Chernozhukov, Victor; Chetverikov, Denis; Kato, Kengo

Working Paper

Central limit theorems and multiplier bootstrap when p is much larger than n

cemmap working paper, No. CWP45/12

Provided in Cooperation with:
The Institute for Fiscal Studies (IFS), London

Suggested Citation: Chernozhukov, Victor; Chetverikov, Denis; Kato, Kengo (2012) : Central limit theorems and multiplier bootstrap when p is much larger than n, cemmap working paper, No. CWP45/12, Centre for Microdata Methods and Practice (cemmap), London, https://doi.org/10.1920/wp.cem.2012.4512

This Version is available at:
https://hdl.handle.net/10419/79523

cemmap
centre for microdata methods and practice

Central limit theorems and multiplier bootstrap when *p* is much larger than *n*

Victor Chernozhukov
Denis Chetverikov
Kengo Kato

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP45/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

CENTRAL LIMIT THEOREMS AND MULTIPLIER BOOTSTRAP WHEN p IS MUCH LARGER THAN n

VICTOR CHERNOZHUKOV, DENIS CHETVERIKOV, AND KENGO KATO

ABSTRACT. We derive a central limit theorem for the maximum of a sum of high dimensional random vectors. More precisely, we establish conditions under which the distribution of the maximum is approximated by the maximum of a sum of the Gaussian random vectors with the same covariance matrices as the original vectors. The key innovation of our result is that it applies even if the dimension of random vectors (p) is much larger than the sample size (n). In fact, the growth of p could be exponential in some fractional power of n. We also show that the distribution of the maximum of a sum of the Gaussian random vectors with unknown covariance matrices can be estimated by the distribution of the maximum of the (conditional) Gaussian process obtained by multiplying the original vectors with i.i.d. Gaussian multipliers. We call this procedure the "multiplier bootstrap". Here too, the growth of p could be exponential in some fractional power of n. We prove that our distributional approximations, either Gaussian or conditional Gaussian, yield a high-quality approximation for the distribution of the original maximum, often with at most a polynomial approximation error. These results are of interest in numerous econometric and statistical applications. In particular, we demonstrate how our central limit theorem and the multiplier bootstrap can be used for high dimensional estimation, multiple hypothesis testing, and adaptive specification testing. All of our results contain non-asymptotic bounds on approximation errors.

1. INTRODUCTION

Consider random variable T_0 defined by

$$T_0 := \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{n} x_{ij}/\sqrt{n},$$

where $(x_i)_{i=1}^n$ is a sequence of independent zero-mean random p-vectors of observations, x_{ij} is the jth component of vector x_i, and $p \geqslant 2$. The distribution of T_0 is of interest in many statistical applications. When p is much smaller than n, this distribution can be approximated by using a classical

Date: First version: June 2012. This version: December 8, 2012.

Key words and phrases. Dantzig selector, Slepian, Stein method, maximum of vector sums, high dimensionality, anti-concentration, spin glasses.

V. Chernozhukov and D. Chetverikov are supported by a National Science Foundation grant. K. Kato is supported by the Grant-in-Aid for Young Scientists (B) (22730179), the Japan Society for the Promotion of Science.

Central Limit Theorem (CLT). In many high-dimensional problems, however, p is comparable or even larger than n, and the classical CLT does not apply. This paper provides tractable approximations to the distribution of T_0 when p is possibly much larger than n.

Specifically, we derive a new Gaussian approximation theorem that gives a bound on the Kolmogorov-Smirnov distance between the distributions of T_0 and its Gaussian analog Z_0:

$$Z_0 := \max_{1 \leqslant j \leqslant p} \sum_{i=1}^{n} y_{ij}/\sqrt{n},$$
$$y_i := (y_{i1}, \ldots, y_{ip})' \sim N(0, \mathrm{E}[x_i x_i']),$$

i.e., random variable Z_0 is the maximum of the normalized sum of Gaussian random vectors y_i having the same covariance structure as x_i's. We show that under mild moment assumptions, there exist some constants $c > 0$ and $C > 0$ such that

$$\rho := \sup_{t \in \mathbb{R}} |\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)| \leqslant C n^{-c} \to 0, \tag{1}$$

as $n \to \infty$, where p could grow as fast as an exponential of some fractional power of n. For example, if x_{ij} are uniformly bounded (i.e., $|x_{ij}| \leqslant C_1$ for some constant $C_1 > 0$ for all i, j), the distance ρ converges to zero at a polynomial rate whenever $(\log p)^7/n \to 0$ at a polynomial rate. Similar results are also obtained when x_{ij} are sub-Gaussian and even non-sub-Gaussian, under some assumptions that restrict the growth of $\max_{1 \leqslant j \leqslant p, 1 \leqslant i \leqslant n} \mathrm{E}[|x_{ij}|^4]$. Figure 1 gives a graphical illustration of the result (1), motivated by the problem of tuning the Dantzig selector of [12] to non-Gaussian settings, an example which we examine in Section 5.

In the process of deriving our results, we employ various tools, including Slepian's "smart path" interpolation (which is related to the solution of Stein's partial differential equation), Stein's leave-one-out method, approximation of maxima by the smooth functions (related to "free energy" in spin glasses), and exponential inequalities for self-normalized sums. See, e.g., [42, 44, 21, 17, 45, 13, 14, 20, 37] for introduction and discussion of some of these tools. The main result also critically relies on the anti-concentration inequality for suprema of Gaussian processes, which is derived in [18] and restated as Lemma 2.1.

Our result (1) has the following innovative features. To the best of our knowledge, our result is the first to show that maxima of sums of random vectors, with general covariance structure, can be approximated in distribution by the maxima of sums of Gaussian random vectors when $p \gg n$, in particular, when p can depend exponentially on a fractional power of n. This condition is weaker than the one that results from the use of Yurinskii's coupling, which also implies (1) but under the rather strong condition $p^5/n \to 0$; see, in particular, Lemma 2.12 in [22] and Example 17 (Section 10) in [38]. Second, note that our analysis specifically covers cases where

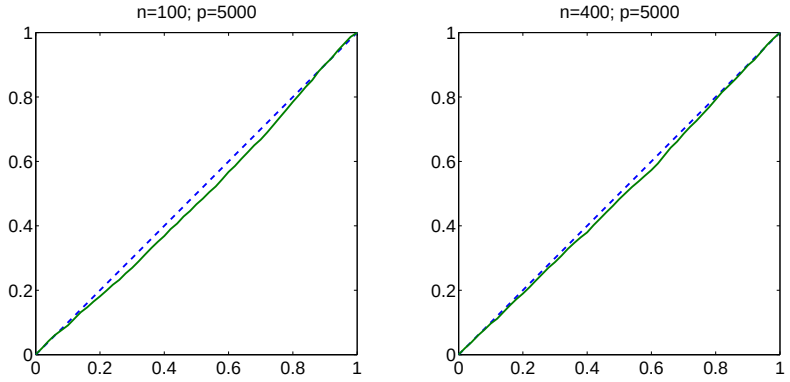

FIGURE 1. P-P plots comparing distributions of T_0 and Z_0 in the example motivated by the Dantzig estimation problem. Here, $x_{ij} = z_{ij}\varepsilon_i$ where $\varepsilon_i \sim t(4)$, (a t-distribution with four degrees of freedom), and z_{ij} are nonstochastic (simulated once using $U[0,1]$ distribution independently across i and j). Dashed line is 45°. The distributions of T_0 and Z_0 are close, as (qualitatively) predicted by the CLT derived in the paper: see Corollaries 2.1 or 2.2. The quality of the Gaussian approximation is particularly good for the tail probabilities, which is most relevant for practical applications.

the process $\{\sum_{i=1}^n x_{ij}/\sqrt{n}, 1 \leqslant j \leqslant p\}$ is not asymptotically equicontinuous and hence is not Donsker. Indeed, otherwise our result would follow from the classical functional central limit theorems for empirical processes, as in [21]. Third, the quality of approximation in (1) is of polynomial order in n, which is better than the logarithmic in n quality that we could obtain in some (though not all) nonparametric applications where the behavior of the maximum T_0 (after a suitable rescaling) could be approximated by the extreme value distribution (as, e.g., in [43] and [10]).

Our result also contributes to the literature on multivariate central limit theorems, which are concerned with conditions under which

$$\left| \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \in A \right) - \mathrm{P}\left(\frac{1}{\sqrt{n}} \sum_{i=1}^n y_i \in A \right) \right| \to 0, \tag{2}$$

uniformly in a collection of sets A, typically *all* convex sets. The results of this kind were developed among others, by [36, 39, 27, 8, 16], under the conditions that $p^c/n \to 0$ or similar conditions (also see [15]). These results rely on the anti-concentration results for Gaussian random vectors on the δ-expansions of boundaries of arbitrary convex sets A (see [5]). Note that our result also establishes (2), but uniformly for all convex sets of the form $A_{\max} = \{a \in \mathbb{R}^p : \max_{1 \leqslant j \leqslant p} a_j \leqslant t\}$ for $t \in \mathbb{R}$. These sets have a rather special structure that allows us to deal with $p \gg n$: in particular,

concentration of measure on the δ-expansion of boundary of $A_{\max}$ is at most of order $\delta\sqrt{\log p}$ for Gaussian random p-vectors with unit variance, as shown in [18] (restated as Lemma 2.1).

Note that the result (1) is immediately useful for inference with statistic T_0, even though $\mathrm{P}(Z_0 \leqslant t)$ needs not converge itself to a well behaved distribution function. Indeed, if the covariance matrix $n^{-1}\sum_{i=1}^n \mathrm{E}[x_i x_i']$ is known, then $c_{Z_0}(1-\alpha) := (1-\alpha)$-quantile of Z_0, can be computed numerically, and we have

$$|\mathrm{P}(T_0 \leqslant c_{Z_0}(1-\alpha)) - (1-\alpha)| \leqslant Cn^{-c} \to 0. \tag{3}$$

A chief application of this kind arises in determination of the penalty level for the Dantzig selector of [12] in the high-dimensional regression with non-Gaussian errors, which we examine in Section 5. There, under the canonical (homoscedastic) noise, the covariance matrix is known, and so quantiles of Z_0 can be easily computed numerically and used for choosing the penalty level. However, if the noise is heteroscedastic, the covariance matrix is no longer known, and this approach is no longer feasible. This motivates our second main result.

The second main result of the paper establishes validity of the multiplier bootstrap for estimating quantiles of Z_0, when the covariance matrix $n^{-1}\sum_{i=1}^n \mathrm{E}[x_i x_i']$ is unknown. More precisely, we define the Gaussian-symmetrized version W_0 of T_0 by multiplying x_i with i.i.d. standard Gaussian random variables $e_1, \ldots, e_n$:

$$W_0 := \max_{1\leqslant j\leqslant p} \sum_{i=1}^n x_{ij} e_i/\sqrt{n}. \tag{4}$$

We show that the conditional quantiles of W_0 given data $(x_i)_{i=1}^n$ consistently estimate the quantiles of Z_0 and hence those of T_0 (where the notion of consistency used is the one that guarantees asymptotically valid inference). Here the primary factor driving the bootstrap estimation error is the maximum difference between the empirical and population covariance matrices:

$$\Delta := \max_{1\leqslant j,k\leqslant p} \left| \frac{1}{n}\sum_{i=1}^n (x_{ij}x_{ik} - \mathrm{E}[x_{ij}x_{ik}]) \right|,$$

which can converge to zero even when p is much larger than n. For example, when x_{ij} are uniformly bounded, the multiplier bootstrap is valid for inference if $(\log p)^7/n \to 0$. Earlier related results on bootstrap in the "$p \to \infty$ but $p/n \to 0$" regime were studied in [35]; interesting results for the case $p \gg n$ based on concentration inequalities and symmetrization are studied in [3, 4], albeit the approach and results are quite different from those given here. In particular, in [3], either Gaussianity or symmetry in distribution is imposed on the data.

As a part of establishing the results on the multiplier bootstrap, we derive a bound on the Kolmogorov-Smirnov distance between distributions of

maxima of two finite dimensional Gaussian random vectors, which again depends on the maximum difference between the covariance matrices of the vectors. The key property of our bound again is that it depends on the dimension p of random vectors only via $\log p$. This result is of independent interest, and extends and complements the work of [14] that derived an explicit Sudakov-Fernique type bound on the difference of expected values of the same quantities; see also [1], Chapter 2.

The key motivating example for our results is the high-dimensional sparse regression model. In this model, [12] and [9] assume Gaussian errors to analyze the Dantzig selector and Lasso. Our results show that Gaussianity is not necessary and the Gaussian-like conclusions hold approximately, with just the fourth moment of the regression errors being bounded. Moreover, our approximation allows to take into account correlations between different components of x_i's. This results in a better choice of the penalty level, and in establishing sharper bounds on performance than those that had been available previously. Note that some of the same goals had been accomplished using moderate deviations for self-normalized sums, combined with the union bound [7]. The limitation, however, is that the union bound does not take into account correlations between different components of x_i's, and so it may be overly conservative in some applications.

Our results have a broad range of other statistical applications. In addition to the high-dimensional estimation example, we show how to apply our result in the multiple hypothesis testing framework of multivariate linear regression. We prove the validity of the stepdown procedure developed in [41] when the critical value is obtained through the multiplier bootstrap. Notably, the number of hypotheses to be tested can be much larger than the sample size. Finally, in the third example, we develop a new specification test for the null hypothesis of the linear regression model and a general nonparametric alternative, which uses a number of moment conditions that is much larger than the sample size. Lastly, in a companion work, [18] and [19], we are exploring the strong coupling for suprema of general empirical processes, based on the methods developed here and maximal inequalities. These results represent a useful complement to the results based on the Hungarian coupling developed by [33, 11, 31, 40] for the entire empirical process. These results have applications to uniform confidence bands in nonparametric regression; see, e.g., [26].

The rest of the paper is organized as follows. In Section 2 we give the results on Gaussian approximation and associated comparison theorems. In Section 3 we give the results on Gaussian comparison and associated multiplier comparison theorems. In Section 4 we provide the results on the multiplier bootstrap. In Sections 5, 6, and 7 we consider the three substantive applications. Appendices contain proofs for each of the sections, with Appendix A stating auxiliary tools and lemmas.

1.1. **Notation.** Throughout the paper, $\mathbb{E}_n[\cdot]$ denotes the average over index $1 \leqslant i \leqslant n$, i.e., it simply abbreviates the notation $n^{-1}\sum_{i=1}^n[\cdot]$. For example, $\mathbb{E}_n[x_{ij}^2] = n^{-1}\sum_{i=1}^n x_{ij}^2$. In addition, $\bar{\mathrm{E}}[\cdot] = \mathbb{E}_n[\mathrm{E}[\cdot]]$. For example, $\bar{\mathrm{E}}[x_{ij}^2] = n^{-1}\sum_{i=1}^n \mathrm{E}[x_{ij}^2]$. For a function $f : \mathbb{R} \to \mathbb{R}$, we write $\partial^k f(x) = \partial^k f(x)/\partial x^k$ for nonnegative integer k; for a function $f : \mathbb{R}^p \to \mathbb{R}$, we write $\partial_j f(x) = \partial f(x)/\partial x_j$ for $j = 1, \ldots, p$, where $x = (x_1, \ldots, x_p)'$. Denote by $C^k(\mathbb{R})$ the class of k times continuously differentiable functions from $\mathbb{R}$ to itself, and denote by $C_b^k(\mathbb{R})$ the class of all functions $f \in C^k(\mathbb{R})$ such that $\sup_{z\in\mathbb{R}} |\partial^j f(z)| < \infty$ for $j = 0, \ldots, k$. We write $a \lesssim b$ if a is smaller than or equal to b up to a universal positive constant. For a given parameter q, we also write $a \lesssim_q b$ if there is a constant $C = C(q)$ depending only on q such that $a \leqslant Cb$. The same rule applies when there are multiple parameters. For example, we will sometime write "$a \lesssim_{c_1,C_1} b$", and this means that there exists a constant $C > 0$ depending only on c_1 and C_1 such that $a \leqslant Cb$.

2. Central Limit Theorems for Maxima of Non-Gaussian Sums

2.1. **Comparison Theorems and Non-Asymptotic Gaussian Approximations.** Let $x_i = (x_{ij})_{j=1}^p$ be a zero-mean p-vector of random variables and let $y_i = (y_{ij})_{j=1}^p$ be a p-vector of Gaussian random variables such that

$$y_i \sim N(0, \mathrm{E}[x_i x_i']).$$

Consider sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ of *independent* vectors, where independence holds across the i index. Let

$$X := (X_1, \ldots, X_p)' := \frac{1}{\sqrt{n}}\sum_{i=1}^n x_i \quad \text{and} \quad Y := (Y_1, \ldots, Y_p)' := \frac{1}{\sqrt{n}}\sum_{i=1}^n y_i.$$

The purpose of this section is to compare and bound the difference between the expectations and distribution functions of the non-Gaussian to the Gaussian maxima:

$$\max_{1\leqslant j\leqslant p} X_j \quad \text{and} \quad \max_{1\leqslant j\leqslant p} Y_j.$$

This problem is of intrinsic difficulty since the maximum function $z = (z_1, \ldots, z_p)' \mapsto \max_{1\leqslant j\leqslant p} z_j$ is non-differentiable. To circumvent the problem, we use a smooth approximation of the maximum function.

For $z = (z_1, \ldots, z_p)' \in \mathbb{R}^p$, consider the function:

$$F_\beta(z) := \beta^{-1}\log\left(\sum_{j=1}^p \exp(\beta z_j)\right),$$

which approximates the maximum function, where $\beta > 0$ is the smoothing parameter that controls the level of approximation (we call this function the

"smooth max function"). Indeed, an elementary calculation shows that for all $z \in \mathbb{R}^p$,

$$0 \leqslant F_\beta(z) - \max_{1 \leqslant j \leqslant p} z_j \leqslant \beta^{-1} \log p. \tag{5}$$

This smooth max function arises in the definition of "free energy" in spin glasses, see, e.g. [45].

We start with the following "warm-up" theorem that conveys main features and ideas of the proof. Here and in what follows, for a smooth function $g : \mathbb{R} \to \mathbb{R}$, write

$$G_k := \sup_{z \in \mathbb{R}} |\partial^k g(z)|, \ k \geqslant 0.$$

Theorem 2.1 (Comparison of Gaussian to Non-Gaussian Maxima, I)**.** *For every $g \in C_b^3(\mathbb{R})$ and every $\beta > 0$,*

$$\begin{aligned} &|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \\ &\quad \lesssim (G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[\max_{1 \leqslant j \leqslant p}(|x_{ij}|^3 + |y_{ij}|^3)]/\sqrt{n}, \end{aligned}$$

and hence

$$\begin{aligned} &|\mathrm{E}[g(\max_{1 \leqslant j \leqslant p} X_j) - g(\max_{1 \leqslant j \leqslant p} Y)]| \\ &\quad \lesssim (G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}[\max_{1 \leqslant j \leqslant p}(|x_{ij}|^3 + |y_{ij}|^3)]/\sqrt{n} + \beta^{-1} G_1 \log p. \end{aligned}$$

The theorem first bounds the difference between the expectations of functions of the smooth max functions, and then bounds the difference between the expectations of functions of the original maxima.

To state the next result we need the following anti-concentration lemma.

Lemma 2.1 (Anti-Concentration)**.** *Let $\xi_1, \ldots, \xi_p$ be (not necessarily independent) $N(0, \sigma_j^2)$ random variables ($p \geqslant 2$) with $\sigma_j^2 > 0$ for all $1 \leqslant j \leqslant p$. Let $\underline{\sigma} = \min_{1 \leqslant j \leqslant p} \sigma_j$ and $\bar{\sigma} = \max_{1 \leqslant j \leqslant p} \sigma_j$. Then for every $u \in (0, 1)$,*

$$\sup_{z \in \mathbb{R}} \mathrm{P}\left(|\max_{1 \leqslant j \leqslant p} \xi_j - z| \leqslant u \right) \lesssim_{\underline{\sigma}, \bar{\sigma}} u\sqrt{\log(p/u)}.$$

When σ_j are all equal, $\sqrt{\log(p/u)}$ on the right side can be replaced by $\sqrt{\log p}$.

The lemma is a special case of Theorem 1 in [18]. Combining Theorem 2.1 and Lemma 2.1 leads to a bound on the Kolmogorov-Smirnov distance between the distribution functions of $\max_{1 \leqslant j \leqslant p} X_j$ and $\max_{1 \leqslant j \leqslant p} Y_j$.

Corollary 2.1 (Central Limit Theorem, I)**.** *Suppose that there are some constants $c_1 > 0$ and $C_1 > 0$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$.*

Then

$$\rho := \sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\max_{1\leqslant j\leqslant p} X_j \leqslant t \right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p} Y_j \leqslant t \right) \right| \lesssim_{c_1,C_1} (\log(pn))^{7/8} \bar{\mathrm{E}} \left[\max_{1\leqslant j\leqslant p} (|x_{ij}|^3 + |y_{ij}|^3)/\sqrt{n} \right]^{1/4}.$$

Comment 2.1 (Main qualitative feature: logarithmic dependence on p)**.** It is well known since the work of [22] that the discrete empirical process can be approximated by the Gaussian process with the same covariance functions with the error bound depending *polynomially* on p. In many cases, however, the main interest is in approximating the supremum of the process, and the approximation of the whole process is not required. Theorem 2.1 and Corollary 2.1 imply that the error of approximating the supremum of the empirical process by the supremum of the Gaussian process depends on p only through slowly growing, *logarithmic* term $\log p$. This is the main qualitative feature of all results presented in this paper. □

While Theorem 2.1 and Corollary 2.1 convey an important qualitative aspect of the problem and admit easy-to-grasp proofs, an important disadvantage of these results is that the bounds depend on the maximum over $1 \leqslant j \leqslant p$ inside the expectation:

$$\bar{\mathrm{E}} \left[\max_{1\leqslant j\leqslant p} (|x_{ij}|^3 + |y_{ij}|^3)/\sqrt{n} \right],$$

which may be unnecessarily large in some applications. Using a truncation method in conjunction with the proof strategy of Theorem 2.1, we show in Theorem 2.2 below that we can take the maximum out of the expectation, and hence replace the above term by

$$\max_{1\leqslant j\leqslant p} \bar{\mathrm{E}} \left[(|x_{ij}|^3 + |y_{ij}|^3)/\sqrt{n} \right].$$

This greatly improves the bound appearing in Theorem 2.1 and Corollary 2.1 in some applications. The improvement here comes at a cost of a more involved statement and more delicate regularity conditions involving parameters used in the truncation method.

The truncation method we employ is described as follows. Given a threshold level $u > 0$, define a truncated version of x_{ij} by

$$\tilde{x}_{ij} = x_{ij} 1\left\{ |x_{ij}| \leqslant u\sqrt{\mathrm{E}\left[x_{ij}^2\right]} \right\} - \mathrm{E}\left[x_{ij} 1\left\{ |x_{ij}| \leqslant u\sqrt{\mathrm{E}[x_{ij}^2]} \right\} \right]. \tag{6}$$

Let $\varphi(u)$ be the infimum, which is attained, over all numbers $\varphi \geqslant 0$ such that for all $1 \leqslant j \leqslant p$ and $1 \leqslant i \leqslant n$,

$$\left(\mathrm{E}\left[x_{ij}^2 1\left\{ |x_{ij}| > u\sqrt{\mathrm{E}[x_{ij}^2]} \right\} \right] \right)^{1/2} \leqslant \sqrt{\mathrm{E}[(x_{ij})^2]}\varphi. \tag{7}$$

Note that the function $\varphi(u)$ is right-continuous. Finally, given a parameter $\gamma \in (0,1)$, define $\delta(u,\gamma)$ as the infimum, which is attained, over all $\delta \geqslant 0$ such that with probability at least $1-\gamma$, for all $1 \leqslant j \leqslant p$,

$$\sqrt{\mathbb{E}_n[(x_{ij}-\tilde{x}_{ij})^2]} \leqslant \sqrt{\bar{\mathrm{E}}[(x_{ij})^2]}\varphi(u)(1+\delta). \tag{8}$$

The truncation construction is done so as to invoke sub-Gaussian tail inequalities for self-normalized sums (see Lemma A.8).

Comment 2.2 (On truncation)**.** To illustrate the truncation method at work, consider the following trivial examples.

(a) Suppose x_{ij} are zero-mean and bounded from above in absolute value by a positive constant $C_1 > 0$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$. Let $u = C_1/\min_{i,j}\sqrt{\mathrm{E}[x_{ij}^2]}$. Then $\varphi(u) = 0$ and $\delta(u,\gamma) = 0$ for all $\gamma \in (0,1)$.

(b) Suppose that x_{ij} are zero-mean sub-Gaussian with parameter $C_1 > 0$, i.e., $\mathrm{P}(|x_{ij}| \geqslant u) \leqslant \exp(1-u^2/C_1^2)$ for all $u \geqslant 0$, and $\mathrm{E}[x_{ij}^2] > 0$ uniformly over $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$. Take $u = C_2(\log(pn))^{1/2}$ for a sufficiently large constant $C_2 > 0$, and $\gamma = 1/n$. Then by Lemma A.9 in the Appendix, $\varphi(u) \leqslant 1/(pn)^2$. By the union bound, with probability at least $1-\gamma$, $|x_{ij}| \leqslant u(\mathrm{E}[x_{ij}^2])^{1/2}$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$, so that $\delta(u,\gamma) = 0$. □

Define

$$M_2 := \max_{1\leqslant j\leqslant p}(\bar{\mathrm{E}}[x_{ij}^2])^{1/2} \text{ and } M_3 := \max_{1\leqslant j\leqslant p}(\bar{\mathrm{E}}[x_{ij}^3])^{1/3}.$$

Let $\phi(z)$ and $\Phi(z)$ denote the density and distribution functions of the standard Gaussian distribution, respectively. Also define $\varphi_N(u) \geqslant 0$ by

$$\varphi_N^2(u) = \int_{|z|\geqslant u} z^2\phi(z)dz.$$

An elementary calculation leads to

$$\varphi_N^2(u) \leqslant (u^2+2)\exp(1-u^2/2).$$

See Lemma A.9. Here is the main theorem of this section.

Theorem 2.2 (Comparison of Gaussian to Non-Gaussian Maxima, II)**.** *Let $\beta > 0$, $u > 0$ and $\gamma \in (0,1)$ be such that $2\sqrt{2}u\beta/\sqrt{n} \leqslant 1$ and $u \geqslant \sqrt{2\log(2pn/\gamma)}$. Then for every $g \in C_b^3(\mathbb{R})$,*

$$|\mathrm{E}[g(F_\beta(X)) - g(F_\beta(Y))]| \lesssim D_n(g,\beta,u,\gamma),$$

and hence

$$|\mathrm{E}[g(\max_{1\leqslant j\leqslant p} X_j) - g(\max_{1\leqslant j\leqslant p} Y_j)]| \lesssim D_n(g,\beta,u,\gamma) + \beta^{-1}G_1\log p,$$

where

$$D_n(g,\beta,u,\gamma) := (G_3+G_2\beta+G_1\beta^2)M_3^3/\sqrt{n} + (G_2+\beta G_1)M_2^2(\varphi(u)+\varphi_N(u)) + G_1M_2\varphi(u)(1+\delta(u,\gamma))\sqrt{\log(p/\gamma)} + G_0\gamma.$$

Combining Theorem 2.2 and Lemma 2.1 (anti-concentration inequality) leads to a bound on the Kolmogorov-Smirnov distance between the distribution functions of $\max_{1\leqslant j\leqslant p} X_j$ and $\max_{1\leqslant j\leqslant p} Y_j$. To state the bound on the Kolmogorov-Smirnov distance in a clean form, we prepare the additional notation. For a given $\gamma \in (0,1)$, let $u_1 = n^{3/8}M_3^{3/4}/(\log(pn/\gamma))^{5/8}$, $u_2 = \sqrt{2\log(2pn/\gamma)}$, and let $u_3 \geqslant 0$ be the smallest number $u \geqslant 0$ such that

$$\sqrt{n}(\varphi(u) + \varphi_N(u))^{1/3} \leqslant u(\log(pn/\gamma))^{5/6}. \tag{9}$$

Note that u_3 is well-defined because both $\varphi(u)$ and $\varphi_N(u)$ are right-continuous and decreasing.

Corollary 2.2 (Central Limit Theorem, II)**.** *For a given $\gamma \in (0,1)$, determine u_1, u_2 and u_3 as described above. Let $u \geqslant u_1 \vee u_2 \vee u_3$. Suppose that there are some constants $0 < c_1 < C_1$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$, and $\delta(u,\gamma) \leqslant C_1$. Then*

$$\rho := \sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\max_{1\leqslant j\leqslant p} X_j \leqslant t\right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p} Y_j \leqslant t\right)\right| \lesssim_{c_1,C_1} u(\log(pn/\gamma))^{3/2}/\sqrt{n} + \gamma.$$

2.2. **Examples of Applications.** The purpose of this subsection is to obtain bounds on ρ for various leading examples frequently encountered in applications. Under primitive conditions, it will be shown that the error ρ converges to zero at least at a polynomial rate with respect to the sample size n.

Let $c_1 > 0, c_2 > 0$ and $C_1 > 0$ be some constants, and let B_n be a sequence of positive constants. We allow for the case where $B_n \to \infty$ as $n \to \infty$. Suppose that one of the following conditions is satisfied for x_{ij} *uniformly in* $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, and $n \geqslant 1$:

(E.1) $\mathrm{P}(|x_{ij}| \geqslant u) \leqslant \exp(1 - u^2/C_1)$ for all $u \geqslant 0$, and $\mathrm{E}[x_{ij}^2] \geqslant c_1$;
(E.2) $\bar{\mathrm{E}}[\max_{1\leqslant j\leqslant p} x_{ij}^4] \leqslant C_1$ and $\mathrm{E}[x_{ij}^2] \geqslant c_1$;
(E.3) $|x_{ij}| \leqslant B_n$ and $\mathrm{E}[x_{ij}^2] \geqslant c_1$;
(E.4) $x_{ij} = z_{ij}\varepsilon_i$ with $\mathrm{P}(|\varepsilon_i| \geqslant u) \leqslant \exp(1 - u^2/C_1)$, z_{ij} are nonstochastic, $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_i^2] \geqslant c_1$;
(E.5) $x_{ij} = z_{ij}\varepsilon_i$ with $\bar{\mathrm{E}}[\varepsilon_i^4] \leqslant C_1$, z_{ij} are nonstochastic, $|z_{ij}| \leqslant B_n$, $\mathbb{E}_n[z_{ij}^2] = 1$, and $\mathrm{E}[\varepsilon_i^2] \geqslant c_1$.

The last two cases cover examples that arise in high-dimensional regression, e.g. [12], which we shall revisit later in the paper. Interestingly, these cases are also connected to spin glasses, see e.g., [45] and [37] (z_{ij} can be interpreted as generalized products of "spins" and ε_i as their random "interactions").

Corollary 2.3 (Central Limit Theorem in Leading Examples with Polynomial Error Bound)**.** *Suppose that $\mathrm{E}[x_{ij}] = 0$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$, and $\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Moreover, suppose that*

one of the conditions E.1-5 is satisfied, where under conditions E.1-2, suppose that $(\log(pn))^7/n \leqslant C_1 n^{-c_2}$*; and under conditions E.3-5, suppose that* $(\log(pn))^7 B_n^2/n \leqslant C_1 n^{-c_2}$*. Then there exist constants* $c > 0$ *and* $C > 0$ *depending only on* c_1, c_2 *and* C_1 *such that*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\max_{1\leqslant j\leqslant p} X_j \leqslant t \right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p} Y_j \leqslant t \right) \right| \leqslant Cn^{-c}.$$

3. Gaussian Comparisons and Multiplier Theorems

3.1. **A Gaussian-to-Gaussian Comparison Theorem.** Let V and Y be zero-mean Gaussian random p-vectors with covariance matrices Σ^V and Σ^Y, respectively. The purpose of this section is to give an error bound on the difference of the expectations of smooth functions and the distribution functions of

$$\max_{1\leqslant j\leqslant p} V_j \quad \text{and} \quad \max_{1\leqslant j\leqslant p} Y_j$$

in terms of p and

$$\Delta_0 := \max_{1\leqslant j,k\leqslant p} \left| \Sigma^V_{jk} - \Sigma^Y_{jk} \right|.$$

Recall that for a smooth function $g : \mathbb{R} \to \mathbb{R}$, we write $G_k := \sup_{z\in\mathbb{R}} |\partial^k g(z)|$. Let F_β be the smooth max function defined in the previous section.

Theorem 3.1 (Comparison of Gaussian Maxima)**.** *For every* $g \in C_b^2(\mathbb{R})$ *and every* $\beta > 0$*,*

$$|\mathrm{E}[g(F_\beta(V)) - g(F_\beta(Y))]| \leqslant (G_2/2 + \beta G_1)\Delta_0,$$

and hence

$$\left| \mathrm{E}\left[g\left(\max_{1\leqslant j\leqslant p} V_j \right) - g\left(\max_{1\leqslant j\leqslant p} Y_j \right) \right] \right| \leqslant (G_2/2 + \beta G_1)\Delta_0 + 2\beta^{-1} G_1 \log p.$$

Comment 3.1. Minimizing the second bound in β gives

$$\left| \mathrm{E}\left[g\left(\max_{1\leqslant j\leqslant p} V_j \right) - g\left(\max_{1\leqslant j\leqslant p} Y_j \right) \right] \right| \leqslant G_2\Delta_0/2 + 2G_1\sqrt{2\Delta_0 \log p}.$$

This result extends the work of [14], which derived the following Sudakov-Fernique type bound on the difference of the expectations of the Gaussian maxima:

$$\left| \mathrm{E}\left[\max_{1\leqslant j\leqslant p} V_j \right] - \mathrm{E}\left[\max_{1\leqslant j\leqslant p} Y_j \right] \right| \leqslant 2\sqrt{2\Delta_0 \log p}.$$

Here we give bounds on the expectations of functions of Gaussian maxima, which can be converted into bounds on the difference of the distribution functions of Gaussian maxima, by taking g as a smooth approximation to the indicator function. A slightly finer bound can be also obtained; see equation (24). □

Corollary 3.1 (Comparison of Distributions of Gaussian Maxima)**.** *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $c_1 \leqslant \Sigma^Y_{jj} \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$*. Moreover, suppose that* $\Delta_0 \leqslant 1$*. Then*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(\max_{1\leqslant j\leqslant p} V_j \leqslant t \right) - \mathrm{P}\left(\max_{1\leqslant j\leqslant p} Y_j \leqslant t \right) \right| \lesssim_{c_1,C_1} \Delta_0^{1/3} (\log(p/\Delta_0))^{2/3}.$$

Comment 3.2. The main contribution of this corollary is that the bound on the Kolmogorov-Smirnov distance between the maxima of Gaussian random p-vectors depends on p only through slowly growing $\log p$. □

3.2. **Gaussian Multiplier Theorem.** Let $(x_i)_{i=1}^n$ be a sequence of zero-mean p-vector of random variables where $x_i = (x_{ij})_{j=1}^p$, and let $(e_i)_{i=1}^n$ be a sequence of $N(0,1)$ random variables independent of $(x_i)_{i=1}^n$. Recall that we have defined

$$T_0 = \max_{1\leqslant j\leqslant p} \sum_{i=1}^n x_{ij}/\sqrt{n} \quad \text{and} \quad W_0 = \max_{1\leqslant j\leqslant p} \sum_{i=1}^n x_{ij} e_i/\sqrt{n}.$$

In this section, we derive the Kolmogorov-Smirnov distance between T_0 and W_0 where the distribution of W_0 is taken conditional on $(x_i)_{i=1}^n$. We show that the key quantity driving the difference between the two quantities is the maximum difference between empirical and population covariance matrices:

$$\Delta = \max_{1\leqslant j,k\leqslant p} \left| \mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}] \right|.$$

Another key quantity is

$$\rho = \sup_{t\in\mathbb{R}} \left| \mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t) \right|,$$

where $Z_0 = \max_{1\leqslant j\leqslant p} \sum_{i=1}^n y_{ij}/\sqrt{n}$ and $y = (y_i)_{i=1}^n$ is a sequence of independent $N(0, \mathrm{E}[x_i x_i'])$ vectors.

Corollary 3.2 (Comparison of conditional Gaussian to non-Gaussian maxima)**.** *Suppose that there are some constants* $0 < c_1 < C_1$ *such that* $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ *for all* $1 \leqslant j \leqslant p$*. Then there exists a constant* $C > 0$ *depending only on* c_1 *and* C_1 *such that for every* $\vartheta \in (0, e^{-2})$*, with probability at least* $\mathrm{P}(\Delta \leqslant \vartheta)$*,*

$$\sup_{t\in\mathbb{R}} \left| \mathrm{P}\left(T_0 \leqslant t\right) - \mathrm{P}_e\left(W_0 \leqslant t\right) \right| \leqslant \rho + C\vartheta^{1/3} (\log(p/\vartheta))^{2/3}.$$

Comment 3.3. Corollary 3.2 can be viewed as a set of new symmetrization inequalities with an explicit error bound. In the asymptotic regime where we can make the error bound asymptotically negligible, Corollary 3.2 provides a multiplier central limit theorem. □

4. Multiplier Bootstrap

4.1. **Multiplier Bootstrap Theorems.** Suppose that we have a dataset $(x_i)_{i=1}^n$ consisting of n independent zero-mean random p-vectors x_i. In this section we are interested in approximating quantiles of

$$T_0 = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n x_i / \sqrt{n} \tag{10}$$

using the multiplier bootstrap method. More precisely, let $(e_i)_{i=1}^n$ be an i.i.d. sequence of $N(0,1)$ random variables independent of $(x_i)_{i=1}^n$, and let

$$W_0 = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n x_{ij} e_i / \sqrt{n}. \tag{11}$$

Then we define the multiplier bootstrap estimator of the α-quantile of T_0 as the conditional α-quantile of W_0 given $(x_i)_{i=1}^n$, i.e.,

$$c_{W_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}_e(W_0 \leqslant t) \geqslant \alpha\},$$

where P_e is the probability measure induced by the multiplier variables $(e_i)_{i=1}^n$ holding $(x_i)_{i=1}^n$ fixed (i.e., $\mathrm{P}_e(W_0 \leqslant t) = \mathrm{P}(W_0 \leqslant t \mid (x_i)_{i=1}^n)$). The multiplier bootstrap theorem below provides a non-asymptotic bound on the bootstrap estimation error:

$$\left|\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) - \alpha\right|.$$

Before presenting the theorem, we first give a simple lemma that is used in the proof of the theorem. The lemma is also useful for power analysis. Define

$$c_{Z_0}(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}(Z_0 \leqslant t) \geqslant \alpha\},$$

where $Z_0 = \max_{1 \leqslant j \leqslant p} \sum_{i=1}^n y_{ij}/\sqrt{n}$ and $y = (y_i)_{i=1}^n$ is a sequence of independent $N(0, \mathrm{E}[x_i x_i'])$ vectors. Recall that

$$\Delta = \max_{1 \leqslant j,k \leqslant p} \left|\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]\right|.$$

Lemma 4.1 (Comparison of Quantiles, I)**.** *Suppose that there are some constants $0 < c_1 < C_1$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Then for every $\vartheta \in (0, e^{-2})$ and $\alpha \in (0,1)$,*

$$\mathrm{P}(c_{W_0}(\alpha) \leqslant c_{Z_0}(\alpha + v(\vartheta))) \geqslant \mathrm{P}(\Delta \leqslant \vartheta),$$
$$\mathrm{P}(c_{Z_0}(\alpha) \leqslant c_{W_0}(\alpha + v(\vartheta))) \geqslant \mathrm{P}(\Delta \leqslant \vartheta),$$

where $v(\vartheta) := C_2 \vartheta^{1/3} (\log(p/\vartheta))^{2/3}$ and $C_2 > 0$ is a constant depending only on c_1 and C_1.

Recall that

$$\rho = \sup_{t \in \mathbb{R}} \left|\mathrm{P}(T_0 \leqslant t) - \mathrm{P}(Z_0 \leqslant t)\right|.$$

We are now in position to state the main theorem of this section.

Theorem 4.1 (Validity of Multiplier Bootstrap, I)**.** *Suppose that there are some constants $0 < c_1 < C_1$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Then for every $\vartheta \in (0, e^{-2})$,*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) - \alpha| \leqslant \rho + C_2 \vartheta^{1/3} (\log(p/\vartheta))^{2/3} + \mathrm{P}(\Delta > \vartheta),$$

where $C_2 > 0$ is a constant that appears in Lemma 4.1.

Theorem 4.1 provides a useful result for the case where the statistics are maxima of exact averages. There are many applications, however, where the relevant statistics arise as maxima of approximate averages. The following result shows that the theorem continues to apply if the approximation error of the relevant statistic by a maximum of an exact average can be suitably controlled. Specifically, suppose that a statistic of interest, say $T = T(x_1 \ldots, x_n)$ which may not be of the form (10), can be approximated by T_0 of the form (10), and that the multiplier bootstrap is performed on a statistic $W = W(x_1, \ldots, x_n, e_1, \ldots, e_n)$, which may be different from (11) but still can be approximated by W_0 of the form (11). The "approximation" here is the following sense: there exist $\zeta_1 \geqslant 0$ and $\zeta_2 \geqslant 0$, depending on n (and typically $\zeta_1 \to 0, \zeta_2 \to 0$ as $n \to \infty$), such that

$$\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2, \tag{12}$$

$$\mathrm{P}(\mathrm{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2. \tag{13}$$

We use the α-quantile of $W = W(x_1, \ldots, x_n, e_1, \ldots, e_n)$, computed conditional on $(x_i)_{i=1}^n$:

$$c_W(\alpha) := \inf\{t \in \mathbb{R} : \mathrm{P}_e(W \leqslant t) \geqslant \alpha\},$$

as an estimate of the α-quantile of T. We are interested in establishing that the bootstrap estimation error approaches zero.

The following lemma will be helpful. The lemma is also useful for power analysis.

Lemma 4.2 (Comparison of Quantiles, II)**.** *Suppose that condition (13) is satisfied. Then for every $\alpha \in (0, 1)$,*

$$\mathrm{P}(c_W(\alpha) \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2,$$
$$\mathrm{P}(c_{W_0}(\alpha) \leqslant c_W(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2.$$

Theorem 4.2 (Validity of Multiplier Bootstrap, II)**.** *Suppose that there are some constants $0 < c_1 < C_1$ such that $c_1 \leqslant \bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$. Moreover, suppose that conditions (12) and (13) are satisfied. Then for every $\vartheta \in (0, e^{-2})$,*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha|$$
$$\leqslant \rho + C_2 \vartheta^{1/3} (\log(p/\vartheta))^{2/3} + \mathrm{P}(\Delta > \vartheta) + C_3 \zeta_1 \sqrt{\log(p/\zeta_1)} + \zeta_2,$$

where $C_2 > 0$ is a constant that appears in Lemma 4.1, and $C_3 > 0$ is a constant depending only on c_1 and C_1.

4.2. **Examples of Applications: Revisited.** Here we revisit the examples in Section 2.2 and see how the multiplier bootstrap works for these leading examples. Let, as before, $c_1 > 0, c_2 > 0$ and $C_1 > 0$ be some constants, and let B_n be a sequence of positive constants. Recall conditions E.1-5 in Section 2.2.

Corollary 4.1 (Multiplier Bootstrap in Leading Examples with Polynomial Error Bound)**.** *Suppose that $\mathrm{E}[x_{ij}] = 0$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant j \leqslant p$; $\bar{\mathrm{E}}[x_{ij}^2] \leqslant C_1$ for all $1 \leqslant j \leqslant p$; and conditions (12) and (13) are satisfied with $\zeta_1\sqrt{\log p} + \zeta_2 \leqslant C_1 n^{-c_2}$. Moreover, suppose that one of the conditions E.1-5 is satisfied, where under conditions E.1-2, suppose that $(\log(pn))^7/n \leqslant C_1 n^{-c_2}$; under conditions E.3-5, suppose that $(\log(pn))^7 B_n^2/n \leqslant C_1 n^{-c_2}$; and finally, under condition E.5, suppose that $(\log p)^6 B_n^4/n \leqslant C_1 n^{-c_2}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, c_2 and C_1 such that*

$$\sup_{\alpha \in (0,1)} |\mathrm{P}(T \leqslant c_W(\alpha)) - \alpha| \leqslant Cn^{-c}.$$

Comment 4.1. This corollary shows that the multiplier bootstrap is valid with a polynomial rate of accuracy for the significance level under very weak conditions. This is in contrast with the extremal theory of Gaussian processes that provides only a logarithmic rate of approximation (see, for example, [34]). □

5. Application: Dantzig Selector in the Non-Gaussian Model

The purpose of this section is to demonstrate the case with which the CLT and the multiplier bootstrap theorem given in Corollaries 2.3 and 4.1 can be applied in important problems, dealing with a high-dimensional inference and estimation. We consider the Dantzig selector previously studied in the path-breaking works of [12] and [9] in a Gaussian setting and of [32] in a sub-Gaussian setting. Here we consider the non-Gaussian case, where the errors have only four bounded moments, and derive the performance bounds that are approximately as sharp as in the Gaussian model. We give results for both homoscedastic and heteroscedastic models.

5.1. **Homoscedastic case.** Let $(z_i, y_i)_{i=1}^n$ be a sample of independent observations where z_i is a nonstochastic p-vector of regressors. We consider the model

$$y_i = z_i'\beta + \varepsilon_i, \quad \mathrm{E}[\varepsilon_i] = 0, \ i = 1, \dots, n, \ \mathbb{E}_n[z_{ij}^2] = 1, \ j = 1, \dots, p,$$

where y_i is a random scalar dependent variable, and the regressors are normalized to have unitary second moments. Here we consider the homoscedastic case:

$$\mathrm{E}[\varepsilon_i^2] = \sigma^2, \ i = 1, \dots, n,$$

where σ is assumed to be known (for simplicity). We allow p to be substantially larger than n. It is well known that a condition that gives a good performance for the Dantzig selector is that β is sparse, namely $\|\beta\|_0 \leqslant s \ll n$ (although this assumption will not be invoked below explicitly).

The aim is to estimate the vector β in some semi-norms of interest: $\|\cdot\|_I$. For example, given an estimator $\widehat{\beta}$ the prediction semi-norm for $\delta = \widehat{\beta} - \beta$ is

$$\|\delta\|_{\mathrm{pr}} = \sqrt{\mathbb{E}_n[(z_i'\delta)^2]},$$

or the j-th component seminorm for δ is

$$\|\delta\|_{\mathrm{jc}} = |\delta_j|,$$

and so on. The label I designates the name of a norm of interest.

The Dantzig selector is the estimator defined by

$$\widehat{\beta} \in \arg\min_{b\in\mathbb{R}^p} \|b\|_{\ell_1} \text{ subject to } \sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'b)]| \leqslant \lambda, \tag{14}$$

where $\|\beta\|_{\ell_1} = \sum_{j=1}^p |\beta_j|$ is the ℓ_1-norm. An ideal choice of the penalty level λ is meant to ensure that

$$T_0 := \sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}\varepsilon_i]| \leqslant \lambda$$

with a prescribed probability $1-\alpha$. Hence we would like to set penalty level λ equal to

$$c_{T_0}(1-\alpha) := (1-\alpha)\text{-quantile of } T_0,$$

(note that z_i are treated as fixed). Indeed, this penalty would take into account the correlation amongst the regressors, thereby adapting the performance of the estimator to the design condition. We can approximate this quantity using the central limit theorems derived in Section 2. Specifically, let

$$Z_0 := \sigma\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}e_i]|,$$

where e_i are i.i.d. $N(0,1)$ random variables independent of the data. We then estimate $c_{T_0}(1-\alpha)$ by

$$c_{Z_0}(1-\alpha) := \ (1-\alpha)\text{-quantile of } Z_0.$$

Note that we can calculate $c_{Z_0}(1-\alpha)$ numerically with any specified precision by the simulation. (In a Gaussian model, design-adaptive penalty level $c_{Z_0}(1-\alpha)$ was proposed in [6], but its extension to non-Gaussian cases was not available up to now).

An alternative choice of the penalty level is given by

$$c_0(1-\alpha) := \sigma\Phi^{-1}(1-\alpha/2p),$$

which is the canonical choice; see [12] and [9]. Note that canonical choice $c_0(1-\alpha)$ disregards the correlation amongst the regressors, and is therefore more conservative than $c_{Z_0}(1-\alpha)$. Indeed, by the union bound, we see that

$$c_{Z_0}(1-\alpha) \leqslant c_0(1-\alpha).$$

Our first result below shows that the *either* of the two penalty choices, $\lambda = c_{Z_0}(1-\alpha)$ or $\lambda = c_0(1-\alpha)$, are approximately valid under non-Gaussian noise–under the mild moment assumption $\mathrm{E}[\varepsilon_i^4] \leqslant$ const. replacing the canonical Gaussian noise assumption. To derive this result we apply our CLT to T_0 to establish that the difference between distribution functions of T_0 and Z_0 approaches zero at polynomial speed. Indeed T_0 can be represented as a maximum of averages, $T_0 = \max_{1\leqslant k\leqslant 2p} n^{-1/2}\sum_{i=1}^n \tilde{z}_{ik}\varepsilon_i$, for $\tilde{z}_i = (z_i', -z_i')'$, and therefore our CLT applies.

To derive the bound on estimation error $\|\delta\|_I$ in a seminorm of interest, we employ the following identifiability factor:

$$\kappa_I(\beta) := \inf_{\delta\in\mathbb{R}^p}\left\{\max_{1\leqslant j\leqslant p}\frac{|\mathbb{E}_n[z_{ij}(z_i'\delta)]|}{\|\delta\|_I} : \delta\in\mathcal{R}(\beta), \|\delta\|_I \neq 0\right\},$$

where $\mathcal{R}(\beta) := \{\delta\in\mathbb{R}^p : \|\beta+\delta\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}\}$ is the restricted set; $\kappa_I(\beta)$ is defined as ∞ if $\mathcal{R}(\beta) = \{0\}$ (this happens if $\beta = 0$). The factors summarize the impact of sparsity of true parameter value β and the design on the identifiability of β with respect to the norm $\|\cdot\|_I$.

Comment 5.1 (A comment on the identifiability factor $\kappa_I(\beta)$)**.** The identifiability factors $\kappa_I(\beta)$ depend on the true parameter value β. This is not the main focus of this section, but we note that these factors represent a modest generalization of the cone invertibility factors and sensitivity characteristics defined in [46] and [25], which are known to be quite general. The main difference perhaps is the use of a norm of interest $\|\cdot\|_I$ instead of the ℓ_q norms and the use of smaller (non-conic) restricted set $\mathcal{R}(\beta)$ in the definition. It is useful to note for later comparisons that in the case of prediction norm $\|\cdot\|_I = \|\cdot\|_{\mathrm{pr}}$ and under the exact sparsity assumption $\|\beta\|_0 \leqslant s$, we have

$$\kappa_{\mathrm{pr}}(\beta) \geqslant 2^{-1}s^{-1/2}\kappa(s,1), \tag{15}$$

where $\kappa(s,1)$ is the restricted eigenvalue defined in [9]. □

The following result states bounds on the estimation error for the Dantzig selector $\widehat{\beta}^{(0)}$ with canonical penalty level $\lambda = \lambda^{(0)} := c_0(1-\alpha)$ and the Dantzig selector $\widehat{\beta}^{(1)}$ with design-adaptive penalty level $\lambda = \lambda^{(1)} := c_{Z_0}(1-\alpha)$.

Theorem 5.1 (Performance of Dantzig Selector in Non-Gaussian Model)**.** *Suppose that there are some constants $c_1 > 0, C_1 > 0$ and $\sigma^2 > 0$, and a sequence B_n of positive constants such that for all $1\leqslant i\leqslant n$, $1\leqslant j\leqslant p$, and $n \geqslant 1$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\mathrm{E}[\varepsilon_i^2] = \sigma^2$; (iv) $\mathrm{E}[|\varepsilon_i|^4] \leqslant C_1$; and (v) $(\log(pn))^7 B_n^2/n \leqslant C_1 n^{-c_1}$. Then there exist constants $c > 0$ and $C > 0$ depending only on c_1, C_1 and σ^2 such that, with probability at least $1-\alpha-Cn^{-c}$, for either $k = 0$ or 1,*

$$\|\widehat{\beta}^{(k)} - \beta\|_I \leqslant \frac{2\lambda^{(k)}}{\sqrt{n}\kappa_I(\beta)}.$$

The most important feature of this result is that it provides Gaussian-like conclusions (as explained below) in a model with non-Gaussian noise, having bounded fourth moment. Note however that the probabilistic guarantee is not $1-\alpha$ as, e.g., in [9], but rather $1-\alpha-Cn^{-c}$, which reflects the cost of non-Gaussianity. In what follows we discuss details of this result. Note that the bound above holds for any semi-norm of interest $\|\cdot\|_I$.

Comment 5.2 (Improved Performance from Design-Adaptive Penalty Level)**.** The use of the design-adaptive penalty level implies a better performance guarantee for $\widehat{\beta}^{(1)}$ over $\widehat{\beta}^{(0)}$. Indeed, we have

$$\frac{2c_{Z_0}(1-\alpha)}{\sqrt{n}\kappa_I(\beta)} \leqslant \frac{2c_0(1-\alpha)}{\sqrt{n}\kappa_I(\beta)}.$$

For example, in some designs, we can have $\sqrt{n}\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[z_{ij}g_i]| = O_{\mathrm{P}}(1)$, so that $c_{Z_0}(1-\alpha) = O(1)$, whereas $c_0(1-\alpha)$ grows at the rate of $\sqrt{\log p}$. Thus, the performance guarantee provided by $\widehat{\beta}^{(1)}$ can be much better than that of $\widehat{\beta}^{(0)}$. □

Comment 5.3 (Relation to the previous results under Gaussianity)**.** To compare to the previous results obtained for the Gaussian settings, let us focus on the prediction norm and on estimator $\widehat{\beta}^{(1)}$ with penalty level $\lambda = c_{Z_0}(1-\alpha)$. In this case, with probability at least $1-\alpha-Cn^{-c}$,

$$\|\widehat{\beta}^{(1)}-\beta\|_{\mathrm{pr}} \leqslant \frac{2c_{Z_0}(1-\alpha)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)} \leqslant \frac{4\sqrt{s}c_0(1-\alpha)}{\sqrt{n}\kappa(s,1)} \leqslant \frac{4\sqrt{s}\sqrt{2\log(\alpha/(2p))}}{\sqrt{n}\kappa(s,1)}, \tag{16}$$

where the last bound is the same as in [9], Theorem 7.1, obtained for the Gaussian case. We recover the same (or tighter) upper bound without making the Gaussianity assumption on the errors. However, the probabilistic guarantee is not $1-\alpha$ as in [9], but rather $1-\alpha-Cn^{-c}$, which is the cost of non-Gaussianity. □

Comment 5.4 (Other refinements)**.** Unrelated to the main theme of this paper, we can see from (16), that there is some tightening of the performance bound due to the use of the identifiability factor $\kappa_{\mathrm{pr}}(\beta)$ in place of the restricted eigenvalue $\kappa(s,1)$; for example, if $p=2$ and $s=1$ and the two regressors are identical, then $\kappa_{\mathrm{pr}}(\beta)>0$, whereas $\kappa(1,1)=0$. There is also some tightening due to the use of $c_{Z_0}(1-\alpha)$ instead of $c_0(1-\alpha)$ as penalty level, as mentioned above. □

5.2. **Heteroscedastic case.** We consider the same model as above, except now the assumption on the error becomes

$$\sigma_i^2 := \mathrm{E}[\varepsilon_i^2] \leqslant \sigma^2, \quad i=1,\dots,n,$$

i.e., σ^2 is the upper bound on the conditional variance, and we assume that this bound is known (for simplicity). As before, ideally we would like to set penalty level λ equal to

$$c_{T_0}(1-\alpha) := (1-\alpha)\text{-quantile of } T_0,$$

(where T_0 is defined above, and we note that z_i are treated as fixed). The CLT applies as before, namely the difference of the distribution functions of T_0 and its Gaussian analog Z_0 converges to zero. In this case, the Gaussian analog can be represented as

$$Z_0 := \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}\sigma_i e_i]|.$$

Unlike in the homoscedastic case, the covariance structure is no longer known, since σ_i are unknown and we can no longer calculate the quantiles of Z_0. However, we can estimate them using the following multiplier bootstrap procedure.

First, we estimate the residuals $\widehat{\varepsilon}_i = y_i - z_i'\widehat{\beta}_0$ obtained from a preliminary Dantzig selector $\widehat{\beta}^{(0)}$ with the conservative penalty level $\lambda = \lambda_0 := c_0(1 - 1/n) := \sigma\Phi^{-1}(1 - 1/(2pn))$, where σ^2 here is the upper bound on the error variance assumed to be known. Let $(e_i)_{i=1}^n$ be an i.i.d. sequence of $N(0,1)$ random variables, and let

$$W := \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}\widehat{\varepsilon}_i e_i]|.$$

Then we estimate $c_{Z_0}(1-\alpha)$ by

$$c_W(1-\alpha) := (1-\alpha)\text{-quantile of } W,$$

defined conditional on data $(z_i, y_i)_{i=1}^n$. Note that $c_W(1-\alpha)$ can be calculated numerically with any specified precision by the simulation. Then we apply program (14) with $\lambda = \lambda_1 = c_W(1-\alpha)$ to obtain $\widehat{\beta}^{(1)}$.

Theorem 5.2 (Performance of Dantzig in Non-Gaussian Model with Bootstrap Penalty Level)**.** *Suppose that there are some constants $c_1 > 0, C_1 > 0, \underline{\sigma}^2 > 0$ and $\sigma^2 > 0$, and a sequence B_n of positive constants such that for all $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, and $n \geqslant 1$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_i^2] \leqslant \sigma^2$; (iv) $\mathrm{E}[|\varepsilon_i|^4] \leqslant C_1$; (v) $(\log(pn))^7 B_n^4/n \leqslant C_1 n^{-c_1}$; and (vi) $(\log p)B_n c_0(1-1/n)/(\sqrt{n}\kappa_{pr}(\beta)) \leqslant C_1 n^{-c_1}$. Then there exist constants $c > 0$ and $C > 0$ depending only on $c_1, C_1, \underline{\sigma}^2$ and σ^2 such that, with probability at least $1 - \alpha - \nu_n$ where $\nu_n = Cn^{-c}$, we have*

$$\|\widehat{\beta}^{(1)} - \beta\|_I \leqslant \frac{2\lambda^{(1)}}{\sqrt{n}\kappa_I(\beta)}. \tag{17}$$

Moreover, with probability at least $1 - \nu_n$,

$$\lambda^{(1)} = c_W(1-\alpha) \leqslant c_{Z_0}(1 - \alpha + \nu_n),$$

where $c_{Z_0}(1-a) :=$ $(1-a)$-quantile of Z_0; in particular $c_{Z_0}(1-a) \leqslant c_0(1-a)$.

5.3. **Some Extensions.** Here we comment on some additional potential applications.

Comment 5.5 (Confidence Sets)**.** Note that bounds given in the preceding theorems can be used for inference on β or components of β, given the

assumption $\kappa_I(\beta) \geqslant \kappa$, where κ is a known constant. For example, consider inference on the j-th component β_j of β. In this case, we set the norm of interest $\|\delta\|_I$ to be $\|\delta\|_{\mathrm{jc}} = |\delta_j|$ on $\mathbb{R}^p$, and consider the corresponding identifiability factor $\kappa_{\mathrm{jc}}(\beta)$. Suppose it is known that $\kappa_{\mathrm{jc}}(\beta) \geqslant \kappa$. Then a $(1 - \alpha - Cn^{-c})$-confidence interval for β_j is given by

$$\{b \in \mathbb{R} : |\widehat{\beta}_j^{(1)} - b| \leqslant 2\lambda^{(1)}/(\sqrt{n}\kappa)\}.$$

This confidence set is of interest, but it does require the investigator to make a stance on what a plausible κ should be. We refer to [25] for possible ways of computing lower bounds on κ; there is also a work by [30], which provides computable lower bounds on related quantities. □

Comment 5.6 (Generalization of Dantzig Selector)**.** There are many interesting applications where the results given above apply. There are, for example, interesting works by [2] and [24] that consider related estimators that minimize a convex penalty subject to the multiresolution screening constraints. In the context of the regression problem studied above, such estimators may be defined as:

$$\widehat{\beta} \in \arg\min_{b \in \mathbb{R}^p} J(b) \text{ subject to } \sqrt{n} \max_{1 \leqslant j \leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'b)]| \leqslant \lambda,$$

where J is a convex penalty, and the constraint is used for multiresolution screening. For example, the Lasso estimator is nested by the above formulation by using $J(b) = \|b\|_{\mathrm{pr}}$, and the previous Dantzig selector by using $J(b) = \|b\|_{\ell_1}$. The estimators can be interpreted as a point in confidence set for β, which lies closest to zero under J-discrepancy. Our results on choosing λ apply to this class of estimators, and the previous analysis also applies by redefining the identifiability factor $\kappa_I(\beta)$ relative to the new restricted set $\mathcal{R}(\beta) := \{\delta \in \mathbb{R}^p : J(\beta + \delta) \leqslant J(\beta)\}$; where $\kappa_I(\beta)$ is defined as ∞ if $\mathcal{R}(\beta) = \{0\}$. □

6. Application: Multiple Hypothesis Testing

In this subsection, we study the problem of multiple hypothesis testing in the framework of multiple linear regressions. (Note that the problem of testing multiple means is a special case of testing multiple regressions.) We combine a general stepdown procedure described in [41] with the multiplier bootstrap developed in this paper. In contrast with [41], our results do not require weak convergence arguments, and, thus, can be applied to models with increasing numbers of both parameters and regressions. Notably, the number of regressions can be exponentially large in the sample size.

Let $(z_i, y_i)_{i=1}^n$ be a sample of independent observations where z_i is a p-vector of nonstochastic covariates and y_i is a K-vector of dependent random variables. For each $k = 1, \ldots, K$, let $I_k \subset \{1, \ldots, p\}$ be a subset of covariates used in the k-th regression. Denote by $|I_k| = p_k$ the number of covariates in the k-th regression, and let $\bar{p} = \max_{1 \leqslant k \leqslant K} p_k$. Let v_{ik} be a subvector of z_i consisting of those elements of z_i whose indices appear in I_k: $v_{ik} =$

$(z_{ij})_{j\in I_k}$. We denote components of v_{ik} by v_{ikj}, $j = 1,\dots,p_k$. Without loss of generality, we assume that $I_k \cap I_{k'} = \varnothing$ for all $k \neq k'$ and $\sum_{1\leqslant k\leqslant K} p_k = p$.

For each $k = 1,\dots,K$, consider the linear regression model

$$y_{ik} = v_{ik}'\beta_k + \varepsilon_{ik},\ i = 1,\dots,n.$$

where $\beta_k \in \mathbb{R}^{p_k}$ is an unknown parameter of interest, and $(\varepsilon_{ik})_{i=1}^n$ is a sequence of independent zero-mean unobservable scalar random variables. We allow for triangular array asymptotics so that everything in the model, and, in particular, the number of regressions K and the dimensions of the parameters β_k and p_k, may depend on n. For brevity, however, we omit index n. We are interested in simultaneous testing the set of null hypotheses $H_{kj} : \beta_{kj} = 0$ against the alternatives $H_{kj}' : \beta_{kj} \neq 0$, $(k,j) \in \mathcal{W}_0$ for some set of pairs $\mathcal{W}_0$ where β_{kj} denotes the jth component of β_k, with the strong control of the family-wise error rate. In other words, we seek a procedure that would reject at least one true null hypothesis with probability not greater than $1-\alpha+o(1)$ uniformly over the set of true null hypotheses. More formally, let Ω be a set of all data generating processes, and ω be the true process. Each null hypothesis H_{kj} is equivalent to $\omega \in \Omega_{kj}$ for some subset Ω_{kj} of Ω. Let $\mathcal{W}$ denote the set of all pairs (k,j) with $k = 1,\dots,K$ and $j = 1,\dots,p_k$:

$$\mathcal{W} = \{(k,j) : k = 1,\dots,K; j = 1,\dots,p_k\}.$$

For a subset $w \subset \mathcal{W}$ let $\Omega^w = (\cap_{(k,j)\in w}\Omega_{kj}) \cap (\cap_{(k,j)\notin w}\Omega_{kj}^c)$ where $\Omega_{kj}^c = \Omega\backslash\Omega_{kj}$. The strong control of the family-wise error rate means

$$\sup_{w\subset\mathcal{W}}\sup_{\omega\in\Omega^w} \mathrm{P}\{\text{reject at least one hypothesis among } H_{kj},\ (k,j)\in w\} \leqslant \alpha + o(1). \tag{18}$$

This setting is clearly of interest in many empirical studies.

Our approach is based on the simultaneous analysis of t-statistics for each component β_{kj}. Let $x_{ik} = (\mathbb{E}_n[v_{ik}v_{ik}'])^{-1}v_{ik}$. Then the OLS estimator $\widehat{\beta}_k$ of β_k is given by $\widehat{\beta}_k = \mathbb{E}_n[x_{ik}y_{ik}]$. The corresponding residuals are $\widehat{\varepsilon}_{ik} = y_{ik} - v_{ik}'\widehat{\beta}$, $i = 1,\dots,n$. Since $(x_{ik})_{i=1}^n$ is nonstochastic, the covariance matrix of $\widehat{\beta}_k$ is given by $V(\widehat{\beta}_k) = \mathbb{E}_n[x_{ik}x_{ik}'\sigma_{ik}^2]/n$ where $\sigma_{ik}^2 = \mathrm{E}[\varepsilon_{ik}^2]$, $i = 1,\dots,n$. The t-statistic for testing H_{kj} against H_{kj}' is $t_{kj} := |\widehat{\beta}_{kj}|/\sqrt{\widehat{V}(\widehat{\beta}_k)_{jj}}$ where $\widehat{V}(\widehat{\beta}_k) = \mathbb{E}_n[x_{ik}x_{ik}'\widehat{\varepsilon}_{ik}^2]/n$. Also define

$$t_{kj}^0 := \frac{|\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|}{\sqrt{\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2]}}.$$

Note that $t_{kj} = t_{kj}^0$ under the hypothesis H_{kj}.

The stepdown procedure of [41] is described as follows. For a subset $w \subset \mathcal{W}$, let $c_{1-\alpha,w}$ be some estimator of the $(1-\alpha)$-quantile of $\max_{(k,j)\in w} t_{kj}^0$. On the first step, let $w(1) = \mathcal{W}_0$. Reject all hypotheses H_{kj} satisfying

$t_{kj} > c_{1-\alpha,w(1)}$. If no null hypothesis is rejected, then stop. If some H_{kj} are rejected, then let $w(2)$ be the set of all null hypotheses that were not rejected on the first step. On step $l \geqslant 2$, let $w(l) \subset \mathcal{W}$ be the subset of null hypotheses that were not rejected up to step l. Reject all hypotheses H_{kj}, $(k,j) \in w(l)$, satisfying $t_{kj} > c_{1-\alpha,w(l)}$. If no null hypothesis is rejected, then stop. If some H_{kj} are rejected, then let $w(l+1)$ be the subset of all null hypotheses among $(k,j) \in w(l)$ that were not rejected. Proceed in this way until the algorithm stops.

[41] proved the following result. Suppose that $c_{1-\alpha,w}$ satisfies

$$c_{1-\alpha,w'} \leqslant c_{1-\alpha,w''} \quad \text{whenever } w' \subset w'', \tag{19}$$

$$\sup_{w \subset \mathcal{W}} \sup_{\omega \in \Omega^w} \mathrm{P}\left(\max_{(k,j) \in w} t^0_{kj} > c_{1-\alpha,w} \right) \leqslant \alpha + o(1), \tag{20}$$

then inequality (18) holds. Indeed, let w be the set of true null hypotheses. Suppose that the procedure rejects at least one of these hypotheses. Let l be the step when the procedure rejected a true null hypothesis for the first time, and let $H_{k_0 j_0}$ be this hypothesis. Clearly, we have $w(l) \supset w$. So,

$$\max_{(k,j) \in w} t^0_{kj} \geqslant t^0_{k_0 j_0} = t_{k_0 j_0} > c_{1-\alpha,w(l)} \geqslant c_{1-\alpha,w}.$$

Combining this chain of inequalities with (20) yields (18).

To obtain suitable $c_{1-\alpha,w}$ that satisfies inequalities (19) and (20) above, we can use the multiplier bootstrap method. Let $(e_i)_{i=1}^n$ be an i.i.d. sequence of $N(0,1)$ random variables that are independent of the data. Let $c_{1-\alpha,w}$ be the conditional $(1-\alpha)$-quantile of

$$\max_{(k,j) \in w} \frac{|\sum_{i=1}^n x_{ikj} \widehat{\varepsilon}_{ik} e_i / \sqrt{n}|}{\sqrt{\mathbb{E}_n[x^2_{ikj} \widehat{\varepsilon}^2_{ik}]}} \tag{21}$$

given $(z_i, y_i)_{i=1}^n$. To prove that so defined critical values $c_{1-\alpha,w}$ satisfy inequalities (19) and (20), we will assume the following regularity condition,

(M) There are some constants $c_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$ and a sequence B_n of positive constants such that for $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, $1 \leqslant k \leqslant K$, $1 \leqslant l \leqslant p_k$, and $n \geqslant 1$: (i) $|z_{ij}| \leqslant B_n$ and $B_n \geqslant c_1$; (ii) $\mathbb{E}_n[(z_{ij})^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon^2_{ik}] \leqslant \bar{\sigma}^2$; (iv) the minimum eigenvalue of $\mathbb{E}_n[v_{ik} v'_{ik}]$ is bounded from below by c_1; and (v) $\mathbb{E}_n[x^2_{ikl}] \geqslant c_1$.

Theorem 6.1 (Strong Control of Family-Wise Error Rate)**.** *Let $C_1 > 0$ be some constant and suppose that assumption M is satisfied, and $\bar{p} B_n^2 (\log(pn))^7/n = o(1)$. Moreover, suppose either (a) $\mathrm{E}[\max_{1 \leqslant k \leqslant K} \varepsilon^4_{ik}] \leqslant C_1$ for all $1 \leqslant i \leqslant n$, $\bar{p}^3 B_n^4 (\log p)^4/n = o(1)$, and $\bar{p}^2 B_n^4 (\log p)^6/n = o(1)$ or (b) $\mathrm{P}(|\varepsilon_{ik}| \geqslant u) \leqslant \exp(1 - u^2/C_1)$ for all $1 \leqslant i \leqslant n$ and $1 \leqslant k \leqslant K$ and $\bar{p}^3 B_n^2 (\log p)^3/n = o(1)$. Then the stepdown procedure with the multiplier bootstrap critical values $c_{1-\alpha,w}$ given above satisfies (18).*

Comment 6.1 (Relation to prior results)**.** There is a vast literature on multiple hypothesis testing. Let us consider the simple case where $K =$

$p, p_k = 1$ for all $k = 1, \ldots, K$ and $v_{ik} = 1$, so that the k-th regression reduces to $y_{ik} = \beta_k + \varepsilon_{ik}$ (here β_k is scalar). The problem then reduces to testing multiple means (without stepdown). It is instructive to see the implication of Theorem 6.1 in this simple setting. Denote by t_k^0 the t-statistic for testing $H_k : \beta_k = 0$ against $H_k' : \beta_k \neq 0$, and let $c_{1-\alpha}$ be the conditional $(1-\alpha)$-quantile of

$$\max_{k=1,\ldots,p} \frac{|\sum_{i=1}^n \widehat{\varepsilon}_{ik} e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[\widehat{\varepsilon}_{ik}^2]}},$$

where $\widehat{\varepsilon}_{ik} = y_{ik} - \bar{y}_k$, $\bar{y}_k = \mathbb{E}_n[y_{ik}]$, and $(e_i)_{i=1}^n$ is a sequence of i.i.d. $N(0,1)$ random variables independent of the data. Theorem 6.1 implies that, when H_k are true for all k, $\mathrm{P}(\max_{1 \leqslant k \leqslant p} t_k^0 > c_{1-\alpha}) \leqslant \alpha + o(1)$ (indeed, the inequality "$\leqslant$" can be replaced by the equality "$=$") uniformly in the underlying distribution provided that $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_{ik}^2] \leqslant \bar{\sigma}^2$, $\log p = o(n^{1/7})$ and either (a) $\mathrm{E}[\max_{1 \leqslant k \leqslant p} \varepsilon_{ik}^4] \leqslant C_1$ or (b) $\mathrm{P}(|\varepsilon_{ik}| \geqslant u) \leqslant \exp(1 - u^2/C_1)$. Hence the multiplier bootstrap as described above leads to an *asymptotically exact* testing procedure for the multiple hypothesis testing problem of which the *logarithm* of the number of hypotheses is nearly of order $n^{1/7}$ (subject to the prescribed assumptions). Note here that no assumption on the dependency structure between $y_{i1}, \ldots, y_{ip}$ is made.

The question on how large p can be was studied in [23] but from a conservative perspective. The motivation there is to know how fast p can grow to maintain the size of the simultaneous test when we calculate critical values (conservatively) ignoring the dependency among t_k^0 and assuming that t_k^0 were distributed as, say, $N(0,1)$. This framework is conservative in that correlation amongst statistics is dealt away with union bounds, namely Bonferroni-Holm procedures. On the other hand, our approach takes into account the correlation amongst statistics and hence is asymptotically exact, that is, asymptotically non-conservative. □

7. Application: Adaptive Specification Testing

In this section, we study the problem of adaptive specification testing. Let $(v_i, y_i)_{i=1}^n$ be a sample of independent random pairs where y_i is a scalar dependent random variable, and v_i is a d-vector of nonstochastic covariates. The null hypothesis, H_0, is that there exists $\beta \in \mathbb{R}^d$ such that

$$\mathrm{E}[y_i] = v_i'\beta;\ i = 1, \ldots, n. \tag{22}$$

The alternative hypothesis, H_a, is that there is no β satisfying (22). We allow for triangular array asymptotics so that everything in the model may depend on n. For brevity, however, we omit index n.

Denote $\varepsilon_i = y_i - \mathrm{E}[y_i]$, $i = 1, \ldots, n$. Then $\mathrm{E}[\varepsilon_i] = 0$, and under H_0, $y_i = v_i'\beta + \varepsilon_i$. To test H_0, consider a set of test functions $P_j(v_i)$, $j = 1, \ldots, p$. Let $z_{ij} = P_j(v_i)$. We choose test functions so that $\mathbb{E}_n[z_{ij} v_i] = 0$ and $\mathbb{E}_n[z_{ij}^2] = 1$ for all $j = 1, \ldots, p$. In our analysis, p may be higher or even much higher

than n. Let $\widehat{\beta} = (\mathbb{E}_n[v_i v_i'])^{-1}(\mathbb{E}_n[v_i y_i])$ be an OLS estimator of β, and let $\widehat{\varepsilon}_i = y_i - z_i'\widehat{\beta}$; $i = 1, \ldots, n$ be corresponding residuals. Our test statistic is

$$T := \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

The test rejects H_0 if T is significantly large.

Note that since $\mathbb{E}_n[z_{ij}v_i] = 0$, we have

$$\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i/\sqrt{n} = \sum_{i=1}^n z_{ij}(\varepsilon_i + v_i'(\beta - \widehat{\beta}))/\sqrt{n} = \sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}.$$

Therefore, under H_0,

$$T = \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

This suggests that we can use the multiplier bootstrap to obtain a critical value for the test. More precisely, let $(e_i)_{i=1}^n$ be an i.i.d. sequence of independent $N(0,1)$ random variables that are independent of the data, and let

$$W := \max_{1 \leqslant j \leqslant p} \frac{|\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}|}{\sqrt{\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2]}}.$$

The multiplier bootstrap critical value $c_W(1-\alpha)$ is the conditional $(1-\alpha)$-quantile of W given the data. To prove the validity of multiplier bootstrap, we will impose the following condition:

(S) There are some constants $c_1 > 0, C_1 > 0, \bar{\sigma}^2 > 0, \underline{\sigma}^2 > 0$, and a sequence B_n of positive constants such that for all $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, $1 \leqslant k \leqslant d$, and $n \geqslant 1$: (i) $|z_{ij}| \leqslant B_n$; (ii) $\mathbb{E}_n[z_{ij}^2] = 1$; (iii) $\underline{\sigma}^2 \leqslant \mathrm{E}[\varepsilon_i^2] \leqslant \bar{\sigma}^2$; (iv) $|v_{ik}| \leqslant C_1$; (v) $d \leqslant C_1$; and (vi) the minimum eigenvalue of $\mathbb{E}_n[v_i v_i']$ is bounded from below by c_1.

Theorem 7.1 (Size Control of Adaptive Specification Test)**.** *Let $c_2 > 0$ be some constant. Suppose that assumption S is satisfied, and $(\log(pn))^7 B_n^2/n \leqslant C_1 n^{-c_2}$. Moreover, suppose that either (a) $\mathrm{E}[\varepsilon_i^4] \leqslant C_1$ for all $1 \leqslant i \leqslant n$ and $(\log p)^6 B_n^4/n \leqslant C_1 n^{-c_2}$ or (b) $\mathrm{P}(|\varepsilon_i| \geqslant u) \leqslant \exp(1 - u^2/C_1)$ for all $1 \leqslant i \leqslant n$. Then there exist constants $c > 0$ and $C > 0$, depending only on $c_1, c_2, C_1, \underline{\sigma}^2$ and $\bar{\sigma}^2$, such that under H_0, $|\mathrm{P}(T \leqslant c_W(1-\alpha)) - (1-\alpha)| \leqslant Cn^{-c}$.*

Comment 7.1. The literature on specification testing is large. In particular, [29] and [28] developed adaptive tests that are suitable for inference in L_2-norm. In contrast, our test is most suitable for inference in sup-norm. An advantage of our procedure is that selecting a wide class of test functions leads to a test that can effectively adapt to a wide range of alternatives, including those that can not be well-approximated by Hölder-continuous functions. □

Appendix A. Preliminaries

A.1. **A Useful Maximal Inequality.** The following lemma is a useful variation on standard maximal inequalities.

Lemma A.1 (Maximal Inequality)**.** *Consider a sequence $(x_i)_{i=1}^n$ of independent random p-vectors ($p \geqslant 2$). Let $M = \max_{1\leqslant i\leqslant n} \max_{1\leqslant j\leqslant n} |x_{ij}|$ and $\sigma^2 = \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[x_{ij}^2]$. Then*

$$\mathrm{E}\left[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[x_{ij}] - \bar{\mathrm{E}}[x_{ij}]|\right] \lesssim \sigma\sqrt{(\log p)/n} + \sqrt{\mathrm{E}[M^2]}(\log p)/n.$$

Proof. Throughout the proof, let $(y_i)_{i=1}^n$ be an independent copy of $(x_i)_{i=1}^n$, and let $(\epsilon_i)_{i=1}^n$ be an i.i.d. sequence of Rademacher random variables independent of everything else.

Step 1. We first prove

$$\mathrm{E}[\max_{1\leqslant j\leqslant p} \mathbb{E}_n[|x_{ij}|]] \lesssim \mathrm{E}[M](\log p)/n + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|]. \tag{23}$$

Observe that

$$\begin{aligned}
I := \mathrm{E}[\max_{1\leqslant j\leqslant p} \mathbb{E}_n[|x_{ij}|]] &\leqslant \mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[|x_{ij}|] - \bar{\mathrm{E}}[|x_{ij}|]|] + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|] \\
&= \mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[|x_{ij}|] - \mathrm{E}[\mathbb{E}_n[|y_{ij}|]]|] + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|] \\
&\leqslant \mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[|x_{ij}|] - \mathbb{E}_n[|y_{ij}|]|] + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|] \\
&\leqslant \mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[\epsilon_i(|x_{ij}| - |y_{ij}|)]|] + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|] \\
&\leqslant 2\mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[\epsilon_i|x_{ij}|]|] + \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|],
\end{aligned}$$

Further, using Pisier inequality conditional on $(x_i)_{i=1}^n$, we have

$$\begin{aligned}
\mathrm{E}[\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[\epsilon_i|x_{ij}|]|] &\lesssim \sqrt{(\log p)/n}\mathrm{E}[\max_{1\leqslant j\leqslant p} (\mathbb{E}_n[|x_{ij}|^2])^{1/2}] \\
&\leqslant \sqrt{(\log p)/n}\mathrm{E}[\max_{1\leqslant j\leqslant p} (M\mathbb{E}_n[|x_{ij}|])^{1/2}] \\
&\leqslant \sqrt{((\log p)/n)\mathrm{E}[M]\mathrm{E}[\max_{1\leqslant j\leqslant p} \mathbb{E}_n[|x_{ij}|]]},
\end{aligned}$$

where the last step follows from the Cauchy-Schwarz inequality. Hence $I \lesssim a\sqrt{I} + b$, where $a = \sqrt{\mathrm{E}[M](\log p)/n}$ and $b = \max_{1\leqslant j\leqslant p} \bar{\mathrm{E}}[|x_{ij}|]$. Solving this inequality gives (23).

Step 2. We now prove the claim of the lemma. Observe that

$$\begin{aligned}
\mathrm{E}[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}]-\bar{\mathrm{E}}[x_{ij}]|] &\leqslant \mathrm{E}[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}]-\mathrm{E}[\mathbb{E}_n[y_{ij}]]|] \\
&\leqslant \mathrm{E}[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[x_{ij}-y_{ij}]|] \\
&= \mathrm{E}[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[\epsilon_i(x_{ij}-y_{ij})]|] \\
&\leqslant 2\mathrm{E}[\max_{1\leqslant j\leqslant p}|\mathbb{E}_n[\epsilon_i x_{ij}]|] \\
&\lesssim \sqrt{(\log p)/n}\mathrm{E}[\max_{1\leqslant j\leqslant p}\mathbb{E}_n[x_{ij}^2]]^{1/2}.
\end{aligned}$$

Applying (23) to x_{ij}^2 instead of x_{ij}, we see that the last expression is

$$\begin{aligned}
&\lesssim \sqrt{(\log p)/n}\left(\mathrm{E}[M^2](\log p)/n+\max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[x_{ij}^2]\right)^{1/2} \\
&\leqslant \sqrt{\mathrm{E}[M^2]}(\log p)/n+\sigma\sqrt{(\log p)/n},
\end{aligned}$$

where the last step follows from inequality $\sqrt{a+b}\leqslant\sqrt{a}+\sqrt{b}$. This completes the proof. □

A.2. **Properties of the Smooth Max Function.** We shall use the following properties of the smooth max function. In this section, we assume $g\in C_b^3(\mathbb{R})$, and let $m=g\circ F_\beta$.

Lemma A.2 (Properties of F_β)**.** *We have*

$$\partial_j F_\beta(z)=\pi_j(z),\quad \partial_j\partial_k F_\beta(z)=\beta w_{jk}(z),\quad \partial_j\partial_k\partial_l F_\beta(z)=\beta^2 q_{jkl}(z).$$

where, for $\delta_{jk}:=1\{j=k\}$,

$$\begin{aligned}
\pi_j(z) &:= e^{\beta z_j}/\sum_{m=1}^p e^{\beta z_m}, \\
w_{jk}(z) &:= (\pi_j\delta_{jk}-\pi_j\pi_k)(z), \\
q_{jkl}(z) &:= (\pi_j\delta_{jl}\delta_{jk}-\pi_j\pi_l\delta_{jk}-\pi_j\pi_k(\delta_{jl}+\delta_{kl})+2\pi_j\pi_k\pi_l)(z).
\end{aligned}$$

Moreover,

$$\pi_j(z)\geqslant 0,\quad \sum_{j=1}^p\pi_j(z)=1,\quad \sum_{j,k=1}^p|w_{jk}(z)|\leqslant 2,\quad \sum_{j,k,l=1}^p|q_{jkl}(z)|\leqslant 6.$$

Proof of Lemma A.2. The first property was noted in [14]. The other properties follow from repeated application of the chain rule. □

Lemma A.3 (Three derivatives of $m = g \circ F_\beta$)**.** *Let π_j, w_{jk}, and q_{jkl} be as defined above. For every $z \in \mathbb{R}^p$,*

$$\begin{aligned}
\partial_j m(z) &= (\partial g(F_\beta)\pi_j)(z),\\
\partial_j\partial_k m(z) &= (\partial^2 g(F_\beta)\pi_j\pi_k + \partial g(F_\beta)\beta w_{jk})(z),\\
\partial_j\partial_k\partial_l m(z) &= (\partial^3 g(F_\beta)\pi_j\pi_k\pi_l + \partial^2 g(F_\beta)\beta(w_{jk}\pi_l + w_{jl}\pi_k + w_{kl}\pi_j)\\
&\quad + \partial g(F_\beta)\beta^2 q_{jkl})(z).
\end{aligned}$$

Proof of lemma A.3. The proof follows from the repeated application of the chain rule and by the properties noted in Lemma A.2. □

Lemma A.4 (Bounds on derivatives of $m = g \circ F_\beta$)**.** *We have*

$$|\partial_j\partial_k m(z)| \leqslant U_{jk}(z), \quad |\partial_j\partial_k\partial_l m(z)| \leqslant U_{jkl}(z),$$

where

$$\begin{aligned}
U_{jk}(z) &:= (G_2\pi_j\pi_k + G_1\beta W_{jk})(z),\\
U_{jkl}(z) &:= (G_3\pi_j\pi_k\pi_l + G_2\beta(W_{jk}\pi_l + W_{jl}\pi_k + W_{kl}\pi_j) + G_1\beta^2 Q_{jkl})(z),\\
W_{jk}(z) &:= (\pi_j\delta_{jk} + \pi_j\pi_k)(z),\\
Q_{jkl}(z) &:= (\pi_j\delta_{jl}\delta_{jk} + \pi_j\pi_l\delta_{jk} + \pi_j\pi_k(\delta_{jl} + \delta_{kl}) + 2\pi_j\pi_k\pi_l)(z).
\end{aligned}$$

Moreover,

$$\sum_{j,k=1}^p U_{jk}(z) \leqslant (G_2 + 2G_1\beta), \quad \sum_{j,k,l=1}^p U_{jkl}(z) \leqslant (G_3 + 6G_2\beta + 6G_1\beta^2).$$

Proof of Lemma A.4. The lemma follows from a direct calculation. □

Lemma A.5. *For every $z \in \mathbb{R}^p$, $w \in \mathbb{R}^p$ such that $\max_{j\leqslant p}|w_j|\beta \leqslant 1$, $\tau \in [0,1]$, and every $1 \leqslant j \leqslant p$, we have*

$$\begin{aligned}
\pi_j(z) &\lesssim \pi_j(z + \tau w) \lesssim \pi_j(z),\\
U_{jk}(z) &\lesssim U_{jk}(z + \tau w) \lesssim U_{jk}(z),\\
U_{jkl}(z) &\lesssim U_{jkl}(z + \tau w) \lesssim U_{jkl}(z).
\end{aligned}$$

Proof of Lemma A.5. To show the first claim we note that

$$\begin{aligned}
\pi_j(z + \tau w) &= \frac{\exp(z_j\beta + \tau w_j\beta)}{\sum_{m=1}^p \exp(z_m\beta + \tau w_m\beta)}\\
&\leqslant \frac{\exp(z_j\beta)}{\sum_{m=1}^p \exp(z_m\beta)} \frac{\exp(\tau \max_{j\leqslant p}|w_j|\beta)}{\exp(-\tau \max_{j\leqslant p}|w_j|\beta)} \leqslant \pi_j(z)\exp(2).
\end{aligned}$$

Proceeding similarly, we have $\pi_j(z + \tau w) \geqslant \pi_j(z)\exp(-2)$.

The second and third claims follow from the first, noting that U_{jk} and U_{jkl} are finite sums of products of order up to 3 of terms such as π_j, π_k, π_l, δ_{jk}. □

Lemma A.6 (Lipschitz Property of F_β)**.** *For each $x \in \mathbb{R}^p$ and $z \in \mathbb{R}^p$, we have*

$$|F_\beta(x) - F_\beta(z)| \leqslant \max_{1 \leqslant j \leqslant p} |x_j - z_j|.$$

Proof of Lemma A.6. For some $t \in [0,1]$,

$$\begin{aligned} |F_\beta(x) - F_\beta(z)| &= |\sum_{j=1}^p \partial_j F_\beta(x + t(z-x))(z_j - x_j)| \\ &\leqslant \sum_{j=1}^p \pi_j(x + t(z-x)) \max_{1 \leqslant j \leqslant p} |z_j - x_j| \leqslant \max_{1 \leqslant j \leqslant p} |z_j - x_j|, \end{aligned}$$

where the property $\sum_{j=1}^p \pi_j(x + t(z-x)) = 1$ was used. □

A.3. **Lemmas on Truncation.** Recall that $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$ and $\tilde{X} = n^{-1/2} \sum_{i=1}^n \tilde{x}_i$, where "tilde" denotes the truncation operation defined in Section 2.

Lemma A.7 (Truncation Impact)**.** *For every $1 \leqslant j, k \leqslant p$ and $q \geqslant 1$, (a) $(\mathrm{E}[|\tilde{x}_{ij}|^q])^{1/q} \leqslant 2(\mathrm{E}[|x_{ij}|^q])^{1/q}$; (b) $|\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}] - \mathrm{E}[x_{ij}x_{ik}]| \leqslant (3/2)(\mathrm{E}[x_{ij}^2] + \mathrm{E}[x_{ik}^2])\varphi(u)$; and (c) with probability at least $1 - 5\gamma$, for all $1 \leqslant j \leqslant p$,*

$$|X_j - \tilde{X}_j| \leqslant \sqrt{\bar{\mathrm{E}}[x_{ij}^2]}\varphi(u)\sqrt{2\log(p/\gamma)}(5 + \delta(u,\gamma)).$$

Proof of Lemma A.7. Claim (a). Define $I_{ij} = 1\{|x_{ij}| \leqslant u\sqrt{\mathrm{E}[x_{ij}^2]}\}$, and note that

$$\begin{aligned} (\mathrm{E}[|\tilde{x}_{ij}|^q])^{1/q} &\leqslant (\mathrm{E}[|x_{ij}I_{ij}|^q])^{1/q} + |\mathrm{E}[x_{ij}I_{ij}]| \\ &\leqslant (\mathrm{E}[|x_{ij}I_{ij}|^q])^{1/q} + (\mathrm{E}[|x_{ij}I_{ij}|^q])^{1/q} \leqslant 2(\mathrm{E}[|x_{ij}|^q])^{1/q}, \end{aligned}$$

where the first inequality follows from the triangle inequality, the second from Hölder's inequality, and the last from the monotonicity of the expectation.

Claim (b). Observe that

$$\begin{aligned} |\mathrm{E}[\tilde{x}_{ij}\tilde{x}_{ik}] - \mathrm{E}[x_{ij}x_{ik}]| &\leqslant |\mathrm{E}[(\tilde{x}_{ij} - x_{ij})\tilde{x}_{ik}]| + |\mathrm{E}[x_{ij}(\tilde{x}_{ik} - x_{ik})]| \\ &\leqslant \sqrt{\mathrm{E}[(\tilde{x}_{ij} - x_{ij})^2]}\sqrt{\mathrm{E}[\tilde{x}_{ik}^2]} + \sqrt{\mathrm{E}[(\tilde{x}_{ik} - x_{ik})^2]}\sqrt{\mathrm{E}[x_{ij}^2]} \\ &\leqslant 2\varphi(u)\sqrt{\mathrm{E}[x_{ij}^2]}\sqrt{\mathrm{E}[x_{ik}^2]} + \varphi(u)\sqrt{\mathrm{E}[x_{ik}^2]}\sqrt{\mathrm{E}[x_{ij}^2]} \\ &\leqslant (3/2)\varphi(u)(\mathrm{E}[x_{ij}^2] + \mathrm{E}[x_{ik}^2]), \end{aligned}$$

where the first inequality follows from the triangle inequality, the second from the Cauchy-Schwarz inequality, the third from the definition of $\varphi(u)$ together with claim (a), and the last from inequality $|ab| \leqslant (a^2 + b^2)/2$.

Claim (c). We shall use the following lemma.

Lemma A.8 (Sub-Gaussian Bounds for Arbitrary Self-Normalized Sums)**.** *Let $\xi_1, \ldots, \xi_n$ be independent real-valued random variables such that $\mathrm{E}[\xi_i] = 0$ and $\mathrm{E}[\xi_i^2] < \infty$ for all $1 \leqslant i \leqslant n$. Let $S_n = \sum_{i=1}^n \xi_i$. Then for every $x > 0$,*

$$\mathrm{P}(|S_n| > x(4B_n + V_n)) \leqslant 4\exp(-x^2/2),$$

where $B_n^2 = \sum_{i=1}^n \mathrm{E}[\xi_i^2]$ and $V_n^2 = \sum_{i=1}^n \xi_i^2$.

Proof of Lemma A.8. See [20], Theorem 2.16. □

Define

$$\Lambda_j := 4\sqrt{\bar{\mathrm{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\mathbb{E}_n[(x_{ij} - \tilde{x}_{ij})^2]}.$$

Then by Lemma A.8, with probability at least $1 - 4\gamma$,

$$|X_j - \tilde{X}_j| \leqslant \Lambda_j \sqrt{2\log(p/\gamma)}, \text{ for all } 1 \leqslant j \leqslant p.$$

By definition of $\varphi(u)$ and $\delta(u, \gamma)$, with probability at least $1 - \gamma$, for all $1 \leqslant j \leqslant p$,

$$\begin{aligned}\Lambda_j &\leqslant 4\sqrt{\bar{\mathrm{E}}[(x_{ij} - \tilde{x}_{ij})^2]} + \sqrt{\bar{\mathrm{E}}[(x_{ij})^2]}\varphi(u)(1 + \delta(u, \gamma)) \\ &\leqslant \sqrt{\bar{\mathrm{E}}[(x_{ij})^2]}\varphi(u)(5 + \delta(u, \gamma)).\end{aligned}$$

This implies claim (c). □

Lemma A.9 (Sub-Gaussian truncation)**.** *For any sub-Gaussian random variable R with parameter $C > 0$, i.e., if $\mathrm{P}(|R| \geqslant u) \leqslant \exp(1 - u^2/C^2)$ for all $u \geqslant 0$, then we have*

$$\mathrm{E}[R^2 1\{|R| \geqslant u\}] \leqslant (u^2 + C^2)\exp(1 - u^2/C^2)$$

for all $u \geqslant 0$.

Proof. The lemma follows from the following calculation:

$$\begin{aligned}\mathrm{E}[R^2 1\{|R| \geqslant u\}] &\leqslant u^2 \mathrm{P}(|R| \geqslant u) + 2\int_u^\infty z\mathrm{P}(|R| \geqslant z)dz \\ &\leqslant u^2 \exp(1 - u^2/C^2) + 2\int_u^\infty z\exp(1 - z^2/C^2)dz \\ &= (u^2 + C^2)\exp(1 - u^2/C^2).\end{aligned}$$

□

Let $\tilde{y}_i = (\tilde{y}_{ij})_{j=1}^p$, where $\tilde{y}_{ij} = y_{ij} 1\{|y_{ij}| \leqslant u(\mathrm{E}[y_{ij}^2])^{1/2}\}$, and let $\tilde{Y} = n^{-1/2}\sum_{i=1}^n \tilde{y}_i$. Note that by the symmetry of the distribution of y_{ij}, $\mathrm{E}[\tilde{y}_{ij}] = 0$.

Lemma A.10 (Normal Truncation Impact)**.** *For every $1 \leqslant j, k \leqslant p$ and $q \geqslant 1$, (a) $(\mathrm{E}[|\tilde{y}_{ij}|^q])^{1/q} \leqslant (\mathrm{E}[|y_{ij}|^q])^{1/q}$; (b) $|\mathrm{E}[\tilde{y}_{ij}\tilde{y}_{ik}] - \mathrm{E}[y_{ij}y_{ik}]| \leqslant (3/2)(\mathrm{E}[y_{ij}^2] + \mathrm{E}[y_{ik}^2])\varphi_N(u)$ where recall that $\varphi_N^2(u) := \int_{|z| \geqslant u} z^2\phi(z)dz$; and (c) for a given $\gamma \in (0, 1)$ and $u \geqslant \Phi^{-1}(1 - \gamma/(2np))$, e.g., $u = \sqrt{2\log(2np/\gamma)}$,*

$$\mathrm{P}(\max_{1 \leqslant j \leqslant p} |Y_j - \tilde{Y}_j| = 0) \geqslant 1 - \gamma.$$

Proof of Lemma A.10. Claim (a) follows from the fact that $|\tilde{y}_{ij}| \leqslant |y_{ij}|$ and the monotonicity of the expectation. Claim (b) follows from claim (b) of Lemma A.7. For claim (c), note that the event $\max_{1\leqslant j\leqslant p}|Y_j - \tilde{Y}_j| = 0$ is implied by the event $\max_{j\leqslant p, i\leqslant n}\left|y_{ij}/\sqrt{\mathrm{E}[y_{ij}^2]}\right| \leqslant u$, which, by the union bound, does not occur with probability $2np(1-\Phi(u)) \leqslant \gamma$ whenever $u \geqslant \Phi^{-1}(1-\gamma/(2np))$. □

Appendix B. Proofs for Section 2

B.1. **Proof of Theorem 2.1.** Without loss of generality we can and will assume that sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent of each other. We will use an interpolation approach of Slepian. For $t \in [0,1]$, define

$$Z(t) := \sqrt{t}X + \sqrt{1-t}Y = \sum_{i=1}^n Z_i(t),\ Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}x_i + \sqrt{1-t}y_i).$$

We shall also employ Stein's leave-one-out expansions to form:

$$Z^{(i)}(t) := (Z_{ij}(t))_{j=1}^p := Z(t) - Z_i(t).$$

Let $\Psi(t) = \mathrm{E}[m(Z(t))]$ for $m := g \circ F_\beta$. Then

$$\begin{aligned}\mathrm{E}[m(X) - m(Y)] = \Psi(1) - \Psi(0) &= \int_0^1 \Psi'(t)dt \\ &= \frac{1}{2}\sum_{j=1}^p\sum_{i=1}^n \int_0^1 \mathrm{E}[\partial_j m(Z(t))\dot{Z}_{ij}(t)]dt = \frac{1}{2}(I + II + III),\end{aligned}$$

where

$$\dot{Z}_{ij}(t) = \frac{d}{dt}Z_{ij}(t) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{t}}x_{ij} - \frac{1}{\sqrt{1-t}}y_{ij}\right),$$

and

$$\begin{aligned}I &= \sum_{j=1}^p\sum_{i=1}^n \int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))\dot{Z}_{ij}(t)]dt, \\ II &= \sum_{j,k=1}^p\sum_{i=1}^n \int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))\dot{Z}_{ij}(t)Z_{ik}(t)]dt, \\ III &= \sum_{j,k,l=1}^p\sum_{i=1}^n \int_0^1\int_0^1 (1-\tau)\mathrm{E}[\partial_j\partial_k\partial_l m(Z^{(i)}(t) + \tau Z_i(t))\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)]d\tau dt.\end{aligned}$$

Since $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)$ are independent of each other and $\mathrm{E}[\dot{Z}_{ij}(t)] = 0$, we have

$$I = \sum_{j=1}^p\sum_{i=1}^n \int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))]\mathrm{E}[\dot{Z}_{ij}(t)]dt = 0.$$

Second, since $\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)] = n^{-1}\mathrm{E}[x_{ij}x_{ik} - y_{ij}y_{ik}] = 0$, we also have

$$II = \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))]\mathrm{E}[\dot{Z}_{ij}(t)Z_{ik}(t)]dt = 0.$$

Consider the third term III. We have

$$\begin{aligned}|III| &\lesssim (G_3 + G_2\beta + G_1\beta^2)\int_0^1 n\bar{\mathrm{E}}\left[\max_{1\leqslant j,k,l\leqslant p}|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|\right]dt \\ &\lesssim (G_3 + G_2\beta + G_1\beta^2)\bar{\mathrm{E}}\left[\max_{1\leqslant j\leqslant p}(|x_{ij}|^3 + |y_{ij}|^3)\right]/\sqrt{n}.\end{aligned}$$

where the first inequality follows from Lemma A.4 and the second inequality from Hölder's inequality. The first claim of the theorem now follows. The second claim follows directly from property (5) of the smooth max function. □

B.2. **Proof of Corollary 2.1.** Set $S := \bar{\mathrm{E}}[\max_{1\leqslant j\leqslant p}(|x_{ij}|^3 + |y_{ij}|^3)]/\sqrt{n}$, $\beta = (\log(pn))^{5/8}S^{-1/4}$ and $\psi = (\log(pn))^{-3/8}S^{-1/4}$. Consider a function $g_0 : \mathbb{R} \to [0,1]$ belonging to the class $C^3(\mathbb{R})$ such that $g_0(s) = 1$ for $s \leqslant 0$ and $g_0(s) = 0$ for $s \geqslant 1$. Fix any $t \in \mathbb{R}$, and define $g(s) = g_0(\psi(s - t - e_\beta))$. For this function g we have

$$G_0 = 1,\ G_1 \leqslant C\psi,\ G_2 \leqslant C\psi^2,\ G_3 \leqslant C\psi^3,$$

for some constant $C > 0$ depending only on g_0. Write $e_\beta = \beta^{-1}\log p$. Then by property (5) of the smooth max function and the definition of g,

$$\mathrm{P}(\max_{1\leqslant j\leqslant p} X_j \leqslant t) \leqslant \mathrm{P}(F_\beta(X) \leqslant t + e_\beta) \leqslant \mathrm{E}[g(F_\beta(X))].$$

Here by Theorem 2.1,

$$\mathrm{E}[g(F_\beta(X))] - \mathrm{E}[g(F_\beta(Y))] \lesssim (\psi^3 + \beta\psi^2 + \beta^2\psi)S.$$

We now bound $\mathrm{E}[g(F_\beta(Y))]$ as follows. By the definition of g and property (5) of the smooth max function, we have

$$\mathrm{E}[g(F_\beta(Y))] \leqslant \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t + e_\beta + \psi^{-1}).$$

This is where Lemma 2.1 plays its role. Let $u := e_\beta + \psi^{-1}$. Then $u \geqslant \psi^{-1} \geqslant S^{1/4} \gtrsim_{c_1} n^{-1/8}$, so that $\log(p/u) \lesssim_{c_1} \log(pn)$. Hence by Lemma 2.1,

$$\mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t + e_\beta + \psi^{-1}) - \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t) \lesssim_{c_1,C_1} (e_\beta + \psi^{-1})\sqrt{\log(pn)},$$

by which we conclude that

$$\begin{aligned}&\mathrm{P}(\max_{1\leqslant j\leqslant p} X_j \leqslant t) - \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t) \\ &\qquad\lesssim_{c_1,C_1} (\psi^3 + \beta\psi^2 + \beta^2\psi)S + (e_\beta + \psi^{-1})\sqrt{\log(pn)}.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t) - \mathrm{P}(\max_{1\leqslant j\leqslant p} X_j \leqslant t) \\ \lesssim_{c_1,C_1} (\psi^3 + \beta\psi^2 + \beta^2\psi)S + (e_\beta + \psi^{-1})\sqrt{\log(pn)}.\end{aligned}$$

Substituting $\beta = (\log(pn))^{5/8}S^{-1/4}$ and $\psi = (\log(pn))^{-3/8}S^{-1/4}$ leads to the desired result. □

B.3. **Proof of Theorem 2.2.** The second claim of the theorem follows from property (5) of the smooth max function. Hence we shall prove the first claim. The proof strategy is similar to the proof of Theorem 2.1. However, to control effectively the third order terms in the leave-one-out expansions we need to invoke truncation; for this purpose we shall replace X and Y by their truncated versions $\tilde{X}$ and $\tilde{Y}$, defined as follows: Let $\tilde{x}_i = (\tilde{x}_{ij})_{j=1}^p$, where $\tilde{x}_{ij}$ was defined before the statement of the theorem, and define the truncated version of X as $\tilde{X} = n^{-1/2}\sum_{i=1}^n \tilde{x}_i$. Also, recall that $\tilde{y}_i = (\tilde{y}_{ij})_{j=1}^p$, where $\tilde{y}_{ij} = y_{ij}1\{|y_{ij}| \leqslant u(\mathrm{E}[y_{ij}^2])^{1/2}\}$, and $\tilde{Y} = n^{-1/2}\sum_{i=1}^n \tilde{y}_i$. Without loss of generality, we will assume that sequences $(x_i)_{i=1}^n$ and $(y_i)_{i=1}^n$ are independent.

The proof consists of four steps. Step 1 will show that we can replace X by $\tilde{X}$ and Y by $\tilde{Y}$. Step 2 will bound the difference of the expectations of the relevant functions of $\tilde{X}$ and $\tilde{Y}$. This is the main step of the proof. Steps 3 and 4 will carry out supporting calculations. The steps of the proof will also call on various technical lemmas collected in section A.

Step 1. Let $m := g \circ F_\beta$. The main goal is to bound $\mathrm{E}[m(X) - m(Y)]$. Define

$$\mathcal{I} = 1\left\{\max_{1\leqslant j\leqslant p} |X_j - \tilde{X}_j| \leqslant \Delta(\gamma, u) \text{ and } \max_{1\leqslant j\leqslant p} |Y_j - \tilde{Y}_j| = 0\right\},$$

where

$$\Delta(\gamma, u) = M_2\varphi(u)\sqrt{2\log(p/\gamma)}(5 + \delta(u, \gamma)).$$

By Lemmas A.7 and A.10 we have $\mathrm{E}[\mathcal{I}] \geqslant 1 - 6\gamma$. Hence

$$\begin{aligned}|\mathrm{E}[m(X) - m(\tilde{X})]| &\leqslant |\mathrm{E}[(m(X) - m(\tilde{X}))\mathcal{I}]| + |\mathrm{E}[(m(X) - m(\tilde{X}))(1 - \mathcal{I})]| \\ &\leqslant G_1\Delta(\gamma, u) + 12G_0\gamma,\end{aligned}$$

where we invoked Lemma A.6 and

$$\begin{aligned}|\mathrm{E}[m(Y) - m(\tilde{Y})]| &\leqslant \mathrm{E}[(m(Y) - m(\tilde{Y}))\mathcal{I}]| + |\mathrm{E}[(m(Y) - m(\tilde{Y}))(1 - \mathcal{I})] \\ &\leqslant 12G_0\gamma.\end{aligned}$$

Therefore,

$$|\mathrm{E}[m(X) - m(Y)]| \leqslant |\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| + G_1\Delta(\gamma, u) + 24G_0\gamma.$$

Step 2 (Main Step) The purpose of this step is to establish the bound:

$$|\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| \lesssim (G_3 + G_2\beta + G_1\beta^2)M_3^3/\sqrt{n} + (G_2 + \beta G_1)M_2^2(\varphi(u) + \varphi_N(u)).$$

Define, as in the proof of Theorem 2.1,

$$Z(t) := \sqrt{t}\tilde{X} + \sqrt{1-t}\tilde{Y} = \sum_{i=1}^{n} Z_i(t), \quad Z_i(t) := \frac{1}{\sqrt{n}}(\sqrt{t}\tilde{x}_i + \sqrt{1-t}\tilde{y}_i).$$

and

$$Z^{(i)}(t) := Z(t) - Z_i(t), \ \dot{Z}_{ij}(t) = \frac{1}{\sqrt{n}}\left(\frac{1}{\sqrt{t}}\tilde{x}_{ij} - \frac{1}{\sqrt{1-t}}\tilde{y}_{ij}\right).$$

Arguing as in the proof of Theorem 2.1, we have

$$|\mathrm{E}[m(\tilde{X}) - m(\tilde{Y})]| = \frac{1}{2}\sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z(t))\dot{Z}_{ij}(t)]dt = \frac{1}{2}(I + II + III),$$

where

$$\begin{aligned}
I &= \sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))\dot{Z}_{ij}(t)]dt,\\
II &= \sum_{j,k=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z^{(i)}(t))\dot{Z}_{ij}(t)Z_{ik}(t)]dt,\\
III &= \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 (1-\tau)\mathrm{E}[\partial_j\partial_k\partial_l m(Z^{(i)}(t) + \tau Z_i(t))\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)]d\tau dt.
\end{aligned}$$

By independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)$together with the fact that $\mathrm{E}[\dot{Z}_{ij}(t)] = 0$, we have

$$I = \sum_{j=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[\partial_j m(Z^{(i)}(t))]\mathrm{E}[\dot{Z}_{ij}(t)]dt = 0.$$

Moreover, in steps 3 and 4 below, we will show that

$$\begin{aligned}
|II| &\lesssim (G_2 + \beta G_1)M_2^2(\varphi(u) + \varphi_N(u)),\\
|III| &\lesssim (G_3 + G_2\beta + G_1\beta^2)M_3^3/\sqrt{n}.
\end{aligned}$$

The claim of this step now follows.

Step 3. (Bound on II) Observe first that by the independence of $Z^{(i)}(t)$ and $Z_i(t)$,

$$\begin{aligned}
|II| &= \left| \sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathrm{E}[\partial_j \partial_k m(Z^{(i)}(t))] \mathrm{E}[\dot{Z}_{ij}(t) Z_{ik}(t)] dt \right| \\
&\leqslant \sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathrm{E}[|\partial_j \partial_k m(Z^{(i)}(t))|] \cdot |\mathrm{E}[\dot{Z}_{ij}(t) Z_{ik}(t)]| dt \\
&\leqslant \sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathrm{E}[U_{jk}(Z^{(i)}(t))] \ \cdot |\mathrm{E}[\dot{Z}_{ij}(t) Z_{ik}(t)]| dt.
\end{aligned}$$

where the last step follows from Lemma A.4. Now since $(\sqrt{t}\tilde{x}_{ij} + \sqrt{1-t}\tilde{y}_{ij}) \leqslant 2\sqrt{2}u$ and so that $\beta(\sqrt{t}\tilde{x}_{ij} + \sqrt{1-t}\tilde{y}_{ij})/\sqrt{n} \leqslant 1$ (which is satisfied by the assumption on β, u and γ), by Lemma A.5, the last expression is bounded by

$$\begin{aligned}
&\sum_{j,k=1}^{p} \sum_{i=1}^{n} \int_0^1 \mathrm{E}[U_{jk}(Z(t))] \cdot |\mathrm{E}[\dot{Z}_{ij}(t) Z_{ik}(t)]| dt \\
&\leqslant \sum_{j,k=1}^{p} \sum_{i=1}^{n} \left\{ \int_0^1 \mathrm{E}[U_{jk}(Z(t))] dt \right\} \left(\mathrm{E}[x_{ij}^2] + \mathrm{E}[x_{ik}^2] \right) n^{-1} (\varphi(u) + \varphi_N(u)) \\
&= \sum_{j,k=1}^{p} \left\{ \int_0^1 \mathrm{E}[U_{jk}(Z(t))] dt \right\} \cdot \left(\bar{\mathrm{E}}[x_{ij}^2] + \bar{\mathrm{E}}[x_{ik}^2] \right) (\varphi(u) + \varphi_N(u)) \\
&\lesssim (G_2 + G_1 \beta) M_2^2 (\varphi(u) + \varphi_N(u)).
\end{aligned}$$

Here the first inequality is due to Lemmas A.7 (b) and A.10 (b), and the last inequality is due to $\sum_{j,k=1}^{p} U_{jk}(z) \leqslant G_2 + 2G_1\beta$ established in Lemma A.4.

Step 4. (Bound on III) Observe that

$$
\begin{aligned}
|III| &\leqslant_{(1)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t)+\tau Z_i(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]d\tau dt \\
&\lesssim_{(2)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&=_{(3)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jkl}(Z^{(i)}(t))]\cdot\mathrm{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&\lesssim_{(4)} \sum_{j,k,l=1}^{p}\sum_{i=1}^{n}\int_0^1 \mathrm{E}[U_{jkl}(Z(t))]\cdot\mathrm{E}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&=_{(5)} \sum_{j,k,l=1}^{p}\int_0^1 \mathrm{E}[U_{jkl}(Z(t))]\cdot n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&\leqslant_{(6)} \int_0^1\left(\sum_{j,k,l=1}^{p}\mathrm{E}[U_{jkl}(Z(t))]\right)\max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&\lesssim_{(7)} (G_3+G_2\beta+G_1\beta^2)\int_0^1 \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt.
\end{aligned}
$$

where (1) follows from $|\partial_j\partial_k\partial_l m(z)| \leqslant U_{jkl}(z)$ (see Lemma A.4), (2) from Lemma A.5, (3) from the independence of $Z^{(i)}(t)$ and $\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)$, (4) from Lemma A.5, (5) from the definition of $\bar{\mathrm{E}}$, (6) from a trivial inequality, (7) from Lemma A.4. We have to bound the integral on the last line. Let $\omega(t) = 1/(\sqrt{t}\wedge\sqrt{1-t})$, and observe that

$$
\begin{aligned}
&\int_0^1 \max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)Z_{ik}(t)Z_{il}(t)|]dt \\
&= \int_0^1 \omega(t)\max_{1\leqslant j,k,l\leqslant p} n\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)(\sqrt{t}\wedge\sqrt{1-t})Z_{ik}(t)Z_{il}(t)|]dt \\
&\leqslant_{(1)} \int_0^1 \omega(t)\max_{1\leqslant j,k,l\leqslant p}\left(\bar{\mathrm{E}}[|\dot{Z}_{ij}(t)(\sqrt{t}\wedge\sqrt{1-t})|^3]\bar{\mathrm{E}}[|Z_{ik}(t)|^3]\bar{\mathrm{E}}[|Z_{il}(t)|^3]\right)^{1/3}dt \\
&\leqslant_{(2)} n^{-1/2}\left\{\int_0^1\omega(t)dt\right\}\max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[(|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)^3] \\
&\lesssim_{(3)} n^{-1/2}\max_{1\leqslant j\leqslant p}[(\bar{\mathrm{E}}[|\tilde{x}_{ij}|^3])^{1/3}+(\bar{\mathrm{E}}[|\tilde{y}_{ij}|^3])^{1/3}]^3 \\
&\lesssim_{(4)} n^{-1/2}\max_{1\leqslant j\leqslant p}[(\bar{\mathrm{E}}[|x_{ij}|^3])^{1/3}+(\bar{\mathrm{E}}[|y_{ij}|^3])^{1/3}]^3 \\
&\lesssim_{(5)} n^{-1/2}\max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[|x_{ij}|^3],
\end{aligned}
$$

where (1) follows from Hölder's inequality, (2) from the fact that $|\dot{Z}_{ij}(t)(\sqrt{t}\wedge\sqrt{1-t})| \leqslant (|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)/\sqrt{n}$, $|Z_{ik}(t)| \leqslant (|\tilde{x}_{ik}|+|\tilde{y}_{ik}|)/\sqrt{n}$, and $|Z_{il}(t)| \leqslant (|\tilde{x}_{il}|+|\tilde{y}_{il}|)/\sqrt{n}$, and that $\bar{\mathrm{E}}[(|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)^3]^{1/3}\bar{\mathrm{E}}[(|\tilde{x}_{ik}|+|\tilde{y}_{ik}|)^3]^{1/3}\bar{\mathrm{E}}[(|\tilde{x}_{il}|+|\tilde{y}_{il}|)^3]^{1/3} \leqslant \max_{1\leqslant j\leqslant p}\bar{\mathrm{E}}[(|\tilde{x}_{ij}|+|\tilde{y}_{ij}|)^3]$, (3) from $\int_0^1\omega(t)dt \lesssim 1$, (4) from Lemmas A.7 (a) and A.10 (a), and (5) from the normality of y_{ij} with $\mathrm{E}[y_{ij}^2]=\mathrm{E}[x_{ij}^2]$, so that $\bar{\mathrm{E}}[|y_{ij}|^3]^{1/3} \lesssim \bar{\mathrm{E}}[y_{ij}^2]^{1/2} = \bar{\mathrm{E}}[|x_{ij}^2|]^{1/2} \leqslant \bar{\mathrm{E}}[|x_{ij}|^3]^{1/3}$. □

B.4. **Proof of Corollary 2.2.** Following the argument in the proof of Corollary 2.1 and applying Theorem 2.2, for $u \geqslant u_1 \vee u_2 \vee u_3$ (note that $u \geqslant \sqrt{2\log(2pn/\gamma)}$), we have

$$\begin{aligned}
&\sup_{t\in\mathbb{R}}|\mathrm{P}(\max_{1\leqslant j\leqslant p}X_j \leqslant t) - \mathrm{P}(\max_{1\leqslant j\leqslant p}Y_j \leqslant t)| \\
&\quad\lesssim_{c_1,C_1} (\psi^3+\psi^2\beta+\psi\beta^2)M_3^3/\sqrt{n} + (\psi^2+\psi\beta)M_2^2(\varphi(u)+\varphi_N(u)) \\
&\quad+ \psi M_2\varphi(u)(1+\delta(u,\gamma))\sqrt{\log(p/\gamma)} + \beta^{-1}(\log(pn))^{3/2} + \psi^{-1}\sqrt{\log(pn)} + \gamma.
\end{aligned}$$

Let $u^\star$ be the smallest number $u \geqslant 0$ such that:

$$\sqrt{n}(\varphi(u)+\varphi_N(u))^{4/3} \leqslant M_3^3(\log(pn/\gamma))^{5/6}.$$

Consider the case $u > u^\star$. Let

$$\psi = n^{1/8}(\log(pn/\gamma))^{-3/8}M_3^{-3/4} \text{ and } \beta = \sqrt{n}/(2\sqrt{2}u).$$

Note that $\beta \lesssim \sqrt{n}/u \leqslant \sqrt{n}/u_1 \leqslant n^{1/8}(\log(pn/\gamma))^{5/8}M_3^{-3/4} =: \bar{\beta}$. Hence

$$\begin{aligned}
(\psi^3+\psi^2\beta+\psi\beta^2)M_3^3/\sqrt{n} &\lesssim \psi\bar{\beta}^2M_3^3/\sqrt{n} \\
&\lesssim n^{-1/8}(\log(pn/\gamma))^{7/8}M_3^{3/4} \\
&\leqslant u_1(\log(pn/\gamma))/n^{1/2} \\
&\leqslant u(\log(pn/\gamma))^{3/2}/n^{1/2}.
\end{aligned}$$

Since $M_2 \leqslant C_1$ and $\delta(u,\gamma) \leqslant C_1$,

$$\begin{aligned}
(\psi^2+\psi\beta)M_2^2(\varphi(u)+\varphi_N(u)) &\lesssim_{C_1} \psi\bar{\beta}(\varphi(u)+\varphi_N(u)) \\
&\leqslant \psi\bar{\beta}(\varphi(u^\star)+\varphi_N(u^\star)) \\
&= n^{-1/8}(\log(pn/\gamma))^{7/8}M_3^{3/4} \\
&\leqslant u(\log(pn/\gamma))^{3/2}/\sqrt{n},
\end{aligned}$$

and

$$\begin{aligned}
\psi M_2\varphi(u)(1+\delta(u,\gamma))\sqrt{\log(p/\gamma)} &\lesssim_{C_1} \psi\varphi(u)\sqrt{\log(pn/\gamma)} \\
&\leqslant \psi\bar{\beta}(\varphi(u)+\varphi_N(u))\sqrt{\log(pn/\gamma)}/\bar{\beta} \\
&\leqslant u(\log(pn/\gamma))^{3/2}/\sqrt{n},
\end{aligned}$$

where the last inequality follows from the calculations above and assuming that $\sqrt{\log(pn/\gamma)}/\bar{\beta} \lesssim 1$ (otherwise, the claim of the corollary is trivial).

Finally, by the definition of β and ψ, $(\log(pn/\gamma))^{3/2}/\beta + \sqrt{\log(pn/\gamma)}/\psi \lesssim u(\log(pn/\gamma))^{3/2}$. Hence the claim of the corollary for the case $u > u^\star$ follows.

Consider the case $u \leqslant u^\star$. Let

$$\psi = (\log(pn/\gamma))^{-1/6}(\varphi(u) + \varphi_N(u))^{-1/3} \text{ and } \beta = \sqrt{n}/(2\sqrt{2}u).$$

Note that $\psi \leqslant (\log(pn/\gamma))^{-1/6}(\varphi(u^\star)+\varphi_N(u^\star))^{-1/3} = n^{1/8}(\log(pn/\gamma))^{-3/8}M_3^{-3/4}$, which is the value of ψ used for the case $u > u^\star$. So, $(\psi^3 + \psi^2\beta + \psi\beta^2)M_3^3/\sqrt{n} \lesssim u(\log(pn/\gamma))^{3/2}$ by the same argument as above. Moreover,

$$\begin{aligned}
\psi\beta M_2^2(\varphi(u) + \varphi_N(u)) &\lesssim_{C_1} \beta(\varphi(u) + \varphi_N(u))^{2/3}(\log(pn/\gamma))^{-1/6} \\
&\leqslant \beta(\varphi(u_3) + \varphi_N(u_3))^{2/3}(\log(pn/\gamma))^{-1/6} \\
&= \beta u_3^2(\log(pn/\gamma))^{5/3-1/6}/n \\
&\leqslant u_3(\log(pn/\gamma))^{5/3-1/6}/\sqrt{n} \\
&\leqslant u(\log(pn/\gamma))^{3/2}/\sqrt{n},
\end{aligned}$$

and

$$\begin{aligned}
\psi^2 M_2^2(\varphi(u) + \varphi_N(u)) &\lesssim_{C_1} (\varphi(u) + \varphi_N(u))^{1/3}(\log(pn/\gamma))^{-1/3} \\
&\leqslant (\varphi(u_3) + \varphi_N(u_3))^{1/3}(\log(pn/\gamma))^{-1/3} \\
&\leqslant u_3\sqrt{\log(pn/\gamma)}/\sqrt{n} \\
&\leqslant u(\log(pn/\gamma))^{3/2}/\sqrt{n}.
\end{aligned}$$

By the definition of β,

$$(\log(pn))^{3/2}/\beta \lesssim u(\log(pn/\gamma))^{3/2}/\sqrt{n}.$$

Moreover, by the same argument as that used for the case $u > u^\star$,

$$\begin{aligned}
\psi M_2\varphi(u)(1 + \delta(u,\gamma))\sqrt{\log(p/\gamma)} &\lesssim_{C_1} \psi\varphi(u)\sqrt{\log(pn/\gamma)} \\
&\leqslant (\varphi(u) + \varphi_N(u))^{2/3}(\log(pn/\gamma))^{1/3} \\
&\leqslant (\varphi(u_3) + \varphi_N(u_3))^{2/3}(\log(pn/\gamma))^{1/3} \\
&\leqslant (u\log(pn/\gamma)/\sqrt{n})^2 \leqslant (u(\log(pn/\gamma))^{3/2}/\sqrt{n})^2 \\
&\lesssim u(\log(pn/\gamma))^{3/2}/\sqrt{n},
\end{aligned}$$

where on the last step we assume that $u(\log(pn/\gamma))^{3/2}/\sqrt{n} \lesssim 1$ because otherwise the claim of the corollary is trivial. Finally, by the same argument as above,

$$\sqrt{\log(pn/\gamma)}/\psi = (\varphi(u) + \varphi_N(u))^{1/3}(\log(pn/\gamma))^{2/3} \leqslant u(\log(pn/\gamma))^{3/2}/\sqrt{n}.$$

The claim of the corollary for the case $u \leqslant u^\star$ follows. □

B.5. **Proof of Corollary 2.3.** In this proof, let $c > 0$ and $C > 0$ denote generic constants depending only on c_1, c_2, C_1 and the their values may change from place to place. Assume that $\gamma \geqslant n^{-c}$ for some $c > 0$. In all cases, we will choose γ so that this assumption holds. Recall the definitions of u_1, u_2, and u_3 given before the statement of corollary 2.2. Let $u = u_0 + u_1 + u_2 + u_3$ where $u_0 > 0$ is to be chosen later.

By Corollary 2.2,

$$\rho \leqslant C \left\{ (u_0 + u_1 + u_2 + u_3)(\log(pn/\gamma))^{3/2}/\sqrt{n} + \gamma \right\},$$

provided that $\delta(u, \gamma) \leqslant C$. Under conditions E.1-2, $M_3^3 \leqslant C$ and under conditions E.3-5, $M_3^3 \leqslant CB_n$. Therefore, under either conditions of the Corollary, $u_1(\log(pn/\gamma))^{3/2}/\sqrt{n} \leqslant Cn^{-c}$ since $u_1 = n^{3/8} M_3^{3/4}/(\log(pn/\gamma))^{5/8}$. Moreover, $u_2(\log(pn/\gamma))^{3/2}/\sqrt{n} \leqslant Cn^{-c}$ since $u_2 = \sqrt{2\log(2pn/\gamma)}$. Hence, it suffices to choose u_0 and γ in such a way that $(u_0+u_3)(\log(pn/\gamma))^{3/2}/\sqrt{n}+\gamma \leqslant Cn^{-c}$ and $\delta(u, \gamma) \leqslant C$.

Case E.1. Set $u_0 = C\sqrt{\log(pn)}$ for sufficiently large $C > 0$ and $\gamma = 1/n$. Then by Comment 2.2, $\delta(u, \gamma) \leqslant \delta(u_0, \gamma) = 0$. To derive a bound on u_3, let

$$f_l(t) = \sqrt{n}(\varphi(t) + \varphi_N(t))^{1/3} \text{ and } f_r(t) = t(\log(pn/\gamma))^{5/6}.$$

By equation (9), u_3 is the smallest number $u \geqslant 0$ such that $f_l(u) \leqslant f_r(u)$. Note that $f_l(\cdot)$ is decreasing while $f_r(\cdot)$ is increasing. Moreover, $f_l(u_0) \leqslant 1/(p^{2/3}n^{1/6}) \leqslant 1$ because $\varphi(u_0) \leqslant 1/(pn)^2$ by Comment 2.2 and $\varphi_N(u_0) \leqslant 1/(pn)^2$ by Theorem 2.2 (C is sufficiently large). On the other hand, $f_r(u_0) \geqslant 1$. Therefore, $f_l(u_0) \leqslant f_r(u_0)$, and so $u_3 \leqslant u_0$, from which we conclude that that $(u_0 + u_3)(\log(pn/\gamma))^{3/2}/\sqrt{n} \leqslant Cn^{-c}$.

Case E.4. This case is similar to case E.1. Indeed, when $x_{ij} = z_{ij}\varepsilon_i$ and z_{ij} are nonstochastic, equation (7), which defines $\varphi(u)$, becomes

$$\sqrt{\mathrm{E}[\varepsilon_i^2 1\{|\varepsilon_i| > u\sqrt{\mathrm{E}[\varepsilon_i^2]}\}]} \leqslant \sqrt{\mathrm{E}[\varepsilon_i^2]}\varphi.$$

Moreover, $x_{ij} - \tilde{x}_{ij} = z_{ij}(\varepsilon_i - \tilde{\varepsilon}_i)$ where

$$\tilde{\varepsilon}_i = \varepsilon_i 1\left\{ |\varepsilon_i| \leqslant u\sqrt{\mathrm{E}[\varepsilon_i^2]} \right\} - \mathrm{E}\left[\varepsilon_i 1\left\{ |\varepsilon_i| \leqslant u\sqrt{\mathrm{E}[\varepsilon_i^2]} \right\} \right].$$

Therefore, setting $u_0 = C\sqrt{\log(pn)}$ for sufficiently large $C > 0$ and $\gamma = 1/n$ gives $\delta(u, \gamma) = 0$ and $u_3 \leqslant u_0$. The conclusion follows as in case E.1.

Case E.3. Set $u_0 = CB_n$ for sufficiently large $C > 0$ and $\gamma = 1/n$. Then by comment 2.2, $\delta(u, \gamma) = 0$ and calculations like those used above show that $u_3 \leqslant u_0$. Hence $(u_0 + u_3)(\log(pn/\gamma))^{3/2}/\sqrt{n} + \gamma \leqslant Cn^{-c}$.

Case E.2. Set $u_0 = n^{2/7}/\sqrt{C_1}$ and $\gamma = C_1 n^{-1/7}$. Then

$$\begin{aligned}\mathrm{P}\left(\max_{i,j}(|x_{ij}| - u\sqrt{\mathrm{E}[x_{ij}^2]}) > 0\right) &\leqslant \mathrm{P}\left(\max_{i,j}|x_{ij}| > u\sqrt{C_1}\right)\\ &\leqslant \sum_{i=1}^n \mathrm{E}[\max_{1\leqslant j\leqslant p} x_{ij}^4]/(u\sqrt{C_1})^4\\ &\leqslant n/(u^4 C_1)\\ &\leqslant n/(u_0^4 C_1) = \gamma,\end{aligned}$$

and hence $\delta(u,\gamma) = 0$. Moreover,

$$\mathrm{E}\left[x_{ij}^2 1\{|x_{ij}| > u\sqrt{\mathrm{E}[x_{ij}^2]}\}\right] \leqslant \mathrm{E}[x_{ij}^4]/(u^2\mathrm{E}[x_{ij}^2]) \lesssim_{c_1,C_1} 1/u^2,$$

so that $\varphi(u) \lesssim_{c_1,C_1} 1/u$. Hence $f_l(\tilde{u}) \lesssim_{c_1,C_1} f_r(\tilde{u})$ with $\tilde{u} = n^{3/8}/(\log(pn/\gamma))^{5/8}$ and $u_3 \leqslant C\tilde{u}$, from which we have $(u_0+u_3)(\log(pn/\gamma))^{3/2}/\sqrt{n} + \gamma \leqslant Cn^{-c}$.

Case E.5. This case is similar to case E.2 and hence we omit the detail. □

Appendix C. Proof for Section 3

C.1. **Proof of Theorem 3.1.** We shall use the following form of Stein's identity.

Lemma C.1 ([45])**.** *Let (W,U) be a zero-mean Gaussian random vector, where W is scalar and U is a p-vector. Let $f : \mathbb{R}^p \to \mathbb{R}$ be a function (of moderate growth at infinity). Then*

$$\mathrm{E}[Wf(U)] = \sum_{j=1}^p \mathrm{E}[WU_j]\mathrm{E}[\partial_j f(U)].$$

Proof of Lemma C.1. See Section A.6 of [45], and also [44]. □

Proof of Theorem 3.1. Without loss of generality, we can and will assume that V and Y are independent, so that $\mathrm{E}[V_j Y_k] = 0$ for all $j,k = 1,\ldots,p$. For $t \in [0,1]$, define the following Slepian interpolant:

$$Z(t) := \sqrt{t}V + \sqrt{1-t}Y.$$

Let $m := g \circ F_\beta$ and $\Psi(t) := \mathrm{E}[m(Z(t))]$. Then

$$|\mathrm{E}[m(V) - m(Y)]| = |\Psi(1) - \Psi(0)| = \left|\int_0^1 \Psi'(t)dt\right|.$$

Here we have

$$\begin{aligned}\Psi'(t) &= \frac{1}{2}\sum_{j=1}^p \mathrm{E}\left[\partial_j m(Z(t))(t^{-1/2}V_j - (1-t)^{-1/2}Y_j)\right]\\ &= \frac{1}{2}\sum_{j=1}^p\sum_{k=1}^p (\Sigma_{jk}^V - \Sigma_{jk}^Y)\mathrm{E}[\partial_j\partial_k m(Z(t))],\end{aligned}$$

where the second equality holds by application of Lemma C.1 to $W = t^{-1/2}V_j - (1-t)^{-1/2}Y_j$ and $f(U) = \partial_j m(U)$ with $U = Z(t)$. Hence,

$$\begin{aligned}\left|\int_0^1 \Psi'(t)dt\right| &\leqslant \frac{1}{2}\sum_{1\leqslant j,k\leqslant p}\left|\Sigma_{jk}^V - \Sigma_{jk}^Y\right|\left|\int_0^1 \mathrm{E}[\partial_j\partial_k m(Z(t))]dt\right| \\ &\leqslant \frac{1}{2}\max_{1\leqslant j,k\leqslant p}\left|\Sigma_{jk}^V - \Sigma_{jk}^Y\right|\int_0^1 \sum_{1\leqslant j,k\leqslant p}\left|\mathrm{E}[\partial_j\partial_k m(Z(t))]\right| dt \\ &= \frac{\Delta_0}{2}\int_0^1 \sum_{1\leqslant j,k\leqslant p}\left|\mathrm{E}[\partial_j\partial_k m(Z(t))]\right| dt.\end{aligned}$$

In view of Lemmas A.2 and A.3,

$$\sum_{1\leqslant j,k\leqslant p}|\partial_j\partial_k m(Z(t))| \leqslant |\partial^2 g(F_\beta(Z(t)))| + 2\beta|\partial g(F_\beta(Z(t))|.$$

Hence we obtain

$$\begin{aligned}&|\mathrm{E}[g(F_\beta(V)) - g(F_\beta(Y))]| \\ &\leqslant \Delta_0 \times \left(2^{-1}\int_0^1 \mathrm{E}|\partial^2 g(F_\beta(Z(t)))|dt + \beta\int_0^1 \mathrm{E}|\partial g(F_\beta(Z(t))|dt\right) \\ &\leqslant \Delta_0(2^{-1}G_2 + \beta G_1),\end{aligned} \tag{24}$$

which leads to the first claim. The second claim follows from property (5) of the smooth max function. □

C.2. **Proof of Corollary 3.1.** Recall that $\Delta_0 := \max_{1\leqslant j,k\leqslant p}|\Sigma_{jk}^V - \Sigma_{jk}^Y|$. For $\beta > 0$, define $e_\beta = \beta^{-1}\log p$. Set $\beta = (\log(p/\Delta_0))^{5/6}\Delta_0^{-1/3}$ and $\psi = (\log(p/\Delta_0))^{-1/6}\Delta_0^{-1/3}$. Fix any $t \in \mathbb{R}$. Using the same argument as in the proof of Corollary 2.1 and applying Theorem 3.1, we have

$$\mathrm{P}(\max_{1\leqslant j\leqslant p} X_j \leqslant t) - \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t + e_\beta + \psi^{-1}) \lesssim (\psi^2 + \beta\psi)\Delta_0.$$

Note that $e_\beta + \psi^{-1} \geqslant \psi^{-1} \geqslant \Delta_0^{1/3} \geqslant \Delta_0$. Hence application of Lemma 2.1 leads to

$$\mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t + e_\beta + \psi^{-1}) - \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t) \lesssim_{c_1,C_1} (e_\beta + \psi^{-1})\sqrt{\log(p/\Delta_0)}.$$

Combining these inequalities and substituting the values of β and ψ leads to

$$\mathrm{P}(\max_{1\leqslant j\leqslant p} X_j \leqslant t) - \mathrm{P}(\max_{1\leqslant j\leqslant p} Y_j \leqslant t) \lesssim_{c_1,C_1} \Delta_0^{1/3}(\log(p/\Delta_0))^{2/3}.$$

This gives one half of the asserted claim. The second half is similar. □

C.3. **Proof of Corollary 3.2.** It is easy to check that the map $\vartheta \mapsto \vartheta^{1/3}(\log(p/\vartheta))^{2/3}$ is increasing in ϑ for $\vartheta \in (0, e^{-2})$. So, by Corollary 3.1, on the event $\{(x_i)_{i=1}^n : \Delta \leqslant \vartheta\}$, we have $|\mathrm{P}(Z_0 \leqslant t) - \mathrm{P}_e(W_0 \leqslant t)| \lesssim_{c_1,C_1} \vartheta^{1/3}(\log(p/\vartheta))^{2/3}$ for all $t \in \mathbb{R}$. Therefore, the claim of the corollary follows from the triangle inequality. □

Appendix D. Proofs for Section 4

D.1. **Proof of Lemma 4.1.** Recall that $\Delta = \max_{1\leqslant j,k\leqslant p}|\mathbb{E}_n[x_{ij}x_{ik}] - \bar{\mathrm{E}}[x_{ij}x_{ik}]|$. By Corollary 3.1, on the event $\{(x_i)_{i=1}^n : \Delta \leqslant \vartheta\}$, we have $|\mathrm{P}(Z_0 \leqslant t) - \mathrm{P}_e(W_0 \leqslant t)| \leqslant v(\vartheta)$ for all $t \in \mathbb{R}$, and so on this event

$$\mathrm{P}_e(W_0 \leqslant c_{Z_0}(\alpha+v(\vartheta))) \geqslant \mathrm{P}(Z_0 \leqslant c_{Z_0}(\alpha+v(\vartheta)))-v(\vartheta) \geqslant \alpha+v(\vartheta)-v(\vartheta) = \alpha,$$

implying the first claim of the lemma. The second claim follows similarly. □

D.2. **Proof of Lemma 4.2.** By equation (13), the probability of the event $\{(x_i)_{i=1}^n : \mathrm{P}_e(|W - W_0| > \zeta_1) \leqslant \zeta_2\}$ is at least $1 - \zeta_2$. On this event,

$$\mathrm{P}_e(W \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant \mathrm{P}_e(W_0 \leqslant c_{W_0}(\alpha + \zeta_2)) - \zeta_2 \geqslant \alpha + \zeta_2 - \zeta_2 = \alpha,$$

implying that $\mathrm{P}(c_W(\alpha) \leqslant c_{W_0}(\alpha + \zeta_2) + \zeta_1) \geqslant 1 - \zeta_2$. The second claim of the lemma follows similarly. □

D.3. **Proof of Theorem 4.1.** For $\vartheta > 0$, let $v(\vartheta) := C_2\vartheta^{1/3}(\log(p/\vartheta))^{2/3}$ with $C_2 > 0$ as in Lemma 4.1. Then

$$\begin{aligned}\mathrm{P}(T_0 \leqslant c_{W_0}(\alpha)) &\leqslant_{(1)} \mathrm{P}(T_0 \leqslant c_{Z_0}(\alpha + v(\vartheta))) + \mathrm{P}(\Delta > \vartheta)\\ &\leqslant_{(2)} \alpha + v(\vartheta) + \mathrm{P}(\Delta > \vartheta) + \rho,\end{aligned}$$

where (1) follows from Lemma 4.1 and (2) follows from the definition of ρ and the fact that Z_0 has no point masses. The upper bound in the claim of theorem follows by substituting $v(\vartheta)$. The lower bound follows from a similar argument.

D.4. **Proof of Theorem 4.2.** For $\vartheta > 0$, let $v(\vartheta) := C_2\vartheta^{1/3}(\log(p/\vartheta))^{2/3}$ with $C_2 > 0$ as in Lemma 4.1. Then

$$\begin{aligned}&\mathrm{P}(T \leqslant c_W(\alpha)) \leqslant_{(1)} \mathrm{P}(T_0 \leqslant c_W(\alpha) + \zeta_1) + \zeta_2\\ &\leqslant_{(2)} \mathrm{P}(T_0 \leqslant c_{W_0}(\alpha + \zeta_2) + 2\zeta_1) + 2\zeta_2\\ &\leqslant_{(3)} \mathrm{P}(T_0 \leqslant c_{Z_0}(\alpha + \zeta_2 + v(\vartheta)) + 2\zeta_1) + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta)\\ &\leqslant_{(4)} \mathrm{P}(T_0 \leqslant c_{Z_0}(\alpha + \zeta_2 + v(\vartheta) + C_3\zeta_1\sqrt{\log(p/\zeta_1)})) + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta)\\ &\leqslant_{(5)} \alpha + \zeta_2 + v(\vartheta) + C_3\zeta_1\sqrt{\log(p/\zeta_1)} + 2\zeta_2 + \mathrm{P}(\Delta > \vartheta) + \rho\end{aligned}$$

where $C_3 > 0$ depends on c_1 and C_1 only and where (1) follows from equation (12), (2) from Lemma 4.2, (3) from Lemma 4.1, (4) from Lemma 2.1, and (5) from the definition of ρ and the fact that Z_0 has no point masses. The upper bound in the claim of theorem follows by substituting $v(\vartheta)$. The lower bound follows from a similar argument. □

D.5. **Proof of Corollary 4.1.** Let $c > 0$ and $C > 0$ denote generic constants depending only on c_1, c_2, C_1, and their value may change from place to place. Corollary 2.3, in every case, $\rho \leqslant Cn^{-c}$. Moreover, since it is assumed throughout the paper that $p \geqslant 2$, $\zeta_1\sqrt{\log p} \leqslant C_1 n^{-c_2}$ implies that $\zeta_1 \lesssim n^{-c_2}$, and so $\zeta_1\sqrt{\log(p/\zeta_1)} \leqslant Cn^{-c}$. Also, $\zeta_2 \leqslant Cn^{-c}$ by assumption.

By Markov's inequality, $\mathrm{P}(\Delta > \vartheta) \leqslant \mathrm{E}[\Delta]/\vartheta$. In addition, $\vartheta^{1/3}(\log(p/\vartheta))^{2/3}$ converges to zero at least at a polynomial rate if $\vartheta(\log p)^2$ converges to zero at a polynomial rate. Therefore, we can find $\vartheta > 0$ such that $\vartheta^{1/3}(\log(p/\vartheta))^{2/3} + \mathrm{P}(\Delta > \vartheta) \leqslant Cn^{-c}$ if $\mathrm{E}[\Delta](\log p)^2 \leqslant Cn^{-c}$ (with possibly different C and c). The last bound follows by applying Lemma A.1. This gives the first claim of the corollary. The second claim follows similarly. □

Appendix E. Proofs for Section 5

E.1. **Proof of Theorem 5.1.** The proof proceeds in three steps. In the proof $(\widehat{\beta}, \lambda)$ denote $(\widehat{\beta}^{(k)}, \lambda^{(k)})$ with k either 0 or 1.

Step 1. Here we show that there exist some constant $c > 0$ and $C > 0$ (depending only c_1, C_1 and σ^2) such that for either $k \in \{0, 1\}$,

$$\mathrm{P}(T_0 \leqslant \lambda^{(k)}) \geqslant 1 - \alpha - \nu_n, \tag{25}$$

with $\nu = Cn^{-c}$. We first note that $T_0 = \sqrt{n}\max_{1\leqslant k\leqslant 2p} \mathbb{E}_n[\tilde{z}_{ik}\varepsilon_i]$, where $\tilde{z}_i = (z_i', -z_i')'$. Application of Corollary 2.3, case E.5, gives

$$|\mathrm{P}(T_0 \leqslant \lambda) - \mathrm{P}(Z_0 \leqslant \lambda)| \leqslant Cn^{-c},$$

where $c > 0$ and $C > 0$ are constants depending only on c_1, C_1 and σ^2. Since $\lambda \geqslant c_{Z_0}(1-\alpha)$ the claim follows. Indeed, $\lambda^{(1)} = c_{Z_0}(1-\alpha)$, and $\lambda^{(1)} \leqslant \lambda^{(0)} = c_0(1-\alpha) := \sigma\Phi^{-1}(1-\alpha/(2p))$, since by the union bound $\mathrm{P}(Z_0 \geqslant c_0(1-\alpha)) \leqslant 2p\mathrm{P}(\sigma N(0,1) \geqslant c_0(1-\alpha)) = \alpha$.

Step 2. We claim that with probability at least $1 - \alpha - \nu_n$, $\widehat{\delta} = \widehat{\beta} - \beta$ obeys:

$$\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]| \leqslant 2\lambda.$$

Indeed, by definition of $\widehat{\beta}$, $\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(y_i - z_i'\widehat{\beta})]| \leqslant \lambda$, which by the triangle inequality for the maximum norm implies that $\sqrt{n}\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]| \leqslant T_0 + \lambda$. The claim follows from Step 1.

Step 3. By Step 1, with probability at least $1 - \alpha - \nu_n$, the true parameter value β obeys the constraint in optimization problem (14), so that by definition of $\widehat{\beta}$ we must have $\|\widehat{\beta}\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}$. Therefore, with the same probability $\widehat{\delta} \in \mathcal{R}(\beta) = \{\delta \in \mathbb{R}^d : \|\beta + \delta\|_{\ell_1} \leqslant \|\beta\|_{\ell_1}\}$. By definition of $\kappa_I(\beta)$ we have that

$$\kappa_I(\beta)\|\widehat{\delta}\|_I \leqslant \max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}(z_i'\widehat{\delta})]|.$$

Combining this inequality with that in Step 2 gives the claim of the theorem. □

E.2. **Proof of Theorem 5.2.** The proof has four steps. In the proof, we let $\varrho_n = Cn^{-c}$ for sufficiently small $c > 0$ and sufficiently large $C > 0$ depending only on $c_1, C_1, \underline{\sigma}^2, \sigma^2$, where c and C (and hence ϱ_n) may change from place to place.

Step 0. The same argument as in the previous proof applies to $\widehat{\beta}^{(0)}$ with $\lambda = \lambda^{(0)} := c_0(1 - 1/n)$, noting that now σ^2 is the upper bound on $\mathrm{E}[\varepsilon_i^2]$. Thus, we conclude that with probability at least $1 - \varrho_n$,

$$\|\widehat{\beta}_0 - \beta\|_{\mathrm{pr}} \leqslant \frac{2c_0(1 - 1/n)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)}.$$

Step 1. We claim that with probability at least $1 - \varrho_n$,

$$\max_{1 \leqslant j \leqslant p} \left(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2]\right)^{1/2} \leqslant B_n \frac{2c_0(1 - 1/n)}{\sqrt{n}\kappa_{\mathrm{pr}}(\beta)} =: \iota_n.$$

Application of Hölder's inequality and identity $\varepsilon_i - \widehat{\varepsilon}_i = z_i'(\widehat{\beta}_0 - \beta)$ give

$$\max_{1 \leqslant j \leqslant p} \left(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2]\right)^{1/2} \leqslant B_n(\mathbb{E}_n[z_i'(\widehat{\beta}_0 - \beta)]^2)^{1/2} \leqslant B_n\|\widehat{\beta}_0 - \beta\|_{\mathrm{pr}}.$$

The claim follows from Step 0.

Step 2. In this step, we apply Corollary 4.1 to

$$\begin{aligned} T = T_0 &= \sqrt{n} \max_{1 \leqslant j \leqslant 2p} \mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i], \\ W &= \sqrt{n} \max_{1 \leqslant j \leqslant 2p} \mathbb{E}_n[\tilde{z}_{ij}\widehat{\varepsilon}_i e_i], \\ W_0 &= \sqrt{n} \max_{1 \leqslant j \leqslant 2p} \mathbb{E}_n[\tilde{z}_{ij}\varepsilon_i e_i], \end{aligned}$$

where $\tilde{z}_i = (z_i', -z_i')'$, to conclude that uniformly in $\alpha \in (0, 1)$

$$\mathrm{P}(T_0 \leqslant c_W(1 - \alpha)) \geqslant 1 - \alpha - \varrho_n. \tag{26}$$

To show applicability of Corollary 4.1 , we note that for any $\zeta_1 > 0$, we have

$$\begin{aligned} \mathrm{P}_e(|W - W_0| > \zeta_1) &\leqslant \mathrm{E}_e[|W - W_0|]/\zeta_1 \\ &\leqslant \sqrt{n}\mathrm{E}_e[\max_j |\mathbb{E}_n[z_{ij}(\widehat{\varepsilon}_i - \varepsilon_i)e_i]|]/\zeta_1 \\ &\lesssim \sqrt{\log p} \max_j (\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2}/\zeta_1, \end{aligned}$$

where the third inequality is due to Pisier's inequality. The last quantity is bounded by $(\iota_n^2 \log p)^{1/2}/\zeta_1$ with probability $\geqslant 1 - \varrho_n$ by Step 1.

Since $\iota_n \log p \leqslant C_1 n^{-c_1}$ by assumption (vi) of the theorem, we can set ζ_1 in such a way that $\zeta_1(\log p)^{1/2} \leqslant \varrho_n$ and $(\iota_n^2 \log p)^{1/2}/\zeta_1 \leqslant \varrho_n$. Then all conditions of Corollary 4.1 with so defined ζ_1 and $\zeta_2 = \varrho_n \vee ((\iota_n^2 \log p)^{1/2}/\zeta_1)$ hold, and hence application of the corollary then gives that uniformly in $\alpha \in (0, 1)$

$$|\mathrm{P}(T_0 \leqslant c_W(1 - \alpha)) - 1 - \alpha| \leqslant \varrho_n, \tag{27}$$

which implies the claim of this step.

Step 3. In this step we claim that with probability at least $1-\varrho_n$,

$$c_W(1-\alpha) \leqslant c_{Z_0}(1-\alpha+2\varrho_n).$$

Combining Step 2 and Lemma 4.2 gives that with probability at least $1-\zeta_2$, $c_W(1-\alpha) \leqslant c_{W_0}(1-\alpha+\zeta_2)+\zeta_1$, where ζ_1 and ζ_2 are chosen as in Step 2. In addition, Lemma 4.1 shows that $c_{W_0}(1-\alpha+\zeta_2) \leqslant c_{Z_0}(1-\alpha+\varrho_n)$. Finally, Lemma 2.1 yields $c_{Z_0}(1-\alpha+\varrho_n)+\zeta_1 \leqslant c_{Z_0}(1-\alpha+2\varrho_n)$. Combining presented bounds gives the claim of this step.

Step 4. Given (26), the rest of the proof is identical to Steps 2-3 in the preceding proof of Theorem 5.1, using $c = c_W(1-\alpha)$. The result follows for $\nu_n = 2\varrho_n$. □

Appendix F. Proofs for Section 6

F.1. **Proof of Theorem 6.1.** The multiplier bootstrap critical value $c_{1-\alpha,w}$ clearly satisfies $c_{1-\alpha,w} \leqslant c_{1-\alpha,w'}$ whenever $w \subset w'$, so that inequality (19) holds. Therefore, it suffices to prove (20). We will only consider $w = \mathcal{W}$, and note that the same $o(1)$ in equation (20) applies when $w \subset \mathcal{W}$. Also, we will only consider the case with four bounded moments. The sub-Gaussian case is similar. The following direct corollaries of assumptions will be used repeatedly in the sequel,

$$\begin{aligned} \max_{i,k,j} |x_{ikj}| &\leqslant \max_{i,k} \|x_{ik}\| \leqslant_{(1)} c_1^{-1} \max_{i,k} \|v_{ik}\| \\ &\leqslant c_1^{-1}\sqrt{\bar{p}} \max_{i,k,j} |v_{ikj}| \leqslant_{(2)} c_1^{-1}\sqrt{\bar{p}} B_n, \end{aligned} \tag{28}$$

$$\begin{aligned} \max_{k,j} \mathbb{E}_n[x_{ikj}^2] &\leqslant \max_{k} \mathbb{E}_n[\|x_{ik}\|^2] \\ &\leqslant_{(3)} c_1^{-2} \max_{k} \mathbb{E}_n[\|v_{ik}\|^2] \leqslant_{(4)} c_1^{-2}\bar{p}, \end{aligned} \tag{29}$$

where (1) and (3) follow from assumption M-(iv) and the definition of x_{ik}, (2) is from M-(i) since v_{ik} is a subvector of z_i, and (4) is due to M-(ii). We start with several lemmas. In all the lemmas below, we will assume the same conditions as in the theorem.

Lemma F.1. $\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n} = O_{\mathrm{P}}(r_{n1})$ *uniformly over* $k = 1,\dots,K$ *and* $j = 1,\dots,p_k$ *where* $r_{n1} = \sqrt{\bar{p}\log p}$.

Proof. Applying Lemma A.1 combined with inequalities (28) and (29) gives

$$\mathrm{E}[\max_{k,j} |\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}|] = (\sqrt{\bar{p}}B_n(\log p)/n^{1/4} + \sqrt{\bar{p}\log p}) = O(\sqrt{\bar{p}\log p}),$$

where the second step follows because $B_n\sqrt{\log p}/n^{1/4} = o(1)$. The asserted claim follows from Markov's inequality. □

Lemma F.2. $\mathbb{E}_n[x_{ikj}^2(\widehat{\varepsilon}_{ik}^2 - \sigma_{ik}^2)] = O_{\mathrm{P}}(r_{n2})$ *uniformly over* $k = 1,\dots,K$ *and* $j = 1,\dots,p_k$ *where* $r_{n2} = \bar{p}B_n^2\log p/\sqrt{n}$.

Proof. We have

$$\begin{aligned}\mathbb{E}_n[x_{ikj}^2(\widehat{\varepsilon}_{ik}^2-\sigma_{ik}^2)] &= \mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2-\sigma_{ik}^2)]+\mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k-\beta_k))^2]\\ &\quad -2\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}v_{ik}'(\widehat{\beta}_k-\beta_k)]\\ &=: I_{jk}+II_{jk}+III_{jk}.\end{aligned}$$

All three terms are bounded below in three steps. Summing up gives the result because $\bar{p}/\sqrt{n}\to 0$.

Step 1. We prove that $I_{jk}=\mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2-\sigma_{ik}^2)]=O_{\mathrm{P}}(r_{n21})$ uniformly over $k=1,\dots,K$ and $j=1,\dots,p_k$ where $r_{n21}=\bar{p}B_n^2\log p/\sqrt{n}$.

Applying Lemma A.1 combined with inequalities (28) and (29) gives

$$\begin{aligned}\mathrm{E}[\max_{k,j}|\mathbb{E}_n[x_{ikj}^2(\varepsilon_{ik}^2-\sigma_{ik}^2)]|] &= O(\bar{p}B_n^2\log p/\sqrt{n}+\bar{p}B_n\sqrt{(\log p)/n})\\ &= O(\bar{p}B_n^2\log p/\sqrt{n}),\end{aligned}$$

where the second step holds because B_n is bounded away from zero. The claim of this step follows from Markov's inequality.

Step 2. We prove that $II_{jk}=\mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k-\beta_k))^2]=O_{\mathrm{P}}(r_{n22})$ uniformly over $k=1,\dots,K$ and $j=1,\dots,p_k$ where $r_{n22}=\bar{p}^2B_n^2(\log p)/n$. We have

$$\begin{aligned}\max_{k,j}\mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k-\beta_k))^2] &\leqslant_{(1)} \bar{p}B_n^2\max_k\mathbb{E}_n[(v_{ik}'(\widehat{\beta}_k-\beta_k))^2]\\ &= \bar{p}B_n^2\max_k\mathbb{E}_n[\varepsilon_{ik}v_{ik}']\mathbb{E}_n[v_{ik}v_{ik}']^{-1}\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\\ &\leqslant_{(2)} c_1^{-1}\bar{p}B_n^2\max_k\|\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\|^2\\ &\leqslant c_1^{-1}\bar{p}^2B_n^2\max_{k,j}|\mathbb{E}_n[v_{ikj}\varepsilon_{ik}]|^2\\ &=_{(3)} O_{\mathrm{P}}(\bar{p}^2B_n^2(\log p)/n),\end{aligned}$$

where (1) follows from inequality (28), (2) is by Assumption M-(iv), and (3) follows by applying Lemma A.1. The claim of this step follows.

Step 3. We prove that $III_{jk}=\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}(v_{ik}'(\widehat{\beta}_k-\beta_k))]=O_{\mathrm{P}}(r_{n23})$ uniformly over $k=1,\dots,K$ and $j=1,\dots,p_k$ where $r_{n23}=\bar{p}^2B_n^2(\log p)/n$. We have

$$\begin{aligned}\max_{k,j}|\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}(v_{ik}'(\widehat{\beta}_k-\beta_k))]| &\leqslant \max_{k,j}\|\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}v_{ik}']\|\|\widehat{\beta}_k-\beta_k\|\\ &\leqslant \max_{k,j,l}\sqrt{\bar{p}}|\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}v_{ikl}]|\|\widehat{\beta}_k-\beta_k\|.\end{aligned}$$

Then

$$\begin{aligned}&\max_k\|\widehat{\beta}_k-\beta_k\| = \max_k\|\mathbb{E}_n[v_{ik}v_{ik}']^{-1}\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\| \leqslant_{(1)} c_1^{-1}\max_k\|\mathbb{E}_n[v_{ik}\varepsilon_{ik}]\|\\ &\leqslant c_1^{-1}\sqrt{\bar{p}}\max_{k,j}|\mathbb{E}_n[v_{ikj}\varepsilon_{ik}]| =_{(2)} O_{\mathrm{P}}(\sqrt{\bar{p}(\log p)/n})\end{aligned}$$

where (1) follows from Assumption M-(iv) and (2) is as in step 2. In addition, Lemma A.1 combined with inequalities (28) and (29) gives

$$\begin{aligned}\mathrm{E}[\max_{k,j,l}|\mathbb{E}_n[x_{ikj}^2\varepsilon_{ik}v_{ikl}]|] &= O(\bar{p}B_n^3(\log p)/n^{3/4} + \bar{p}B_n^2\sqrt{(\log p)/n}) \\ &= O(\bar{p}B_n^2\sqrt{(\log p)/n})\end{aligned}$$

because $B_n\sqrt{\log p}/n^{1/4} = o(1)$. Combining presented bounds yields the claim of this step. □

Lemma F.3. *$\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n} = O_{\mathrm{P}}(r_{n1})$ uniformly over $k = 1,\dots,K$ and $j = 1,\dots,p_k$. Recall that $r_{n1} = \sqrt{\bar{p}\log p}$.*

Proof. We have

$$\begin{aligned}\mathrm{E}_e[\max_{k,j}|\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}|] &\lesssim_{(1)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2])^{1/2} \\ &=_{(2)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2] + O_{\mathrm{P}}(r_{n2}))^{1/2} \\ &\leqslant_{(3)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2])^{1/2} + O_{\mathrm{P}}(r_{n2}\sqrt{\log p}) \\ &\leqslant_{(4)} \sigma\sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2])^{1/2} + O_{\mathrm{P}}(r_{n2}\sqrt{\log p}) \\ &=_{(5)} O_{\mathrm{P}}(\sqrt{\bar{p}\log p}),\end{aligned}$$

where (1) follows from Pisier's inequality, (2) is from lemma F.2, (3) follows by applying Taylor expansion since $r_{n2} = o(1)$ and $\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]$ bounded away from zero, which is because of Assumptions M-(iii) and M-(v), (4) follows from Assumption M-(iii), and (5) is due to equation (29) and $r_{n2} = o(1)$ again. The claim of the lemma follows. □

Lemma F.4. *$\sum_{i=1}^n x_{ikj}(\widehat{\varepsilon}_{ik} - \varepsilon_{ik})e_i/\sqrt{n} = O_{\mathrm{P}}(r_{n3})$ uniformly over $k = 1,\dots,K$ and $j = 1,\dots,p_k$ where $r_{n3} = \bar{p}B_n\log p/\sqrt{n}$.*

Proof. We have

$$\begin{aligned}\mathrm{E}_e[|\sum_{i=1}^n x_{ikj}(\widehat{\varepsilon}_{ik} - \varepsilon_{ik})e_i/\sqrt{n}|] &\lesssim_{(1)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2(\widehat{\varepsilon}_{ik} - \varepsilon_{ik})^2])^{1/2} \\ &=_{(2)} \sqrt{\log p}\max_{k,j}(\mathbb{E}_n[x_{ikj}^2(v_{ik}'(\widehat{\beta}_k - \beta_k))^2])^{1/2} \\ &=_{(3)} O_{\mathrm{P}}(\bar{p}B_n\log p/\sqrt{n})\end{aligned}$$

where (1) follows from Pisier's inequality, (2) is by the definition of $\widehat{\varepsilon}_{ik}$, and (3) is by step 2 in the proof of lemma F.2. The result follows. □

We now complete the proof of the theorem by applying Theorem 4.2 with T, T_0, W, and W_0 defined below. Let

$$T_0 := \max_{k,j}\frac{\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]}}.$$

Then

$$\begin{aligned} T &:= \max_{k,j} \frac{\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2]}} = \max_{k,j} \frac{\sum_{i=1}^n x_{ikj}\varepsilon_{ik}/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2] + O_{\mathrm{P}}(r_{n2})}} \\ &= T_0 + O_{\mathrm{P}}(r_{n1}r_{n2}) = T_0 + o_{\mathrm{P}}(1/\sqrt{\log p}), \end{aligned}$$

where we used lemmas F.1 and F.2 and the facts that $\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]$ is bounded away from zero and that $\bar{p}^3 B_n^4 (\log p)^4/n = o(1)$. Similarly,

$$\begin{aligned} W &:= \max_{k,j} \frac{\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\widehat{\varepsilon}_{ik}^2]}} = \max_{k,j} \frac{\sum_{i=1}^n x_{ikj}\widehat{\varepsilon}_{ik}e_i/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]}} + O_{\mathrm{P}}(r_{n1}r_{n2}) \\ &= W_0 + O_{\mathrm{P}}(r_{n1}r_{n2} + r_{n3}) = W_0 + o_{\mathrm{P}}(1/\sqrt{\log p}), \end{aligned}$$

where

$$W_0 := \max_{k,j} \frac{\sum_{i=1}^n x_{ikj}\varepsilon_{ik}e_i/\sqrt{n}}{\sqrt{\mathbb{E}_n[x_{ikj}^2\sigma_{ik}^2]}},$$

and where we additionally used lemmas F.3 and F.4 and the fact that $\bar{p}B_n(\log p)^{3/2}/\sqrt{n} = o(1)$. It follows that equations (12) and (13) in Section 4 hold for some ζ_1 and ζ_2 satisfying $\zeta_1\sqrt{\log p} + \zeta_2 = o(1)$ for chosen T, T_0, W, and W_0. Therefore, the theorem follows from Theorem 4.2. □

Appendix G. Proofs for Section 7

G.1. **Proof of Theorem 7.1.** We only consider the case with four bounded moments. The sub-Gaussian case is similar. In this proof, let $c > 0$ and $C > 0$ denote generic constants depending only on $c_1, C_1, \underline{\sigma}^2, \bar{\sigma}^2$ and their values may change from place to place. Let

$$T_0 := \max_{1\leqslant j\leqslant p} \frac{\sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n}}{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}} \text{ and } W_0 := \max_{1\leqslant j\leqslant p} \frac{\sum_{i=1}^n z_{ij}\varepsilon_i e_i/\sqrt{n}}{\sqrt{\mathbb{E}_n[z_{ij}^2\sigma_i^2]}}.$$

Step 1. We show that $\mathrm{P}(|T - T_0| > \zeta_1) < \zeta_2$ for some ζ_1 and ζ_2 satisfying $\zeta_1\sqrt{\log p} + \zeta_2 \leqslant Cn^{-c}$.

Under the conditions of the theorem, applying Corollary 2.2 followed by Pisier's inequality for Gaussian random vectors, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} \sum_{i=1}^n z_{ij}\varepsilon_i/\sqrt{n} > C\sqrt{\log p}\right) \leqslant n^{-c}. \tag{30}$$

for sufficiently large C. Moreover,

$$\begin{aligned} \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)] &= \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2] + \mathbb{E}_n[z_{ij}^2(\varepsilon_i^2 - \sigma_i^2)] + 2\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i - \varepsilon_i)] \\ &=: I + II + III. \end{aligned}$$

Consider I. We have

$$I \leqslant_{(1)} \max_{1\leqslant i\leqslant n} (\widehat{\varepsilon}_i - \varepsilon_i)^2 \leqslant_{(2)} C\|\widehat{\beta} - \beta\|^2 \leqslant_{(3)} C\|\mathbb{E}_n[v_i\varepsilon_i]\|^2$$

where (1) follows from assumption S-(ii), (2) from S-(iv) and S-(v), and (3) holds because of S-(vi). Since $\mathrm{E}[\|\mathbb{E}_n[v_i\varepsilon_i]\|^2] \leqslant C/n$, Markov's inequality implies that for every $\psi > 0$,

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2] > \psi\right) \leqslant C/(n\psi). \tag{31}$$

Consider II. Combining Lemma A.1 and Markov's inequality, we have

$$\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}^2(\varepsilon_i^2 - \sigma_i^2)]| > \psi\right) \leqslant CB^2 \log p/(\sqrt{n}\psi). \tag{32}$$

Consider III. We have $|III| \leqslant 2|\mathbb{E}_n[z_{ij}^2 v_i'(\beta - \widehat{\beta})\varepsilon_i]| \leqslant 2\|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\|\|\widehat{\beta} - \beta\|$. Using the same arguments as above, we have

$$\begin{aligned}\mathrm{P}\left(\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i - \varepsilon_i)]| > \psi\right) &\leqslant \mathrm{P}\left(\max_{1\leqslant j\leqslant p} \|\mathbb{E}_n[z_{ij}^2\varepsilon_i v_i]\| > \psi/2\right) + C/n \\ &\leqslant C\{B_n^2 \log p/(\sqrt{n}\psi) + 1/n\}. \end{aligned} \tag{33}$$

Combining (30), (31), (32), and (33) and using Taylor's expansion leads to

$$\mathrm{P}\left(|T - T_0| > C\psi\sqrt{\log p}\right) \leqslant B_n^2 \log p/(\sqrt{n}\psi) + 1/(n\psi) + n^{-c}$$

for sufficiently small ψ. By setting $\psi = (\log p)^{-1}n^{-c}$ for sufficiently small $c > 0$, we obtain the claim of this step.

Step 2. We show that $\mathrm{P}(\mathrm{P}_e(|W - W_0| > \zeta_1) > \zeta_2) < \zeta_2$ for some ζ_1 and ζ_2 satisfying $\zeta_1\sqrt{\log p} + \zeta_2 \leqslant Cn^{-c}$.

For $\psi > 0$, consider the event such that $\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)]| \leqslant \psi$, $\max_{1\leqslant j\leqslant p} |\mathbb{E}_n[z_{ij}^2\varepsilon_i(\widehat{\varepsilon}_i - \varepsilon_i)]| \leqslant \psi$, and $\max_{1\leqslant i\leqslant p}(\widehat{\varepsilon}_i - \varepsilon_i)^2 \leqslant \psi^2$. By calculations in step 1, this event has probability at least $1 - C(B^2 \log p/(\sqrt{n}\psi) + 1/(n\psi^2) + 1/n)$. Moreover, on this event,

$$\mathrm{P}_e\left(\max_{1\leqslant j\leqslant p} |\sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n}| > C\sqrt{(1+\psi)\log p}\right) \leqslant 1/n,$$

$$\mathrm{P}_e\left(\left|\max_{1\leqslant j\leqslant p} \sum_{i=1}^n z_{ij}\widehat{\varepsilon}_i e_i/\sqrt{n} - \max_{1\leqslant j\leqslant p} \sum_{i=1}^n z_{ij}\varepsilon_i e_i/\sqrt{n}\right| > C\psi\sqrt{\log p}\right) \leqslant 1/n$$

for sufficiently large $C > 0$ because $\mathbb{E}_n[z_{ij}^2\widehat{\varepsilon}_i^2] = \mathbb{E}_n[z_{ij}^2\sigma_i^2] + \mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i^2 - \sigma_i^2)] \leqslant C + \psi$ and $(\mathbb{E}_n[z_{ij}^2(\widehat{\varepsilon}_i - \varepsilon_i)^2])^{1/2} \leqslant \max_{1\leqslant i\leqslant n} |\widehat{\varepsilon}_i - \varepsilon_i| \leqslant \psi$. Therefore, on this event, using Taylor's expansion, we have

$$\mathrm{P}_e\left(|W - W_0| > C\psi\sqrt{\log p}\right) \lesssim 1/n$$

for sufficiently small ψ. By setting $\psi = (\log p)^{-1}n^{-c}$ for sufficiently small $c > 0$, we obtain the claim of this step.

Step 3. Steps 1 and 2 verified conditions (12) and (13) of section 4. The claim of the theorem now follows by applying Corollary 4.1. □

Acknowledgments

The authors are grateful for David Gamarnik, Qi-Man Shao, and Enno Mammen for helpful discussions.

References

[1] Adler, R. and Taylor, J. (2007). *Random Fields and Geometry*. Springer.
[2] Alquier, P. and Hebiri, M. (2011). Generalization of ℓ_1 constraints for high dimensional regression problems. *Statist. Probab. Lett.* **81** 1760-1765.
[3] Arlot, S., Blanchard, G. and Roquain, E. (2010a). Some non-asymptotic results on resampling in high dimension I: confidence regions. *Ann. Statist.* **38** 51-82.
[4] Arlot, S., Blanchard, G. and Roquain, E. (2010b). Some non-asymptotic results on resampling in high dimension II: multiple tests. *Ann. Statist.* **38** 83-99.
[5] Ball, K. (1993). The reverse isoperimetric problem for Gaussian measure. *Discrete Comput. Geom.* **10** 411-420.
[6] Belloni, A. and Chernozhukov, V. (2009). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, to appear.
[7] Belloni, A., Chernozhukov, V. and Wang, L. (2011). Square-root Lasso: pivotal recovery of sparse signals via conic programming. *Biometrika* **98** 791-806.
[8] Bentkus, V. (2003). On the dependence of the Berry-Esseen bound on dimension. *J. Statist. Plann. Infer.* **113** 385-402.
[9] Bickel, P., Ritov, Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.
[10] Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071-1095.
[11] Bretagnolle, J. and Massart, P. (1989). Hungarian construction from the non asymptotic viewpoint. *Ann. Probab.* **17** 239-256.
[12] Candès, E.J. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35** 2313-2351.
[13] Chatterjee, S. (2005a). A simple invariance theorem. arXiv:math/0508213.
[14] Chatterjee, S. (2005b). An error bound in the Sudakov-Fernique inequality. arXiv:math/0510424.
[15] Chatterjee, S. and Meckes, E. (2008). Multivariate normal approximation using exchangeable pairs. *ALEA Lat. Am. J. Probab. Math. Stat.* **4** 257-283.
[16] Chen, L., and Fang, X. (2011). Multivariate normal approximation by Stein's method: the concentration inequality approach. arXiv: 1111.4073.
[17] Chen, L., Goldstein, L. and Shao, Q.-M. (2011). *Normal Approximation by Stein's Method*. Springer.

[18] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012a). Anti-concentration and honest adaptive confidence bands. Working paper.
[19] Chernozhukov, V., Chetverikov, D. and Kato, K. (2012b). Gaussian approximation of suprema of empirical processes. ArXiv.
[20] de la Peña, V., Lai, T. and Shao, Q.-M. (2009). *Self-Normalized Processes: Limit Theory and Statistical Applications*. Springer.
[21] Dudley, R.M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.
[22] Dudley, R.M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Warhsch. Verw. Gabiete* **62** 509-552.
[23] Fan, J., Hall, P. and Yao, Q. (2007). To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied. *J. Amer. Stat. Assoc.* **102** 1282-1288.
[24] Frick, K., Marnitz, P. and and Munk, A. (2012). Shape-constrained regularization by statistical multiresolution for inverse problems: asymptotic analysis. *Inverse Problems* **28** 065006.
[25] Gautier, E. and Tsybakov, A. (2011). High-dimensional istrumental variables regression and confidence sets. arXiv: 1105.2454.
[26] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122-1170.
[27] Götze, F. (1991). On the rate of convergence in the multivariate CLT. *Ann. Probab.* **19** 724-739.
[28] Guerre, E. and Lavergne, P. (2005). Data-driven rate-optimal specification testing in regression models. *Ann. Statist.* **33** 840-870.
[29] Horowitz, J. L. and Spokoiny, V.G. (2001). An adaptive, rate-optimal test of a parametric mean-regression model against a nonparametric alternative. *Econometrica* **69** 599-631.
[30] Juditsky, A. and A. Nemirovski. (2011). On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Math. Program. Ser. B* **127** 57-88.
[31] Koltchinskii, V.I. (1994). Komlós-Major-Tusnády approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoret. Probab.* **7** 73-118.
[32] Koltchinskii, V. (2009). The Dantzig selector and sparsity oracle inequalities. *Bernoulli* **15** 799-828.
[33] Komlós, J., Major, P., and Tusnády, G. (1975). An approximation for partial sums of independent rv's and the sample df I. *Z. Warhsch. Verw. Gabiete* **32** 111-131.
[34] Leadbetter, M., Lindgren, G. and Rootzen, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer.
[35] Mammen, E. (1993). Bootstrap and wild bootstrap for high dimensional linear models. *Ann. Statist.* **21** 255-285.
[36] Nagaev, S. (1976). An estimate of the remainder term in the multidimensional central limit theorem. *Proc. Third Japan-USSR Symp.*

Probab. Theory. Lecture Notes in Math. pp. 419-438.
[37] Panchenko, D. (2013). *The Sherrington-Kirkpatrick Model*. Springer.
[38] Pollard, D. (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.
[39] Portnoy, S. (1986). On the central limit theorem in $\mathbb{R}^p$ when $p \to \infty$. *Probab. Theory Related Fields* **73** 571-583.
[40] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Related Fields* **98** 21-45.
[41] Romano, J., and Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *J. Amer. Stat. Assoc.* **100** 94-108.
[42] Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell Syst. Tech. J.* **41** 463-501.
[43] Smirnov, N. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR* **74** 189-191.
[44] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135-1151.
[45] Talagrand, M. (2003). *Spin Glasses: A Challenge for Mathematicians*. Springer.
[46] Ye, F. and Zhang, C. (2010). Rate minimaxity of the Lasso and Dantzig selector for the ℓ_q loss in ℓ_r balls. *J. Machine Learning Research* **11** 3519-3540.

(V. Chernozhukov) Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142, USA.
E-mail address: vchern@mit.edu

(D. Chetverikov) Department of Economics, MIT, 50 Memorial Drive, Cambridge, MA 02142, USA.
E-mail address: dchetver@mit.edu

(K. Kato) Department of Mathematics, Graduate School of Science, Hiroshima University, 1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8526, Japan.
E-mail address: kkato@hiroshima-u.ac.jp