Distributions of Maximum Spurious Correlations and Their Statistical Implications

Wen-Xin Zhou

Department of Mathematics and Statistics
The University of Melbourne

Joint work with Jianqing Fan and Qi-Man Shao

Jan 7 2015, Hong Kong

Outline

1. Introduction
2. Distributions of maximum spurious correlations
3. Extension to sparse linear models
4. Main ideas of the proof

1. Introduction

Massive and complex data and high dimensionality characterize contemporary statistical problems in many frontiers of sciences and engineering. Various statistical methods and algorithms have been proposed to find a small group of covariate variables that are associated with given responses such as biological and clinical outcomes. They have been very successfully applied to genomics, genetics, neurosciences, economics and finance.

▶ Background

Consider the problem of estimating a p-vector of parameters $\boldsymbol{\beta}$ from the linear model

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where

- $\mathbb{Y} = (Y_1, \ldots, Y_n)^T$ is an n-vector of responses;
- $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ is an $n \times p$ random design matrix with i.i.d. rows;
- $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a p-vector of parameters;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-vector of i.i.d. random errors.

Background

Consider the problem of estimating a p-vector of parameters $\boldsymbol{\beta}$ from the linear model

$$\mathbb{Y} = \mathbb{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon},$$

where

- $\mathbb{Y} = (Y_1, \ldots, Y_n)^T$ is an n-vector of responses;
- $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ is an $n \times p$ random design matrix with i.i.d. rows;
- $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^T$ is a p-vector of parameters;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is an n-vector of i.i.d. random errors.

When $p > n$ or $p \gg n$, assume that the true model $S_0 = \{j : \beta_j^* \neq 0\}$ is sparse, i.e. the number of non-zero coefficients $s = |S_0|$ is small.

Under various sparsity assumptions and regularity conditions, the most popular variable selection tools such as

- LASSO (Tibshirani, 1996)
- SCAD (Fan and Li, 2001)
- adaptive LASSO (Zou, 2006)
- Dantzig selector (Candes and Tao, 2007), etc.

possess various good properties regarding model selection consistency.

Theoretically, under suitable regularity conditions, all aforementioned model selection tools can achieve model consistency, i.e. they can exactly pick out the true sparse model with probability tending to one. However, in practice, these conditions are impossible to check and hard to meet. Hence, it is very difficult to extract the exact subset of significant variables among a huge set of covariates. One of the reasons is the spurious correlation, as illustrated by Fan, Guo and Hao (2012).

Behind machine learning, data-mining, and high-dimensional statistics techniques, there are many model assumptions and even heuristics arguments. For sample, LASSO and SCAD are based on the assumption of exogeneity:

$$E(\varepsilon X_j) = 0 \quad \text{for all } j = 1, \ldots, p. \tag{0.1}$$

Behind machine learning, data-mining, and high-dimensional statistics techniques, there are many model assumptions and even heuristics arguments. For sample, LASSO and SCAD are based on the assumption of exogeneity:

$$E(\varepsilon X_j) = 0 \quad \text{for all } j = 1, \ldots, p. \tag{0.1}$$

Fan and Liao (2014) provides evidences that such an ideal assumption might not be valid, yet such an ideal assumption is a necessary condition for model selection consistency. Despite of its fundamental importance to high-dimensional statistics, there are no available tools for validating (0.1). Regarding (0.1) as a null hypothesis, it is instructive to consider the following test statistic:

$$\widehat{T}_{n,p} = \max_{1 \le j \le p} \left| \widehat{\mathrm{corr}}_n(X_j, \varepsilon) \right|.$$

Behind machine learning, data-mining, and high-dimensional statistics techniques, there are many model assumptions and even heuristics arguments. For sample, LASSO and SCAD are based on the assumption of exogeneity:

$$E(\varepsilon X_j) = 0 \quad \text{for all } j = 1, \ldots, p. \tag{0.1}$$

Fan and Liao (2014) provides evidences that such an ideal assumption might not be valid, yet such an ideal assumption is a necessary condition for model selection consistency. Despite of its fundamental importance to high-dimensional statistics, there are no available tools for validating (0.1). Regarding (0.1) as a null hypothesis, it is instructive to consider the following test statistic:

$$\widehat{T}_{n,p} = \max_{1 \le j \le p} \left| \widehat{\text{corr}}_n(X_j, \varepsilon) \right|.$$

To conduct statistical inference, we need to derive the limiting distributions of the maximum spurious correlation.

- Spurious correlation, conditions and notation

Let $\varepsilon, \varepsilon_1, \ldots, \varepsilon_n$ be i.i.d. random variables with

$$E(\varepsilon) = 0, \quad E(\varepsilon^2) = \sigma^2 < \infty.$$

Let $\mathbf{X}, \mathbf{X}_1, \ldots, \mathbf{X}_n$ be i.i.d. p-dimensional random vectors with

$$E(\mathbf{X}_i) = \mathbf{0}, \quad \mathbf{\Sigma} = E(\mathbf{X}\mathbf{X}^T) = (\sigma_{jk})_{1 \le j,k \le p}.$$

Assume that the two samples $\{\varepsilon_i\}_{i=1}^n$ and $\{\mathbf{X}_i\}_{i=1}^n$ are independent. The maximum spurious correlation is defined as

$$\widehat{R}_n(s,p) = \max_{\boldsymbol{\alpha} \in \mathcal{S}^{p-1}:\, |\boldsymbol{\alpha}|_0 = s} \widehat{\text{corr}}_n(\varepsilon, \boldsymbol{\alpha}^T \mathbf{X}),$$

where $\widehat{\text{corr}}_n(\cdot,\cdot)$ denotes the Pearson sample correlation coefficient and $\mathcal{S}^{p-1} := \{\boldsymbol{\alpha} \in \mathbb{R}^p : |\boldsymbol{\alpha}|_2 = 1\}$ is the unit sphere of $\mathbb{R}^p$.

Let $\mathbf{D} = \mathrm{diag}(\mathbf{\Sigma}) = \mathrm{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. We can rewrite $\widehat{R}_n(s, p)$ as

$$\max_{\substack{S \subseteq [p] \\ |S| = s}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^s \setminus \{\mathbf{0}\}} \frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n) \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S} \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n)^2 \cdot \sum_{i=1}^n \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S} \right\rangle^2}}$$

$$= \max_{\substack{S \subseteq [p] \\ |S| = s}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^s \setminus \{\mathbf{0}\}} \frac{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*) \left\langle \mathbf{D}_{SS}^{1/2} \boldsymbol{\alpha}, \mathbf{D}_{SS}^{-1/2} (\mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S}) \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*)^2 \cdot \sum_{i=1}^n \left\langle \mathbf{D}_{SS}^{1/2} \boldsymbol{\alpha}, \mathbf{D}_{SS}^{-1/2} (\mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S}) \right\rangle^2}}$$

$$= \max_{\substack{S \subseteq [p] \\ |S| = s}} \max_{\boldsymbol{\alpha} \in \mathcal{S}^{s-1}} \frac{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*) \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S}^* - \bar{\mathbf{X}}_{n,S}^* \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*)^2 \cdot \sum_{i=1}^n \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S}^* - \bar{\mathbf{X}}_{n,S}^* \right\rangle^2}},$$

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 $\bar{\varepsilon}_n^* = n^{-1} \sum_{i=1}^n \varepsilon_i^*$, $\bar{\mathbf{X}}_n^* = n^{-1} \sum_{i=1}^n \mathbf{X}_i^*$ with $\varepsilon_i^* = \sigma^{-1} \varepsilon_i$ and $\mathbf{X}_i^* = \mathbf{D}^{-1/2} \mathbf{X}_i$.

Let $\mathbf{D} = \text{diag}(\mathbf{\Sigma}) = \text{diag}(\sigma_{11}, \ldots, \sigma_{pp})$. We can rewrite $\widehat{R}_n(s,p)$ as

$$\max_{\substack{S \subseteq [p] \\ |S|=s}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^s \setminus \{\mathbf{0}\}} \frac{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n) \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S} \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon}_n)^2 \cdot \sum_{i=1}^n \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S} \right\rangle^2}}$$

$$= \max_{\substack{S \subseteq [p] \\ |S|=s}} \max_{\boldsymbol{\alpha} \in \mathbb{R}^s \setminus \{\mathbf{0}\}} \frac{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*) \left\langle \mathbf{D}_{SS}^{1/2} \boldsymbol{\alpha}, \mathbf{D}_{SS}^{-1/2} (\mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S}) \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*)^2 \cdot \sum_{i=1}^n \left\langle \mathbf{D}_{SS}^{1/2} \boldsymbol{\alpha}, \mathbf{D}_{SS}^{-1/2} (\mathbf{X}_{i,S} - \bar{\mathbf{X}}_{n,S}) \right\rangle^2}}$$

$$= \max_{\substack{S \subseteq [p] \\ |S|=s}} \max_{\boldsymbol{\alpha} \in \mathcal{S}^{s-1}} \frac{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*) \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S}^* - \bar{\mathbf{X}}_{n,S}^* \right\rangle}{\sqrt{\sum_{i=1}^n (\varepsilon_i^* - \bar{\varepsilon}_n^*)^2 \cdot \sum_{i=1}^n \left\langle \boldsymbol{\alpha}, \mathbf{X}_{i,S}^* - \bar{\mathbf{X}}_{n,S}^* \right\rangle^2}},$$

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 $\bar{\varepsilon}_n^* = n^{-1} \sum_{i=1}^n \varepsilon_i^*$, $\bar{\mathbf{X}}_n^* = n^{-1} \sum_{i=1}^n \mathbf{X}_i^*$ with $\varepsilon_i^* = \sigma^{-1} \varepsilon_i$ and $\mathbf{X}_i^* = \mathbf{D}^{-1/2} \mathbf{X}_i$.

Assume $\sigma^2 = 1$ and $\mathbf{\Sigma}$ is such that $\mathbf{D} = \text{diag}(\mathbf{\Sigma}) = \mathbf{I}_p$.

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

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

respectively. 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).

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

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

respectively. 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 ε and $\mathbf{X}$ are imposed.

Condition (1)

There exists a random vector $\mathbf{U}$ *such that* $E(\mathbf{U}) = \mathbf{0}$, $E(\mathbf{U}\mathbf{U}^T) = \mathbf{I}_p$,

$$\mathbf{X} = (X_1, \ldots, X_p)^T = \boldsymbol{\Sigma}^{1/2}\mathbf{U}, \quad K_1 := \sup_{\boldsymbol{\alpha} \in \mathcal{S}^{p-1}} \| \langle \boldsymbol{\alpha}, \mathbf{U} \rangle \|_{\psi_2}.$$

The random variable ε *has zero mean and unit variance and is sub-Gaussian with* $K_0 := \|\varepsilon\|_{\psi_2} < \infty$.

For $1 \leq s \leq p$, the s-sparse minimal and maximal eigenvalues of the covariance matrix $\mathbf{\Sigma}$ are defined as

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

respectively, where $|\mathbf{u}|_{\mathbf{\Sigma}} = (\mathbf{u}^T \mathbf{\Sigma} \mathbf{u})^{1/2}$ and $|\mathbf{u}|_2 = (\mathbf{u}^T \mathbf{u})^{1/2}$ is the ℓ_2-norm of $\mathbf{u}$. We impose the following restricted eigenvalues assumption on the covariance matrix $\mathbf{\Sigma}$.

Condition (2)

For $1 \leq s \leq p$, the s-sparse condition number of $\mathbf{\Sigma}$ is finite; that is,

$$\gamma_s = \gamma_s(\mathbf{\Sigma}) = \sqrt{\frac{\phi_{\max}(s)}{\phi_{\min}(s)}} \in [1, \infty).$$

▶ Asymptotic distribution of the maximum spurious correlation

Rewrite $\widehat{R}_n(s,p)$ as

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

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

$$\mathcal{F} = \mathcal{F}(s,p) = \left\{\mathbf{x} \mapsto f_{\boldsymbol{\alpha}}(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{x}\rangle : \boldsymbol{\alpha} \in \mathcal{S}^{p-1}, |\boldsymbol{\alpha}|_0 = s\right\}$$

is a class of linear functions $\mathbb{R}^p \mapsto \mathbb{R}$.

We regard $\boldsymbol{\alpha} \in \mathcal{F}$ as the linear map $\mathbf{x} \mapsto \langle \boldsymbol{\alpha}, \mathbf{x}\rangle$ induced by $\boldsymbol{\alpha} \in \mathcal{S}^{p-1}$ with $|\boldsymbol{\alpha}|_0 = s$.

Let $\mathbf{Z} = (Z_1, \ldots, Z_p)^T$ be a p-dimensional Gaussian random vector with mean zero and covariance matrix $\boldsymbol{\Sigma}$, i.e. $\mathbf{Z} \stackrel{d}{=} N(\mathbf{0}, \boldsymbol{\Sigma})$, and denote by $Z_{(1)}^2 \leq Z_{(2)}^2 \leq \cdots \leq Z_{(p)}^2$ the order statistics of $\{Z_1^2, \ldots, Z_p^2\}$.

The following theorem shows that, under certain moment conditions, 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 the function class $\mathcal{F}$.

Theorem (1)

Assume that Conditions (1), (2) hold, $n, p \geq 2$ and $1 \leq s \leq p$. Then there exists a constant $C > 0$ depending only on K_0, K_1 such that

$$\sup_{t \geq 0} \left| P\{\sqrt{n}\,\widehat{R}_n(s,p) \leq t\} - P\{R^*(s,p) \leq t\} \right| \leq Cn^{-1/8}\{s\,b_n(s,p)\}^{7/8},$$

where $b_n(s,p) := \log \frac{\gamma_s p}{s} \vee \log n$ *and*

$$R^*(s,p) := \sup_{\boldsymbol{\alpha} \in \mathcal{F}} \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}^T \mathbf{Z} = \sup_{\boldsymbol{\alpha} \in \mathcal{F}} \frac{\boldsymbol{\alpha}^T \mathbf{Z}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}}}.$$

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

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

where $a_p = 2\log p - \log(\log p)$.

Remark (1)

The independence assumption of ε and $\mathbf{X}$ can be relaxed as

$$E(\varepsilon\mathbf{X}) = 0, \;\; E\big(\varepsilon^2|\mathbf{X}\big) = \sigma^2 \;\; \text{and} \;\; E\big(\varepsilon^4|\mathbf{X}\big) \leq C, \text{ a.s.}$$

where $C > 0$ is an absolute constant. In addition, the above result indicates that the increment $n\{\widehat{R}_n(s,p)^2 - \widehat{R}_n(s-1,p)^2\}$ is approximately the same as $Z^2_{(p-s+1)}$.

Remark (1)

The independence assumption of ε and $\mathbf{X}$ can be relaxed as

$$E(\varepsilon \mathbf{X}) = 0, \ \ E\big(\varepsilon^2 | \mathbf{X}\big) = \sigma^2 \ \ \text{and} \ \ E\big(\varepsilon^4 | \mathbf{X}\big) \leq C, \text{ a.s.}$$

where $C > 0$ is an absolute constant. In addition, the above result indicates that the increment $n\{\widehat{R}_n(s,p)^2 - \widehat{R}_n(s-1,p)^2\}$ is approximately the same as $Z^2_{(p-s+1)}$.

When $s = 1$, it is straightforward to verify that, for any $t \in \mathbb{R}$,

$$P\big\{Z^2_{(p)} - 2\log p + \log(\log p) \leq t\big\} \to \exp\left(-\frac{1}{\sqrt{\pi}}\, e^{-t/2} \right)$$

as $p \to \infty$.

Proposition (1)

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

$$P\{Z_{(p)}^2 + \cdots + Z_{(p-s+1)}^2 - sa_p \leq t\}$$
$$\to \frac{\pi^{(1-s)/2}}{(s-1)!\,\Gamma(s-1)} \int_{-\infty}^{t/s} \left\{ \int_0^{(t-sv)/2} u^{s-2}\, e^{-u}\, du \right\} e^{-(s-1)v/2}\, g(v)\, dv$$

where $a_p = 2\log p - \log(\log p)$,

$$G(t) = \exp\left(-\frac{1}{\sqrt{\pi}}\, e^{-t/2} \right) \quad \textit{and} \quad g(t) = G'(t) = \frac{e^{-t/2}}{2\sqrt{\pi}}\, G(t).$$

In particular, when $s = 2$,

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

- Multiplier bootstrap approximation

  Since the covariance matrix $\boldsymbol{\Sigma}$ of $\mathbf{X}$ is unspecified, the distribution of $R^*(s,p)$ is unknown and thus can not be used for statistical inference. In the following, we consider to use a Monte Carlo method to simulate a process that mimics

$$\left\{ \frac{\boldsymbol{\alpha}^T \mathbf{Z}}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma} \boldsymbol{\alpha}}} : \boldsymbol{\alpha} \in \mathcal{F} \right\},$$

  now known as the multiplier (wild) bootstrap method that is similar to that used in Hansen (1996), Barrett and Donald (2003) and Chernozhukov, Chetverikov and Kato (2013), among others.

Let $\mathbf{\Sigma}_n$ be the sample covariance matrix:

$$\mathbf{\Sigma}_n = n^{-1} \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}}_n)(\mathbf{X}_i - \bar{\mathbf{X}}_n)^T.$$

Let $\xi_1, \ldots, \xi_n$ be i.i.d. $N(0,1)$ random variables that are independent of $\{\varepsilon_i\}_{i=1}^n$ and $\{\mathbf{X}_i\}_{i=1}^n$, and write

$$\mathbf{Z}_n = n^{-1/2} \sum_{i=1}^{n} \xi_i (\mathbf{X}_i - \bar{\mathbf{X}}_n).$$

Conditional on $\{\mathbf{X}_i\}_{i=1}^n$, $\mathbf{Z}_n$ is a p-variate Gaussian random vector with mean zero and covariance matrix $\mathbf{\Sigma}_n$. Let $\widehat{\mathbb{G}}$ be the Gaussian process induced by $\mathbf{Z}_n$, i.e.

$$\widehat{\mathbb{G}}_{\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}^T \mathbf{Z}_n}{\sqrt{\boldsymbol{\alpha}^T \mathbf{\Sigma}_n \boldsymbol{\alpha}}}, \quad \boldsymbol{\alpha} \in \mathcal{F}.$$

Theorem (2)

Under Conditions (1) and (2), if the triplet (s, p, n) satisfies $1 \leq s \leq p$ and $s \log(\gamma_s pn) = o(n^{1/5})$, then as $n \to \infty$,

$$\sup_{t \geq 0} \left| P\{R^*(s,p) \leq t\} - P\{R^*_{MB}(s,p) \leq t \,|\mathbf{X}_1, \ldots, \mathbf{X}_n\} \right| \xrightarrow{P} 0,$$

where $R^*_{MB}(s,p) := \sup_{\boldsymbol{\alpha} \in \mathcal{F}} \widehat{\mathbb{G}}_{\boldsymbol{\alpha}}$.

Theorem (2)

Under Conditions (1) and (2), if the triplet (s, p, n) *satisfies* $1 \le s \le p$ *and* $s \log(\gamma_s pn) = o(n^{1/5})$, *then as* $n \to \infty$,

$$\sup_{t \ge 0} \left| P\left\{R^*(s,p) \le t\right\} - P\left\{R^*_{MB}(s,p) \le t \,\middle|\, \mathbf{X}_1, \ldots, \mathbf{X}_n\right\} \right| \xrightarrow{P} 0,$$

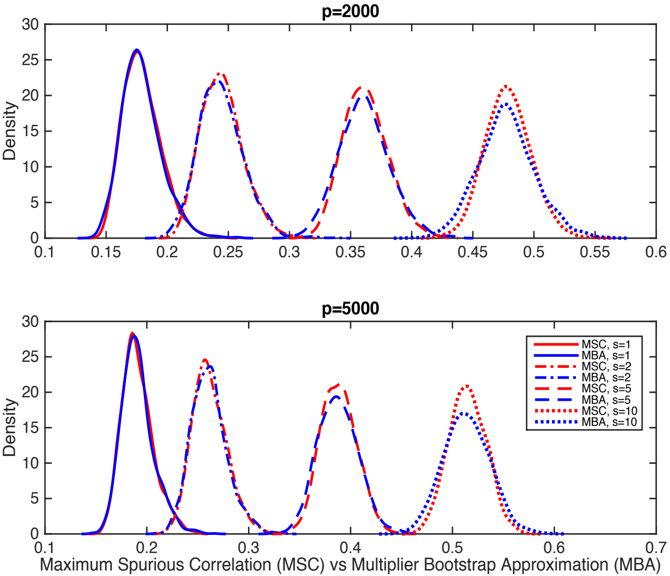
where $R^*_{MB}(s,p) := \sup_{\boldsymbol{\alpha} \in \mathcal{F}} \widehat{\mathbb{G}}_{\boldsymbol{\alpha}}$.

$R^*_{\mathrm{MB}}(s,p)$ is data-driven and its quantiles can be computed via Monte Carlo simulations with arbitrary precision.

For the simulation, consider the case where the noise ε follows the uniform distribution standardized so that $E(\varepsilon) = 0$ and $E(\varepsilon^2) = 1$. Independent of ε, the p-vector $\mathbf{X}$ of covariates has i.i.d. $N(0, 1)$ components.

We report in the following figure the distributions of the maximum spurious correlations and their multiplier bootstrap approximations conditional on a given data set $\{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ based on 1600 simulations when $p \in \{2000, 5000\}$, $s \in \{1, 2, 5, 10\}$ and $n = 400$.

The figure shows that the multiplier bootstrap method provides a fairly good approximation to the (unknown) distribution of the maximum spurious correlation.



p=2000
Density
p=5000
Density
MSC, s=1
MBA, s=1
MSC, s=2
MBA, s=2
MSC, s=5
MBA, s=5
MSC, s=10
MBA, s=10
Maximum Spurious Correlation (MSC) vs Multiplier Bootstrap Approximation (MBA)

3. Extension to sparse linear models

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

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

where the regression coefficient $\boldsymbol{\beta}^*$ is sparse. The sparsity is typically explored by the LASSO or the SCAD.

Now it is well-known that, under suitable conditions, the SCAD, among other folded concave penalized least square estimators, also enjoys the unbiasedness and the (strong) oracle properties (Fan and Li, 2001; Fan and Lv, 2011).

For a given random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, recall that

- $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ is the $n \times p$ design matrix;
- $\mathbb{Y} = (Y_1, \ldots, Y_n)^T$ is the n-dimensional response vector;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the n-dimensional noise vector.

For a given random sample $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$, recall that

- $\mathbb{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_n)^T$ is the $n \times p$ design matrix;
- $\mathbb{Y} = (Y_1, \ldots, Y_n)^T$ is the n-dimensional response vector;
- $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the n-dimensional noise vector.

W.L.O.G., assume that $\boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ with each component of $\boldsymbol{\beta}_1 \in \mathbb{R}^s$ non-zero and $\boldsymbol{\beta}_2 = \mathbf{0}$, such that

$$S_0 := \text{supp}(\boldsymbol{\beta}^*) = \{1, \ldots, s\}$$

is the true underlying sparse model of the indices with $s = |\boldsymbol{\beta}^*|_0$.

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.

Note that $\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbb{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and the oracle estimator $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ has an explicit form of

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

Note that $\mathbb{Y} = \mathbb{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} = \mathbb{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}$ and the oracle estimator $\widehat{\boldsymbol{\beta}}^{\text{oracle}}$ has an explicit form of

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

Denote by $\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}} = (\widehat{\varepsilon}_1^{\text{oracle}}, \ldots, \widehat{\varepsilon}_n^{\text{oracle}})^T$ the residuals after the oracle fit:

$$\widehat{\varepsilon}_i^{\text{oracle}} = Y_i - \mathbf{X}_i^T\widehat{\boldsymbol{\beta}}^{\text{oracle}}, \quad i = 1, \ldots, n.$$

Construct the maximum spurious correlation as before except that $\{\varepsilon_i\}$ is now replaced by $\{\widehat{\varepsilon}_i^{\text{oracle}}\}$, i.e.

$$\widehat{R}_n^{\text{oracle}}(1,p) = \max_{1\le j\le p} \frac{|\sum_{i=1}^n (\widehat{\varepsilon}_i^{\text{oracle}} - n^{-1}\mathbf{e}_n^T\,\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}})(X_{ij} - \bar{X}_j)|}{\sqrt{\sum_{i=1}^n (\widehat{\varepsilon}_i^{\text{oracle}} - n^{-1}\mathbf{e}_n^T\,\widehat{\boldsymbol{\varepsilon}}^{\text{oracle}})^2} \cdot \sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2}}.$$

where $\mathbf{e}_n = (1, \ldots, 1)^T \in \mathbb{R}^n$ and $\bar{X}_j = n^{-1}\sum_{i=1}^n X_{ij}$.

▶ Assumptions:

Condition (3)

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

Condition (4)

$\boldsymbol{\Sigma} = E(\mathbf{X}_i\mathbf{X}_i^T)$ is such that diag $(\boldsymbol{\Sigma}) = \mathbf{I}_p$ and can be expressed in a block-wise form as follows:

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

where $d = p - s$. Let $\widetilde{\boldsymbol{\Sigma}} = (\widetilde{\sigma}_{jk})_{1\le j,k\le d} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ be the Schur complement of $\boldsymbol{\Sigma}_{11}$ in $\boldsymbol{\Sigma}$ satisfying $\widetilde{\sigma}_{\min} = \min_{1\le j\le d} \widetilde{\sigma}_{jj} > 0$.

- Asymptotic distribution of $\widehat{R}_n^{\text{oracle}}(1,p)$:

Theorem (3)

Assume that Conditions (3) and (4) hold, and that the triplet (s,p,n) satisfies $s \log p = o(\sqrt{n})$ and $\log p = o(n^{1/7})$. Then, as $n \to \infty$,

$$\sup_{t \geq 0} \left| P\{\sqrt{n}\,\widehat{R}_n^{\text{oracle}}(1,p) \leq t\} - P\big(|\widetilde{\mathbf{Z}}|_\infty \leq t\big) \right| \to 0,$$

where $\widetilde{\mathbf{Z}}$ is a d-dimensional centered Gaussian random vector with covariance matrix $\widetilde{\mathbf{\Sigma}}$.

For $\lambda > 0$ and $t \geq 0$, let $p_\lambda(t)$ be the SCAD penalty function whose derivative is given by

$$p'_\lambda(t) = \lambda\left\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right\} \quad \text{for some } a > 2.$$

SCAD exploits the sparsity by p_λ-regularization, which solves

$$\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T \in \mathbb{R}^p} (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

For $\lambda > 0$ and $t \geq 0$, let $p_\lambda(t)$ be the SCAD penalty function whose derivative is given by

$$p'_\lambda(t) = \lambda\left\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right\} \quad \text{for some } a > 2.$$

SCAD exploits the sparsity by p_λ-regularization, which solves

$$\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T \in \mathbb{R}^p} (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

This is a non-convex optimization problem which has multiple local minimizers. 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).

For $\lambda > 0$ and $t \geq 0$, let $p_\lambda(t)$ be the SCAD penalty function whose derivative is given by

$$p'_\lambda(t) = \lambda\left\{I(t \leq \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda)\right\} \quad \text{for some } a > 2.$$

SCAD exploits the sparsity by p_λ-regularization, which solves

$$\min_{\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)^T \in \mathbb{R}^p} (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|).$$

This is a non-convex optimization problem which has multiple local minimizers. 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).

Fan, Xue and Zou (2014) proved 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 (e.g. LASSO).

Let $\widehat{\beta}^{\text{lla}}$ be the estimator computed via the one-step LLA algorithm initiated by the LASSO estimator

$$\widehat{\beta}^{\text{lasso}} = \arg\min_{\beta} (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta)^2 + \lambda_{\text{lasso}} |\beta|_1.$$

Accordingly, denote by $\widehat{R}_n^{\text{lla}}(1, p)$ the maximum spurious correlation with $\widehat{\varepsilon}_i^{\text{oracle}}$ replaced by $\widehat{\varepsilon}_i^{\text{lla}} = Y_i - \mathbf{X}_i^T \widehat{\beta}^{\text{lla}}$.

Let $\widehat{\beta}^{\text{lla}}$ be the estimator computed via the one-step LLA algorithm initiated by the LASSO estimator

$$\widehat{\beta}^{\text{lasso}} = \arg\min_{\beta} (2n)^{-1} \sum_{i=1}^{n} (Y_i - \mathbf{X}_i^T \beta)^2 + \lambda_{\text{lasso}} |\beta|_1.$$

Accordingly, denote by $\widehat{R}_n^{\text{lla}}(1, p)$ the maximum spurious correlation with $\widehat{\varepsilon}_i^{\text{oracle}}$ replaced by $\widehat{\varepsilon}_i^{\text{lla}} = Y_i - \mathbf{X}_i^T \widehat{\beta}^{\text{lla}}$.

- Restricted eigenvalue condition (Bickel, Ritov and Tsybakov, 2009):

Definition (1)

For some integer s_0 such that $1 \le s_0 \le p$ and a positive number c_0, we say that a $p \times p$ matrix $\mathbf{A}$ satisfies RE (s_0, c_0) condition if

$$\kappa(s_0, c_0, \mathbf{A}) := \min_{S \subseteq [p]: |S| \le s_0} \min_{\delta \ne 0: |\delta_{S^c}|_1 \le c_0 |\delta_S|_1} \frac{\delta^T \mathbf{A} \delta}{|\delta_S|_2^2} > 0.$$

Condition (5)

$\boldsymbol{\Sigma}$ *satisfies the RE* $(s, 3+\epsilon)$ *condition for some* $\epsilon > 0$, $s = |\mathrm{supp}(\beta^*)|$.

Theorem (4)

Assume that Conditions (3), (4) and (5) hold, and that

- $\min_{j \in S_0} |\beta_j| > (a+1)\lambda$;
- $s \log p = o(\sqrt{n})$, $\log p = o(n^{1/7})$;
- $\lambda \geq \frac{8\sqrt{s}\,\lambda_{\text{lasso}}}{\kappa(s,3,\boldsymbol{\Sigma})}$, $\lambda_{\text{lasso}} \geq C K_0\, n^{-1/2}\sqrt{\log p}$ *for* $C > 0$ *large enough.*

Then, as $n \to \infty$,

$$\sup_{t \geq 0} \left| P\left\{\sqrt{n}\,\widehat{R}_n^{\text{lla}}(1,p) \leq t\right\} - P\left(|\widetilde{\mathbf{Z}}|_\infty \leq t\right) \right| \to 0,$$

where $\widetilde{\mathbf{Z}}$ *is a d-dimensional centered Gaussian random vector with covariance matrix* $\widetilde{\boldsymbol{\Sigma}}$.

- Step 1.

Observe that

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

and consider the standardized counterpart $\widehat{R}_n(s,p)$:

$$R_n(s,p) = \sup_{\boldsymbol{\alpha}\in\mathcal{F}} n^{-1}\sum_{i=1}^{n}\frac{\langle \boldsymbol{\alpha}, \varepsilon_i \mathbf{X}_i\rangle}{\sqrt{\boldsymbol{\alpha}^T \boldsymbol{\Sigma}\boldsymbol{\alpha}}} = \sup_{\boldsymbol{\alpha}\in\mathcal{F}} n^{-1}\sum_{i=1}^{n}\langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{y}_i\rangle,$$

where $\mathbf{y}_i = \varepsilon_i \mathbf{X}_i$ are i.i.d. random vectors with mean zero and covariance matrix $\boldsymbol{\Sigma}$.
First we show that

$$\widehat{L}_n = \sqrt{n}\,\widehat{R}_n(s,p) \quad \text{and} \quad L_n = \sqrt{n}\,R_n(s,p)$$

are close.

- Step 2. (Gaussian approximation of L_n)
  - Step 2.1 (Discretization). The induced metric on the space of all linear functions $\mathbf{x} \mapsto f_{\boldsymbol{\alpha}}(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{x} \rangle$ is $\rho(f_{\boldsymbol{\alpha}}, f_{\boldsymbol{\beta}}) = |\boldsymbol{\alpha} - \boldsymbol{\beta}|_2$. Let $N(\mathcal{F}, \rho, \epsilon)$ be the ϵ-covering number of $(\mathcal{F}, d)$. It is known that $N(\mathcal{S}^{p-1}, \rho, \epsilon) \leq (1 + \frac{2}{\epsilon})^p$.

- Step 2. (Gaussian approximation of L_n)
  - Step 2.1 (Discretization). The induced metric on the space of all linear functions $\mathbf{x} \mapsto f_{\boldsymbol{\alpha}}(\mathbf{x}) = \langle \boldsymbol{\alpha}, \mathbf{x} \rangle$ is $\rho(f_{\boldsymbol{\alpha}}, f_{\boldsymbol{\beta}}) = |\boldsymbol{\alpha} - \boldsymbol{\beta}|_2$. Let $N(\mathcal{F}, \rho, \epsilon)$ be the ϵ-covering number of $(\mathcal{F}, d)$. It is known that $N(\mathcal{S}^{p-1}, \rho, \epsilon) \leq (1 + \frac{2}{\epsilon})^p$.

    By the decomposition

$$\{\boldsymbol{\alpha} \in \mathcal{S}^{p-1} : |\boldsymbol{\alpha}|_0 = s\} = \bigcup_{S \subseteq [p] : |S| = s} \{\boldsymbol{\alpha} \in \mathcal{S}^{p-1} : \ \mathrm{supp}(\boldsymbol{\alpha}) = S\}$$

    and the binomial coefficient bound $\binom{p}{s} \leq (\frac{ep}{s})^s$, we have

$$N(\mathcal{F}, \rho, \epsilon) \leq \binom{p}{s} (1 + 2/\epsilon)^s \leq \left\{ \frac{(2+\epsilon)ep}{s\epsilon} \right\}^s.$$

- Step 2.1 (Continued). For $\epsilon \in (0,1)$ and $S \subseteq [p]$ fixed, let $\mathcal{N}_{S,\epsilon}$ be an ϵ-net of the unit ball in $(\mathbb{R}^S, \rho)$ with $|\mathcal{N}_{S,\epsilon}| \leq (1+\frac{2}{\epsilon})^s$. Then

$$\mathcal{N}_\epsilon := \bigcup_{S\subseteq[p]} \mathcal{N}_{S,\epsilon} = \{\mathbf{x} \mapsto \langle \boldsymbol{\alpha}, \mathbf{x} \rangle : \boldsymbol{\alpha} \in \mathcal{N}_{S,\epsilon}, S \subseteq [p]\}$$

forms an ϵ-net of $(\mathcal{F}, \rho)$, satisfying $d = |\mathcal{N}_\epsilon| \leq \binom{p}{s}(1+\frac{2}{\epsilon})^s$.

- Step 2.1 (Continued). For $\epsilon \in (0,1)$ and $S \subseteq [p]$ fixed, let $\mathcal{N}_{S,\epsilon}$ be an ϵ-net of the unit ball in $(\mathbb{R}^S, \rho)$ with $|\mathcal{N}_{S,\epsilon}| \leq (1+\frac{2}{\epsilon})^s$. Then

$$\mathcal{N}_\epsilon := \bigcup_{S\subseteq[p]} \mathcal{N}_{S,\epsilon} = \{\mathbf{x} \mapsto \langle \boldsymbol{\alpha}, \mathbf{x}\rangle : \boldsymbol{\alpha} \in \mathcal{N}_{S,\epsilon}, S \subseteq [p]\}$$

 forms an ϵ-net of $(\mathcal{F}, \rho)$, satisfying $d = |\mathcal{N}_\epsilon| \leq \binom{p}{s}(1+\frac{2}{\epsilon})^s$. For every $\boldsymbol{\alpha} \in \mathcal{F}$ with $\mathrm{supp}(\boldsymbol{\alpha}) = S$, there exists $\boldsymbol{\alpha}' \in \mathcal{N}_{S,\epsilon}$ satisfying (i) $\mathrm{supp}(\boldsymbol{\alpha}') = \mathrm{supp}(\boldsymbol{\alpha})$; (ii) $|\boldsymbol{\alpha} - \boldsymbol{\alpha}'|_2 \leq \epsilon$; and (iii)

$$|\boldsymbol{\alpha}_{\boldsymbol{\Sigma}} - \boldsymbol{\alpha}'_{\boldsymbol{\Sigma}}|_{\boldsymbol{\Sigma}} \leq \sqrt{\frac{\phi_{\max}(s)}{\phi_{\min}(s)}}\,|\boldsymbol{\alpha} - \boldsymbol{\alpha}'|_2.$$

- Step 2.1 (Continued). For $\epsilon \in (0,1)$ and $S \subseteq [p]$ fixed, let $\mathcal{N}_{S,\epsilon}$ be an ϵ-net of the unit ball in $(\mathbb{R}^S, \rho)$ with $|\mathcal{N}_{S,\epsilon}| \leq (1+\frac{2}{\epsilon})^s$. Then

$$\mathcal{N}_\epsilon := \bigcup_{S\subseteq[p]} \mathcal{N}_{S,\epsilon} = \{\mathbf{x} \mapsto \langle \boldsymbol{\alpha}, \mathbf{x}\rangle : \boldsymbol{\alpha} \in \mathcal{N}_{S,\epsilon}, S \subseteq [p]\}$$

forms an ϵ-net of $(\mathcal{F}, \rho)$, satisfying $d = |\mathcal{N}_\epsilon| \leq \binom{p}{s}(1+\frac{2}{\epsilon})^s$. For every $\boldsymbol{\alpha} \in \mathcal{F}$ with $\mathrm{supp}(\boldsymbol{\alpha}) = S$, there exists $\boldsymbol{\alpha}' \in \mathcal{N}_{S,\epsilon}$ satisfying (i) $\mathrm{supp}(\boldsymbol{\alpha}') = \mathrm{supp}(\boldsymbol{\alpha})$; (ii) $|\boldsymbol{\alpha} - \boldsymbol{\alpha}'|_2 \leq \epsilon$; and (iii)

$$|\boldsymbol{\alpha}_{\boldsymbol{\Sigma}} - \boldsymbol{\alpha}'_{\boldsymbol{\Sigma}}|_{\boldsymbol{\Sigma}} \leq \sqrt{\frac{\phi_{\max}(s)}{\phi_{\min}(s)}}\, |\boldsymbol{\alpha} - \boldsymbol{\alpha}'|_2.$$

For any $\epsilon \in (0, \gamma_s^{-1})$,

$$\max_{\boldsymbol{\alpha}\in\mathcal{N}_\epsilon} \langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{W}_n\rangle \leq L_n \leq (1-\gamma_s\epsilon)^{-1} \cdot \max_{\boldsymbol{\alpha}\in\mathcal{N}_\epsilon} \langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{W}_n\rangle,$$

where $W_n = n^{-1}\sum_{i=1}^n \mathbf{y}_i$.

- Step 2.2 (Coupling inequality). Write $\mathcal{F}_\epsilon = \{\boldsymbol{\alpha}_j : j = 1, \ldots, d\}$ and define the d-dimensional Gaussian random vector $\mathbf{G} = (G_1, \ldots, G_d)^T$, where

$$G_j = \langle \boldsymbol{\alpha}_{j,\boldsymbol{\Sigma}}, \mathbf{Z} \rangle = \langle \boldsymbol{\alpha}_j / |\boldsymbol{\alpha}_j|_{\boldsymbol{\Sigma}}, \mathbf{Z} \rangle, \ \ j = 1, \ldots, d.$$

- Step 2.2 (Coupling inequality). Write $\mathcal{F}_\epsilon = \{\boldsymbol{\alpha}_j : j = 1, \ldots, d\}$ and define the d-dimensional Gaussian random vector $\mathbf{G} = (G_1, \ldots, G_d)^T$, where

$$G_j = \langle \boldsymbol{\alpha}_{j,\boldsymbol{\Sigma}}, \mathbf{Z} \rangle = \langle \boldsymbol{\alpha}_j / |\boldsymbol{\alpha}_j|_{\boldsymbol{\Sigma}}, \mathbf{Z} \rangle, \ \ j = 1, \ldots, d.$$

Based on a recent result of Chernozhukov, Chetverikov and Kato (2014), we show that there exists a random variable $T_\epsilon^* \overset{d}{=} \max_{1 \le j \le d} G_j$ such that, for every $\delta > 0$ and $\epsilon \in (0,1)$,

$$P\left(\left| \max_{\boldsymbol{\alpha} \in \mathcal{N}_\epsilon} \langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{W}_n \rangle - T_\epsilon^* \right| \ge 16\delta \right)$$
$$\lesssim v_4^{1/2} K_1^2 \frac{c_{n,\epsilon}^{3/2}(s,p)}{\delta^2 \sqrt{n}} + v_3 K_1^3 \frac{c_{n,\epsilon}^2(s,p)}{\delta^3 \sqrt{n}} + v_4 K_1^4 \frac{c_{n,\epsilon}^5(s,p)}{\delta^4 n}$$
$$+ (K_0 K_1)^2 \frac{c_{n,\epsilon}^4(s,p)}{\delta^2 n} + (K_0 K_1)^3 \frac{c_{n,\epsilon}^6(s,p)}{\delta^3 n^{3/2}} + \frac{\log n}{n},$$

where $c_{n,\epsilon}(s,p) := s \log \frac{ep}{s\epsilon} \vee \log n$ and $v_q = (E|\varepsilon|^q)^{1/q}$.

- Step 2.3 (Gaussian supremum). There exits a random variable $T^* \overset{d}{=} \sup_{\boldsymbol{\alpha}\in\mathcal{F}} \langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{Z} \rangle$ such that

$$P\Big\{ \big| T^* - T^*_\epsilon \big| > C\,\gamma_s\,\epsilon\,c_n^{1/2}(s,p) \Big\} \leq n^{-1}.$$

- Step 2.3 (Gaussian supremum). There exits a random variable $T^* \stackrel{d}{=} \sup_{\boldsymbol{\alpha}\in\mathcal{F}} \langle \boldsymbol{\alpha}_{\boldsymbol{\Sigma}}, \mathbf{Z}\rangle$ such that

$$P\left\{\left|T^* - T^*_\epsilon\right| > C\,\gamma_s\,\epsilon\,c_n^{1/2}(s,p)\right\} \leq n^{-1}.$$

▶ Step 3. (Concentration and anti-concentration)

Lemma

Let $R^*(s,p) = \sup_{\boldsymbol{\alpha}\in\mathcal{F}} \frac{\langle \boldsymbol{\alpha},\mathbf{Z}\rangle}{\sqrt{\boldsymbol{\alpha}^T\boldsymbol{\Sigma}\boldsymbol{\alpha}}}$, where $\mathbf{Z} \stackrel{d}{=} N(\mathbf{0},\boldsymbol{\Sigma})$. Under Condition (2), there exists an absolute constant $C > 0$ such that, for every $p \geq 2$, $1 \leq s \leq p$ and $t > 0$,

$$P\left\{R^*(s,p) \geq C\sqrt{s\log\frac{\gamma_s p}{s}} + t\right\} \leq e^{-t^2/2}$$

and

$$\sup_{x\geq 0} P\left\{|R^*(s,p) - x| \leq t\right\} \leq C\,t\sqrt{s\log\frac{\gamma_s p}{s}}.$$

THANK YOU !!!