation and sample size, we conduct a similar but more extensive simulation study based on a combination of the stepwise addition and branch-and-bound algorithms. We simulate $\mathbf{X}$ from $N(\mathbf{0}, \mathbf{I}_p)$ and $N(\mathbf{0}, \mathbf{\Sigma}_0)$, where $\mathbf{\Sigma}_0$ is block diagonal with the first block being a 500×500 equi-correlation matrix with a correlation 0.8 and the second block being the $(p-500) \times (p-500)$ identity matrix. Y is simulated independently of $\mathbf{X}$ and follows the standard normal distribution. Figure 2 depicts the simulation results for $n = 50$, 100 and 200. Clearly, the distributions depend on (s, p, n) and $\mathbf{\Sigma}$, the covariance matrix of $\mathbf{X}$, although the dependence on $\mathbf{\Sigma}$ does not seem very strong. However, the theoretical result of Fan, Guo and Hao (2012) covers only the very specific case where $s = 1$ and $\mathbf{\Sigma} = \mathbf{I}_p$.

There are several challenges to deriving the asymptotic distribution of the statistic $\widehat{R}_n(s, p)$, as it involves combinatorial optimization. Further technical complications are added by the dependence among the covariates $\mathbf{X}$. Nevertheless, under the restricted eigenvalue condition [Bickel, Ritov and Tsybakov (2009)] on $\mathbf{\Sigma}$, in this paper, we derive the asymptotic distribution of such a spurious correlation statistic for both a fixed s and a diverging s, using the empirical process and Gaussian approximation techniques given in Chernozhukov, Chetverikov and Kato (2014). As expected, such distributions depend on the unknown covariance matrix $\mathbf{\Sigma}$. To provide a consistent estimate of the distributions of the spurious correlations, we consider the use of a multiplier bootstrap method and demonstrate its consistency under mild conditions. The multiplier bootstrap procedure has been widely used due to its good numerical performance. Its theoretical validity is guaranteed by the multiplier central limit theorem [van der Vaart and Wellner (1996)]. For the most advanced recent results, we refer to Chatterjee and Bose (2005), Arlot, Blanchard and Roquain (2010) and Chernozhukov, Chetverikov and Kato (2013). In particular, Chernozhukov, Chetverikov and Kato (2013) developed a number of

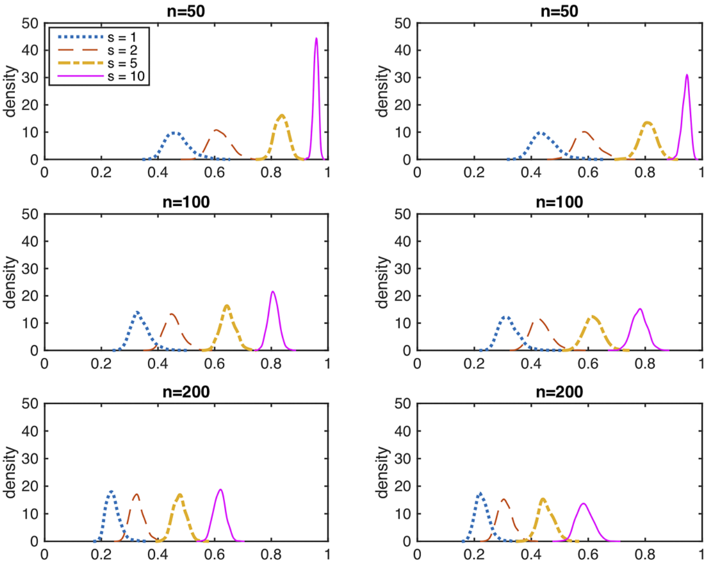

FIG. 2. *Distributions of maximum spurious correlations for* $p = 1000$ *and* $s = 1, 2, 5$ *and* 10 *when* $\mathbf{\Sigma}$ *is the identity matrix* (*left panel*) *or block diagonal* (*right panel*) *with the first block being a* 500×500 *equi-correlation matrix with a correlation* 0.8 *and the second block being the* 500×500 *identity matrix. From top to bottom*: $n = 50$, 100 *and* 200.

nonasymptotic results on a multiplier bootstrap for the maxima of empirical mean vectors in high dimensions with applications to multiple hypothesis testing and parameter choice for the Dantzig selector. The use of multiplier bootstrapping enables us to empirically compute the upper confidence limit of $\widehat{R}_n(s, p)$, and hence decide whether discoveries by statistical machine learning techniques are any better than spurious correlations.

The rest of this paper is organized as follows. Section 2 discusses the concept of spurious correlation and introduces the main conditions and notation. Section 3 presents the main results of the asymptotic distributions of spurious correlations and their bootstrap approximations, which are further extended in Section 4. Section 5 identifies three important applications of our results to high-dimensional statistical inference. Section 6 presents the numerical studies. The proof of Theorem 3.1 is provided in Section 7, and the proofs for the remaining theoretical results are provided in the Supplementary Material [Fan, Shao and Zhou (2018)].

2. Spurious correlation, conditions and notation. Let $\varepsilon, \varepsilon_1, \dots, \varepsilon_n$ be i.i.d. random variables with a mean of zero and a finite variance $\sigma^2 > 0$, and let $\mathbf{X}, \mathbf{X}_1, \dots, \mathbf{X}_n$ be i.i.d. p-dimensional random vectors with a mean of zero and a covariance matrix $\mathbf{\Sigma} = \mathbb{E}(\mathbf{X}\mathbf{X}^{\mathrm{T}}) = (\sigma_{jk})_{1 \le j,k \le p}$. Write

$$\mathbf{X} = (X_1, \dots, X_p)^{\mathrm{T}}, \qquad \mathbf{X}_i = (X_{i1}, \dots, X_{ip})^{\mathrm{T}}, \qquad i = 1, \dots, n.$$

Assume that the two samples $\{\varepsilon_i\}_{i=1}^n$ and $\{\mathbf{X}_i\}_{i=1}^n$ are independent. Then the spurious correlation (1.1) can be written as

$$\widehat{R}_n(s,p) = \max_{\boldsymbol{\alpha}\in\mathbb{S}^{p-1}:|\boldsymbol{\alpha}|_0=s} \widehat{\operatorname{corr}}_n\big(\varepsilon, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}\big), \tag{2.1}$$

where the dimension p and sparsity s are allowed to grow with the sample size n. Here, $\widehat{\operatorname{corr}}_n(\cdot,\cdot)$ denotes the sample Pearson correlation coefficient and $\mathbb{S}^{p-1} := \{\boldsymbol{\alpha}\in\mathbb{R}^p : |\boldsymbol{\alpha}|_2 = 1\}$ is the unit sphere of $\mathbb{R}^p$. Due to the anti-symmetric property of the sample correlation under the sign transformation of $\boldsymbol{\alpha}$, we have also

$$\widehat{R}_n(s,p) = \max_{\boldsymbol{\alpha}\in\mathbb{S}^{p-1}:|\boldsymbol{\alpha}|_0=s} \big|\widehat{\operatorname{corr}}_n\big(\varepsilon, \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{X}\big)\big|. \tag{2.2}$$

More specifically, we can express $\widehat{R}_n(s,p)$ as

$$\max_{S\subseteq[p]:|S|=s}\max_{\boldsymbol{\alpha}\in\mathbb{S}^{s-1}} \frac{\sum_{i=1}^n(\varepsilon_i-\bar{\varepsilon}_n)\langle\boldsymbol{\alpha}, \mathbf{X}_{i,S}-\bar{\mathbf{X}}_{n,S}\rangle}{\sqrt{\sum_{i=1}^n(\varepsilon_i-\bar{\varepsilon}_n)^2\cdot\sum_{i=1}^n\langle\boldsymbol{\alpha}, \mathbf{X}_{i,S}-\bar{\mathbf{X}}_{n,S}\rangle^2}}. \tag{2.3}$$

By the scale-invariance property of $\widehat{R}_n(s,p)$, we assume without loss of generality that $\sigma^2 = 1$ and $\boldsymbol{\Sigma}$ is a correlation matrix, so that $\operatorname{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_p$.

For a random variable X, the sub-Gaussian norm $\|X\|_{\psi_2}$ and sub-exponential norm $\|X\|_{\psi_1}$ of X are defined, respectively, as

$$\|X\|_{\psi_2} = \sup_{q\geq 1} q^{-1/2}\big(\mathbb{E}|X|^q\big)^{1/q} \quad \text{and} \quad \|X\|_{\psi_1} = \sup_{q\geq 1} q^{-1}\big(\mathbb{E}|X|^q\big)^{1/q}.$$

A random variable X that satisfies $\|X\|_{\psi_2} < \infty$ (resp., $\|X\|_{\psi_1} < \infty$) is called a sub-Gaussian (resp., sub-exponential) random variable [Vershynin (2012)].

The following moment conditions for $\varepsilon\in\mathbb{R}$ and $\mathbf{X}\in\mathbb{R}^p$ are imposed.

CONDITION 2.1. *There exists a random vector* $\mathbf{U}$ *such that* $\mathbf{X} = \boldsymbol{\Sigma}^{1/2}\mathbf{U}$, $\mathbb{E}(\mathbf{U}) = \mathbf{0}$, $\mathbb{E}(\mathbf{U}\mathbf{U}^{\mathrm{T}}) = \mathbf{I}_p$ *and* $K_1 := \sup_{\boldsymbol{\alpha}\in\mathbb{S}^{p-1}}\|\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{U}\|_{\psi_2} < \infty$. *The random variable* ε *has a zero mean and unit variance, and is sub-Gaussian with* $K_0 := \|\varepsilon\|_{\psi_2} < \infty$. *Moreover, write* $v_q = \mathbb{E}(|\varepsilon|^q)$ *for* $q \geq 3$.

The following is our assumption for the sampling process.

CONDITION 2.2. $\{\varepsilon_i\}_{i=1}^n$ *and* $\{\mathbf{X}_i\}_{i=1}^n$ *are independent random samples from the distributions of* ε *and* $\mathbf{X}$, *respectively.*

For $1 \leq s \leq p$, the s-sparse minimal and maximal eigenvalues [Bickel, Ritov and Tsybakov (2009)] of the covariance matrix $\boldsymbol{\Sigma}$ are defined as

$$\phi_{\min}(s) = \min_{\mathbf{u}\in\mathbb{R}^p:1\leq|\mathbf{u}|_0\leq s}\big(|\mathbf{u}|_{\boldsymbol{\Sigma}}/|\mathbf{u}|_2\big)^2, \qquad \phi_{\max}(s) = \max_{\mathbf{u}\in\mathbb{R}^p:1\leq|\mathbf{u}|_0\leq s}\big(|\mathbf{u}|_{\boldsymbol{\Sigma}}/|\mathbf{u}|_2\big)^2,$$

where $|\mathbf{u}|_{\boldsymbol{\Sigma}} = (\mathbf{u}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{u})^{1/2}$ and $|\mathbf{u}|_2 = (\mathbf{u}^{\mathrm{T}}\mathbf{u})^{1/2}$ is the ℓ_2-norm of $\mathbf{u}$. Consequently, for $1 \le s \le p$, the s-sparse condition number of $\boldsymbol{\Sigma}$ is given by

$$\gamma_s = \gamma_s(\boldsymbol{\Sigma}) = \sqrt{\phi_{\max}(s)/\phi_{\min}(s)}. \tag{2.4}$$

The quantity γ_s plays an important role in our analysis.

The following notation is used. For the two sequences $\{a_n\}$ and $\{b_n\}$ of positive numbers, we write $a_n = O(b_n)$ or $a_n \lesssim b_n$ if there exists a constant $C > 0$ such that $a_n/b_n \le C$ for all sufficiently large n; we write $a_n \asymp b_n$ if there exist constants $C_1, C_2 > 0$ such that, for all n large enough, $C_1 \le a_n/b_n \le C_2$; and we write $a_n \sim b_n$ and $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 1$ and $\lim_{n\to\infty} a_n/b_n = 0$, respectively. For $a, b \in \mathbb{R}$, we write $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$. For every vector $\mathbf{u}$, we denote by $|\mathbf{u}|_q = (\sum_{i\ge 1} |u_i|^q)^{1/q}$ for $q > 0$ and $|\mathbf{u}|_0 = \sum_{i\ge 1} I\{u_i \ne 0\}$. We use $\langle \mathbf{u}, \mathbf{v}\rangle = \mathbf{u}^{\mathrm{T}}\mathbf{v}$ to denote the inner product of two vectors $\mathbf{u}$ and $\mathbf{v}$ with the same dimension and $\|\mathbf{M}\|$ to denote the spectral norm of a matrix $\mathbf{M}$. For every positive integer ℓ, we write $[\ell] = \{1, 2, \ldots, \ell\}$, and for any set S, we use S^{c} to denote its complement and $|S|$ for its cardinality. For each p-dimensional vector $\mathbf{u}$ and $p \times p$ positive semidefinite matrix $\mathbf{A}$, we write $|\mathbf{u}|_{\mathbf{A}} = (\mathbf{u}^{\mathrm{T}}\mathbf{A}\mathbf{u})^{1/2}$. In particular, put

$$\boldsymbol{\alpha}_{\boldsymbol{\Sigma}} = \boldsymbol{\alpha}/|\boldsymbol{\alpha}|_{\boldsymbol{\Sigma}} \tag{2.5}$$

for every $\boldsymbol{\alpha} \in \mathbb{R}^p$ and set $\mathbf{0}_{\boldsymbol{\Sigma}} = \mathbf{0}$ as the convention.

3. Distributions of maximum spurious correlations.

In this section, we first derive the asymptotic distributions of the maximum spurious correlation $\widehat{R}_n(s, p)$. The analytic form of such asymptotic distributions can be obtained in the isotropic case. As the asymptotic distributions of $\widehat{R}_n(s, p)$ depend on the unknown covariance matrix $\boldsymbol{\Sigma}$, we provide a bootstrap estimate and demonstrate its consistency.

3.1. *Asymptotic distributions of maximum spurious correlations.*

In view of (2.3), we can rewrite $\widehat{R}_n(s, p)$ as

$$\widehat{R}_n(s, p) = \sup_{f\in\mathcal{F}} \frac{n^{-1}\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n) f(\mathbf{X}_i - \bar{\mathbf{X}}_n)}{\sqrt{n^{-1}\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n)^2} \cdot \sqrt{n^{-1}\sum_{i=1}^n f^2(\mathbf{X}_i - \bar{\mathbf{X}}_n)}}, \tag{3.1}$$

where $\bar{\varepsilon}_n = n^{-1}\sum_{i=1}^n \varepsilon_i$, $\bar{\mathbf{X}}_n = n^{-1}\sum_{i=1}^n \mathbf{X}_i$ and

$$\mathcal{F} = \mathcal{F}(s, p) = \{\mathbf{x} \mapsto f_{\boldsymbol{\alpha}}(\mathbf{x}) := \langle \boldsymbol{\alpha}, \mathbf{x}\rangle : \boldsymbol{\alpha} \in \mathcal{V}\} \tag{3.2}$$

is a class of linear functions $\mathbb{R}^p \mapsto \mathbb{R}$, where $\mathcal{V} = \mathcal{V}(s, p) = \{\boldsymbol{\alpha} \in \mathbb{S}^{p-1} : |\boldsymbol{\alpha}|_0 = s\}$. The dependence of $\mathcal{F}$ and $\mathcal{V}$ on (s, p) is suppressed.

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^{\mathrm{T}}$ be a p-dimensional Gaussian random vector with a mean of zero and the covariance matrix $\boldsymbol{\Sigma}$, that is, $\mathbf{Z} \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Sigma})$. Denote by $Z_{(1)}^2 \le Z_{(2)}^2 \le \cdots \le Z_{(p)}^2$ the order statistics of $\{Z_1^2, \ldots, Z_p^2\}$. The following theorem shows that the distribution of the maximum absolute multiple correlation $\widehat{R}_n(s, p)$ can be approximated by that of the supremum of a centered Gaussian process $\mathbb{G}^*$ indexed by $\mathcal{F}$.

THEOREM 3.1. *Let Conditions 2.1 and 2.2 hold, $n, p \geq 2$ and $1 \leq s \leq p$. Then there exists a constant $C > 0$ independent of (s, p, n) such that*

$$\begin{aligned} &\sup_{t\geq 0}\big|\mathbb{P}\{\sqrt{n}\widehat{R}_n(s,p) \leq t\} - \mathbb{P}\{R^*(s,p) \leq t\}\big| \\ &\quad \leq C(K_0K_1)^{3/4}n^{-1/8}\{sb_n(s,p)\}^{7/8}, \end{aligned} \tag{3.3}$$

where K_0 and K_1 are defined in Condition 2.1, $b_n(s,p) := \log(\gamma_s p/s) \vee \log n$ for γ_s as in (2.4), $R^*(s,p) := \sup_{f\in\mathcal{F}} \mathbb{G}^* f$ and $\mathbb{G}^* = \{\mathbb{G}^* f\}_{f\in\mathcal{F}}$ is a centered Gaussian process indexed by $\mathcal{F}$ defined as, for every $f_{\boldsymbol{\alpha}} \in \mathcal{F}$,

$$\mathbb{G}^* f_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}^{\mathrm{T}}\mathbf{Z} = \frac{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Z}}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}}}. \tag{3.4}$$

In particular, if $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $s\log(pn) = o(n^{1/7})$, then as $n \to \infty$,

$$\sup_{t\geq 0}\big|\mathbb{P}\{n\widehat{R}_n^2(s,p) \leq t\} - \mathbb{P}\{Z_{(p)}^2 + \cdots + Z_{(p-s+1)}^2 \leq t\}\big| \to 0. \tag{3.5}$$

REMARK 3.1. The Berry–Esseen bound given in Theorem 3.1 depends explicitly on the triplet (s, p, n), and it depends on the covariance matrix $\boldsymbol{\Sigma}$ only through its s-sparse condition number γ_s, defined in (2.4). The proof of (3.3) builds on a number of technical tools including a standard covering argument, maximal and concentration inequalities for the suprema of unbounded empirical processes and Gaussian processes as well as a coupling inequality for the maxima of sums of random vectors derived in Chernozhukov, Chetverikov and Kato (2014). Instead, if we directly resort to the general framework in Theorem 2.1 of Chernozhukov, Chetverikov and Kato (2014), the function class of interest is $\mathcal{F} = \{\mathbf{x} \mapsto \frac{\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}}{(\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha})^{1/2}} : \boldsymbol{\alpha} \in \mathbb{S}^{p-1}, |\boldsymbol{\alpha}|_0 = s\}$. Checking high-level conditions in Theorem 2.1 can be rather complicated and less intuitive. Also, dealing with the (uniform) entropy integral that corresponds to the class $\mathcal{F}$ relies on verifying various VC-type properties and thus can be fairly tedious. Following a strategy similar to that used to prove Theorem 2.1, we provide a self-contained proof of Theorem 3.1 in Section 7.2 by making the best use of the specific structure of $\mathcal{F}$. The proof is more intuitive and straightforward. More importantly, it leads to an explicit nonasymptotic bound under transparent conditions.

REMARK 3.2. In Theorem 3.1, the independence assumption of ε and $\mathbf{X}$ can be relaxed as $\mathbb{E}(\varepsilon\mathbf{X}) = 0$, $\mathbb{E}(\varepsilon^2|\mathbf{X}) = \sigma^2$ and $\mathbb{E}(\varepsilon^4|\mathbf{X}) \leq C$ almost surely, where $C > 0$ is a constant.

Expression (3.5) indicates that the increment $n\{\widehat{R}_n^2(s,p) - \widehat{R}_n^2(s-1,p)\}$ is approximately the same as $Z_{(p-s+1)}^2$. This can simply be seen from the asymptotic joint distribution of $(\widehat{R}_n(1,p), \widehat{R}_n(2,p), \ldots, \widehat{R}_n(s,p))$. The following proposition establishes the approximation of the joint distributions when both the dimension p and sparsity s are allowed to diverge with the sample size n.

PROPOSITION 3.1. *Let Conditions 2.1 and 2.2 hold with $\mathbf{\Sigma} = \mathbf{I}_p$. Assume that the triplet (s, p, n) satisfies $1 \le s < n \le p$ and $s^2 \log p = o(n^{1/7})$. Then as $n \to \infty$,*

$$\sup_{0 \equiv t_0 < t_1 < t_2 < \cdots < t_s < 1} \Bigg| \mathbb{P}\Bigg[\bigcap_{k=1}^{s}\{\widehat{R}_n(k, p) \le t_k\}\Bigg] - \mathbb{P}\Bigg[\bigcap_{k=1}^{s}\{Z^2_{(p-k+1)} \le n(t_k^2 - t_{k-1}^2)\}\Bigg]\Bigg| \to 0.$$

REMARK 3.3. When $s = 1$ and if (n, p) satisfies $\log p = o(n^{1/7})$, it is straightforward to verify that, for any $t \in \mathbb{R}$,

$$\mathbb{P}\{Z^2_{(p)} - 2\log p + \log(\log p) \le t\} \to \exp(-\pi^{-1/2}e^{-t/2}) \qquad \text{as } p \to \infty. \tag{3.6}$$

This result is similar in nature to (5) in Fan, Guo and Hao (2012). In fact, it is proved in Shao and Zhou (2014) that the extreme-value statistic $\widehat{R}_n(1, p)$ is sensitive to heavy-tailed data in the sense that, under the ultra-high dimensional scheme, even the law of large numbers for the maximum spurious correlation requires exponentially light tails of the underlying distribution. We refer readers to Theorem 2.1 in Shao and Zhou (2014) for details. Therefore, we believe that the exponential-type moment assumptions required in Theorem 3.1 cannot be weakened to polynomial-type ones as long as $\log p$ is allowed to be as large as n^c for some $c \in (0, 1)$. However, it is worth mentioning that the factor $1/7$ in Proposition 3.1 may not be optimal, and according to the results in Shao and Zhou (2014), $1/3$ is the best possible factor to ensure that the asymptotic theory is valid. To close this gap in theory, a significant amount of additional work and new probabilistic techniques are needed. We do not pursue this line of research in this paper.

For a general $s \ge 2$, we establish in the following proposition the limiting distribution of the sum of the top s order statistics of i.i.d. chi-square random variables with degree of freedom 1.

PROPOSITION 3.2. *Assume that $s \ge 2$ is a fixed integer. For any $t \in \mathbb{R}$, we have as $p \to \infty$,*

$$\begin{aligned} &\mathbb{P}\{Z^2_{(p)} + \cdots + Z^2_{(p-s+1)} - s a_p \le t\} \\ &\quad \longrightarrow \frac{\pi^{(1-s)/2}}{(s-1)!\Gamma(s-1)} \int_{-\infty}^{t/s} \bigg\{\int_0^{(t-sv)/2} u^{s-2} e^{-u}\, du\bigg\} \\ &\quad\quad \times e^{-(s-1)v/2} g(v)\, dv, \end{aligned} \tag{3.7}$$

where $a_p = 2\log p - \log(\log p)$, $G(t) = \exp(-\pi^{-1/2}e^{-t/2})$ *and* $g(t) = G'(t) = \frac{e^{-t/2}}{2\sqrt{\pi}}G(t)$. *The above integral can further be expressed as*

$$\begin{aligned} & G(t/s) + \frac{\pi^{1-s/2}e^{-t/2}}{(s-1)!}\int_{-\infty}^{t/s} e^u g(u)\,du + \frac{\pi^{(1-s)/2}e^{-t/2}}{(s-1)!} \\ & \times \sum_{j=1}^{s-2}\Biggl\{G(t/s)e^{(j+1)t/(2s)}\pi^{j/2}\prod_{\ell=1}^{j}(s-\ell) \\ & - \frac{1}{j!2^j}\int_{-\infty}^{t/s}(t-sv)^j e^{v/2}g(v)\,dv\Biggr\}. \end{aligned} \tag{3.8}$$

In particular, when $s = 2$, *the last term on the right-hand side of* (3.8) *vanishes so that, as* $p \to \infty$,

$$\mathbb{P}\{Z_{(p)}^2 + Z_{(p-1)}^2 - 2a_p \le t\} \to G(t/2) + \frac{e^{-t/2}}{2\sqrt{\pi}}\int_{-\infty}^{t/2} e^{u/2}G(u)\,du.$$

The proofs of Propositions 3.1 and 3.2 are placed in the Supplementary Material [Fan, Shao and Zhou (2018)].

3.2. *Multiplier bootstrap approximation.* The distribution of $R^*(s,p) = \sup_{f\in\mathcal{F}}\mathbb{G}^* f$ for $\mathbb{G}^*$ in (3.4) depends on the unknown $\boldsymbol{\Sigma}$ and thus cannot be used for statistical inference. In the following, we consider the use of a Monte Carlo method to simulate a process that mimics $\mathbb{G}^*$, now known as the multiplier (wild) bootstrap method, which is similar to that used in Hansen (1996), Barrett and Donald (2003) and Chernozhukov, Chetverikov and Kato (2013), among others.

Let $\widehat{\boldsymbol{\Sigma}}_n$ be the sample covariance matrix based on the data $\{\mathbf{X}_i\}_{i=1}^n$ and $\xi_1,\ldots,\xi_n$ be i.i.d. standard normal random variables that are independent of $\{\varepsilon_i\}_{i=1}^n$ and $\{\mathbf{X}_i\}_{i=1}^n$. Then, given $\{\mathbf{X}_i\}_{i=1}^n$,

$$\mathbf{Z}_n = n^{-1/2}\sum_{i=1}^n \xi_i(\mathbf{X}_i - \bar{\mathbf{X}}_n) \sim N(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_n). \tag{3.9}$$

The following result shows that the (unknown) distribution of $R^*(s,p) = \sup_{f_{\boldsymbol{\alpha}}\in\mathcal{F}}\frac{f_{\boldsymbol{\alpha}}(\mathbf{Z})}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}}}$ for $\mathbf{Z} \overset{d}{=} N(\mathbf{0},\boldsymbol{\Sigma})$ can be consistently estimated by the conditional distribution of

$$R_n^{\mathrm{MB}}(s,p) := \sup_{f_{\boldsymbol{\alpha}}\in\mathcal{F}}\frac{f_{\boldsymbol{\alpha}}(\mathbf{Z}_n)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\alpha}}}. \tag{3.10}$$

THEOREM 3.2. *Let Conditions* 2.1 *and* 2.2 *hold. Assume that the triplet* (s,p,n) *satisfies* $1 \le s \le p$ *and* $s\log(\gamma_s pn) = o(n^{1/5})$. *Then as* $n \to \infty$,

$$\sup_{t\ge 0}\bigl|\mathbb{P}\{R^*(s,p)\le t\} - \mathbb{P}\{R_n^{\mathrm{MB}}(s,p)\le t|\mathbf{X}_1,\ldots,\mathbf{X}_n\}\bigr| \xrightarrow{\mathbb{P}} 0. \tag{3.11}$$

REMARK 3.4. Together, Theorems 3.1 and 3.2 show that the maximum spurious correlation $\widehat{R}_n(s,p)$ can be approximated in distribution by the multiplier bootstrap statistic $n^{-1/2}R_n^{\mathrm{MB}}(s,p)$. In practice, when the sample size n is relatively small, the value of $n^{-1/2}R_n^{\mathrm{MB}}(s,p)$ may exceed 1, which makes it less favorable as a proxy for spurious correlation. To address this issue, we propose using the following corrected bootstrap approximation:

$$R_n^{\mathrm{CMB}}(s,p) := \sup_{f_{\boldsymbol{\alpha}} \in \mathcal{F}} \frac{\sqrt{n}}{|\boldsymbol{\xi}|_2} \frac{f_{\boldsymbol{\alpha}}(\mathbf{Z}_n)}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}} \widehat{\boldsymbol{\Sigma}}_n \boldsymbol{\alpha}}}, \tag{3.12}$$

where $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)^{\mathrm{T}}$ is used in the definition of $\mathbf{Z}_n$. By the Cauchy–Schwarz inequality, $n^{-1/2}R_n^{\mathrm{CMB}}(s,p)$ is always between 0 and 1. In view of (3.10) and (3.12), $R_n^{\mathrm{CMB}}(s,p)$ differs from $R_n^{\mathrm{MB}}(s,p)$ only up to a multiplicative random factor $n^{-1/2}|\boldsymbol{\xi}|_2$, which in theory is concentrated around 1 with exponentially high probability. Thus, R_n^{MB} and R_n^{CMB} are asymptotically equivalent, and (3.11) remains valid with R_n^{MB} replaced by R_n^{CMB}.

4. Extension to sparse linear models.

Suppose that the observed response Y and p-dimensional covariate $\mathbf{X}$ follows the sparse linear model

$$Y = \mathbf{X}^{\mathrm{T}} \boldsymbol{\beta}^* + \varepsilon, \tag{4.1}$$

where the regression coefficient $\boldsymbol{\beta}^*$ is sparse. The sparsity is typically explored by the LASSO [Tibshirani (1996)], the SCAD [Fan and Li (2001)], or the MCP [Zhang (2010)]. Now it is well known that, under suitable conditions, the SCAD and the MCP, among other folded concave penalized least-square estimators, also enjoy the unbiasedness property and the (strong) oracle properties. For simplicity, we focus on the SCAD. For a given random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, the SCAD exploits the sparsity by p_λ-regularization, which minimizes

$$(2n)^{-1} \sum_{i=1}^n (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})^2 + \sum_{j=1}^p p_\lambda(|\beta_j|; a) \tag{4.2}$$

over $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}} \in \mathbb{R}^p$, where $p_\lambda(\cdot; a)$ denotes the SCAD penalty function [Fan and Li (2001)], that is, $p_\lambda'(t;a) = \lambda I(t \le \lambda) + \frac{(a\lambda - t)_+}{a-1} I(t > \lambda)$ for some $a > 2$, and $\lambda = \lambda_n \ge 0$ is a regularization parameter.

Denote by $\mathbb{X} = (\mathbf{X}_1, \dots, \mathbf{X}_n)^{\mathrm{T}}$ the $n \times p$ design matrix, $\mathbb{Y} = (Y_1, \dots, Y_n)^{\mathrm{T}}$ the n-dimensional response vector, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$, the n-dimensional noise vector. Without loss of generality, we assume that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$ with each component of $\boldsymbol{\beta}_1 \in \mathbb{R}^s$ being nonzero and $\boldsymbol{\beta}_2 = \mathbf{0}$, such that $S_0 := \mathrm{supp}(\boldsymbol{\beta}^*) = \{1, \dots, s\}$ is the true underlying sparse model of the indices with $s = |\boldsymbol{\beta}^*|_0$. Moreover, write $\mathbb{X} = (\mathbb{X}_1, \mathbb{X}_2)$, where $\mathbb{X}_1 \in \mathbb{R}^{n \times s}$ consists of the columns of $\mathbb{X}$ indexed by S_0. In this notation, $\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbb{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and the oracle estimator $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ has an explicit form of

$$\widehat{\boldsymbol{\beta}}_1^{\text{oracle}} = (\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1)^{-1}\mathbb{X}_1^{\mathrm{T}}\mathbb{Y} = \boldsymbol{\beta}_1 + (\mathbb{X}_1^{\mathrm{T}}\mathbb{X}_1)^{-1}\mathbb{X}_1^{\mathrm{T}}\boldsymbol{\varepsilon}, \qquad \widehat{\boldsymbol{\beta}}_2^{\text{oracle}} = \mathbf{0}. \tag{4.3}$$

In other words, the oracle estimator is the unpenalized estimator that minimizes $\sum_{i=1}^n (Y_i - \mathbf{X}_{i,S_0}^{\mathrm{T}}\boldsymbol{\beta}_{S_0})^2$ over the true support set S_0.

Denote by $\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}} = (\widehat{\varepsilon}_1^{\text{oracle}}, \ldots, \widehat{\varepsilon}_n^{\text{oracle}})^{\mathrm{T}} = \mathbf{Y} - \mathbb{X}^{\mathrm{T}}\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ the residuals after the oracle fit. Then we can construct the maximum spurious correlation as in (2.2), except that $\{\varepsilon_i\}_{i=1}^n$ is now replaced by $\{\widehat{\varepsilon}_i^{\text{oracle}}\}_{i=1}^n$, that is,

$$\begin{aligned} &\widehat{R}_n^{\text{oracle}}(1, p) \\ &= \max_{j\in[p]} \frac{|\sum_{i=1}^n (\widehat{\varepsilon}_i^{\text{oracle}} - \mathbf{e}_n^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}})(X_{ij} - \bar{X}_j)|}{\sqrt{\sum_{i=1}^n (\widehat{\varepsilon}_i^{\text{oracle}} - \mathbf{e}_n^{\mathrm{T}}\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}})^2} \cdot \sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}, \end{aligned} \tag{4.4}$$

where $\mathbf{e}_n = (1/n, \ldots, 1/n)^{\mathrm{T}} \in \mathbb{R}^n$ and $\bar{X}_j = n^{-1}\sum_{i=1}^n X_{ij}$. We here deal with the specific case of a spurious correlation of size 1, as this is what is needed for testing the exogeneity assumption (1.2).

To establish the limiting distribution of $\widehat{R}_n^{\text{oracle}}(1, p)$, we make the following assumptions.

CONDITION 4.1. $\mathbb{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ *with* $\operatorname{supp}(\boldsymbol{\beta}^*) = \{1, \ldots, s\}$ *and* $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ *being i.i.d. centered sub-Gaussian satisfying that* $K_0 = \|\varepsilon_i\|_{\psi_2} < \infty$. *The rows of* $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^{\mathrm{T}}$ *are i.i.d. sub-Gaussian random vectors as in Condition* 2.1.

As before, we can assume that $\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{X}_i\mathbf{X}_i^{\mathrm{T}})$ is a correlation matrix with $\operatorname{diag}(\boldsymbol{\Sigma}) = \mathbf{I}_p$. Set $d = p - s$ and partition

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \qquad \text{with } \boldsymbol{\Sigma}_{11} \in \mathbb{R}^{s\times s}, \boldsymbol{\Sigma}_{22} \in \mathbb{R}^{d\times d}, \boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma}_{12}^{\mathrm{T}}. \tag{4.5}$$

Let $\boldsymbol{\Sigma}_{22.1} = (\widetilde{\sigma}_{jk})_{1\le j,k\le d} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ be the Schur complement of $\boldsymbol{\Sigma}_{11}$ in $\boldsymbol{\Sigma}$.

CONDITION 4.2. $\widetilde{\sigma}_{\min} = \min_{1\le j\le d} \widetilde{\sigma}_{jj}$ *is bounded away from zero.*

THEOREM 4.1. *Assume that Conditions* 4.1 *and* 4.2 *hold, and that the triplet* (s, p, n) *satisfies* $s\log p = o(\sqrt{n})$ *and* $\log p = o(n^{1/7})$. *Then the maximum spurious correlation* $\widehat{R}_n^{\text{oracle}}(1, p)$ *in* (4.4) *satisfies that, as* $n \to \infty$,

$$\sup_{t\ge 0} \bigl|\mathbb{P}\{\sqrt{n}\widehat{R}_n^{\text{oracle}}(1, p) \le t\} - \mathbb{P}(|\widetilde{\mathbf{Z}}|_\infty \le t)\bigr| \to 0, \tag{4.6}$$

where $\widetilde{\mathbf{Z}} \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Sigma}_{22.1})$ *is a* d*-variate centered Gaussian random vector with covariance matrix* $\boldsymbol{\Sigma}_{22.1}$.

As p_λ is a folded-concave penalty function, (4.2) is a nonconvex optimization problem. The local linear approximation (LLA) algorithm can be applied to produce a certain local minimum for any fixed initial solution [Zou and Li (2008), Fan, Xue and Zou (2014)]. In particular, Fan, Xue and Zou (2014) prove that the LLA algorithm can deliver the oracle estimator in the folded concave penalized problem with overwhelming probability if it is initialized by some appropriate initial estimator.

Let $\widehat{\boldsymbol{\beta}}^{\mathrm{LLA}}$ be the estimator computed via the one-step LLA algorithm initiated by the LASSO estimator [Tibshirani (1996)]. That is,

$$\widehat{\boldsymbol{\beta}}^{\mathrm{LLA}} = \arg\min_{\boldsymbol{\beta}} \left\{ (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda' (|\widehat{\beta}_j^{\mathrm{LASSO}}|) |\beta_j| \right\}, \tag{4.7}$$

where p_λ is a folded concave penalty, such as the SCAD and MCP penalties, and $\widehat{\boldsymbol{\beta}}^{\mathrm{LASSO}} = \arg\min_{\boldsymbol{\beta}} \{(2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^{\mathrm{T}} \boldsymbol{\beta})^2 + \lambda |\boldsymbol{\beta}|_1\}$. Accordingly, denote by $\widehat{R}_n^{\mathrm{LLA}}(1, p)$ the maximum spurious correlation as in (4.4) with $\widehat{\varepsilon}_i^{\mathrm{oracle}}$ replaced by $\widehat{\varepsilon}_i^{\mathrm{LLA}} = Y_i - \mathbf{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{\mathrm{LLA}}$. Applying Theorem 4.1, we derive the limiting distribution of $\widehat{R}_n^{\mathrm{LLA}}(1, p)$ under suitable conditions. First, let us recall the *Restricted Eigenvalue* concept formulated by Bickel, Ritov and Tsybakov (2009).

DEFINITION 4.1. For any integer $s_0 \in [p]$ and positive number c_0, the RE(s_0, c_0) parameter $\kappa(s_0, c_0, \mathbf{A})$ of a $p \times p$ matrix $\mathbf{A}$ is defined as

$$\kappa(s_0, c_0, \mathbf{A}) := \min_{S \subseteq [p]: |S| \le s_0} \min_{\boldsymbol{\delta} \neq 0: |\boldsymbol{\delta}_{S^c}|_1 \le c_0 |\boldsymbol{\delta}_S|_1} \frac{\boldsymbol{\delta}^{\mathrm{T}} \mathbf{A} \boldsymbol{\delta}}{|\boldsymbol{\delta}_S|_2^2}. \tag{4.8}$$

THEOREM 4.2. *Assume that Conditions 4.1 and 4.2 hold, the minimal signal strength of $\boldsymbol{\beta}^*$ satisfies $\min_{j \in S_0} |\beta_j| > (a+1)\lambda$ for a, λ as in (4.2), and that the triplet (s, p, n) satisfies $s \log p = o(\sqrt{n})$, $\frac{s \log p}{\kappa(s, 3+\epsilon, \boldsymbol{\Sigma})} = o(n)$ for some $\epsilon > 0$ and $\log p = o(n^{1/7})$. If the regularization parameters $(\lambda, \lambda_{\mathrm{LASSO}})$ are such that $\lambda \ge \frac{8\sqrt{s}}{\kappa(s,3,\boldsymbol{\Sigma})} \lambda_{\mathrm{LASSO}}$ and $\lambda_{\mathrm{LASSO}} \ge C K_0 \sqrt{(\log p)/n}$ for $C > 0$ large enough, then as $n \to \infty$,*

$$\sup_{t \ge 0} \left| \mathbb{P}\{\sqrt{n} \widehat{R}_n^{\mathrm{LLA}}(1, p) \le t\} - \mathbb{P}(|\widetilde{\mathbf{Z}}|_\infty \le t) \right| \to 0, \tag{4.9}$$

where $\widetilde{\mathbf{Z}} \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Sigma}_{22.1})$.

5. Applications to high-dimensional inferences. This section outlines three applications in high-dimensional statistics. The first determines whether discoveries by machine learning and data mining techniques are any better than those reached by chance. Second, we show that the distributions of maximum spurious correlations can also be applied to model selection. In the third application, we validate the fundamental assumption of exogeneity (1.2) in high dimensions.

5.1. *Spurious discoveries.* Let $q_\alpha^{\mathrm{CMB}}(s,p)$ be the upper α-quantile of the random variable $R_n^{\mathrm{CMB}}(s,p)$ defined by (3.12). Then an approximate $1-\alpha$ upper confidence limit of the spurious correlation is given by $q_\alpha^{\mathrm{CMB}}(s,p)$. In view of Theorems 3.1 and 3.2, we claim that

$$\mathbb{P}\{\widehat{R}_n(s,p) \le q_\alpha^{\mathrm{CMB}}(s,p)\} \to 1-\alpha. \tag{5.1}$$

To see this, recall that $R_n^{\mathrm{CMB}} = \sqrt{n} R_n^{\mathrm{MB}}/|\boldsymbol{\xi}|_2$ for $\boldsymbol{\xi} = (\xi_1,\ldots,\xi_n)^{\mathrm{T}}$ as in (3.12), and given $\{\mathbf{X}_i\}_{i=1}^n$, R_n^{MB} is the supremum of a Gaussian process. Let $F_n^{\mathrm{MB}}(t) = \mathbb{P}\{R_n^{\mathrm{MB}}(s,p) \le t|\mathbf{X}_1,\ldots,\mathbf{X}_n\}$ be the (conditional) distribution function of R_n^{MB} and define $t_0 = \inf\{t : F_n^{\mathrm{MB}}(t) > 0\}$. By Theorem 11.1 of Davydov, Lifshits and Smorodina (1998), F_n^{MB} is absolutely continuous with respect to the Lebesgue measure and is strictly increasing on (t_0,∞), indicating that $\mathbb{P}\{R_n^{\mathrm{CMB}} \le q_\alpha^{\mathrm{CMB}}(s, p)|\mathbf{X}_1,\ldots,\mathbf{X}_n\} = \alpha$ almost surely. This, together with (3.3) and (3.11), proves (5.1) under Conditions 2.1, 2.2 and when $s\log(\gamma_s pn) = o(n^{1/7})$.

Let $\widehat{Y}_i$ be fitted values using s predictors indexed by $\widehat{S}$ selected by a data-driven technique and Y_i be the associated response value. They are denoted in the vector form by $\widehat{\mathbb{Y}}$ and $\mathbb{Y}$, respectively. If

$$|\widehat{\mathrm{corr}}_n(\mathbb{Y},\widehat{\mathbb{Y}})| \le q_\alpha^{\mathrm{CMB}}(s,p), \tag{5.2}$$

then the discovery of variables $\widehat{S}$ can be regarded as spurious, that is, no better than by chance. Therefore, the multiplier bootstrap quantile $q_\alpha^{\mathrm{CMB}}(s,p)$ provides an important critical value and yardstick for judging whether the discovery is spurious, or whether the selected set $\widehat{S}$ includes too many spurious variables. This yardstick is independent of the method used in the fitting.

REMARK 5.1. The problem of judging whether the discovery is spurious is intrinsically different from that of testing the global null hypothesis $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$, which itself is an important problem in high-dimensional statistical inference and has been well studied in the literature since the seminal work of Goeman, van de Geer and van Houwelingen (2006). For example, the global null hypothesis $H_0 : \boldsymbol{\beta}^* = \mathbf{0}$ can be rejected by a test; still, the correlation between Y and the variables $\widehat{S}$ selected by a statistical method can be smaller than the maximum spurious correlation, and we should interpret the findings of $\widehat{S}$ with caution. We need either more samples or more powerful variable selection methods. This motivates us to derive the distribution of the maximum spurious correlation $\widehat{R}_n(s,p)$. This distribution serves as an important benchmark for judging whether the discovery (of s features from p explanatory variables based on a sample of size n) is spurious. The magnitude of $\widehat{R}_n(s,p)$ gives statisticians an idea of how big a spurious correlation can be and, therefore, an idea of how much the covariates really contribute to the regression for a given sample size.

5.2. *Model selection.* In the previous section, we consider the reference distribution of the maximum spurious correlation statistic $\widehat{R}_n(s, p)$ as a benchmark for judging whether the discovery of s significant variables (among all of the p variables using a random sample of size n) is impressive, regardless of which variable selection tool is applied. In this section, we show how the distribution of $\widehat{R}_n(s, p)$ can be used to select a model. Intuitively, we would like to select a model that fits better than the spurious fit. This limits the candidate sets of models and provides an upper bound on the model size. In our experience, this upper bound itself provides a model selector.

We now use LASSO as an illustration of the above idea. Owing to spurious correlation, almost all of the variable selection procedures will, with high probability, select a number of spurious variables in the model so that the selected model is over-fitted. For example, the LASSO method with the regularization parameter selected by cross-validation typically selects a far larger model size, as the bias caused by the ℓ_1 penalty forces the cross-validation procedure to choose a smaller value of λ. Thus, it is important to stop the LASSO path earlier and the quantiles of $\widehat{R}_n(s, p)$ provide useful guards.

Specifically, consider the LASSO estimator $\widehat{\boldsymbol{\beta}}_\lambda$ for the sparse linear model (4.1) with $\widehat{s}_\lambda = |\operatorname{supp}(\widehat{\boldsymbol{\beta}}_\lambda)|$, where $\lambda > 0$ is the regularization parameter. We consider the LASSO solution path with the largest knot $\lambda_{\text{ini}} := |\mathbb{X}^{\mathrm{T}}\mathbb{Y}|_\infty$ and the smallest knot λ_{cv} selected by tenfold cross-validation. To avoid over-fitting, we propose using q_α^{CMB} as a guide to choose the regularization parameter that guards us from selecting too many spurious variables. For each λ in the path, we compute $\widehat{\operatorname{corr}}_n(\widehat{\mathbb{Y}}_\lambda, \mathbb{Y})$, the sample correlation between the post-LASSO fitted and observed responses, and $q_\alpha^{\text{CMB}}(\widehat{s}_\lambda, p)$. Let $\widehat{\lambda}_\alpha$ be the largest λ such that the sign of $\widehat{\operatorname{corr}}_n(\widehat{\mathbb{Y}}_\lambda, \mathbb{Y}) - q_\alpha^{\text{CMB}}(\widehat{s}_\lambda, p)$ is nonnegative and then flips in the subsequent knot. The selected model is given by $\widehat{S}_\alpha = \operatorname{supp}(\widehat{\boldsymbol{\beta}}_{\widehat{\lambda}_\alpha})$. As demonstrated by the simulation studies in Section 6.4, this procedure selects a much smaller model size that is closer to the real data.

5.3. *Validating exogeneity.* Fan and Liao (2014) show that the exogenous condition (1.2) is necessary for penalized least-squares to achieve a model selection consistency. They question the validity of such an exogeneous assumption, as it imposes too many equations. They argue further that even when the exogenous model holds for important variables $\mathbf{X}_S$, that is,

$$Y = \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S^* + \varepsilon, \qquad \mathbb{E}(\varepsilon \mathbf{X}_S) = \mathbf{0}, \tag{5.3}$$

the extra variables $\mathbf{X}_N$ (with $N = S^{\mathrm{c}}$) are collected in an effort to cover the unknown set S—but no verification of the conditions

$$\mathbb{E}(\varepsilon \mathbf{X}_N) = \mathbb{E}\{(Y - \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S^*)\mathbf{X}_N\} = \mathbf{0} \tag{5.4}$$

has ever been made. The equality $\mathbb{E}\{(Y - \mathbf{X}_S^{\mathrm{T}}\boldsymbol{\beta}_S^*)X_j\} = 0$ in (5.4) holds by luck for some covariate X_j, but it cannot be expected that this holds for all $j \in N$. They propose a focussed generalized method of moment (FGMM) to avoid the unreasonable assumption (5.4). Recognizing (5.3) is not identifiable in high-dimensional linear models, they impose additional conditions such as $\mathbb{E}(\varepsilon \mathbf{X}_S^2) = \mathbf{0}$.

Despite its fundamental importance to high-dimensional statistics, there are no available tools for validating (1.2). Regarding (1.2) as a null hypothesis, an asymptotically α-level test can be used to reject assumption (1.2) when

$$\widehat{T}_{n,p} = \max_{j \in [p]} \bigl| \sqrt{n} \widehat{\mathrm{corr}}_n (X_j, \varepsilon) \bigr| \geq q_\alpha^{\mathrm{CMB}}(1, p). \tag{5.5}$$

By Theorems 3.1 and 3.2, the test statistic has an approximate size α. The p-value of the test can be computed via the distribution of the Gaussian multiplier process $R_n^{\mathrm{CMB}}(1, p)$.

As pointed out in the Introduction, when the components of $\mathbf{X}$ are weakly correlated, the distribution of the maximum spurious correlation does not depend very sensitively on $\mathbf{\Sigma}$; see also Lemma 6 in Cai, Liu and Xia (2014). In this case, we can approximate it by the identity matrix, and hence one can compare the renormalized test statistic

$$J_{n,p} = \widehat{T}_{n,p}^2 - 2 \log p + \log(\log p) \tag{5.6}$$

with the limiting distribution in (3.6). The critical value for test statistic $J_{n,p}$ is

$$J_\alpha = -2 \log\bigl\{ -\sqrt{\pi} \log(1 - \alpha) \bigr\}, \tag{5.7}$$

and the associated p-value is given by

$$\exp\bigl(-\pi^{-1/2} e^{-J_{n,p}/2} \bigr). \tag{5.8}$$

Expressions (5.7) and (5.8) provide analytic forms for a quick validation of the exogenous assumption (1.2) under weak dependence. In general, we recommend using the wild bootstrap, which takes into account the correlation effect and provides more accurate estimates especially when the dependence is strong; see Chang et al. (2017) for more empirical evidences.

In practice, ε is typically unknown to us. Therefore, $\widehat{T}_{n,p}$ in (5.5) is calculated using the fitted residuals $\{\widehat{\varepsilon}_i^{\mathrm{LLA}}\}_{i=1}^n$. In view of Theorem 4.2, we need to adjust the null distribution according to (4.9). By Theorem 3.2, we adjust the definition of the process $\mathbf{Z}_n$ in (3.9) by

$$\mathbf{Z}_n^{\mathrm{LLA}} = n^{-1/2} \sum_{i=1}^n \xi_i \bigl(\mathbf{X}_i^{\mathrm{LLA}} - \overline{\mathbf{X}}_n^{\mathrm{LLA}} \bigr) \in \mathbb{R}^{p - |\widehat{S}|}, \tag{5.9}$$

where $\mathbf{X}_i^{\mathrm{LLA}} = \mathbf{X}_{i,\widehat{N}} - \widehat{\mathbf{\Sigma}}_{\widehat{N}\widehat{S}} \widehat{\mathbf{\Sigma}}_{\widehat{S}\widehat{S}}^{-1} \mathbf{X}_{i,\widehat{S}}$ is the residuals of $\mathbf{X}_{\widehat{N}}$ regressed on $\mathbf{X}_{\widehat{S}}$, where $\widehat{S}$ is the set of selected variables, $\widehat{N} = [p] \setminus \widehat{S}$, and $\widehat{\mathbf{\Sigma}}_{SS'}$ denotes the sub-matrix of $\widehat{\mathbf{\Sigma}}_n$ containing entries indexed by $(k, \ell) \in S \times S'$. From (5.9), the multiplier bootstrap approximation of $|\widetilde{\mathbf{Z}}|_\infty$ is $R_n^{\mathrm{MB,LLA}}(1, p) = |\widehat{\mathbf{D}}^{-1/2} \mathbf{Z}_n^{\mathrm{LLA}}|_\infty$, where $\widehat{\mathbf{D}} =$ diagonal matrix of the sample covariance matrix of $\{\mathbf{X}_i^{\mathrm{LLA}}\}_{i=1}^n$. Consequently, we reject (1.2) if $\widehat{T}_{n,p} > q_\alpha^{\mathrm{MB,LLA}}(1, p)$, where $q_\alpha^{\mathrm{MB,LLA}}(1, p)$ is the (conditional) upper α-quantile of $R_n^{\mathrm{MB,LLA}}(1, p)$ given $\{\mathbf{X}_i\}_{i=1}^n$.

REMARK 5.2. To the best of our knowledge, this is the first paper to consider testing the exogenous assumption (1.2), for which we use the maximum correlation between covariates and fitted residuals as the test statistic. A referee kindly informed us in his/her review report that in the context of specification testing, Chernozhukov, Chetverikov and Kato (2013) propose a similar extreme value statistic and use the multiplier bootstrap to compute a critical value for the test. To construct marginal test statistics, they use self-normalized covariances between generated regressors and fitted residuals obtained via ordinary least squares, whereas we use sample correlations between the covariates and fitted residuals obtained by the LLA algorithm. We refer readers to Appendix M in the Supplementary Material of Chernozhukov, Chetverikov and Kato (2013) for more details.

6. Numerical studies.

In this section, Monte Carlo simulations are used to examine the finite-sample performance of the bootstrap approximation (for a given data set) of the distribution of the maximum spurious correlation (MSC).

6.1. *Computation of spurious correlation.* First, we observe that $\widehat{R}_n(s,p)$ in (2.2) can be written as $\widehat{R}_n^2(s,p) = \widehat{\sigma}_\varepsilon^{-2} \max_{S\subseteq[p]:|S|=s} \mathbf{v}_{n,S}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{SS}^{-1} \mathbf{v}_{n,S}$, where $\widehat{\sigma}_\varepsilon^2 = n^{-1}\sum_{i=1}^n(\varepsilon_i - \bar{\varepsilon}_n)^2$ and $\mathbf{v}_n = n^{-1}\sum_{i=1}^n(\varepsilon_i - \bar{\varepsilon}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)$. Therefore, the computation of $\widehat{R}_n(s,p)$ requires solving the combinatorial optimization problem

$$\widehat{S} = \arg\max_{S\subseteq[p]:|S|=s} \mathbf{v}_{n,S}^{\mathrm{T}} \widehat{\mathbf{\Sigma}}_{SS}^{-1} \mathbf{v}_{n,S}. \tag{6.1}$$

It is computationally intensive to obtain $\widehat{S}$ for large values of p and s as one essentially needs to enumerate all $\binom{p}{s}$ possible subsets of size s from p covariates. A fast and easily implementable approach is to use the stepwise addition (forward selection) algorithm as in Fan, Guo and Hao (2012), which results in some value that is no larger than $\widehat{R}_n(s,p)$ but avoids computing all $\binom{p}{s}$ multiple correlations in (6.1). Note that the optimization (6.1) is equivalent to finding the best subset regression of size s. When p is relatively small, say if p ranges from 20 to 40, the branch-and-bound procedure is commonly used for finding the best subset of a given size that maximizes multiple R^2 [Brusco and Stahl (2005)]. However, this approach becomes computationally infeasible very quickly when there are hundreds or thousands of potential predictors. As a trade-off between approximation accuracy and computational intensity, we propose using a two-step procedure that combines the stepwise addition and branch-and-bound algorithms. First, we use the forward selection to pick the best d variables, say $d = 40$, which serves as a prescreening step. Second, across the $\binom{d}{s}$ subsets of size s, the branch-and-bound procedure is implemented to select the best subset that maximizes the multiple-R^2. This subset is used as an approximate solution to (6.1). Note that when $s > 40$, which is rare in many applications, we only use the stepwise addition to reduce the computational cost.

TABLE 1
(*Isotropic case*) *The mean of* 200 *empirical sizes* $c^{\mathrm{MB}}(\cdot,\alpha)\times 100$, *with its estimate of SD in the parenthesis, when* $p=2000$, $s=1,2,5,10$, $n=400,800,1200$ *and* $\alpha=0.1,0.05$

	$s=1$		$s=2$		$s=5$		$s=10$	
n	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$
400	9.54 (0.643)	4.68 (0.294)	9.13 (0.568)	4.38 (0.284)	9.08 (0.480)	3.78 (0.245)	8.67 (0.506)	4.44 (0.291)
800	9.43 (0.444)	4.93 (0.296)	9.47 (0.474)	4.42 (0.296)	9.73 (0.488)	4.73 (0.294)	9.94 (0.557)	5.62 (0.331)
1200	9.09 (0.507)	4.32 (0.261)	9.00 (0.542)	4.46 (0.278)	9.42 (0.543)	4.87 (0.322)	9.97 (0.579)	5.15 (0.318)

6.2. *Accuracy of the multiplier bootstrap approximation.* For the first simulation, we consider the case where the random noise ε follows the uniform distribution standardized so that $\mathbb{E}(\varepsilon)=0$ and $\mathbb{E}(\varepsilon^2)=1$. Independent of ε, the p-variate vector $\mathbf{X}$ of covariates has i.i.d. $N(0,1)$ components. In the results reported in Table 1, the ambient dimension $p=2000$, the sample size n takes a value in $\{400,800,1200\}$, and s takes a value in $\{1,2,5,10\}$. For a given significance level $\alpha\in(0,1)$, let $q_\alpha(s,p)$ be the upper α-quantile of $\widehat{R}_n(s,p)$ in (2.1). For each data set $\mathcal{X}_n=\{\mathbf{X}_1,\dots,\mathbf{X}_n\}$, a direct application of Theorems 3.1 and 3.2 is that

$$c^{\mathrm{MB}}(\mathcal{X}_n,\alpha):=\mathbb{P}\{R_n^{\mathrm{CMB}}(s,p)\ge q_\alpha(s,p)|\mathcal{X}_n\}\to\alpha \qquad \text{as } n\to\infty.$$

The difference $c^{\mathrm{MB}}(\mathcal{X}_n,\alpha)-\alpha$, however, characterizes the extent of the size distortions and the finite-sample accuracy of the multiplier bootstrap approximation (MBA). Table 1 summarizes the mean and the standard deviation (SD) of $c^{\mathrm{MB}}(\mathcal{X}_n,\alpha)$ based on 200 simulated data sets with $\alpha\in\{0.05,0.1\}$. The α-quantile $q_\alpha(s,p)$ is calculated from 1600 replications, and $c^{\mathrm{MB}}(\mathcal{X}_n,\alpha)$ for each data set is simulated based on 1600 bootstrap replications. In addition, we report in Figure 3 the distributions of the maximum spurious correlations and their multiplier bootstrap approximations conditional on a given data set $\mathcal{X}_n$ when $p\in\{2000,5000\}$, $s\in\{1,2,5,10\}$ and $n=400$. Together, Table 1 and Figure 3 show that the multiplier bootstrap method indeed provides a quite good approximation to the (unknown) distribution of the maximum spurious correlation.

For the second simulation, we focus on an anisotropic case where the covariance matrix $\mathbf{\Sigma}$ of $\mathbf{X}$ is nonidentity, and the condition number of $\mathbf{\Sigma}$ is well controlled. Specifically, we assume that ε follows the centered Laplace distribution rescaled so that $\mathbb{E}(\varepsilon)=0$ and $\mathbb{E}(\varepsilon^2)=1$. To introduce dependence among covariates, first we denote with $\mathbf{A}$ a 10×10 symmetric positive definite matrix with a prespecified condition number $c>1$ and let $\rho\in(0,1)$. Then the p-dimensional vector $\mathbf{X}$ of the covariates is generated according to $\mathbf{X}=\mathbf{G}_1(\rho)\mathbf{Z}_1+\mathbf{G}_2(\rho)\mathbf{Z}_2$, where $\mathbf{Z}_1\overset{d}{=}$

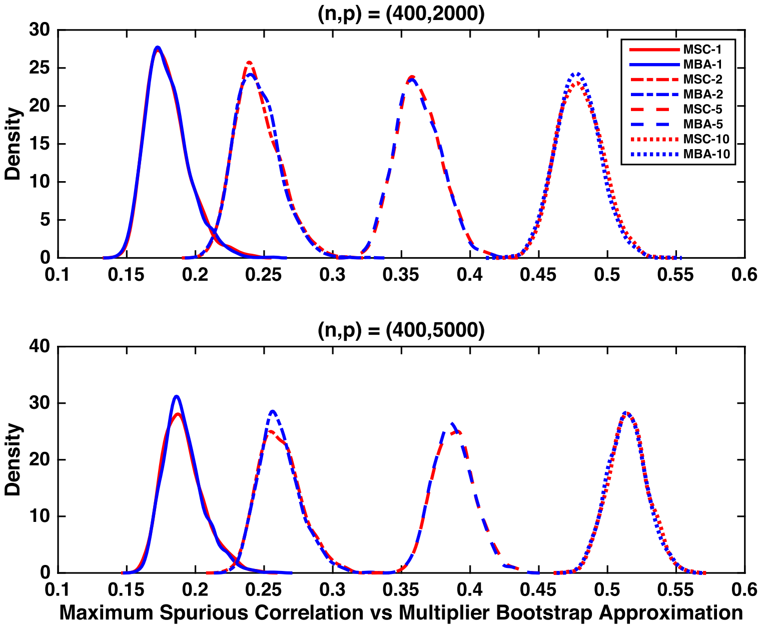

FIG. 3. *Distributions of maximum spurious correlations* (*blue*) *and multiplier bootstrap approximations* (*for a given data set*; *red*) *based on* 1600 *simulations with combinations of* $p = 2000, 5000$, $s = 1, 2, 5, 10$ *and* $n = 400$ *when* $\mathbf{\Sigma}$ *is an identity matrix.*

$N(\mathbf{0}, \mathbf{A})$, $\mathbf{Z}_2 \overset{d}{=} N(\mathbf{0}, \mathbf{I}_{p-10})$ and $\mathbf{G}_1(\rho) \in \mathbb{R}^{p\times 10}$, $\mathbf{G}_2(\rho) \in \mathbb{R}^{p\times(p-10)}$ are given respectively by $\mathbf{G}_1(\rho)^{\mathrm{T}} = (\mathbf{I}_{10}, \frac{\rho}{\sqrt{1+\rho^2}}\mathbf{I}_{10}, \mathbf{G}_{11}(\rho)^{\mathrm{T}})$ with

$$\mathbf{G}_{11}(\rho) = \frac{1-\rho}{\sqrt{1+(1-\rho)^2}} \begin{pmatrix} 1 & 0 & \dots & 0 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix} \in \mathbb{R}^{(p-20)\times 10}$$

and

$$\mathbf{G}_2(\rho) = \begin{pmatrix} \mathbf{0}_{10\times 10} & \mathbf{0}_{10\times(p-20)} \\ \frac{1}{\sqrt{1+\rho^2}}\mathbf{I}_{10} & \mathbf{0}_{10\times(p-20)} \\ \mathbf{0}_{(p-20)\times 10} & \frac{1}{\sqrt{1+(1-\rho)^2}}\mathbf{I}_{p-20} \end{pmatrix}.$$

In particular, we take $c = 5$ and $\rho = 0.8$ in the simulations reported in Table 2, which summarizes the mean and the standard deviation (SD) of the size $c^{\mathrm{MB}}(\mathcal{X}_n, \alpha)$ based on 200 simulated data sets with $\alpha \in \{0.05, 0.1\}$. Comparing the

TABLE 2
(*Anisotropic case*) *Mean of* 200 *empirical sizes* $c^{\mathrm{MB}}(\cdot,\alpha)\times 100$, *with its estimate of SD in the parenthesis, when* $p=2000$, $s=1,2,5,10$, $n=400,800,1200$ *and* $\alpha=0.1,0.05$

	$s=1$		$s=2$		$s=5$		$s=10$	
n	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.1$	$\alpha=0.05$
400	9.83 (0.426)	4.39 (0.222)	9.04 (0.402)	4.75 (0.208)	9.27 (0.492)	4.65 (0.273)	9.34 (0.557)	4.53 (0.291)
800	10.18 (0.556)	5.19 (0.296)	10.48 (0.519)	5.12 (0.272)	9.98 (0.576)	4.86 (0.220)	9.21 (0.474)	4.73 (0.269)
1200	9.42 (0.500)	4.41 (0.233)	9.60 (0.543)	5.71 (0.339)	9.90 (0.553)	4.85 (0.333)	10.11 (0.606)	5.19 (0.337)

simulation results shown in Tables 1 and 2, we find that the bootstrap approximation is fairly robust against heterogeneity in the covariance structure of the covariates.

6.3. *Detecting spurious discoveries.* To examine how the multiplier bootstrap quantile $q_\alpha^{\mathrm{CMB}}(s,p)$ (see Section 5.1) serves as a benchmark for judging whether the discovery is spurious, we compute the Spurious Discovery Probability (SDP) by simulating 200 data sets from (4.1) with $n=100,120,160$, $p=400$, $\boldsymbol\beta^*=(1,0,-0.8,0,0.6,0,-0.4,0,\dots,0)^{\mathrm{T}}\in\mathbb{R}^p$, and standard Gaussian noise $\varepsilon\overset{d}{=}N(0,1)$. For some integer $s\le r\le p$, we let $\mathbf{x}\overset{d}{=}N(\mathbf{0},\mathbf{I}_r)$ be an r-dimensional Gaussian random vector. Let $\boldsymbol\Gamma_r$ be a $p\times r$ matrix satisfying $\boldsymbol\Gamma_r^{\mathrm{T}}\boldsymbol\Gamma_r=\mathbf{I}_r$. The rows of the design matrix $\mathbb{X}$ are sampled as i.i.d. copies from $\boldsymbol\Gamma_r\mathbf{x}\in\mathbb{R}^p$, where r takes a value in $\{120,160,200,240,280,320,360\}$. To save space, we give the numerical results for the case of non-Gaussian design and noise in the Supplementary Material [Fan, Shao and Zhou (2018)].

Put $\mathbb{Y}=(Y_1,\dots,Y_n)^{\mathrm{T}}$ and let $\widehat{\mathbb{Y}}=\mathbb{X}_{\widehat S}\widehat{\boldsymbol\beta}^{\mathrm{pLASSO}}$ be the n-dimensional vector of fitted values, where $\widehat{\boldsymbol\beta}^{\mathrm{pLASSO}}=(\mathbb{X}_{\widehat S}^{\mathrm{T}}\mathbb{X}_{\widehat S})^{-1}\mathbb{X}_{\widehat S}^{\mathrm{T}}\mathbb{Y}$ is the post-LASSO estimator using covariates selected by the tenfold cross-validated LASSO estimator. Let $\widehat s=|\widehat S|_0$ be the number of variables selected. For $\alpha\in(0,1)$, the level-α SDP is defined as $\mathbb{P}\{|\widehat{\mathrm{corr}}_n(\mathbb{Y},\widehat{\mathbb{Y}})|\le q_\alpha^{\mathrm{CMB}}(\widehat s,p)\}$. As the simulated model is not null, this SDP is indeed a type II error. Given $\alpha=5\%$ and for each simulated data set, $q_\alpha^{\mathrm{CMB}}(s,p)$ is computed based on 1000 bootstrap replications. Then we compute the empirical SDP based on 200 simulations. The results are given in Table 3.

In this study, the design matrix is chosen so that there is a low-dimensional linear dependency in the high-dimensional covariates. The collected covariates are highly correlated when r is much smaller than p. It is known that collinearity and high dimensionality add difficulty to the problem of variable selection and deteriorate the performance of the LASSO. The smaller the r is, the more severe

TABLE 3

Empirical α-level spurious discovery probability (ESDP) based on 200 *simulations when* $p = 400$, $n = 100, 120, 160$ *and* $\alpha = 5\%$

	$r = 120$	$r = 160$	$r = 200$	$r = 240$	$r = 280$	$r = 320$	$r = 360$
$n = 100$	0.6950	0.6650	0.6000	0.5200	0.5100	0.4500	0.4000
$n = 120$	0.6600	0.5350	0.3350	0.3800	0.2500	0.2850	0.1950
$n = 160$	0.1950	0.1300	0.0500	0.0400	0.0550	0.0700	0.0250

the problem of collinearity becomes. As reflected in Table 3, the empirical SDP increases as r decreases, indicating that the correlation between fitted and observed responses is more likely to be smaller than the spurious correlation.

6.4. *Model selection.* We demonstrate the idea in Section 5.2 through the following toy example. Consider the linear model (4.1) with $(n, p) = (160, 400)$ and $\boldsymbol{\beta}^* = (1, 0, -0.8, 0, 0.6, 0, -0.4, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$. The covariate vector is taken to be $\mathbf{X} = \boldsymbol{\Gamma}\mathbf{x}$ with $\mathbf{x} = (x_1, \ldots, x_{200})^{\mathrm{T}}$, where $x_1, \ldots, x_{200}$ are i.i.d. random variables following the continuous uniform distribution on $[-1, 1]$ and $\boldsymbol{\Gamma}$ is a 400×200 matrix satisfying $\boldsymbol{\Gamma}^{\mathrm{T}}\boldsymbol{\Gamma} = \mathbf{I}_{200}$. The noise variable ε follows a standardized t-distribution with 4 degrees of freedom. Moreover, let $S_0 = \{j : \beta_j^* \neq 0\}$ be the true model.

Applying tenfold cross-validated LASSO selects 35 variables. Along the solution path, we compute the number of correctly selected variables $|\widehat{S} \cap S_0|$, the fitted correlation and the upper 5%-quantile of the multiplier bootstrap approximation of the maximum spurious correlation based on 1000 bootstrap samples. The results are provided in Table 4, from which we see that the cross-validation procedure under the guidance of MSC selects 15 variables including all of the signal covariates.

6.5. *Gene expression data.* In this section, we extend the previous study in Section 6.3 to an analysis of a real life data set. To further address the question that for a given data set, whether the discoveries based on certain data-mining technique are any better than spurious correlation, we consider again the gene expression data from 90 individuals (45 Japanese and 45 Chinese, JPT-CHB) from the international "HapMap" project [Thorisson et al. (2005)] discussed in the Introduction.

The gene *CHRNA6* is thought to be related to the activation of dopamine-releasing neurons with nicotine and, therefore, has been the subject of many nicotine addiction studies [Thorgeirsson et al. (2010)]. We took the expressions of *CHRNA6* as the response Y and the remaining $p = 47{,}292$ expressions of probes as covariates $\mathbf{X}$. For a given $\lambda > 0$, LASSO selects $\widehat{s}_\lambda$ probes indexed by $\widehat{S}_\lambda$. In particular, using tenfold cross-validation to select the tuning parameter gives $\widehat{s}_{\lambda_0} = 25$ probes with $\lambda_0 = 0.0674$. Define fitted vectors $\widehat{\mathbb{Y}}_\lambda^{\mathrm{LASSO}} = \mathbb{X}\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{LASSO}}$ and

TABLE 4

Number of true positive results, the sample correlation between fitted and observed responses, and the upper 5%*-quantile of the multiplier bootstrap approximation based on* 1000 *bootstrap samples*

	$\|\widehat{S}_\lambda \cap S_0\|$	$\widehat{\mathrm{corr}}_n(\mathbb{Y}, \widehat{\mathbb{Y}}_\lambda^{\mathrm{pLASSO}})$	$q_{0.05}^{\mathrm{CMB}}(\widehat{s}_\lambda, p)$
$\lambda = 0.3410$ ($\widehat{s}_\lambda = 1$)	1	0.3314	0.3040
$\lambda = 0.2703$ ($\widehat{s}_\lambda = 2$)	2	0.4802	0.3870
$\lambda = 0.2580$ ($\widehat{s}_\lambda = 3$)	3	0.5255	0.4435
$\lambda = 0.2351$ ($\widehat{s}_\lambda = 4$)	3	0.5536	0.4907
$\lambda = 0.2142$ ($\widehat{s}_\lambda = 5$)	3	0.5791	0.5297
$\lambda = 0.2044$ ($\widehat{s}_\lambda = 6$)	3	0.5971	0.5608
$\lambda = 0.1952$ ($\widehat{s}_\lambda = 8$)	3	0.6205	0.6131
$\lambda = 0.1778$ ($\widehat{s}_\lambda = 9$)	3	0.6377	0.6365
$\lambda = 0.1697$ ($\widehat{s}_\lambda = 11$)	4	0.6953	0.6758
$\lambda = 0.1620$ ($\widehat{s}_\lambda = 14$)	4	0.7380	0.7208
$\lambda = 0.1409$ ($\widehat{s}_\lambda = 15$)	4	**0.7490**	**0.7346**
$\lambda = 0.1345$ ($\widehat{s}_\lambda = 19$)	4	0.7685	0.7799
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\lambda = 0.0885$ ($\widehat{s}_\lambda = 35$)	4	0.8428	0.8847

$\widehat{\mathbb{Y}}_\lambda^{\mathrm{pLASSO}} = \mathbb{X}_{\widehat{S}_\lambda}\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{pLASSO}}$, where $\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{LASSO}}$ is the LASSO estimator and $\widehat{\boldsymbol{\beta}}_\lambda^{\mathrm{pLASSO}}$ is the post-LASSO estimator, which is the least-square estimator based on the LASSO selected set.

We depict the observed correlations between the fitted value and the response as well as the median and upper α-quantile of the multiplier bootstrap approximation with $\alpha = 10\%$ based on 1000 bootstrap replications in Table 5. Even though $\widehat{\mathrm{corr}}_n(\mathbb{Y}, \widehat{\mathbb{Y}}^{\mathrm{LASSO}}) = 0.8991$ and $\widehat{\mathrm{corr}}_n(\mathbb{Y}, \widehat{\mathbb{Y}}^{\mathrm{pLASSO}}) = 0.9214$, the discoveries appear to be no better than chance. We therefore increase λ, which decreases the size of discovered probes. From Table 4, only the discovery of three probes is above chance results at $\alpha = 10\%$. The three probes are BBS1—Homo sapiens Bardet-Biedl syndrome 1, POLE2—Homo sapiens polymerase (DNA directed), epsilon 2 (p59 subunit) and TG737—Homo sapiens Probe hTg737 (polycystic kidney disease, autosomal recessive), transcript variant 2. Figure 4 shows the observed correlations of the fitted values and observed values compared to the reference null distribution.

We now use the test statistic (5.5) to test whether the null hypothesis (1.2) holds. We take $\lambda_0 = 0.0674$ and compute the observed test statistic $\widehat{T}_{n,p}^{\mathrm{obs}} = 4.6318$. This corresponds to $\sqrt{n}$ times the maximum correlation presented in Figure 1. Using the null distribution provided by (4.9), which can be estimated via the multiplier bootstrap, yields the p-value 0.001. Further, using the SCAD gives $\widehat{T}_{n,p}^{\mathrm{obs}} = 4.1324$ and a p-value 0.0164. Both calculations are based on 5000 bootstrap replications. Therefore, the evidence against the exogeneity assumption is very strong. Figure 4 depicts the observed test statistics relative to the null distribution.

TABLE 5

Sample correlations between fitted and observed responses, and the empirical median and upper α-quantile of the multiplier bootstrap approximation based on 1200 *bootstrap samples when* $\alpha = 10\%$

	$\widehat{\mathbf{corr}}_n(\mathbb{Y}, \widehat{\mathbb{Y}}^{\mathbf{LASSO}}_{\lambda})$	$\widehat{\mathbf{corr}}_n(\mathbb{Y}, \widehat{\mathbb{Y}}^{\mathbf{pLASSO}}_{\lambda})$	$q^{\mathbf{CMB}}_{0.5}(\widehat{s}_\lambda, p)$	$q^{\mathbf{CMB}}_{0.1}(\widehat{s}_\lambda, p)$
$\lambda = 0.1789$ $(\widehat{s}_\lambda = 2)$	0.6813	0.6879	0.5585	0.5988
$\lambda = 0.1708$ $(\widehat{s}_\lambda = 3)$	0.6915	0.7010	0.6555	0.6904
$\lambda = 0.1630$ $(\widehat{s}_\lambda = 4)$	0.7059	0.7260	0.7252	0.7554
$\lambda = 0.1556$ $(\widehat{s}_\lambda = 5)$	0.7141	0.7406	0.7797	0.8044
$\lambda = 0.1292$ $(\widehat{s}_\lambda = 8)$	0.7454	0.7641	0.8828	0.8988
$\lambda = 0.1177$ $(\widehat{s}_\lambda = 14)$	0.7714	0.8307	0.9658	0.9724
$\lambda = 0.1073$ $(\widehat{s}_\lambda = 17)$	0.8026	0.8739	0.9817	0.9860
$\lambda = 0.0933$ $(\widehat{s}_\lambda = 21)$	0.8451	0.9019	0.9915	0.9945
$\lambda = 0.0891$ $(\widehat{s}_\lambda = 23)$	0.8561	0.9109	0.9937	0.9966
$\lambda = 0.0674$ $(\widehat{s}_\lambda = 25)$	0.8991	0.9214	0.9953	0.9979

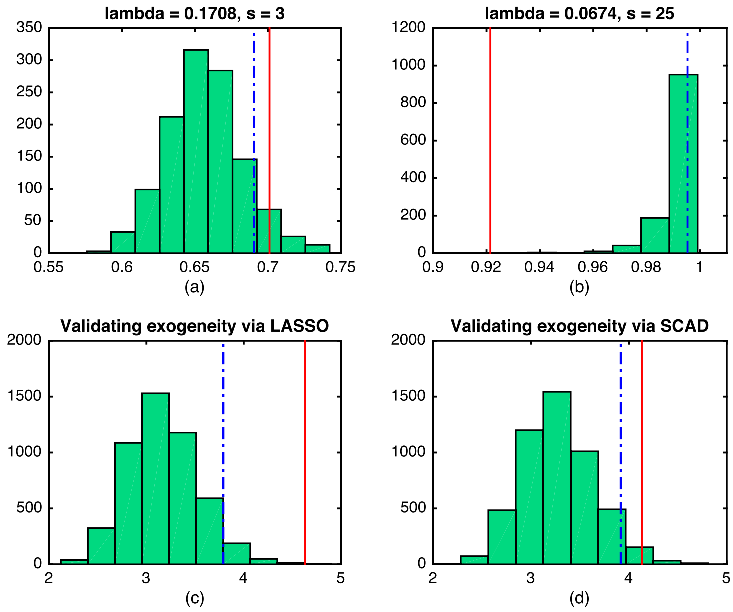

FIG. 4. *Top panel: Distributions of the spurious correlation $\widehat{R}_n(s, p)$ estimated by the bootstrap approximation for* (a) $s = 3$ *and* (b) $s = 25$ *and the sample correlation between fitted and observed responses (see Table* 5*). Red solid lines are observed correlations and blue dash-dot lines mark the* 90*th percentile in* (a) *the median and* (b) *the distributions of the median. Bottom panel: Null distributions for testing exogeneity* (1.2) *and its* 95*th percentile (indicated by dash blue line) using bootstrap approximation* (4.9) *and observed test statistics $\widehat{T}^{\mathrm{obs}}_{n,p}$ (indicated by solid red line) based on the residuals of the LASSO and SCAD.*

7. Proofs. We first collect several technical lemmas in Section 7.1 before proving our main result, Theorem 3.1 in Section 3. The proofs of Theorems 3.2, 4.1 and 4.2 are given in the Supplemental Material [Fan, Shao and Zhou (2018)], where the proofs of Propositions 3.1 and 3.2 and Lemmas 7.2–7.6 can also be found. Throughout, the letters $C, C_1, C_2, \dots$ and $c, c_1, c_2, \dots$ denote generic positive constants that are independent of (s, p, n), whose values may change from line to line.

7.1. *Technical lemmas.* The following lemma combines Propositions 5.10 and 5.16 in Vershynin (2012).

LEMMA 7.1. *Let $X_1, \dots, X_n$ be independent centered random variables and write $\mathbf{x}_n = (X_1, \dots, X_n)^{\mathrm{T}} \in \mathbb{R}^n$. Then for every $\mathbf{a} = (a_1, \dots, a_n)^{\mathrm{T}} \in \mathbb{R}^n$ and every $t \geq 0$, we have*

$$\mathbb{P}(|\mathbf{a}^{\mathrm{T}}\mathbf{x}_n| \geq t) \leq 2\exp\biggl\{-c_{\mathrm{B}} \min\biggl(\frac{t^2}{B_1^2|\mathbf{a}|_2^2}, \frac{t}{B_1|\mathbf{a}|_\infty}\biggr)\biggr\} \tag{7.1}$$

and

$$\mathbb{P}(|\mathbf{a}^{\mathrm{T}}\mathbf{x}_n| \geq t) \leq 2\exp\biggl(-c_{\mathrm{H}}\frac{t^2}{B_2^2|\mathbf{a}|_2^2}\biggr), \tag{7.2}$$

where $B_v = \max_{1\leq i\leq n} \|X_i\|_{\psi_v}$ for $v = 1, 2$ and $c_{\mathrm{B}}, c_{\mathrm{H}} > 0$ are absolute constants.

LEMMA 7.2. *Let Conditions 2.1 and 2.2 be fulfilled. Write*

$$D_n = D_n(s, p) := \sup_{\boldsymbol{\alpha}\in\mathcal{V}} |\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\alpha}/\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha} - 1| \quad \textit{and} \quad \widehat{\sigma}_\varepsilon^2 = n^{-1}\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n)^2,$$

where $\widehat{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^{\mathrm{T}}$ and $\mathcal{V}$ is as in (3.2). Then there exists a constant $C_1 > 0$ such that, for every $t \geq 1$,

$$D_n \leq C_1 K_1^2\biggl[\sqrt{\frac{s}{n}\log(\gamma_s ep/s)} + \max\biggl\{\sqrt{\frac{t}{n}}, c_n(s, p)\frac{t}{n}\biggr\}\biggr] \tag{7.3}$$

holds with probability at least $1 - 8e^{-t}$, where $c_n(s, p) := s\log(\gamma_s ep/s) \vee \log n$. Moreover, for every $t > 0$,

$$|\widehat{\sigma}_\varepsilon^2 - 1| \leq K_0^2 n^{-1}t + 4K_0^2 \max(n^{-1/2}\sqrt{t}, n^{-1}t) \tag{7.4}$$

holds with probability greater than $1 - 2\exp(-c_{\mathrm{B}}t) - 2\exp(-c_{\mathrm{H}}t)$, where $c_{\mathrm{B}}, c_{\mathrm{H}} > 0$ are absolute constants as in (7.1) and (7.2).

The following results address the concentration and anti-concentration phenomena of the supremum of the Gaussian process $\mathbb{G}^*$ indexed by $\mathcal{F}$ [see (3.4)]. In line with Chernozhukov, Chetverikov and Kato (2013), inequalities (7.5) and (7.6) below are referred to as the concentration and anti-concentration inequalities, respectively.

LEMMA 7.3. *Let $R^*(s,p) = \sup_{f_{\boldsymbol{\alpha}} \in \mathcal{F}} f_{\boldsymbol{\alpha}}(\mathbf{Z})/|\boldsymbol{\alpha}|_{\boldsymbol{\Sigma}}$ for $\mathcal{F} = \mathcal{F}(s,p)$ given in* (3.2) *and $\mathbf{Z} \overset{d}{=} N(\mathbf{0}, \boldsymbol{\Sigma})$. Then there exists an absolute constant $C > 0$ such that, for every $p \ge 2$, $1 \le s \le p$ and $t > 0$,*

$$\mathbb{P}\{R^*(s,p) \ge C\sqrt{s\log(\gamma_s ep/s)} + t\} \le e^{-t^2/2} \quad \text{and} \tag{7.5}$$

$$\sup_{x \ge 0} \mathbb{P}\{|R^*(s,p) - x| \le t\} \le Ct\sqrt{s\log(\gamma_s ep/s)}, \tag{7.6}$$

where $\gamma_s = \sqrt{\phi_{\max}(s)}/\sqrt{\phi_{\min}(s)}$.

LEMMA 7.4. *Suppose that $a \ge 1$ and $b_j, c_j > 0$ for $j = 1, \dots, m$ are positive constants. Let $X_1, \dots, X_m$ be real-valued random variables that satisfy*

$$\mathbb{P}(|X_j| \ge t) \le a\exp\{-t^2/(2b_j)\} \qquad \text{for } t > 0,\ j = 1, \dots, m.$$

Then, for all $m \ge 4/a$, we have $\mathbb{E}(\max_{1 \le j \le m} |X_j|) \le 2\sqrt{\log(am)\max_{1 \le j \le m} b_j}$. Furthermore, suppose that $\mathbb{P}(|X_j| \ge t) \le a\exp(-t/c_j)$ holds for all $t > 0$ and $j = 1, \dots, m$. Then, for any $m \ge 4/a$, we have $\mathbb{E}(\max_{1 \le j \le m} |X_j|) \le \{\log(am) + 1\}\max_{1 \le j \le m} c_j$.

To save space, we leave the proofs of Lemmas 7.2–7.4 to Appendix A in the Supplemental Material [Fan, Shao and Zhou (2018)].

7.2. *Proof of Theorem* 3.1. In view of (3.1), we have

$$\widehat{R}_n(s,p) = \sup_{\boldsymbol{\alpha} \in \mathcal{V}} \frac{n^{-1}\sum_{i=1}^n \boldsymbol{\alpha}^{\mathrm{T}}(\varepsilon_i \mathbf{X}_i) - \bar{\varepsilon}_n \boldsymbol{\alpha}^{\mathrm{T}}\bar{\mathbf{X}}_n}{(\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}_n\boldsymbol{\alpha})^{1/2} \cdot \{n^{-1}\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n)^2\}^{1/2}},$$

where $\mathcal{V}$ is as in (3.2).

By Lemma 7.2, instead of dealing with $\widehat{R}_n(s,p)$ directly, we first investigate the asymptotic behavior of its standardized counterpart given by

$$R_n(s,p) = \sup_{\boldsymbol{\alpha} \in \mathcal{V}} n^{-1}\sum_{i=1}^n \frac{\boldsymbol{\alpha}^{\mathrm{T}}(\varepsilon_i \mathbf{X}_i)}{|\boldsymbol{\alpha}|_{\boldsymbol{\Sigma}}} = \sup_{\boldsymbol{\alpha} \in \mathcal{V}} n^{-1}\sum_{i=1}^n \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}^{\mathrm{T}}\mathbf{y}_i, \tag{7.7}$$

where $\mathbf{y}_i = \varepsilon_i \mathbf{X}_i = (Y_{i1}, \dots, Y_{ip})^{\mathrm{T}}$ are i.i.d. random vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$. Let $\mathbb{P}_{\mathbf{y}}$ be the probability measure on $\mathbb{R}^p$ induced by $\mathbf{y} = \varepsilon\mathbf{X}$. Further, define rescaled versions of $\widehat{R}_n(s,p)$ and $R_n(s,p)$ as

$$\widehat{L}_n = \widehat{L}_n(s,p) = \sqrt{n}\widehat{R}_n(s,p), \qquad L_n = L_n(s,p) = \sup_{\boldsymbol{\alpha} \in \mathcal{V}} n^{-1/2}\sum_{i=1}^n \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}^{\mathrm{T}}\mathbf{y}_i. \tag{7.8}$$

The main strategy is to prove the Gaussian approximation of L_n by the supremum of a Gaussian process $\mathbb{G}^*$ indexed by $\mathcal{F}$ with covariance function

$$\mathbb{E}(\mathbb{G}^* f_{\boldsymbol{\alpha}_1}\mathbb{G}^* f_{\boldsymbol{\alpha}_2}) = \frac{\boldsymbol{\alpha}_1^{\mathrm{T}}\boldsymbol{\Sigma}\boldsymbol{\alpha}_2}{|\boldsymbol{\alpha}_1|_{\boldsymbol{\Sigma}} \cdot |\boldsymbol{\alpha}_2|_{\boldsymbol{\Sigma}}}, \qquad \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2 \in \mathcal{V}.$$

Let $\mathbf{Z}$ be a p-variate centered Gaussian random vector with covariance matrix $\mathbf{\Sigma}$. Then the aforementioned Gaussian process $\mathbb{G}^*$ can be induced by $\mathbf{Z}$ in the sense that for every $\boldsymbol{\alpha} \in \mathcal{V}$, $\mathbb{G}^* f_{\boldsymbol{\alpha}} = \boldsymbol{\alpha}_{\mathbf{\Sigma}}^{\mathrm{T}} \mathbf{Z}$. The following lemmas show that, under certain moment conditions, the distribution of $L_n = \sqrt{n} R_n(s, p)$ can be consistently estimated by that of the supremum of the Gaussian process $\mathbb{G}^*$, denoted by $R^*(s, p) = \sup_{\boldsymbol{\alpha} \in \mathcal{V}} \mathbb{G}^* f_{\boldsymbol{\alpha}}$, and $\widehat{L}_n$ and L_n are close. We state them first in the following two lemmas and prove them in Appendix A of the Supplementary Material [Fan, Shao and Zhou (2018)].

LEMMA 7.5. *Under Conditions 2.1 and 2.2, there exists a random variable* $T^* = T^*(s, p) \overset{d}{=} \sup_{\boldsymbol{\alpha} \in \mathcal{V}} \boldsymbol{\alpha}_{\mathbf{\Sigma}}^{\mathrm{T}} \mathbf{Z}$ *for* $\mathbf{Z} \overset{d}{=} N(\mathbf{0}, \mathbf{\Sigma})$ *such that, for any* $\delta \in (0, K_0 K_1]$,

$$|L_n - T^*| \lesssim n^{-1} c_n^{1/2}(s, p) + K_0 K_1 n^{-3/2} c_n^2(s, p) + \delta \tag{7.9}$$

holds with probability at least $1 - C\Delta_n(s, p; \delta)$, *where* $c_n(s, p) = s \log(\gamma_s e p/s) \vee \log n$ *and*

$$\Delta_n(s, p; \delta) = (K_0 K_1)^3 \frac{\{s b_n(s, p)\}^2}{\delta^3 \sqrt{n}} + (K_0 K_1)^4 \frac{\{s b_n(s, p)\}^5}{\delta^4 n}$$

with $b_n(s, p) = \log(\gamma_s p/s) \vee \log n$.

LEMMA 7.6. *Let Conditions 2.1 and 2.2 hold. Assume that the sample size satisfies* $n \geq C_1(K_0 \vee K_1)^4 c_n(s, p)$. *Then, with probability at least* $1 - C_2 n^{-1/2} c_n^{1/2}(s, p)$,

$$|\widehat{L}_n - L_n| \lesssim (K_0 \vee K_1)^2 K_0 K_1 n^{-1/2} c_n(s, p), \tag{7.10}$$

where $c_n(s, p) = s \log(\gamma_s e p/s) \vee \log n$.

Let $b_n(s, p) = \log(\gamma_s p/s) \vee \log n$. Applying Lemmas 7.5 and 7.6 with

$$\delta = \delta_n(s, p) = (K_0 K_1)^{3/4} \min[1, n^{-1/8}\{s b_n(s, p)\}^{3/8}]$$

yields that, with probability at least $1 - C(K_0 K_1)^{3/4} n^{-1/8} \{s b_n(s, p)\}^{7/8}$,

$$|\widehat{L}_n - T^*| \lesssim (K_0 K_1)^{3/4} n^{-1/8} \{s b_n(s, p)\}^{3/8}.$$

Together with the inequality (7.6), this proves (3.3).

Further, using (3.2), (3.4) and the identity $\mathbf{v}^{\mathrm{T}} \mathbf{A}^{-1} \mathbf{v} = \max_{\boldsymbol{\alpha} \in \mathbb{S}^{s-1}} \frac{(\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{v})^2}{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{A} \boldsymbol{\alpha}}$ that holds for any $s \times s$ positive definite matrix $\mathbf{A}$, we find that with probability one,

$$R^*(s, p) = \max_{S \subseteq [p]: |S| = s} \max_{\boldsymbol{\alpha} \in \mathbb{S}^{s-1}} \frac{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{Z}_S}{\sqrt{\boldsymbol{\alpha}^{\mathrm{T}} \mathbf{\Sigma}_{SS} \boldsymbol{\alpha}}} = \max_{S \subseteq [p]: |S| = s} \sqrt{\mathbf{Z}_S^{\mathrm{T}} \mathbf{\Sigma}_{SS}^{-1} \mathbf{Z}_S}, \tag{7.11}$$

where for each $S \subseteq [p]$ fixed, the second maximum over $\boldsymbol{\alpha}$ is achieved when $\boldsymbol{\alpha} = \mathbf{\Sigma}_{SS}^{-1/2} \mathbf{Z}_S / |\mathbf{\Sigma}_{SS}^{-1/2} \mathbf{Z}_S|_2$, as for each $p \geq 1$ fixed, all of the coordinates of $\mathbf{Z}$ are nonzero almost surely. In particular, when $\mathbf{\Sigma} = \mathbf{I}_p$, the right-hand side of (7.11) is reduced to $\max_{S \subseteq [p]: |S| = s} |\mathbf{Z}_S|_2$ and, therefore, $\{R^*(s, p)\}^2 = \max_{S \subseteq [p]: |S| = s} \sum_{j \in S} Z_j^2 = Z_{(p)}^2 + \cdots + Z_{(p-s+1)}^2$ happens with probability one. This and (3.3) complete the proof of (3.5).

Acknowledgements. We thank the Editor, Dr. Edward George, the Past Editor, Dr. Runze Li, the Associate Editor and three reviewers for their careful reviews and constructive comments.

SUPPLEMENTARY MATERIAL

Supplement to "Are discoveries spurious? Distributions of maximum spurious correlations and their applications" (DOI: 10.1214/17-AOS1575SUPP; .pdf). This supplemental material contains additional proofs for all the remaining theoretical results in the main text, including Lemmas 7.2–7.6, Theorems 3.2, 4.1 and 4.2 and Propositions 3.1 and 3.2. A discussion on the moment assumptions is also included.

REFERENCES

ARLOT, S., BLANCHARD, G. and ROQUAIN, E. (2010). Some nonasymptotic results on resampling in high dimension. I. Confidence regions. *Ann. Statist.* **38** 51–82. MR2589316

BARRETT, G. F. and DONALD, S. G. (2003). Consistent tests for stochastic dominance. *Econometrica* **71** 71–104. MR1956856

BICKEL, P. J., RITOV, Y. and TSYBAKOV, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37** 1705–1732. MR2533469

BRUSCO, M. J. and STAHL, S. (2005). *Branch-and-Bound Applications in Combinatorial Data Analysis*. Springer, New York. MR2172059

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

CAI, T., FAN, J. and JIANG, T. (2013). Distributions of angles in random packing on spheres. *J. Mach. Learn. Res.* **14** 1837–1864. MR3104497

CAI, T. T. and JIANG, T. (2011). Limiting laws of coherence of random matrices with applications to testing covariance structure and construction of compressed sensing matrices. *Ann. Statist.* **39** 1496–1525. MR2850210

CAI, T. T., LIU, W. and XIA, Y. (2014). Two-sample test of high dimensional means under dependence. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 349–372. MR3164870

CHANG, J., ZHENG, C., ZHOU, W.-X. and ZHOU, W. (2017). Simulation-based hypothesis testing of high dimensional means under covariance heterogeneity. *Biometrics* **73** 1300–1310. MR3744543

CHATTERJEE, S. and BOSE, A. (2005). Generalized bootstrap for estimating equations. *Ann. Statist.* **33** 414–436. MR2157808

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2013). Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. *Ann. Statist.* **41** 2786–2819. MR3161448

CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597. MR3262461

DAVYDOV, YU. A., LIFSHITS, M. A. and SMORODINA, N. V. (1998). *Local Properties of Distributions of Stochastic Functionals*. *Translations of Mathematical Monographs* **173**. Amer. Math. Soc., Providence, RI. Translated from the 1995 Russian original by V. E. Nazaĭkinskiĭ and M. A. Shishkova. MR1604537

DUDOIT, S. and VAN DER LAAN, M. J. (2008). *Multiple Testing Procedures with Applications to Genomics*. Springer, New York. MR2373771

EFRON, B. (2010). *Large-Scale Inference*: *Empirical Bayes Methods for Estimation*, *Testing*, *and Prediction*. *Institute of Mathematical Statistics* (*IMS*) *Monographs* **1**. Cambridge Univ. Press, Cambridge. MR2724758

FAN, J., GUO, S. and HAO, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **74** 37–65. MR2885839

FAN, J., HAN, F. and LIU, H. (2014). Challenges of big data analysis. *Natl. Sci. Rev.* **1** 293–314.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FAN, J. and LIAO, Y. (2014). Endogeneity in high dimensions. *Ann. Statist.* **42** 872–917. MR3210990

FAN, J. and LV, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statist. Sinica* **20** 101–148. MR2640659

FAN, J., SHAO, Q.-M. and ZHOU, W.-X. (2018). Supplement to "Are discoveries spurious? Distributions of maximum spurious correlations and their applications." DOI:10.1214/17-AOS1575SUPP.

FAN, J., XUE, L. and ZOU, H. (2014). Strong oracle optimality of folded concave penalized estimation. *Ann. Statist.* **42** 819–849. MR3210988

GOEMAN, J. J., VAN DE GEER, S. A. and VAN HOUWELINGEN, H. C. (2006). Testing against a high dimensional alternative. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68** 477–493. MR2278336

HANSEN, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica* **64** 413–430. MR1375740

HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. Springer, New York. MR2722294

SHAO, Q.-M. and ZHOU, W.-X. (2014). Necessary and sufficient conditions for the asymptotic distributions of coherence of ultra-high dimensional random matrices. *Ann. Probab.* **42** 623–648. MR3178469

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. *Nat. Genet.* **39** 1217–1224.

THORGEIRSSON, T. E. et al. (2010). Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior. *Nat. Genet.* **42** 448–453.

THORISSON, G. A., SMITH, A. V., KRISHNAN, L. and STEIN, L. D. (2005). The international HapMap project web site. *Genome Res.* **15** 1592–1593.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58** 267–288. MR1379242

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*: *With Applications to Statistics*. Springer, New York. MR1385671

VERSHYNIN, R. (2012). Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing* (Y. Eldar and G. Kutyniok, eds.) 210–268. Cambridge Univ. Press, Cambridge. MR2963170

ZHANG, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38** 894–942. MR2604701

ZOU, H. and LI, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1509–1533. MR2435443

J. FAN
SCHOOL OF DATA SCIENCE
FUDAN UNIVERSITY
SHANGHAI 200433
CHINA
AND
DEPARTMENT OF OPERATIONS RESEARCH
AND FINANCIAL ENGINEERING
PRINCETON UNIVERSITY
PRINCETON, NEW JERSEY 08544
USA
E-MAIL: jqfan@princeton.edu

Q.-M. SHAO
DEPARTMENT OF STATISTICS
CHINESE UNIVERSITY OF HONG KONG
SHATIN, NT
HONG KONG
E-MAIL: qmshao@cuhk.edu.hk

W.-X. ZHOU
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF CALIFORNIA, SAN DIEGO
LA JOLLA, CALIFORNIA 92093
USA
E-MAIL: wez243@ucsd.edu